RECREATIONAL CANNABIS LEGALIZATION: PREDICTING LOCAL POLICY ADOPTION AND ESTIMATING THE ASSOCIATED EFFECTS ON POPULATION CANNABIS USE

By

Barrett Wallace Montgomery

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Epidemiology – Doctor of Philosophy

2022

ABSTRACT

RECREATIONAL CANNABIS LEGALIZATION: PREDICTING LOCAL POLICY ADOPTION AND ESTIMATING THE ASSOCIATED EFFECTS ON POPULATION CANNABIS USE

By

Barrett Wallace Montgomery

Cannabis is undergoing a remarkable transformation from a regulated drug to a recreationally legal one in the United States (U.S.). Yet, in states that have legalized recreational cannabis, there is substantial geographic variability in actual cannabis policies and the effects of cannabis legalization are still being debated. This dissertation addresses these modern scientific issues of the recreational cannabis landscape.

The population under study primarily includes non-institutionalized U.S. civilian residents, sampled and assessed in successive waves of the National Survey on Drug Use and Health (NSDUH) starting in 2008 through 2019. Estimates on drug use and mental illness prevalences are aggregated to the county level for the first aim, and to the state level for the second and third aims. In the first aim, the county-level data are linked to several other publicly available sources of information on all 3,142 U.S. counties including the 2010 Census, 2012 presidential election, and recreational cannabis sales policies. I then used these data to train a machine learning algorithm to predict which counties allowed for the recreational sale of cannabis in 2014. In the second aim, I used state-level estimates of cannabis incidence in an event study model to estimate the effects of legalizing recreational cannabis on cannabis use onsets for persons under and over the legal minimum age of 21. The final aim focuses specifically on 21 year-olds to better understand the implications for setting a legal minimum age drug policy on age-specific patterns of incidence and proposes a theoretical framework that may help understand these findings.

For the first aim, the model-averaging predictions classified almost 94% of the U.S. counties correctly. The main factors associated with county-level recreational cannabis laws were the prevalences of past-month cannabis use and past-year cocaine use. In the second aim, I found that for those who were legally able to purchase cannabis (21 and older), cannabis legalization did not appear to affect incidence in the first year following legalization. Even so, between two and four years after legalization, the difference in differences modeling disclosed statistically robust increases of 0.6% for this sub-population of adults. After four years, the estimated increase is 1.3%. The corresponding estimates for underage persons who were ineligible to legally purchase cannabis show no appreciable differences in the occurrence in past-year cannabis use incidence. Finally, the age-specific incidence estimates for 21-year-olds show a rise after the passage of recreational cannabis laws (RCL) and are suggestive of the arrival of a new pattern of age-specific incidence.

Taken together, the work and results of this dissertation point toward four potential conclusions. First, cannabis legalization might depend on a predictable process driven in part by prior drug use in each jurisdiction. Second, once implemented, recreational cannabis legalization might not have effects on adolescent onset newly incident cannabis use. Third, for adults permitted to buy cannabis without penalty, the occurrence of newly incident cannabis use might increase. Fourth, a tentative conclusion is that legalization of retail sales to adults removes a barrier for adults who had been interested in trying cannabis, but did not do so, perhaps due to concerns about legal or social consequences faced before legalization.

This dissertation is dedicated to Kyle, Jared, Matt, CT, and Gabe.
May the next generation suffer less.

ACKNOWLEDGEMENTS

This is work supported by a National Institute on Drug Abuse R25 Science Education research training program grant award (R25DA051249) and by Michigan State University. The content is the sole responsibility of the author and does not necessarily represent the official views of Michigan State University, the National Institute on Drug Abuse, the National Institutes of Health, or the United States Substance Abuse and Mental Health Services Administration. In addition, we would like to thank the United States Substance Abuse and Mental Health Services Administration Center for Behavioral Health Statistics and Quality for sponsoring the National Surveys on Drug Use and Health and for making the datasets available for public use to allow research of this nature.

This work is the culmination of the efforts and sacrifices of many. Special acknowledgement should first be given to Olga Vsevolozhskaya, Xiaoran Tong, Meaghan Roberts, and Claire Margerison for their contributions to this work. Many thanks to Debra Furr-Holden and Jim Anthony for believing in me and guiding me, and to the professors of the department who were generous with their time and talents during my education. Many thanks are also of course due to my family and friends across the country for supporting this major decision in my life. Finally, to my wife, Evelyn, for being by my side through every struggle and success.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

KEY TO ABBREVIATIONS xii

1. Introduction and Specific Aims 1

2. History, Background, and Significance 5
2.1 Overview of this chapter 5
2.2 History 5
2.2.1 The Opium Wars 7
2.2.2 The Origins of Domestic Drug Policy 9
2.2.3 Roots in Racism and Class Warfare 12
2.2.4 Arriving at the Controlled Substances Act 15
2.2.5 The Modern Era 17
2.3 An Initial Look at Drug Policy and its Intentions 19
2.4 Related Developments in Social Statistics and Study Design 21
2.4.1 Earliest work 21
2.4.2 Early Sociology and Psychiatric Epidemiology 23
2.4.3 Classical to Modern Statistics and Causal Inference 24
2.4.4 Controlling for Confounding and Policy Analysis 25
2.5 The Current Understanding of the Effects of Cannabis Legalization 28
2.5.1 Cannabis Use in Individuals Under 21 After Legalization 28
2.5.2 Cannabis Use in Individuals Over 21 After Legalization 30
2.6 Significance 30
2.7 Potential Impact on the Field 32

3. Materials and Methods 33
3.1 Overview of this Chapter 33
3.1.1 Details on IRB Approval, Recruitment, and Participation Levels 34
3.2 Aim 1 35
3.2.1 Study population and sample 35
3.2.1.1 National Surveys on Drug Use and Health Small Area Estimates 35
3.2.1.2 Census 36
3.2.1.3 County Presidential Data 36
3.2.1.4 Cannabis Legalization Status 36
3.2.2 Data Management 37
3.2.3 Study Design 38
3.2.3.1 Pre-iteration Modelling and Validation 39
3.2.3.2 Building Ensemble Prediction 40
3.2.3.3 Sensitivity Analyses 40
3.3 Aim 2 42

3.3.1 Study population and sample....42
3.3.2 Outcome....43
3.3.3 Study Design and Statistical Analysis....44
3.3.3.1 Dates of Legalization vs. Dates of Implementation....46
3.3.3.2 Alternative Specifications and Robustness Checks....46
3.4 Aim 3....47
3.4.1 Study population and sample....47
3.4.2 Outcome....47
3.4.3 Study design....49

4. Results....51
4.1 Aim 1....51
4.1.1 Descriptive statistics....51
4.1.2 Predictive model....53
4.1.3 County-level predictions....55
4.1.4 Sensitivity analyses....57
4.2 Aim 2....61
4.2.1 Descriptive statistics....61
4.2.2 Event Study Findings....63
4.2.3 DiD Findings....65
4.2.4 Alternative Specifications and Robustness Checks....66
4.3 Aim 3....66
4.3.1 Panel study approach....66
4.3.2 Stratification at age 21....67

5. Discussion....70
5.1 Aim 1....70
5.2 Aim 2....73
5.3 Aim 3....78

6. Summary....80

APPENDICES....82
Appendix A: Supplemental Figures and Tables....83
Appendix B: Program Code Used to Derive the Constructed Study....95

BIBLIOGRAPHY....538

LIST OF TABLES

Table 1. Sample sizes and participation levels of successive years of the National Surveys on Drug Use and Health.…….....…….....……...…….....……...…….....……... 34

Table 2. Sociodemographic and political compositions and prevalences of mental illness and drug use in counties that allowed for the sale of recreational cannabis and those that did not. ……..…….....……..…….....……..…….....……..…….....……....52

Table 3. Predictors for legal cannabis sales in 2014 as represented by median z score over 1000 model iterations.……..……..……..……..……..……..…….55

Table 4. Sensitivity and Specificity of Models Using Various Weighting Techniques and Hard Cut-off Values.……..……..……..……..……..……..……..60

Table 5. Characteristics of the U.S. Population Under Study. Data from the U.S. National Surveys on Drug Use and Health.……..……..……..……..……..…….62

Table A.1. Predictors for legal cannabis sales in 2014 as represented by median z score over 1000 model iterations when proportions of voters are replaced by a binary indicating the party of the majority.....94

LIST OF FIGURES

Figure 1. Chemist J.J. Pemberton writes on the use of the coca plant in his invention, the recipe for Macalister's Cough Mixture contained cannabis, alcohol, and chloroform, and an original Bayer product containing both aspirin and heroin..... 10

Figure 2. Controlled Substances Act schedules and criteria..... 16

Figure 3. A map of states that give local authorities the right to depart from the state provisions regarding the recreational use of cannabis as of January 1, 2020..... 19

Figure 4. How the causal effect is estimated in the differences-in-differences model..... 27

Figure 5. ROC curves of 1000 predictions of county-level legal cannabis sales in 2014 and the ensemble average..... 54

Figure 6. Ensemble produced county-level probability of allowing recreational cannabis sales in 2014..... 56

Figure 7. Actual cannabis policies by county, 2014 compared to predicted policy outcomes. 57

Figure 8. ROC curves of 1000 models and average profiles for each possibility of coding the outcome in sensitivity analyses..... 58

Figure 9. Distinguishing power of ensemble predictions (weighted or naïve average)..... 59

Figure 10. Estimated effect of time since cannabis legalization on cannabis incidence in the 21 and older age group with 95% confidence intervals..... 64

Figure 11. Estimated effect of time since legalization on incidence in those aged 12 to 20 with 95% confidence intervals..... 65

Figure 12. Trends in past-year cannabis incidence by age in Colorado and Washington vs. all other states in the US, 2010-2017..... 67

Figure 13. Trends in past-year cannabis incidence for 21 year-olds in Colorado and Washington vs. all other states in the US, 2010-2017..... 68

Figure 14. Estimated effect of time since cannabis legalization on cannabis incidence at age 21 with 95% confidence intervals..... 69

Figure A.1. Percent of variance captured from over 1000 census variables in each principal component..... 84

Figure A.2. Cannabis incidence in the 21 and older age group, first wave legalizing states vs untreated states..... 85

Figure A.3. Cannabis incidence in the 21 and older age group, second wave legalizing states vs. untreated states.....86

Figure A.4. Cannabis incidence in the 21 and older age group, third wave legalizing states vs. untreated states..... 87

Figure A.5. Cannabis incidence in the 21 and older age group, first wave legalizing states vs. third wave legalizing states..... 88

Figure A.6. Cannabis incidence in the 21 and older age group, second wave legalizing states vs. third wave legalizing states.....89

Figure A.7. Estimated effect of time since cannabis legalization on past-month cannabis prevalence in the 21 and older age group.....90

Figure A.8. Estimated effect of time since legalization on past-month cannabis prevalence in the 12 to 20 age group.....91

Figure A.9. Estimated placebo effect of time since cannabis legalization on past-year cannabis incidence in the 21 and older age group.....92

Figure A.10. Estimated placebo effect of time since cannabis legalization on past-year cannabis incidence in the 12 to 20 age group..... 93

KEY TO ABBREVIATIONS

ATT	Average Treatment Effect on the Treated
AUC	Area Under the Curve
B.C.E.	Before Common Era
C.E.	Common Era
CI	Confidence Intervals
CSA	Controlled Substances Act
CUD	Cannabis Use Disorder
DiD	Differences-in-Differences
FDA	Food and Drug Administration
FMTA	The Federal Marihuana Tax Act
IRD	Internationally Regulated Drugs
IRS	Internal Revenue Service
LMA	Legal Minimum Age
NSDUH	National Survey on Drug Use and Health
PFDA	Pure Food and Drug Act
RCL	Recreational Cannabis Legalization
R-DAS	Restricted Data Access System

ROC	Receiver Operator Characteristic
SAMHSA	Substance Abuse and Mental Health Services Administration
TNR	True Negative Rate
TPR	True Positive Rate
U.S.	United States

1. Introduction and Specific Aims

Imagine for a moment that you are the mayor of a small city. The state legislature has just announced that a legal referendum on recreational cannabis has just passed by popular vote with 56% of the state population voting for the measure. You know that the legislation allows for some local control over what will happen in your city. What will you do?

Over the course of the past decade, policy and decision-makers across the country have been experiencing this hesitation in droves. It is not a simple decision. The votes passed with popular support, but almost half of the population is likely to disagree with the decision. And what about the constituents of their city? Opinions on the legalization of recreational cannabis can vary dramatically within a state.

Over the past ten years, eighteen states and three United States (U.S.) territories have legalized cannabis for people aged 21 and older. Policy and decision-makers in every municipality within these states, districts, and territories have had to wrestle with how to move forward when there are so many unanswered questions: Do *my* constituents want this? How will this affect them? How could this affect their children?

To date, researchers and social scientists have tried to answer these questions with varying degrees of success and consensus. Questions about the motivations and feelings of voters are based almost exclusively on polling, a practice that has come under fire in recent years for incorrectly predicting the results of presidential campaigns. Meanwhile, questions on how the changing laws may affect adults and children have relied on a variety of state and nationally representative surveys with results pointing in all directions depending on the outcome measured and the type of analysis. However, not one of the many studies has produced estimates on how

the changing law affects the decision making of potential first-time users. The rationale for a concentration of policy research on 'prevalence' of active cannabis use might be based on issues of statistical precision and power because recently active cannabis users are more numerous than newly incident cannabis users. Nonetheless, the recently active cannabis users often are dominated by long-time cannabis users who started their cannabis use many years before the policy change. The focus on 'prevalence' of use ignores the distinction between 'being a cannabis user' versus 'becoming a cannabis user.' A policy analysis failure can occur when investigations do not discriminate the epidemiological processes of prevalence (determined by both duration of use and incidence of use) from the epidemiological processes of becoming a newly incident user (determined solely by incidence of use). Ignoring age-specific estimates of incidence will inevitably leave important relationships between society and cannabis use hiding in the dark. This dissertation research project seeks to shine light to uncover the changing dynamics of these relationships.

In this dissertation, I aim to achieve the following goals:

1. Develop a predictive model of sub-state cannabis legalization using publicly available datasets that are readily available to other investigators, and that can be used in future investigations.

2. Provide evidence on the degree to which the incidence of cannabis use might have increased or decreased after cannabis legalization for two important subgroups of the population: (1) the adults who are permitted to make a retail purchase of a cannabis product in each jurisdiction, and (2) the underage adolescents (<21 years

old) for whom retail purchase of cannabis products remains prohibited in each jurisdiction.

3. Estimate the degree to which the legalization of cannabis might have affected the age of first cannabis use with special attention to the legal minimum age (LMA).

An important point of departure for this dissertation research project is the distinction between prevalence proportions and incidence rates. My dissertation research project is focused on the occurrence of newly incident cannabis use, year by year. Prior studies have focused upon prevalence proportions. Any epidemiologist knows that prevalence varies as a function of both incidence and duration, and the estimated size of the prevalence proportion can be dominated by long-sustained cannabis users who would likely have continued to consume cannabis with, or without, a change in policy. In contrast, the incidence parameter reflects occurrence of newly incident use, given no prior use before the interval under study (e.g., before and after the interval of cannabis policy change).

While the prevalence of cannabis use is an important measure for public health, it is only one piece of the larger puzzle. Criticism of relying so heavily on prevalence and the need to better measure incidence in chronic and mental health disorders has a long tradition (Kramer, 1957; Lapouse, 1967; Wu et al., 2003). The criticism was best espoused by Reema Lapouse:

> "Prevalence rates measure the size of the disease problem and as such are useful in planning services. They are, however, a fallible indicator of the risk of acquiring any chronic disease including psychiatric disorder. Since prevalence is a function of incidence and duration, any factors affecting duration of disease will similarly influence its prevalence rate. Thus, long-term, nonfatal, noncurable diseases which limit migration produce a pile-up of

cases and a rise in the prevalence rates. Survivorship, mobility, and duration may, in turn, be associated with demographic factors. Consequently an association between these factors and prevalence may occur even though demographic factors bear no relationship to the genesis of disease. The only suitable measure applicable to the search for possible causes of disease is the incidence rate." (1967).

In this dissertation research project, my work focuses upon the occurrence of newly incident cannabis use, consistent with the principles set forth by Kramer and Lapouse more than a half-century ago.

2. History, Background, and Significance

2.1 Overview of this chapter

This chapter will provide an overview of the prior theory, concepts, principles, and research approaches used in prior studies of cannabis policy, as well as a substantial body of literature that is relevant to understanding the development and interpretation of the results of the current line of research. The first part of this review of the literature will describe the history of drug policy in the United States (U.S.) and its international origins to inform the current views on cannabis policy in the U.S. The second section will briefly outline the history of the research conducted and methods used, intentions, and/or probable consequences on the various drug policy changes in our national history. The third section will cover the related and coinciding developments in the literature of social statistics which has allowed for the modern methods of analysis and the increasing number of publications on policy effects. The fourth and final section will cover the new wave of state-level cannabis liberalization and the research on possible changes in epidemiological parameters of cannabis use, key discoveries, and the current state of the science of drug policy research. This is not meant to be an exhaustive review of the literature. My review is intended to familiarize the reader with the basic assumptions about epidemiological research and issues that must be considered in studies intended to study non-random policy events that stimulate population-level changes in the occurrence of cannabis or other drug use. This chapter will conclude with a discussion of the significance and potential impact of the results on drug dependence epidemiology in response to policy events.

2.2 History

The legal regulation of cannabis is the most recent example in a long history of how societies have chosen to treat psychoactive drugs. Although the earliest regulation of any

psychoactive drug is believed to date back to Hammurabi's punishment for overcharging tavern patrons for alcohol, this review will focus on drugs that are currently internationally regulated (King, 1915). The current research addresses recent state and sub-state level changes in cannabis legalization in the U.S. The primary focus of this review will be cannabis policy and how the U.S. arrived at the federal law currently in effect, the Controlled Substances Act (CSA). Yet, the history of cannabis policy cannot be fully understood without some understanding of the earlier regulations to other drugs, most importantly, opium. This review will begin with the international socio-political climate which shaped so much of our modern drug policies. The chapter then covers how the U.S. arrived at the current and most conservative iteration of drug policy, the CSA.

In this brief review of the history of drug policy, I will demonstrate that the modern origins of drug policy are intrinsically tied to the economic interests of powerful countries throughout the 19th century as well as intra-national racism and class warfare. This history was expertly summarized in the book *Drug policy and the public good*:

"In its initial stages, the effort to control drugs at an international level was aimed at limiting the reach and effects of colonial empires (Carstairs, 2005). Psychoactive substances were a glue of empires in the period of European colonial expansion from about 1500 until the late 19th century (Courtwright, 2001; Jankowiak and Bradburd, 2003). From the point of view of those seeking to create markets and dependence on trade, psychoactive substances were an obvious choice; once the demand for them has been created, it becomes self-sustaining. Thus, psychoactive substances became a favourite commodity from which to extract revenues for the state, either with excise taxes or through a state-run or farmed-out monopoly. In particular, opium monopolies were an important source of revenue for colonial powers in Asia (e.g.,

Munn, 2000). In the interests of financing their empires, European states had no compunctions about forcing open markets for their psychoactive wares." (Babor et al., 2010).

Modern readers may be surprised to see no discussion of civil liberties, civil rights, the costs and benefits of incarceration, or potential effects on health that characterize the contemporary drug policy conversation. Drug policy originated from the greed of nations warring with each other for the psychoactive drug market. Nowhere in history is this more evident than in The Opium Wars fought between Britain and China in the 1840s and 1850s.

2.2.1 The Opium Wars

Opium was introduced to the Chinese sometime between the 5th and 7th centuries by Arab traders. The drug was praised and widely used to cure diarrhea, induce sleep, and reduce pain (Feige & Miron, 2008). The English, however, did not arrive in China until the 17th century and opened the first legal trading station in China in the 18th. Later, between the late 18th century and the early 19th, the famous East India Trading Company of Britain obtained a monopoly on opium from the two major trading ports of India and began exchanging opium for tea from China (Beeching, 1975).

By 1729, opium use had risen to levels deemed unacceptable by the emperor and he delivered an imperial edict, forbidding the sale of opium for smoking. Throughout the century, the Chinese government became wary of the increasing British influence on their country through the opium trade, and another imperial edict was issued, banning the import of opium. Despite the edicts, opium use was increasing in the country and stricter laws against the sale and import of opium were decreed in both 1814 and 1831. After again seeing no reductions in the use of opium, a series of internal debates were held by the emperor between those favoring legalization

and those who wanted to suppress the trade even further. The tone of these debates would be familiar to modern readers, with one side claiming that legalization and taxation would be beneficial and that black markets and wasted resources were the real crime, while the other side feared that legalizing the drug would set a poor moral standard and would result in even more widespread opium use (Feige & Miron, 2008).

The emperor ended up siding with the moral purists and opium addiction itself became a capital offense while eliminating the trade became a goal. After a few more years of still not seeing the desired reductions in opium use, the emperor appointed a special commissioner, Lin Tseh-Sen, to do anything necessary to stop the opium trade in China. Lin ordered the seizure and destruction of British opium, halted food shipments to British sailors, and poisoned their water supply. In retaliation, British sailors attacked and murdered a Chinese villager, inciting a full-on war between Lin and the British Navy (Feige & Miron, 2008).

The British defeated the Chinese and forced them to sign the treaty of Nanjing in 1842 (Chang, 1970). This is the treaty that gave Hong Kong to the British and established more ports of trade to be open to the British. The emperor, however, succeeded in keeping opium illegal and largely out of the treaty. The Second Opium War began in 1856 when Chinese officials ripped down a British flag. Within two years, the British again won the war and forced the signing of the treaty of Tientsin which gave special trading privileges to the British. It was not until after the treaty when the emperor finally succumbed to the British argument that legalizing opium was the only way to control the epidemic. Opium was legalized in 1858 and taxed at a rate of about 8% (Feige & Miron, 2008).

In this example, we see that opium was the tool of revenue extraction and international political control for the British Empire and a subject of much frustration for the Ching dynasty. The conflicts and the seeming insolubility of the problem made the emperor feel that he appeared as a weak subject of the British. Many believe that the Chinese fought, not for the health of their people, but for the economic prosperity that would come with selling opium from India (Babor *et al.,* 2010).

2.2.2 The Origins of Domestic Drug Policy

Meanwhile, in the United States of the 19th century, products derived from cannabis, opiates, and cocaine were popular and largely unrecognized everyday items in American life (Musto, 1999). The lack of regulation allowed for the proliferation of 'patent medicines' – over-the-counter concoctions often brewed with psychoactive drugs in a proprietary formula.

As medicinal chemistry and pharmaceutical industry advances were made, the most active components of the cannabis, opium, and coca leaf plants were extracted and later were synthesized ‘de novo’ (e.g., heroin, cocaine, morphine, oxycodone, hydrocodone, methedrine). Vocabulary shifted from opiates (derived strictly from opium) toward ‘opioids’ as the new laboratory synthesized products were introduced to the market and regulatory responses were required (Offermanns, 2008).

In the 19th century and the early 20th century, there was no U.S. federal policy designed to regulate whether any ‘patent medicine’ or other product included cannabis, opiates, or cocaine. No special labeling was required. The inclusion of cannabis, opiates, or cocaine in the product required no special label. In the most famous example, the original formulation of Coca-

Cola contained cocaine. At the time, numerous products used psychoactive drugs marketed with a familiar blurriness between medicinal and recreational use (see Figure 1).

Figure 1. Chemist J.J. Pemberton writes on the use of the coca plant in his invention, the recipe for Macalister's Cough Mixture contained cannabis, alcohol, and chloroform, and an original Bayer product containing both aspirin and heroin.

The first federal law governing psychoactive drugs in the U.S. did not occur until 1906 with the Pure Food and Drug Act (PFDA). Despite major opposition from the American Pharmacist Association, this legislation created the Food and Drug Administration (FDA) with its staff of trained physicians, chemists, and pharmacists (Musto, 1999). The FDA did not outlaw any psychoactive drugs or their use in patent medicines. Rather, the PFDA only required truthfulness about ingredients and prohibited false and misleading labels. Later amendments to the PFDA would require that the quantity of each drug be stated on the label and that the drugs meet official standards of purity. Thus, for a time the act served to safeguard consumers of the patent medicines.

Around the same period, the recognition of morphine addiction in the U.S. led to a quick spread of anti-morphine laws in the 1890s (Musto, 1999). However, it was not until the U.S. saw an opportunity in global politics that a national policy on opiates would be adopted. U.S. officials saw this opportunity in the long struggle between the British and Chinese governments over the British marketing of opium in China. The U.S. wanted to take an active role to meet with Chinese officials to create a system of international drug control. To their embarrassment, the Americans realized they had no national law against opium themselves. The result was the sudden efforts that lead to outlawing opium in 1909 (Musto, 1999).

Then U.S. President Theodore Roosevelt began this effort by appointing Dr. Hamilton Wright as the United States Opium Commissioner in 1908. Dr. Wright was an American physician who built his reputation on the discovery of a pathogen which supposedly caused beri-beri (Jonnes, 1999). Of course, this claim did not survive the test of time as beri-beri was determined to be a severe and chronic form of Thiamine deficiency, a discovery which Christian Eijkman and Frederick Hopkins were awarded the 1929 Nobel Prize for Physiology and Medicine (Nobel Prize Outreach AB, 2022). In studying pathogens in the tropics, Wright became interested in the countries' social and economic problems and authored many articles on the topic (Jonnes, 1999). These writings clearly caught the attention of the U.S. policymakers as he was chosen to represent the country at a series of international conventions to facilitate the regulation of opium.

The Americans met with the Chinese at the Shanghai Convention of 1909, and later, when more countries got involved, the Hague Convention of 1912. The legislation that resulted from these conventions created the framework for international drug policy still in use today. The conventions were also the first in setting the precedent of the U.S. being a driving force for these

efforts, and that one American in particular (in this case, Dr. Wright) to be incredibly influential (Babor *et al.,* 2010; Helmer & Vietorisz, 1974).

International drug policy is still largely influenced by the internal conditions in the U.S. (Musto, 1999). The U.S. policymakers of the time were motivated by a mixture of moral leadership, protection of U.S. prosperity, and a desire to assuage the resistance of the Chinese regarding American financial investments. In the creation of international and domestic laws, they were greatly aided by American citizens' prejudice against the Chinese and their association with smoking opium (Helmer & Vietorisz, 1974).

2.2.3 Roots in Racism and Class Warfare

In the late 19th century, opium smoking was largely believed to be a cultural norm of Chinese-Americans which came along with the Chinese labor force which largely built the railroads and other early industries in the American West. As in other examples of race-specific drug use stereotypes, this is likely a false over-simplification. The truth is likely more complex, involving the socio-economic drivers of race and class hierarchies and the economic prospects of White Americans (Helmer & Vietorisz, 1974).

The competition for jobs between working-class whites and the immigrant Chinese led to a campaign of excluding Chinese immigrants from the labor force between 1875 and 1880 (Helmer & Vietorisz, 1974). There is no record or official notice of Chinese-operated opium dens until this large-scale labor exclusion period (Helmer & Vietorisz, 1974). Opium use became a part of this hostile stereotype against the Chinese immigrants and was used to fuel the campaign, leading to the earliest opium legislation in the country enacted in San Francisco in 1875. "It was

its character as a Chinese habit, not as a narcotic, which warranted the earliest legislation against opium in the country" (Helmer & Vietorisz, 1974).

A similar story can be told about cocaine and Black Americans a few decades later. In the early 20th century, racially motivated horror stories regarding the actions of Black men using cocaine were widespread. Newspapers, including the New York Times, reported that "negro cocaine fiends" were raping white women (New York Times, 1914). Such reports emphasized that the local police did not have the means to prevent these violent acts. These reports were supported by important policymakers like Dr. Wright, who stated in 1910: "The use of cocaine by the negroes of the South is one of the most elusive and troublesome questions which confront the enforcement of the law in most of the Southern states…. [Cocaine] is often the direct incentive to the crime of rape by the negroes…" (Wright, 1910). Today, most believe that Wright was reporting unsubstantiated gossip as all data from the time point to extremely low prevalence rates for cocaine or other drug use among the Blacks of the South (Brecher et al., 1972; Musto, 1999; Helmer & Vietorisz, 1974). The fear spread by these news outlets coincided with the "peak of lynchings, legal segregation, and voting laws all designed to remove political and social power from him" (Musto, 1999).

Inpatient psychiatric treatment data, policing data, and import records from the time paint a very different picture. Cocaine use is believed to have peaked in 1907 followed by a sharp decrease and stabilization at low rates during World War I, a period during which cocaine use did not seem to vary appreciably across the U.S. subgroups characterized as 'White' versus 'Negro' (Brecher et al., 1972). Many believe this was due to the PFDA when liquor prohibition produced liquor scarcity. As a result, poor southerners, especially minority group members, were purported to turn to cola drinks or cocaine itself. In response, more state-level laws against

cocaine use followed, using the blueprint of stoking white fear through news outlets. Later, when some minority group members *were* becoming over-represented among drug-using groups, the public health community tended to ignore the evidence (Musto, 1999).

These racial and economic tensions helped contribute to the U.S. Congress enactment of the Harrison Narcotic Act of 1914, largely authored by Dr. Wright. The legislation defined narcotics as any opiate or cocaine (which Congress had erroneously labeled as a narcotic) derivative product (Schaffer Library, n.d.; Brecher et al., 1972). The act required that standard order receipts be issued and kept by any purchaser of narcotics and kept for two years for review by federal revenue agents (IRS). Copies were kept in a permanent file by the IRS. Registered physicians were required only to keep records of drugs dispensed or prescribed, thus protecting physicians prescribing the drugs “in the course of his professional practice only”. Maximum amounts were set for patent medicines containing heroin, opium, cocaine, or morphine, but products could still be sold in general stores or by mail order. Essentially, everyone dealing in narcotics except the consumer would have to be registered. Cannabis was notably omitted from the final version of the law (Schaefer Library of Drug Policy, n.d.).

An important development in the story of the racist and classist roots of drug policy involves cannabis and Mexican-Americans in the 1930s. In this case, Mexican immigrants certainly were using cannabis, a common custom that resembled the use of alcohol among Americans. However, there was almost no awareness of or concern for cannabis use by law enforcement or the community before the 1930s. As in the case of Chinese immigrants and opium, the conflicts began when white working-class jobs and the sustainability of industry were threatened. This situation, however, had the added ingredients of repealing alcohol prohibition laws and one incredibly amoral and powerful figure- Harry Anslinger.

Similar to Dr. Wright, Ansligner was a government appointee who fundamentally shaped drug policy in the U.S. Named as the founding commissioner of the Federal Narcotics Bureau (the precursor to the Drug Enforcement Agency) by Andrew Mellon, Anslinger led a harsh legal campaign that often conflated race with drug use and inferiority (McWilliams, 1990; Smith, 2018). Criticizing Anslinger and his methods has become very popular among scholars and activists and there is no shortage of evidence to put Anslinger's hypocrisy and racism on full display (Brecher et al., 1972; Bonnie & Whitebread, 1970; Smith, 2018). With the help of William Hearst, the prolific media mogul of the early 20th century, Anslinger successfully lobbied congress against the scientific consensus that cannabis was not harmful (McWilliams, 1990). The resulting framework of laws would come to be the foundation of the later war on drugs.

The Federal Marihuana Tax Act (FMTA) was passed in 1937 in response to political pressure from states bordering Mexico. Many members of the House of Representatives famously did not even know what cannabis was, nor what the act was introducing (Bonnie & Whitebread, 1970). A brief explanation for the changed policy, and for the origins of the pressure to change it, is that cannabis was commonly used by Mexican immigrants to the U.S. Hence, the legislation was intended to increase the cost of cannabis with the hypothesis that some Mexican immigrants would return to Mexico if they could no longer afford the drug. The FMTA was directly modeled after the earlier Harrison act authored by Wright, with a few differences. Most importantly, possession of cannabis without a written order would be punishable with a fine of up to $2,000 and no more than five years in prison.

2.2.4 Arriving at the Controlled Substances Act

The FMTA in its various forms regulated cannabis for over 30 years until Timothy Leary, a Harvard professor of psychology and outspoken psychedelics advocate, was arrested for

possession of cannabis in 1969. Leary contended that the FMTA violated his 5th amendment rights against self-incrimination and the case was elevated up to the Supreme Court of the United States (SCOTUS). In its decision, SCOTUS unanimously agreed with Leary's case and declared the FMTA to be unconstitutional. Leaving cannabis unregulated, however, was an unacceptable idea to most of the U.S. government. The removal of the outdated legislation cleared the way for a new, more conservative approach. The Controlled Substances Act (CSA) of 1970 replaced most of the federal regulations regarding psychoactive drugs that came before it, including the FMTA and the Harrison Narcotic Act, reinforcing the illegality of all previously regulated drugs. The CSA clearly stated what the authority of the federal government would be and provided a framework within which all existing and new drugs could be regulated based on the three criteria of abuse potential, safety, and medical utility (see figure 2).

Figure 2. Controlled Substances Act schedules and criteria.

	Schedule I	Schedule II	Schedule III	Schedule IV	Schedule V
Potential for abuse	The drug or other substance has a high potential for abuse	The drug or other substance has a high potential for abuse	The drug or other substance has a potential for abuse less than the drugs or other substances in schedules I and II	The drug or other substance has a low potential for abuse relative to the drugs or other substances in schedule III	The drug or other substance has a low potential for abuse relative to the drugs or other substances in schedule IV
Medical use	The drug or other substance has no currently accepted medical use in treatment in the United States	The drug or other substance has a currently accepted medical use in treatment in the United States or a currently accepted medical use with severe restrictions	The drug or other substance has a currently accepted medical use in treatment in the United States	The drug or other substance has a currently accepted medical use in treatment in the United States	The drug or other substance has a currently accepted medical use in treatment in the United States
Consequences of abuse	There is a lack of accepted safety for use of the drug or other substance under medical supervision	Abuse of the drug or other substance may lead to severe psychological of physical dependence	Abuse of the drug or other substance may lead to moderate or low physical dependence or high psychological dependence	Abuse of the drug or other substance may lead to limited physical dependence or psychological dependence relative to the drugs or other substances in schedule III	Abuse of the drug or other substance may lead to limited physical dependence or psychological dependence relative to the drugs or other substances in schedule IV

An important change in the legal framework between this legislation and prior legislation was the ability to regulate drugs as they were developed using a common structure, without the

need for congressional legislation. This need was made apparent in the period after World War II in which many new synthetic narcotics were developed (Spillane, 2004). Most importantly for our purposes, however, is that cannabis was now classified alongside heroin and hallucinogens as drugs with a high potential for abuse, unsafe to use (even under medical supervision), and with no currently accepted medical use (U.S. Department of Justice, 1970).

The use of psychoactive drugs is a timeless societal practice. Whether for recreational, medicinal, or spiritual use, societies from before the common era until the past century largely accepted and did not think to regulate the use of cannabis. When cannabis did become regulated in the US, the intentions were undoubtedly immoral and facetious. Yet, the intention and consequences of the actions are two very different constructs. Overall, it is not clear whether this misguided process has been beneficial for citizens of the U.S. or detrimental. Judgments of this type cannot be prescribed as the benefits and harms of cannabis regulation are largely value-based and vary widely from person to person. In the history of drug policy in the U.S., there has been some movement from its origins in racism and personal crusades to a more evidence-based approach, yet the legacy of the CSA has not achieved anything close to equality. The evidence regarding cannabis shows a widening gap in how the drug is treated by the law and how it is perceived by society (Pew, 2021). The FMTA, with all its flaws, lasted 40 years. It has now been 50 years since the CSA was passed, and society has demonstrated its impatience with this outdated law through state-level legalization, most often through popular vote on a ballot measure.

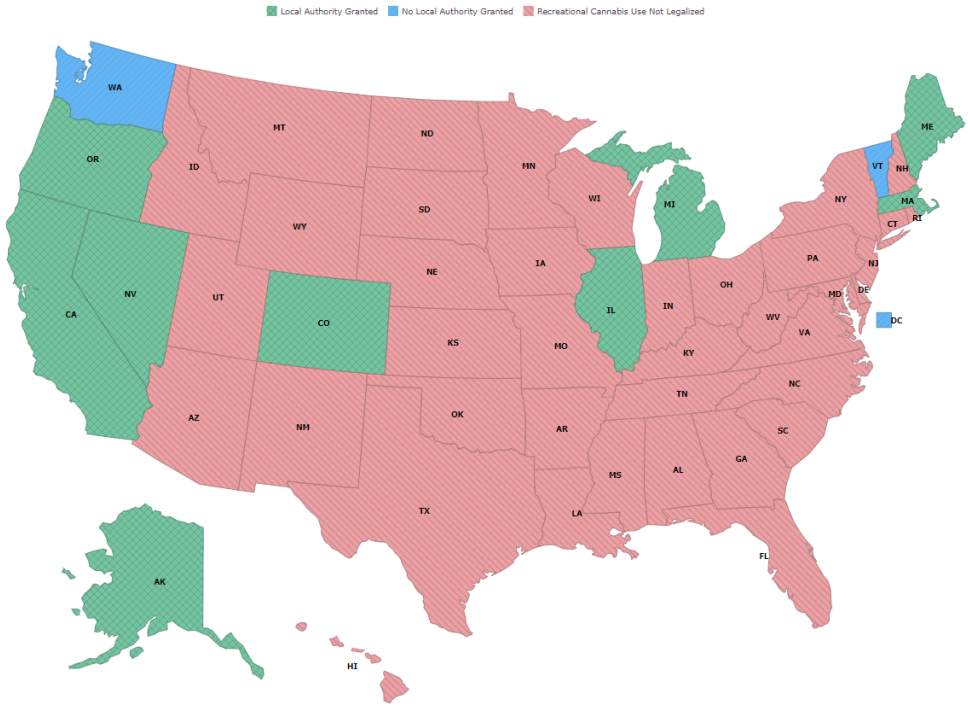
2.2.5 The Modern Era

The era of modern cannabis legalization can be traced back to the medicalization movement in 1990's California but did not truly begin until the first two legalization ballots passed

in 2012. Colorado and Washington were the first two states to legalize the recreational use and sale of cannabis in 2012. Alaska and Oregon followed suit in 2014, legalizing recreational cannabis through ballot measures as well. In 2016, California, Nevada, Maine, and Massachusetts all approved ballot measures to legalize recreational cannabis as well. In a slight deviation, Vermont became the first state to legalize recreational cannabis through the state legislature in 2018, although not allowing for the commercial sale of cannabis. In that same year, voters in Michigan approved a ballot measure to legalize recreational cannabis as well. In 2019, Illinois followed the example set by Vermont and legalized recreational cannabis through the state legislature, and in 2020, Vermont legalized cannabis sales, again through the state legislature, and voters in Arizona, Montana, New Jersey, and South Dakota all approved ballot measures to legalize recreational cannabis. In the most recent movement, voters in South Dakota voted to legalize both recreational and medical cannabis.

Critical to this dissertation, many states have granted sub-state jurisdictions (cities and counties) the authority to make their own decisions regarding the legalization of cannabis. This results in a legal patchwork of many county or sub-state areas with differing policies regarding cannabis. Figure 3 shows a map of the states that allow local authorities to depart from the state provisions for recreational cannabis legalization as of January 2020.

Figure 3. A map of states that give local authorities the right to depart from the state provisions regarding the recreational use of cannabis as of January 1, 2020.



2.3 An Initial Look at Drug Policy and its Intentions

Research into the epidemiological parameters of drug use did not occur until the turn of the 20th century around the time that Hamilton Wright participated in the Shanghai convention. In what is perhaps the first analysis of drug dependence epidemiology and the laws which influence it, Lawrence Kolb and A.G. Du Mez published a review of drug use under different policies. The evidence showed disparate estimates of “the number of addicts” using a number of different methods and data sources in areas with differing policies (1924).

In this section I provide a rough overview of historical drug policy instruments that have been used (sometimes) in attempts to shift epidemiological parameters in directions of public health improvements. The table lists a selection of the drug policy instruments I have found in my review of the literature, along with my description of what I have discerned as the intended

shifts in epidemiological parameters and any epidemiological evidence that does exist. To some extent, there is a vagueness in the epidemiological parameters because the historical records do not always clearly state the intended purpose of implementing each drug policy instrument. In some instances, I have had to infer the intended purpose. In other instances, the intended purpose might have been outside the boundaries of what we think of as epidemiological and public health parameters, as illustrated by apparently racist social control effects of early policies on opium smoking and cannabis use described elsewhere.

Estimates of the number of narcotic (again, defined at the time as both opiates and cocaine) dependents in the U.S. during the time of the first national laws range from 182,215 (1884) to 782,118 (1913), or an estimated 1-2% of the population. Survey research approaches of the time were not nearly as rigorous then as it is today and scholars agree that no one survey can be trusted (Bonnie & Whitebread, 1970). In a 1924 review of the most rigorous studies completed between 1915 and 1922, the breadth of the estimate was found to be even wider. However, after a careful evaluation of the biases in each survey method and sample, the authors arrive at a likely figure of around 215,000 in 1915 and 110,000 in 1922 (Kolb & Du Mez, 1924).

In a masterful work of the time, the authors of *The Opium Problem* come to the similar conclusions on the state of the estimates of opium users. Terry and Pellens cite much of the same work with estimates between “a few thousand individuals to several millions” and conclude that “under present conditions it is impossible to obtain [an accurate estimate of the total number of opiate users]” (1928). Nonetheless, the book presented some of the earliest evidence on the basic epidemiology of opium users including the demographic differences, etiology, pathology, treatments, and symptomology.

The decrease in the estimates reported by Kolb and Du Mez reflects the cultural shift of the time as opposition grew to the use and promotion of psychoactive drugs, partly influenced by religiosity and the temperance movement in Britain and the U.S., but also by indigenous movements among colonized peoples (Babor et al., 2010). In the first international law regarding psychoactive drugs, the Brussels General Act of 1890 regulated distilled spirits for large parts of Africa. Liquor would be prohibited "for the native population" or its sale would be taxed (Babor et al., 2010). The proposed intention of the law was to protect the native people of Africa. Seen through the lens of international power dynamics, the prohibition aspects of the law appear more as a mechanism of control. In this regulation, we also see the earliest framework of the bifurcated system of law governing psychoactive drugs- one mechanism controlling supply while the other controls demand. The taxation aspect of the law is a penalty to the supplier, while prohibition seeks to control demand, an important dynamic in drug policy.

In this line of research, the development of new statistical techniques has been critical. Controlling for changing demographics and other differences in populations before and after a policy change, or between populations with different policy experiences, continues to be the most important issue.

2.4 Related Developments in Social Statistics and Study Design

2.4.1 Earliest work

The earliest instance of probabilistic thinking is commonly believed to have been recorded by Cicero, who, around the year 85 B.C.E., referred to events likely to happen as *probabile* (Gigerenzer, et al., 1990). It was not until the 14th century C.E. that the first known attempt at a systematic calculus for enumerating all possibilities of dice rolls was written, the ancestor of today's concept of permutation (Kendall, 1956). The origin of statistical probability is

commonly believed to have started with the Italian maritime insurance industry of the 15th and 16th centuries. Despite this persistent belief, these insurers kept no data on shipwrecks or other mishaps and the premiums they established were somewhat arbitrary (Gigerenzer et al., 1990). The reality is in the opposite direction; it was the early mathematics of Blaise Pascal and Pierre de Fermat which influenced the insurance industry (Maistrov, 2014). Pascal and Fermat are primarily credited with creating the first formulations of probability in 1654. Their contributions included the fundamental concept of expectation defined as the product of the probability of an event e and its outcome value V:

$$P_{(e)}V = E$$

Once the work of Pascal and Fermat was completed, applications of probability theory spread. One early use of data to draw statistical inferences includes an important milestone for epidemiology, as manifest in John Graunt's Natural and Political Observations Made upon the Bills of Mortality, first published in 1662 (only 8 years after Pascal and Fermat's seminal work). Graunt sought to predict mortality and survivorship of the citizens of London in ten-year intervals (Graunt, 1939). Important developments in probability theory continued alongside statistics, most notably by Huygens, Bernoulli, DeMoivre, Bayes, and LaPlace, until the two were merged to create the important science of inference we use today (Gigerenzer, et al., 1990; Maistrov, 2014). Nevertheless, here I will branch out to the early epidemiological work focused on disease outcomes in populations

It was during the age of enlightenment that these classical probability theorists made it clear that their mathematics should be applied to "civil life", a "social mathematics" as the early social scientist Nicolas de Condorcet put it (Baker, 1975). Condorcet even went as far as arguing to restructure the French judicial system to be based on statistical probability (Boland, 1989).

This movement, of course, was never embraced by society at large as later statisticians found the logic to be too simplistic (Gigerenzer, et al., 1990). The application of statistics to the social sciences for other problems, however, proved to be enormously useful and influential. Especially influential contributions were made by William Farr and Emile Durkheim, including attempts to understand the epidemiological patterns of suicide mortality rates (Durkheim, 1897; Farr, 2000).

2.4.2 Early Sociology and Psychiatric Epidemiology

Farr worked primarily with census data and mortality records, being the first to combine the two to test whether suicide risk varied with other life experiences. In his own words, Farr explained and concluded that "The Importance of this determination will become apparent by enumerating some of the relations the mortality bears to other orders of facts… the difference of external circumstances and sanitary condition exercise a very real influence on life, disease, and death…" (Farr, 2000).

Emile Durkheim's *Rules of the Sociological Method* built on the types of observations Farr had made. Although they were contemporaries, I could find no evidence that Farr influenced Durkheim's work or vice-versa. Durkheim applied his newly developed sociological theory in his seminal work *On Suicide.* He argued that higher rates of suicide were partially due to the absence of shared social values and norms (*anomie*) in the general population and lower rates of suicide with the opposite - more social integration and shared values (Durkheim, 1897).

While great strides were being made on the pioneering topic of psychiatric epidemiology, similar work on drug dependence syndromes was not nearly as advanced as the analytical epidemiology already being conducted on suicide. The earliest descriptive epidemiological estimates on drug dependence often took the form of attempts at estimating the "number of addicts" in the U.S. (Kolb & Du Mez, 1924) (see section 2.2 for estimates). Much work had yet to

be done in the domain of survey methodology to produce reliable population-level estimates, but more important to the current work was the development of modern statistics and the concept of controlling for confounding to create causal estimates.

2.4.3 Classical to Modern Statistics and Causal Inference

The concept of correlation (and the need to control for it in certain cases with a tool he named "regression") was first introduced by Francis Galton in his seminal paper *Typical Laws of Heredity*, published in Nature in 1877 (Galton, 1877). Some of Galton's later work sparked the interest of Karl Pearson, who is largely credited with the development of modern mathematical statistics (Varberg, 1963). Among Pearson's many contributions are the precursors to conventional probability distributions and the P-value. However, most important to the current work are the contributions of Ronald Fischer and his later debates on how to constitute causality with Austin Bradford Hill.

Fisher is credited with developing the foundations of modern statistical science, initiating the original principles of study design, and developing the first randomized trials in his agricultural work as a way of correctly adjusting for random error terms (Hald, 1998). British statistician and epidemiologist Austin Bradford Hill is credited as the first scientist to apply Fisher's concept of randomization in studies of humans (Hill, 1951; Hill, 1952; Hill, 1953). The randomized controlled trial is perhaps the greatest advance towards the framework of causality as it is still considered the gold standard for estimating the counterfactual (what would have happened to these people if an event did *not* occur?). Yet, it was Hill's work with British epidemiologist Richard Doll using non-random case-control studies, that would convince the world that smoking tobacco was the primary cause of lung cancer and other illnesses (Hill, 1965; Hill and Doll, 1956).

Despite the large, estimated associations linking tobacco with morbidity and mortality, Fisher was skeptical. He asked whether causal inference was justified when the data did not have the rigor of a randomized trial (Andersen, 2007). The arguments and debates of this age culminated in the now-famous first Surgeon General's report on tobacco and health of the 1960s. This report laid out a new framework for determining causality and sided with Hill and Doll that smoking indeed was a cause of lung cancer, heart problems, and many other detrimental health outcomes, without the need for experimental evidence (United States Surgeon General, 1964). This new framework opened the door for the "web of causation". Allowing for the consideration of different forms of evidence (strength of association, consistency, specificity, temporality, etc.) in analyzing many possible contributing causes of disease and health behaviors is now widely used, yet still controversial (MacMahon, Pugh, and Ipsen, 1960). The framework was later revisited by Hill who authored what many consider to be the best list of criteria for determining causality. Hill's criteria are still in use today in the study the causes of chronic diseases which have no definitive microorganism or other specific causal agent to which we can point our collective finger (Hill, 1965; Shimonovich et al., 2020). The conceptualization of the "web of causation" was necessary to facilitate our more modern concept of studying proportions of causal attribution in a probabilistic world.

2.4.4 Controlling for Confounding and Policy Analysis

To date, governments have rarely adopted Fisher's random experiment design when its decision-makers seek to change the state of affairs by making new policy decisions. There are some exceptions to this general rule, as in the U.S. federal government's Moving To Opportunity housing voucher experiment (Katz, Kling, & Liebman, 2001; Leventhal & Brooks-Gunn, 2003; Chetty, Hendren, & Katz, 2016). Instead, the implementation of policy is almost always inherently

nonrandom, particularly in a democratic society like the United States. The passage of laws and policies in the U.S. can occur by several mechanisms but are primarily voted on and approved by a majority ballot or precedent can be set by court rulings. In both cases, the beliefs and shared values of the citizens affected by the law drive the change of the law itself - whether directly, by voting on a ballot measure, or indirectly, by voting on political appointments and local judges.

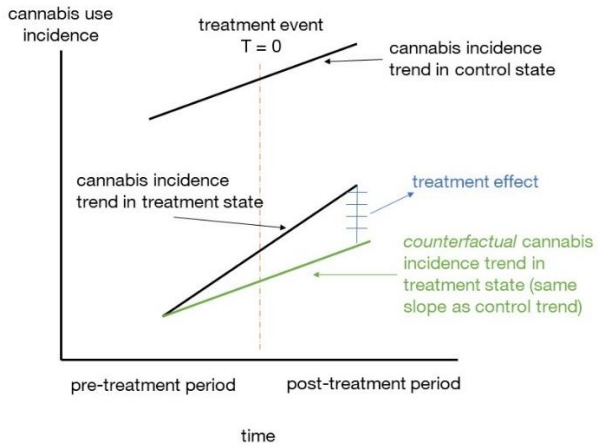
This policy-making process yields inherently confounded relationships when the goal is to estimate the effects of various policies. This topic of study is of the utmost importance to the current work which seeks to estimate the causal effect of recreational cannabis legalization (RCL). The roots of this dissertation project's approach can be seen in early evolution of time series experiments and quasi-experiments in education and psychology (e.g., see Campbell & Stanley , 1963).

The field began with the development of interrupted time series analysis by Campbell and Cook (1979). In their textbook for quasi-experimentation in field settings, Campbell and Cook outline the method of interrupted time series as an evaluation of the change in the level of an outcome over time if the level of the outcome would not have changed if the intervention under study did not occur. Many critics noted that this is a strong assumption to be made, is especially prone to selection bias, and the outcome may vary with many other factors, both known and unknown (Baicker & Svoronos, 2019).

To deal with these criticisms, the differences-in-differences (DiD) approach was devised. DiD is one of the best tools econometrics offers to deal with unobserved confounders with a loose assumption that the trends in both the treatment and control groups are parallel (Angrist and Pischke, 2008). DiD models require panel data or repeated cross-sections and are ideal when the event to be modeled occurs at an aggregate level, such as a state. DiD takes advantage

of the trends, before and after the event, in a group that experienced the event (treatment group) and a group that did not experience the event (control group) (figure 4).

Figure 4. How the causal effect is estimated in the differences-in-differences model.



Recent explorations and analyses by economists have revealed that this estimate of the average treatment effect is a bit of an over-simplification, especially when more than two time periods can be defined for each group (Goodman-Bacon, 2018; Callaway and Sant'Anna, 2020; Cunningham, 2020). The average treatment effect on the treated (ATT) is a weighted average of all the possible two-period estimators, which is problematic as it averages out the treatment effect heterogeneity that can take place over time. When treatment effects change over time, the ATT estimate is biased (Goodman-Bacon, 2021). In the drug policy literature, there is good evidence that effects will change over time due to the so-called policy lag effect (Cheng et al., 2019; Hall & Weier, 2015).

In this dissertation research project, my goal has been to estimate the causal effect of RCL passage on the occurrence of newly incident cannabis use in the United States using an extension of the DiD model that uses treatment leads and lags to dynamically model the changes in cannabis use incidence before and after the law was changed. This model allows for the effect

of time since or before the RCL passage to be estimated while controlling for the fixed effects of the states and time. Using this event study model with leads and lags in treatment timing, I will show that all states were comparable on cannabis incidence dynamics. I then estimate the degree to which the treatment effect changes over time and estimate the effects of RCL passage on cannabis incidence separately for underage persons and for those aged 21 and older.

2.5 The Current Understanding of the Effects of Cannabis Legalization

Research into U.S. drug policy has gained pace in recent years alongside the growing number of states which have legalized cannabis. The legalization movement has been fueled by a growing belief among Americans that cannabis should be legal (Pew, 2019), partially fueled by recognition of the adverse consequences of mass incarceration (Gallup, 2016). At the heart of this research lies the goal of quantifying the changes in cannabis use epidemiology among the constituents of states that have legalized cannabis. Since the beginning of the legalization era in the U.S., much has been written about the potential impacts of the law changes and a few studies have gone as far as to ascribe causal changes to the differences the authors saw. This literature has focused almost exclusively on measures of prevalence and by and large focuses on two distinct populations- individuals under the age of 21, and individuals 21 and over.

2.5.1 Cannabis Use in Individuals Under 21 After Legalization

The evidence on the effects of cannabis legalization on cannabis use among youth is mixed. In one of the earliest studies of its kind, no change in the past 30-day prevalence of cannabis use among Colorado high schoolers was found between 2013 and 2014 in the Healthy Kids Survey (Gruber et al., 2016). According to a more recent analysis of the Healthy Kids Service, this flat trend has remained unchanged through 2019 (Reed, 2021). In Washington,

researchers found an increase in past 30-day prevalence among tenth graders according to Monitoring the Future data (Cerdá et al., 2017). However, another group analyzed a different dataset collected from the same population of high schoolers in Washington State and found decreases among eighth and tenth graders in the Healthy Youth Survey data (Dilley et al., 2019).

More recent analyses of the National Survey on Drug Use and Health (NSDUH) found that 12–17-year-old participants in the states with legalized recreational cannabis (RCL) had an increased prevalence of cannabis use disorder (CUD) (Cerdá et al., 2020). Yet, another analysis found no changes in cannabis use prevalence for any racial or ethnic group among individuals aged 12 to 20 (Martins et al., 2021). In an analysis of the national sample survey data collected for the Youth Risk Behavior Survey, Coley and colleagues found no evidence that RCLs were associated with increased likelihood or level of cannabis use among adolescents and even found a 16% lower prevalence of use among prior cannabis users (2021). Meanwhile, however, Paschall, García-Ramírez, & Grube reported an increase in lifetime cumulative incidence proportion of cannabis use in the California Healthy Kids Survey (2021).

Midgette and Rueter have argued that some of the heterogeneity in results might be attributable to differences in sampling between nationally representative surveys and state representative surveys when investigating state-specific results (2020). However, differences between nationally representative and state representative surveys are not always in the same direction. A 2019 meta-analysis on the topic showed null findings when analyzing the results of studies with the lowest risk of bias, while a small increase was found in prevalence when including all studies (Melchior et al., 2019). In a systematic review published in the same year, the authors concluded that findings among youth were mixed primarily by state, with increased

use prevalence among youth in Washington and Oregon, but not in Colorado (Smart and Pacula, 2019).

2.5.2 Cannabis Use in Individuals Over 21 After Legalization

The evidence on the prevalence of cannabis use after legalization among adults has been more consistent. Excepting some early null findings from the Colorado Behavioral Risk Factor Surveillance System and the National Alcohol Surveys (Reed, 2016; Kerr, Lui, & Ye, 2018), evidence from studies published after 2016 consistently show increases in cannabis use prevalence. Despite the earlier null finding in Colorado, recent cannabis use prevalence increased to 17.5% in 2017 and continued to 19% in 2019 (Reed, 2021). Increases in some adult age groups in the prevalence of frequent users, CUD, and past-year cannabis use prevalence have been consistently reported among adults in the NSDUH (Cerdá et al., 2020). These findings are especially robust among the White and Hispanic sub-groups (Martins et al., 2021). Although the early studies and systematic review of the literature reported no effects for adults (Reed, 2016; Kerr, Lui, & Ye, 2018; Smart and Pacula, 2019), the more recent analyses show consistent evidence of increased use among adults.

2.6 Significance

As proposed, my dissertation research project was designed to address three critical barriers in the current state of policy-guiding evidence that should be produced when cannabis use epidemiology seeks to shed light on recent changes in state-level recreational cannabis laws. First, many major works to date have failed to adequately address the within-state variation that exists in each of these states with regards to municipality and county cannabis laws. Second, none of these analyses have been based on estimates of the occurrence of newly

incident cannabis use. All prior research has focused on prevalence proportions. Third, there has not been a uniform framework to make it possible to control for differences between areas that legalized cannabis and those that have not, or within the same state before and after cannabis retail sales have been permitted.

Ignoring the variation of the law at a sub-state level is an error that could result in a biased view of recreational cannabis legalization by averaging out important subgroup heterogeneity. Many local municipalities and counties have chosen to keep cannabis as a regulated schedule I drug, or to ban the commercial sale of cannabis, after the state legalized the drug. Residents in these areas are included in state-level and nationally representative samples. When analyzing differences between states, these individuals are included in a state where recreational cannabis is “legal” (i.e., differential misclassification).

As stated in Chapter 1, prevalence estimates do not capture the rate at which adolescents and adults are trying the drug for the first time. Age of first use is of particular interest in cannabis epidemiology given that one of the major pillars of cannabis policy is to prevent new users, especially in their adolescent years. Incidence is a critical component in understanding the public health consequences of legalization given the plethora of evidence that associates younger age of first use with a vast array of negative outcomes (Volkow et al., 2014; Fontes et al., 2011; Horwood et al., 2010; Wagner, 2002). Incidence is also an essential component in testing the hypothesis that a sub-group of the population does not use cannabis for the sole reason that it is illegal and for understanding the shifting distributions in the age of first use.

One path toward attributing a causal relationship between cannabis policy change and an epidemiological parameter of interest is to create a framework that can be used to control for

differences in states with and without cannabis legalization. Developing the prediction algorithm is a novel method to determine which facets of different areas vary with the policy changes and which need to be controlled for in non-experimental designs. This framework allows for repeatable experiments and can be used by other researchers in their work on other changes that may have occurred as a result of recreational cannabis legalization.

Addressing these limitations is significant to public health. Targeted prevention campaigns for alcohol and tobacco use have been a major public health success story, partly due to early age-specific targeting (Dobbins et al., 2008) and appropriate messaging (Pierce, White, & Messer, 2009). The results of this study might reveal if the age of first use pattern changes after cannabis policy liberalization and will give us a detailed understanding of the demographics who are experiencing these changes.

2.7 Potential Impact on the Field

If the aims of this project are achieved, a more specific understanding of cannabis initiation and use after recreational cannabis legalization will help guide future policy decisions and initiatives. Technical capacity to project changes in epidemiology will be improved by incorporating the effects of these often-unmeasured local factors and the confounding variables which facilitated legalization in the first place. Successful completion of these aims will increase the accuracy of current models of cannabis use and improve our understanding of the effects of drug policy changes.

3. Materials and Methods

3.1 Overview of this Chapter

To understand how cannabis legalization affects the epidemiology of cannabis use we, of course, must have valid estimates of the main epidemiological parameters at the population level. In addition, we also must understand the differences between states and counties which prefer to legalize cannabis from those that prefer to keep cannabis as a schedule I drug to make valid inferences regarding causation. This chapter will outline the methods, by aim, that I used to:

1. Develop a predictive model of sub-state cannabis legalization using publicly available datasets that are readily available to other investigators, and that can be used in future investigations.

2. Provide evidence on the degree to which the incidence of cannabis use might have increased or decreased after cannabis legalization for two important subgroups of the population: (1) the adults who are permitted to make a retail purchase of a cannabis product in each jurisdiction, and (2) the underage adolescents (<21 years old) for whom retail purchase of cannabis products remains prohibited in each jurisdiction.

3. Estimate the degree to which the legalization of cannabis might have affected the age of first cannabis use with special attention to the legal minimum age (LMA).

3.1.1 Details on IRB Approval, Recruitment, and Participation Levels

The current study was determined by the MSU IRB as not human research on 8/27/2021. Proof: STUDY00006620.

Overall interview participation levels in the NSDUH are between 67%-75%, which is slightly lower than corresponding levels for the 12-22 year-olds in this study's sub-samples. See table 1 for the sample size, response rates, and overall participation rates in the NSDUH for each year under study.

Table 1. Sample sizes and participation levels of successive years of the National Surveys on Drug Use and Health.

Survey Year	Total Sample Size	Weighted Screening Response Rate	Weighted Interview Response Rate	Overall Participation Level*
2008	68736	89%	74%	66%
2009	68700	89%	76%	67%
2010	68487	89%	75%	66%
2011	70109	87%	74%	65%
2012	68309	86%	73%	63%
2013	67838	84%	72%	60%
2014	67901	82%	71%	58%
2015	68073	80%	70%	56%
2016	67942	78%	68%	53%
2017	68032	75%	67%	50%
2018	67791	73%	67%	49%
2019	67625	71%	65%	46%

3.2 Aim 1

3.2.1 Study population and sample

For this predictive study of recreational cannabis policy change, I used data from a variety of publicly available sources to study 3094 counties (including county equivalents) of the United States (U.S.). The data on the counties consist of data collected from individuals aggregated to the county level as well as data that is inherent to the counties themselves. The publicly available data sources I used include the 2010 - 2012 Small Area Estimates from the National Surveys on Drug Use and Health (NSDUH), the 2010 Census, and the 2012 County Presidential Data from the MIT Elections Lab. Each data set used in the analysis is described separately.

3.2.1.1 National Surveys on Drug Use and Health Small Area Estimates

The population in these surveys was specified to include non-institutionalized U.S. civilian residents, sampled and assessed for successive NSDUH surveys in the years 2010, 2011, and 2012. The data used in this project was made available at the substate region level (n=369) and downloaded from SAMHSA's Public Data Access System. These NSDUH cross-sectional surveys were conducted with multistage area probability sampling to draw state-level representative samples and to over-sample 12-to-17-year-olds. In the NSDUH surveys administered in 2010, 2011, and 2012, data were collected from 206,222 individuals with an average interview participation level of 74% (Montgomery, Thompson, & Anthony, 2022). Data from 170,978 of these participants were made available in the public use file and used in this analysis (Montgomery, Thompson, & Anthony, 2022). Home addresses of all participants were collected and used in a statistical model which links the survey outcome variables to local area predictors so that the survey outcome of interest in an area that may have not been chosen in the probability sampling stage can be predicted. The variables I used from this data source

include the prevalences of alcohol use disorder in the past year, alcohol use in the past month, cigarette use in the past month, cocaine use in the past year, serious thoughts of suicide in the past year, illicit drug dependence in the past year, marijuana use in the past month, and serious mental illness in the past year.

3.2.1.2 Census

Data from the 2010 Census was downloaded from the census.gov website (Census Bureau, 2010). The decennial census seeks to count every member of the U.S. population and records basic demographic information such as age, sex, marriage status, race and ethnicity, information about the households and living arrangements, and county-level information including total population, land area, water area, population density, and the number of occupied and vacant housing units. The data are reported at the county level.

3.2.1.3 County Presidential Data

The Massachusetts Institution of Technology Election Data and Science Lab collects and makes available data on U.S. presidential elections, as well as data on U.S. house and senate elections, and state and local elections. The County Presidential Elections Returns 2000-2020 were used in this analysis (MIT Election Data and Science Lab, 2018). The file contains the total number of votes in every U.S. county for each major party presidential candidate in the general election (democrat, republican) as well as the total votes cast for third parties. I used the percent of votes for the republican and democratic candidates as the variables in this analysis.

3.2.1.4 Cannabis Legalization Status

In 2014, four states had legalized recreational cannabis at the state level. Colorado and Washington legalized recreational cannabis in 2012 and Oregon and Alaska legalized in 2014. A responsible government department from each state collected and published lists of cities and

counties that opted out of different aspects of the recreational cannabis laws. Colorado counties were coded according to data published by the State Governments or Municipal League (CML, 2019). Alaskan boroughs were coded according to data from Alaska's Department of Commerce, Community, and Economic Development (ADCCED, 2017). Counties in Oregon were coded according to data collected and published by the Oregon Liquor and Cannabis Commission (OLCC, 2021).

Unlike the other three states, the legislation approved in Washington did not allow substate municipalities local authority over the issue (Colorado Constitution, 2012; Washington State Liquor Control Board, 2012; Oregon legislature, 2014; Alaska State Legislature, 2014). While local authority was not granted to cities and counties explicitly through the legislation, land use laws were used to effectively ban the sale of cannabis in some areas with varying degrees of success (Darnell, 2015; Dilley et al., 2017). I discuss the methods I used to deal with this complication and others in the Sensitivity Analyses section.

Counties that included a city or town where selling recreational cannabis was legal in 2014 were coded as having legal recreational cannabis sales. I used Statsamerica.org's City and county Finder to trace each municipality to its home county (Statsamerica, 2021). In some cases, a city could exist in more than one county, in these cases, all counties were coded as having legal recreational cannabis.

3.2.2 Data Management

To appropriately merge all data at the county level, I first created a proprietary crosswalk dataset to assign every U.S. county to the NSDUH small area estimate regions as defined by the Substance Abuse and Mental Health Service Administration's (SAMHSA) documentation (United States, 2014). This crosswalk allowed me to use the small area estimate of each NSDUH

outcome for each county that existed within its boundaries. Estimates for areas smaller than a county were not included (District of Columbia wards, LA Statistical Areas, Detroit, and Wilmington City) as these areas were all nested within broader county-defined regions. Because of mismatches in documentation, some small area estimates from North Carolina could not be used. Mental health and drug use estimates in Massachusetts and Connecticut are not county-specific. I attempted to mitigate the effect of this missing data by imputing state-level averages of these variables for all counties in their respective states. Similarly, discrepancies in the presidential election data from Alaska and a small village in Hawaii made this data unusable at the county level, the state level averages of presidential voting in Alaska were imputed for all counties.

Including every variable from the census data caused separation and numerical instability in the regression models. Therefore, I used principal components analysis to reduce the number of census variables from over 1000 to 10 principal components (figure A.1). I then used the first two principal components which account for ~80% of variance as predictors.

The NSDUH small area estimates are often suppressed when a county is not large enough or the estimate is below a certain cut-off. Because of this correlation between county size, estimate sizes, and missing data, I concluded that the data missing from the NSDUH small area estimates are not missing at random and would bias the models if used. However, this never occurs in the estimate which uses the entire age range of the surveys (12 and over, or in some cases, 18 and over). Therefore, all NSDUH variables used in the prediction are for the whole survey population for whom the data is available. Estimates by age subgroups were not used.

3.2.3 Study Design

Because I must classify a relatively small number of counties from a larger set (92 RCL

counties, 3011 non-RCL counties), the overabundance of counties where recreational cannabis remains illegal would bias my attempts at classification (Nekooeimehr & Lai-Yuen 2016). This would, in turn, result in the inapplicability of the algorithm to the real world (Zolbanin et al., 2020). As such, I accounted for the imbalance between the policy conditions being examined using a relatively simple minority oversampling technique (Kubat, Holte, & Matwin, 1998). To create probability estimates and 95% confidence intervals (95%CIs) for each county, I used 1000 resamples in the ensemble of logistic regressions (Kittler, 2001). In each iteration, 80% of the RCL counties (n~74 out of 92) and twice as many non-RCL counties (n ~ 148=74x2) were randomly allowed into the predictive model.

I introduced variance to the NSDUH estimates so that the same estimates of drug use and mental health prevalences were not used for the same county in every iteration. Instead, I transformed the estimate logarithmically with a normal distribution and chose a number probabilistically from the distribution about the log transformed estimate. This number was then back-transformed to be interpretable as a prevalence measure. In other words, I am not using the observed data as is, but as a probabilistic realization surrounding it. In this way, I acknowledge the fact that the NSDUH estimates are not absolute truth, but close estimates. The added variation in each iteration also prevents an overfitted regression model completely separating RCL and non-RCL counties just by a few descriptive statistics, which harms the overall specificity of the stacked ensemble predictor by making it overly sensitive to the said statistics.

3.2.3.1 Pre-iteration Modelling and Validation

Modelling and validation were performed using standard techniques in supervised machine learning. In every sample iteration, after ~74 RCL and ~147 non-RCL counties were

selected, the data were split with a random subset of 70% used to train the logistic regression model, after which I saved the regression coefficients and standard errors. The remaining 30% served as testing data. To evaluate the predictions using the test cases, I used a four-step process. First, I used the trained model to evaluate the expected probability of legalization from the assembled county data. Second, I hard-coded the prediction using a cut-point based on the overall proportion of legalized counties in the model, which was 33.3% in this case due to the 1:2 sampling scheme. A county was predicted to have RCL if its expected probability exceeded 33.3%, otherwise it was labeled as non-RCL (for logic on choosing a prevalence based cut-off value, see Gelman & Hill, 2006). Third, I compared the predicted labels to the truth and evaluated the logistic model's performance in terms of true positive rate (TPR, or sensitivity), true negative rate (TNR, or specificity), overall classification accuracy, and the area under the receiver operator characteristic (ROC) curve (Hanley & McNeil, 1982). Finally, I stored the expected probabilities, the predicted labels, and the performance values of the model for future ensemble building.

3.2.3.2 Building Ensemble Prediction

I calculated the ensemble's probability of legalization for each county by averaging and weighting the expected probabilities from the subset of models that made predictions on that county. Both the weighting mechanism and the cut-off value for the prediction label can be changed to suit the needs of the application. In this study, I weighted the probabilities by the overall classification accuracy of the model from which the probability was derived, emphasizing the importance of models that performed better in the 1000 interactions. This was followed by a final call for the binary prediction label with the same 33.3% cut-point.

3.2.3.3 Sensitivity Analyses

Because of differences between state administrative structures and cannabis policies, I

planned to conduct several sensitivity analyses to understand which policy coding schemes would produce the most accurate model. The two states that warranted some re-coding and analysis were Washington state and Alaska. As mentioned previously, the cannabis reform legislation of Washington state is unique in this sample as it is the only state which did not explicitly allow for local bans. However, this is a matter of some controversy, as zoning laws were used to effectively ban the sale and cultivation of cannabis in 11 counties (Darnell, 2015). Though the mechanisms are not the same, the sociodemographic factors which lead to them may be similar. I sought to understand whether including the 39 Washington counties in this analysis would improve the prediction of local cannabis laws. Washington counties in this analysis were coded according to the Washington State Institute for Public Policy's preliminary implementation report (Darnell, 2015).

As explained previously, Alaska is divided using a system of boroughs and not counties. Although they function much the same, unlike county-equivalents in the other 49 states, the boroughs do not cover the entire land area of the state. Along with the inability to code voter information to Alaskan boroughs, no local action was taken to ban cannabis at the county level. Because there is no variance among Alaskan boroughs in this outcome variable, there is reason to believe that Alaska is not representative of the other 49 states. Thus, I also modeled the data with and without Alaska to understand whether including information from this state improves the predictive model.

For illustrative purposes, I also present sensitivities and specificities of the models using several different weighting mechanisms and varying the binary cut-off at every 0.1 interval between 0.1 and 0.9, besides the default 0.333. As I have noted, the weighting mechanisms and cut-offs can be varied to alter the utility of the model to prioritize sensitivity or specificity. I present

this information in the hope that future researchers may use it to inform the application of such models.

3.3 Aim 2

3.3.1 Study population and sample

For this epidemiological study, the population was specified to include non-institutionalized U.S. civilian residents, sampled and assessed for successive NSDUH survey waves, 2008 through 2019. These NSDUH cross-sectional surveys were conducted with multistage area probability sampling to draw state-level representative samples and to over-sample 12–17-year-olds. The total sample size for surveys conducted in this period includes 819,543 respondents with an average overall interview participation level of 58% (Substance Abuse and Mental Health Services Administration, 2021; Montgomery, Thompson, & Anthony, 2022).

In Aim 1, it was possible to focus on the county-level cannabis laws because the NSDUH variables I used from the small area estimate public use files were more common than cannabis incidence and were also aggregated across three years (2010-2012). First time cannabis use becomes a relatively rare event after the teen years. Hence, estimates of incidence for the respondents aged 21 or older are not made available at the county level to protect respondents from possible re-identification. Aims 2 and 3 focus on the state-level incidence for this reason.

Standardized audio computer-assisted self-interview modules assessed the month and year of first cannabis use, from which age-specific incidence rates can be estimated from the NSDUH Restricted Data Access portal (R-DAS). The R-DAS portal provides analysis weights and variance estimate capabilities for state-specific and national estimates and 95% confidence

intervals (CI). The R-DAS portal also allows for state-specific analysis of data but can only be downloaded in year-pairs and not individual years (e.g., 2018 – 2019 vs. 2018, 2019); therefore, I use data from six year-pairs in the analysis, not 12 individual years. I categorized states into different analysis groups according to each state's year of recreational cannabis legalization (RCL) through 2018. Because the 2018-2019 year-pair is the most recent available data in R-DAS at the time of analysis, states that legalized cannabis in 2019 or later were categorized into the illegal group. Washington and Colorado were included in the 2012 group; Oregon, Alaska, and Washington D.C. were in the 2014 group; California, Maine, Massachusetts, and Nevada were included in the 2016 group; and Vermont and Michigan were included in the 2018 group. All other states were categorized into the same illegal cannabis group for this analysis.

3.3.2 Outcome

To test the hypotheses, the primary estimate is past-year cannabis use incidence, calculated as $\psi = X_r/N_r$, where X_r is the number of individuals starting to use cannabis within the one to twelve month interval before assessment, and N_r is all persons who had not started using cannabis before that interval. Estimates described in this report are not readily available in R-DAS. The estimated prevalence rates ($p_1 = X_r/N$, where N is the total projected population size) and the estimated proportion of the population at risk ($p_2 = N_r/N$), with the corresponding standard errors can be obtained. Incidence can then be calculated in terms of p_1 and p_2 as:

$$\psi = \frac{p_1}{p_2} = \frac{X_r/N}{N_r/N}.$$

3.3.3 Study Design and Statistical Analysis

My study design observed changes in annual cannabis incidence in the RCL states relative to non-RCL states before and after the legalization of cannabis at the state level. I estimate this using an event-study model that allowed me to estimate incidence (or other outcomes) in each period relative to legalization while controlling for fixed differences across states and national trends over time. All analyses were performed in SAS version 9.04 and use NSDUH survey weights.

The models can be expressed as:

$$Y_{st} = RCL_s \times \sum_{\substack{y=-5 \\ y\neq -1}}^{4} \beta_y I(t - t_s^* = y) + \beta_t + \beta_s + \epsilon_{st}$$

As described earlier, the data is constructed at the state category (s) by year (*t*) level. In the primary analyses, Y_{st} measures past-year cannabis incidence for each state grouping and pair of years. In this equation, β_s denotes state fixed effects and β_t denotes fixed effects of time in calendar years. These account for general trends in cannabis incidence for each group of states over time. The variable $RCL_s$is set equal to one if the observation is from a state that legalized cannabis and was measured after the date of legalization and is set equal to zero otherwise. The time-event dummy variables $I(t - t_s^* = y)$ indicate the legality of cannabis in each state group by the first year of the R-DAS year pair relative to the year of legalization (t_s^*) and are set equal to zero for all observations from states that did not legalize recreational cannabis during the study period. These variables are referred to in this analysis as leads (indicators of time-event before legalization) and lags (indicators of time-event after legalization). The omitted category is $y = -1$, the year pair before legalization. Therefore, each estimate of

β_y is an estimate of the difference between past-year cannabis incidence in the RCL states relative to the illegal states during year y, as measured from the year pair that immediately preceded legalization. After multiplying the coefficient by 100, these coefficients can be interpreted as the percentage point change in the past-year cannabis incidence in RCL states relative to non-RCL states. Where only one or two categories of states would be included at a specific time point because of the variation in legalization timing across states (≤6 years before legalization and ≥ 4 years after legalization), the indicators are combined to balance the leads and lags and prevent modelling the outcome for only a small subset of the data. This is commonly referred to as balancing the leads and lags of the model.

If past-year cannabis incidence was trending similarly in all the state groups before legalization, I expect that the estimated coefficients for the lead indicators will be too small to represent a true difference. This is a test of the parallel trends assumption built into the regression models. Similarly, if the estimated coefficients for the lag indicators are positive, this indicates an increase in the incidence of past-year cannabis use in the RCL states whereas negative coefficients would indicate a decreasing incidence.

In addition to the event study estimates of change at each time point, I also present a simple 2x2 DiD estimate of the ATT as a summary of the estimated effect across all post-legalization years through 2019. This is estimated using the same equation except that the event study dummy variables are replaced with a single indicator denoting an RCL state post-legalization.

3.3.3.1 Dates of Legalization vs. Dates of Implementation

The best practice in the field has been to analyze the data using the date that cannabis sales began as the divider between pre and post-periods. However, because of the nature of the data as reported in the R-DAS system, using the date of legalization made for a cleaner analysis. I note that the average number of days between the date of legalization and of sales in the states in the sample (except for Washington D.C. where sales have never been legal) is 497 days. Therefore, the T0 period in this analysis is a close approximation of the time between legalization and implementation of the RCLs. The expectation of increased incidence would begin to show in the surveys after this roughly 500 day period when recreational cannabis sales began.

3.3.3.2 Alternative Specifications and Robustness Checks

To ensure the robustness of the analyses, I used two different alternate specifications. The first alternate specification uses the same method to estimate the effect of RCL on cannabis prevalence. The estimate for prevalence has been studied extensively in the literature and I compare the results to prior estimates as a check of face validity for the model. The second robustness check uses a time placebo as a check of robustness. In this model, a random year within the data was selected as the year that states legalized cannabis. The model is then run with the same specifications. If any of the model's coefficients are appreciably different, then this indicates that there may be a problem in the model or that it is over-sensitive to spurious associations.

3.4 Aim 3

3.4.1 Study population and sample

For this study, the population was specified to include non-institutionalized U.S. civilian residents, sampled and assessed for successive National Surveys on Drug Use and Health (NSDUH), 2010 through 2019. These NSDUH cross-sectional surveys were conducted with multistage area probability sampling to draw state-level representative samples and to over-sample 12–17-year-olds, with overall interview participation levels of 67%-75%, slightly lower than corresponding levels for the 12–22-year-olds in this study's sub-samples. Standardized audio computer-assisted self-interview modules assessed the month and year of first cannabis use, from which age-specific incidence rates can be estimated from the NSDUH Restricted Data Access portal (R-DAS). The R-DAS portal provides analysis weights and variance estimate capabilities for state-specific and national estimates and 95% confidence intervals (CI).

3.4.2 Outcome

For this research, the primary estimate is again the first-time cannabis use (incidence), calculated as $\psi = X_r/N_r$, where X_r is the number of individuals starting to use cannabis within the one to twelve month interval before assessment at age 21. N_r is all persons who had not started using cannabis before that interval, stratified by cannabis policy. Estimates described in this report are not readily available in R-DAS. The estimated prevalence rates ($p_1 = X_r/N$, where N is the total projected population size) and the estimated proportion of the population at risk ($p_2 = N_r/N$), with the corresponding standard errors can be obtained. I note that the incidence can be calculated in term of p_1 and p_2 as:

$$\psi = \frac{p_1}{p_2} = \frac{X_r/N}{N_r/N}.$$

The corresponding variance can be calculated using the standard statistical procedures as:

$$Var(\psi) = \frac{1}{p_2^2} Var(p_2) + \frac{p_1^2}{p_2^4} Var(p_1).$$

Furthermore, I discovered that R-DAS estimates can often be produced for the entire population of interest (e.g., age-specific cannabis incidence over all 50 states), and for a sub-population that includes a relatively large, unweighted numerator and denominator (e.g., first-time cannabis use in every state except Colorado and Washington). Nevertheless, estimates for the other subpopulation (e.g., age-specific cannabis incidence in Colorado or Washington) may often be suppressed due to privacy concerns. In the instance when two sub-populations can be considered mutually exclusive, a method for estimating the suppressed output "by hand" was used (Vsevolozhskaya & Anthony, 2014). Specifically, if I let ψ be the incidence of cannabis use in all 50 states, and ψ_{NCW} be the incidence of cannabis use in every state except Colorado and Washington, I can estimate the suppressed output as:

$$\psi_{CW} = \frac{N}{N - N_{NCW}} \psi - \frac{N_{NCW}}{N - N_{NCW}} \psi_{NCW},$$

Where N is the projected population size in all 50 states and N_{NCW} is the projected population size in every state except Colorado and Washington, then the corresponding variances can be calculated as:

$$Var(\psi_{CW}) = \left(\frac{N}{N - N_{NCW}}\right)^2 Var(\psi) - \left(\frac{N_{NCW}}{N - N_{NCW}}\right)^2 Var(\psi_{NCW})$$

3.4.3 Study design

The RCL policies that were implemented by Colorado and Washington State in 2014 were largely modelled after the state's own policies regarding alcohol sales. The age of onset distribution for alcohol has always been different from that of illegal drugs, characterized by the same peak of incidence in adolescence as the other drugs, but with an additional peak in incidence at age 21. A hypothesis for this difference in patterns was offered by Cheng and colleagues that involves distinct sub-group variation in the population with one group willing to try a drug even though it is illegal for them, and another that waits until it is legal (at age 21) to try (2018). I hypothesized that setting the legal minimum age for cannabis purchase at 21 in Colorado and Washington State would cause a distinct shift in the age of onset distribution from one that had only one peak in the adolescent period to one with a second peak at age 21. This shift would be consistent with the theory that a distinct sub-group of the population was interested in using cannabis but would wait until their 21st birthday to do so. The implication being that a subset of the population only tries to use cannabis if it is legal to do so.

To detect this change in the shape of the age of onset distribution, I first looked at the raw incidence rates using two approaches. The first approach is a panel study with sample restriction to participants in the birth cohort born in either 1995 or 1996, successively re-sampled to secure a new sample each year. Using a cohort born in the years 1995 or 1996 allows for tracking the population experience of persons born in these years whose adolescent period occurred prior to the RCL implementation but turned 21 after cannabis had been legalized. Although the panel approach has constrained statistical power, given its focus on that one birth cohort, the strength of this design is that it provides an intuitive look at how incidence patterns changed for this cohort depending on if they lived in Washington or Colorado or any other state.

The second approach is more tightly focused on what happens at age 21. The expectation is that cannabis incidence at age 21 years in Colorado and Washington will show an increase versus the relatively stable cannabis incidence at age 21 years in the other 48 states.

In addition to observing the raw incidence rates, I applied the same event study model used in aim 2 to detect changes specifically at age 21 between the populations of states that legalized cannabis and those that did not. This third method goes beyond simply looking at the Colorado and Washington test cases and uses data in the same way that aim 2 was approached. Again, using data from six year-pairs in the analysis, not 12 individual years, I categorized states into different analysis groups according to each state's year of RCL through 2018. Because the 2018-2019 year-pair is the most recent available data in R-DAS at the time of analysis, states that legalized cannabis in 2019 or later were categorized into the illegal group. Washington and Colorado were included in the 2012 group; Oregon, Alaska, and Washington D.C. were in the 2014 group; California, Maine, Massachusetts, and Nevada were included in the 2016 group; and Vermont and Michigan were included in the 2018 group. All other states were categorized into the same illegal cannabis group for this analysis.

The study design for aim 3 observes changes in annual cannabis incidence specifically at age 21 in the RCL states relative to non-RCL states before and after the legalization of cannabis at the state level. Again, I used an event-study model that allowed me to estimate incidence (or other outcomes) in each period relative to legalization while controlling for fixed differences across states and national trends over time. All analyses were performed in SAS version 9.04 and use NSDUH survey weights.

4. Results

4.1 Aim 1

4.1.1 Descriptive statistics

Although there are 3142 counties in total in the U.S. at this time, this analysis included 3094 counties in 366 sub-state regions defined by the NSDUH 2010-2012 small area estimates. The three sub-state regions and 39 counties in Washington state were not used in this analysis, Washington residents had voted to legalize cannabis in 2012 but were not allowed local authority over the issue. Because of this, these counties could not be considered an adequate exposure or control group and were not included in this analysis. Additionally, nine counties in North Carolina could not be used due to discrepancies in documentation. RCL counties included 42 of the 64 counties in Colorado, 21 of the 36 counties in Oregon, and all 29 counties in Alaska. In Colorado and Oregon, a similar proportion of the counties (34% and 42%, respectively) opted out of legal cannabis sales using local government mechanisms. Nine municipalities banned retail cannabis sales in Alaska, but there were no county-level bans as in Colorado and Oregon.

Sociodemographics, political affiliations, and mental health and drug use prevalences of the counties by policy exposure are presented in table 2. In counties where the sale of cannabis was legal in 2014 there is a slightly higher proportion of males. These countries tend to have more people falling into the 18–64-year-old age range and less under 18 and over 65. These counties are less racially diverse with a higher proportion of white citizens and less black citizens, although counties where recreational cannabis was sold legally have a higher proportion of American Indians and Native Alaskans. Surprisingly, the RCL counties have a higher proportion of republican than democrat voters, at least according to the 2012 presidential race. The use of alcohol, cannabis, and cocaine was more prevalent in RCL counties, as were alcohol use and

other substance use disorders, and serious mental illnesses.

Table 2. Sociodemographic and political compositions and prevalences of mental illness and drug use in counties that allowed for the sale of recreational cannabis and those that did not.

		Sale of cannabis not legal (n=3011)	Sale of cannabis is legal (n=92)
Variable			
Gender			
	Male	49.1%	49.9%
	Female	50.9%	50.1%
Age			
	Under 18 years	24.0%	23.8%
	18 to 34 years	23.2%	24.1%
	35 to 64 years	39.6%	40.6%
	65 and over	13.1%	11.5%
Race			
	White	63.3%	72.6%
	Black or African American	12.7%	3.1%
	Hispanic or Latino	16.5%	15.8%
	American Indian and Alaska Native	0.7%	1.9%
	Asian	4.7%	3.5%
	Native Hawaiian and Other Pacific Islander	0.1%	0.3%
	Some Other Race	0.2%	0.2%
	Two or More Races	1.9%	2.7%
Political party			
	Voted for the republican nominee in 2012 presidential race	38.1%	50.0%
	Voted for the democratic nominee in 2012 presidential race	59.9%	46.5%
Mental Health			
	Past-year prevalence of serious mental illness [a]	4.3%	4.5%
	Past-year prevalence of suicidal thoughts [a]	3.9%	4.0%
Substance use			
	Past-month alcohol use prevalence [b]	48.8%	54.5%

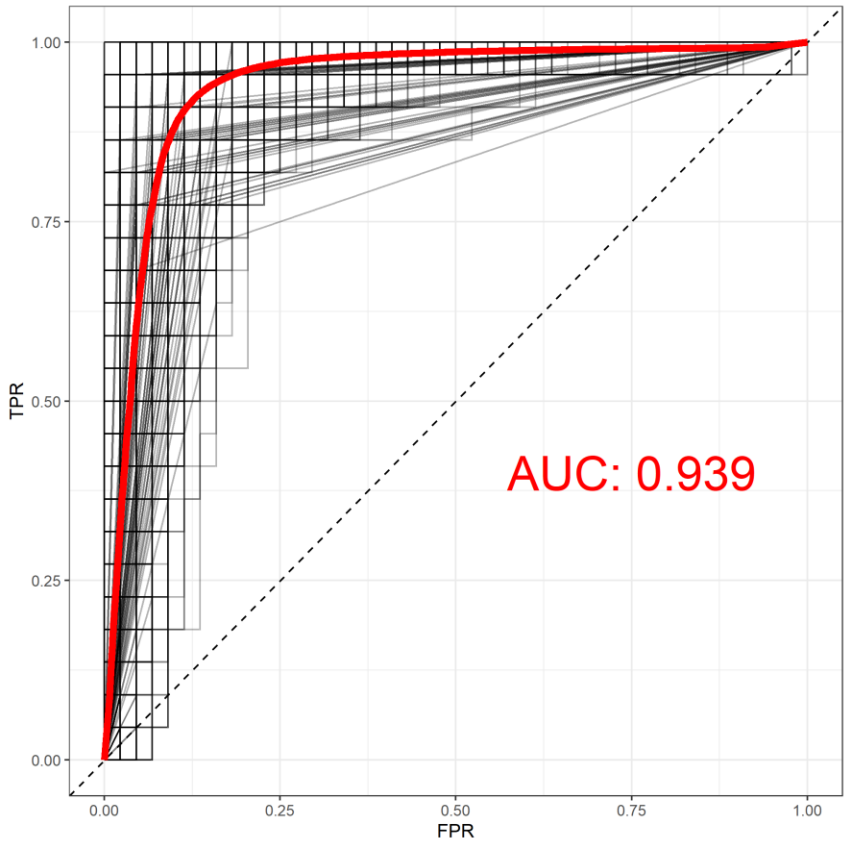
Table 2 (cont'd)

	Past-month cigarette use prevalence [b]	25.2%	24.8%
	Past-month cannabis use prevalence [b]	6.0%	10.9%
	Past-year alcohol use disorder prevalence [b]	6.6%	8.0%
	Past-year cocaine use prevalence [b]	1.4%	2.0%
	Past-year substance use disorder prevalence [b]	1.7%	1.9%
Footnotes [a] Prevalences of mental illnesses for individuals aged 18+ as sampled by the NSDUH [b] Prevalences of substance use for individuals aged 12+ as sampled by the NSDUH			

4.1.2 Predictive model

Figure 5 shows the ROC curves of 1000 iterations of logistic modeling in grey, with their average profile in red. In general, our model demonstrates a high degree of discrimination with an average area under the ROC curve (AUC) of 0.94.

Figure 5. ROC curves of 1000 predictions of county-level legal cannabis sales in 2014 and the ensemble average.



The drivers of the model's predictive power are represented in table 3 by their median Z scores over all iterations. By this metric, the most powerful predictor is past-month cannabis use, followed by past-year cocaine use, and serious mental illness. Although not shown explicitly, the negative predictive power of voting for the Democratic or Republican candidate is relative to the percentage of votes for a third party, therefore a higher prevalence of voting for a different political party was positively predictive of legalizing cannabis sales.

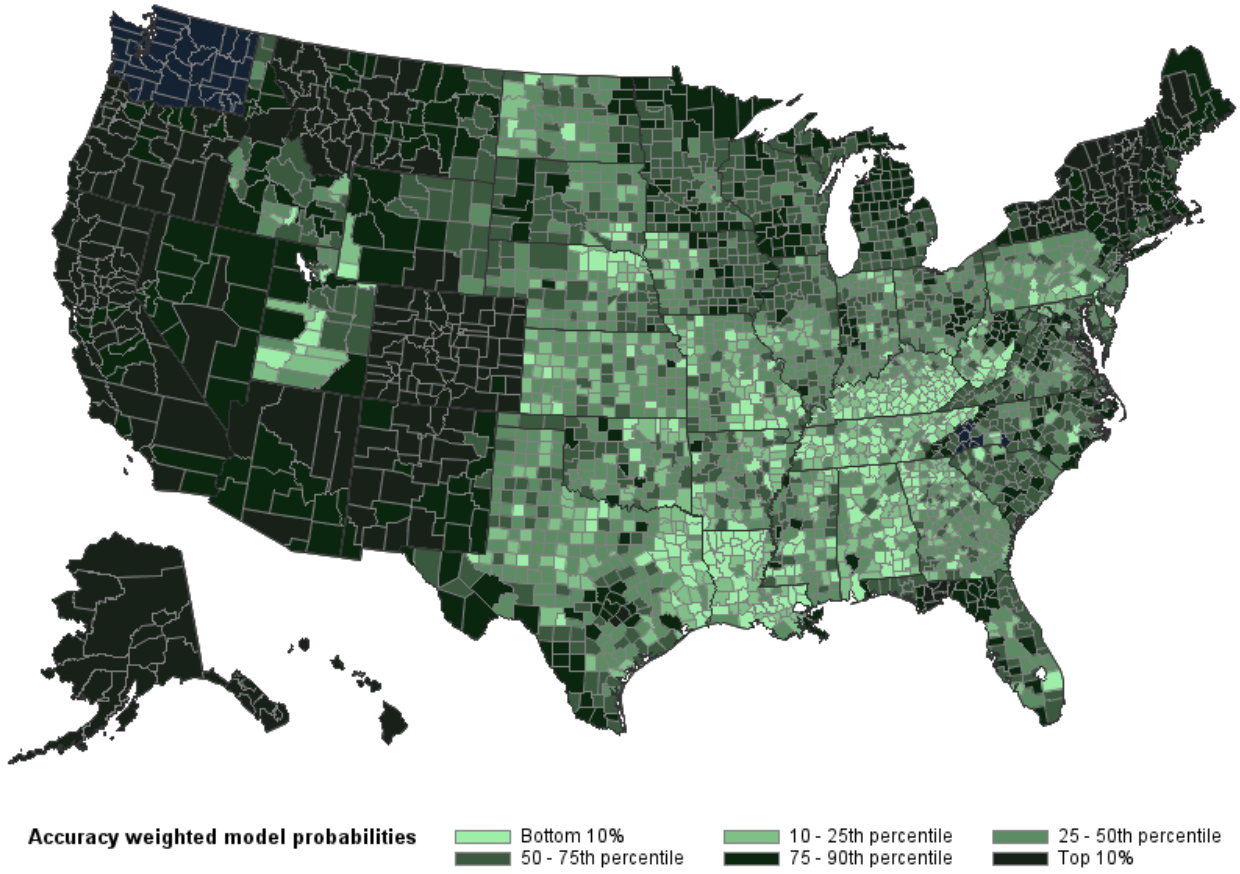
Table 3. Predictors for legal cannabis sales in 2014 as represented by median z score over 1000 model iterations.

Variables	Median z score
Past month cannabis use prevalence [b]	2.871
Past year cocaine use prevalence [b]	2.239
Past year prevalence of serious mental illness [a]	1.351
Land area	1.342
Proportion of votes for a 3rd party candidate in 2012 presidential race	0.912
Proportion of votes for the republican candidate in 2012 presidential race	-1.203
Past month cigarette use prevalence [b]	-0.829
Census principal component 2	-0.802
Census principal component 1	-0.715
Past month alcohol use prevalence [b]	0.618
Past year alcohol use disorder prevalence [b]	0.536
Past year substance use disorder prevalence [b]	-0.481
Area water	0.467
Past year prevalence of suicidal thoughts [a]	-0.326
Footnotes [a] Prevalences of mental illnesses for individuals aged 18+ as sampled by the NSDUH [b] Prevalences of substance use for individuals aged 12+ as sampled by the NSDUH	

4.1.3 County-level predictions

By averaging the weighted expected probabilities derived from 1000 logistic regression models, our ensemble model produced probabilities of legalizing cannabis sales in 2014 for every county. Each probability was weighted by the classification accuracy of the model from which they were derived and averaged across each iteration in which the county appeared. Figure 6 shows every county in the U.S. (except those in Washington state and North Carolina that could not be included) categorized by its percentile in the ensemble predicted probabilities of RCL.

Figure 6. Ensemble produced county-level probability of allowing recreational cannabis sales in 2014.

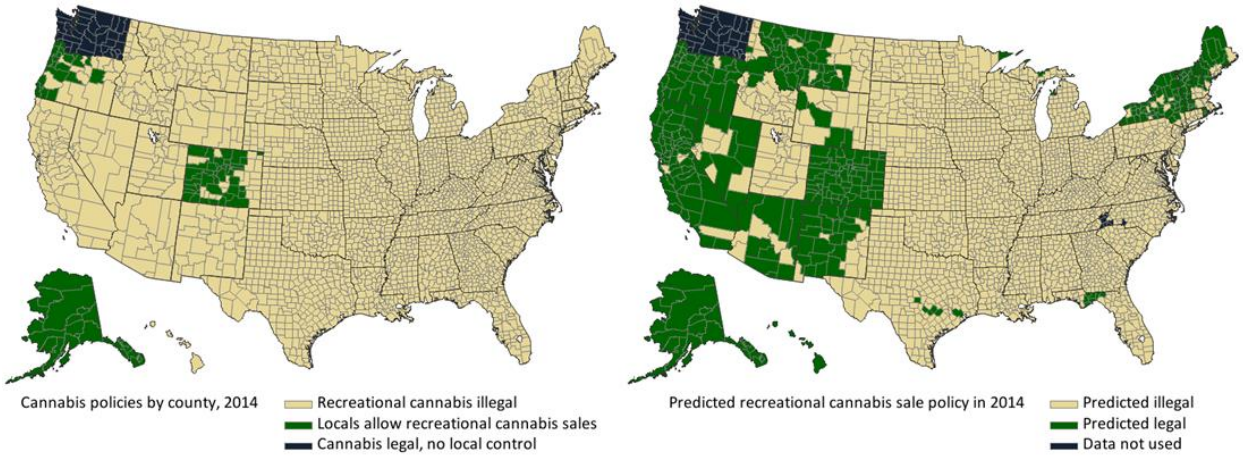


When using accuracy weighted probabilities and a binary cut-off of 0.333, the ensemble correctly categorized all 92 of the counties with legal cannabis and 2721 of the 3002 counties that did not legalize cannabis. For incorrect classifications, 281 counties were predicted to legalize cannabis that did not by 2014 while no counties with legal cannabis sales were incorrectly classified.

Figure 7 compares the actual cannabis policy landscape by county in the U.S. in 2014 to our predicted policy landscape. The figure demonstrates how and where the model performs best and where it lacks specificity as demonstrated by the false positive clusters. Some false

positives appear in the states that legalized cannabis (Oregon and Colorado), but also in Arizona, California, Washington D.C., Florida, Hawaii, Idaho, Maine, Massachusetts, Michigan, Minnesota, Montana, Nevada, New Hampshire, New Mexico, New York, Rhode Island, Texas, Vermont, and Wyoming.

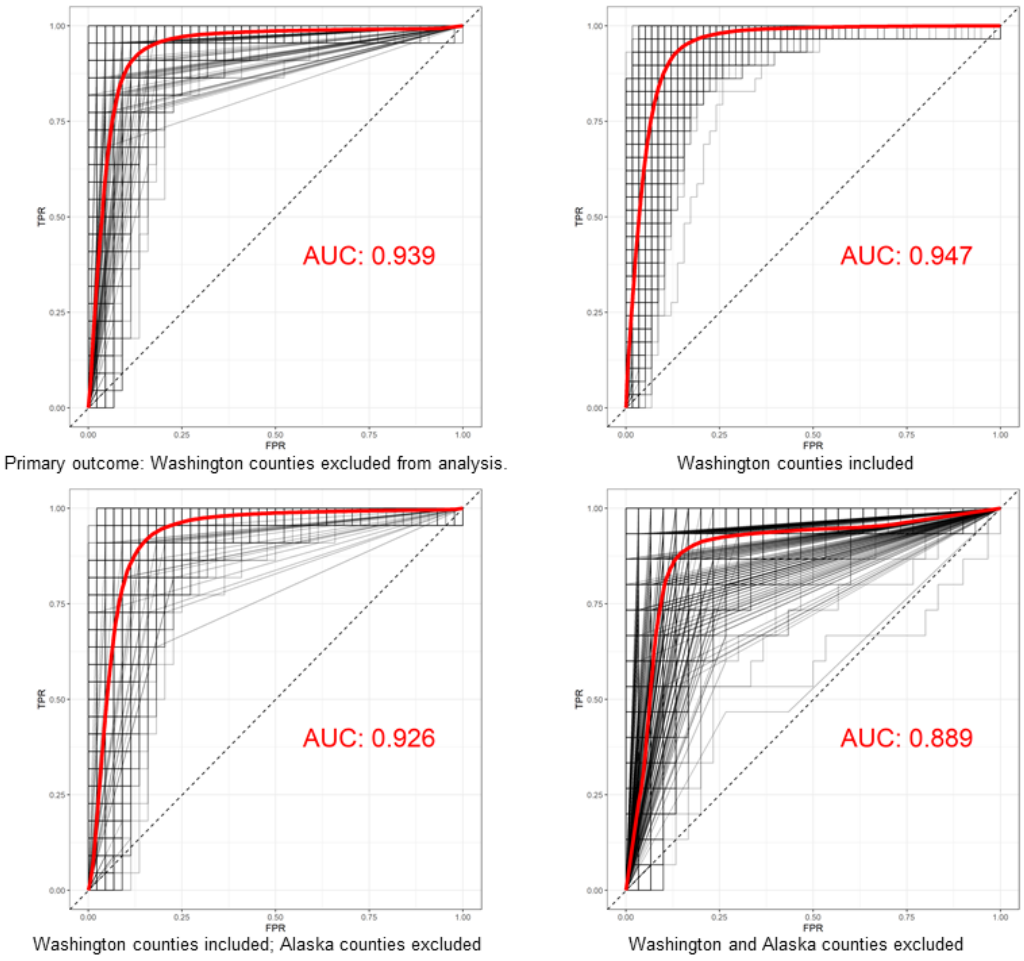
Figure 7. Actual cannabis policies by county, 2014 compared to predicted policy outcomes.



4.1.4 Sensitivity analyses

As previously explained, the sensitivity analyses included different combinations of coding schemes for Washington and Alaska and testing different cut-offs for county-level predictions. Under these different coding schemes, the area under the curve for the ensemble predicted probabilities varies little (0.89 - 0.95 compared to 0.94) (Figure 8). Including the Washington counties performed best while excluding Alaska and Washington counties performed worst (Figure 8). Additionally, including the Washington counties also produced the greatest specificity and sensitivity at the same cut-offs (Table 4).

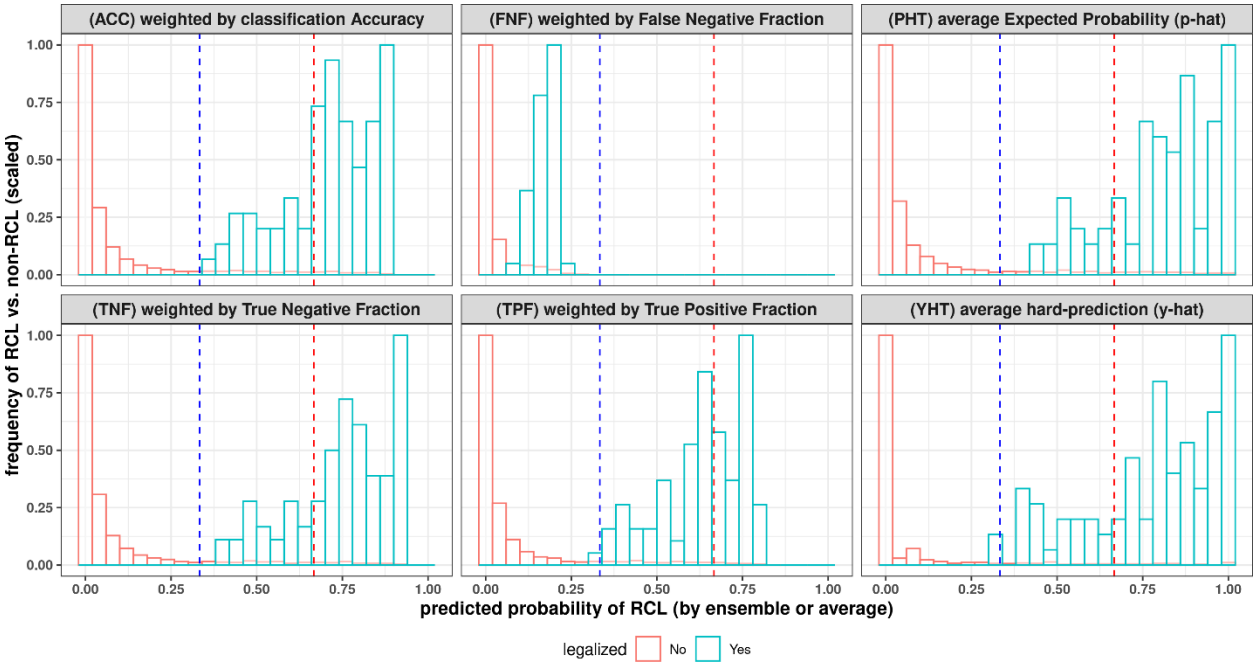
Figure 8. ROC curves of 1000 models and average profiles for each possibility of coding the outcome in sensitivity analyses.



To this point, we have demonstrated the ensemble results of 1000 expected probabilities weighted by each logistic model's classification accuracy. The sensitivity analysis also considers naively averaging the expected probabilities unweighted or averaging the 1000 hard predictions made at a threshold of 0.333 (i.e., the prevalence of RCL in each iteration) instead. We also explored alternative weights derived from quality metrics other than classification accuracy, such as true positive fraction (i.e., sensitivity), true negative fraction (i.e., specificity), and the false negative fraction. Figure 4 shows the frequency of RCL and non-RCL counties (y-axis) against

the probabilities predicted by ensembles of various sorts (x-axis).

Figure 9. Distinguishing power of ensemble predictions (weighted or naïve average).



An effective ensemble prediction should assign distinct probabilities to RCL versus non-RCL counties and separate the predicted probabilities that fall below and above a cut-off that was determined *pre-hoc*. In this case, we chose 0.333 as a liberal cut-off to capture all RCL counties (in favor of sensitivity) and 1 – 0.333 as a conservative cut-off to avoid excessive false positives (in favor of specificity). The default prediction accuracy weighted ensemble (Figure 9, top-left) worked as intended since the vast majority of non-RCL counties (red) were given probabilities lower than the threshold at 0.333 (blue dash). In contrast, most RCL counties (teal) were given probabilities higher than the same threshold, thus providing a reasonable balance between sensitivity and specificity.

If we took the conservative threshold of 0.667=1-0.333 instead (red dash), nearly all non-RCL counites (red) were assigned a probability lower than 0.667 along with a large portion of RCL counties (teal), thus achieving a near 100% specificity at the expense of sensitivity. By the

same standards, the true negative fraction weighted ensemble (TNF, Figure 9 bottom-left), and the ensemble formed by naïvely averaging the 1000 expected probabilities (PHT, Figure 9 top-right) also worked at the threshold of 0.333 (blue dash), with the latter seemingly achieved optimal balance between sensitivity and specificity.

In contrast, the false negative fraction (FNF, Figure 9 top-middle) weighted ensemble was unacceptable due to largely overlapping probabilities assigned to both types of counties and the inability to clearly separate the two groups by the pre-defined threshold at 0.333 (blue dash); as for the conservative threshold of 0.667 (red dash), the 100% specificity could only be achieved by labeling all counties as non-RCL, essentially sacrificing sensitivity entirely. The true positive fraction (TPF, Figure 9 bottom-middle) weighted ensemble and the unweighted average of hard-predictions (YHT, Figure 9 bottom-right) were mediocre because they did not separate the two types of counties at 0.333 (blue dash) as clearly as accuracy weighted ensemble (ACC, Figure 9 top-left).

Finally, table 4 shows the sensitivity and specificity at all cut-offs between .1 and .9 for each type of ensemble where Washington counties were included, and where they were excluded.

Table 4. Sensitivity and Specificity of Models Using Various Weighting Techniques and Hard Cut-off Values.

Excluding Washington Counties										
Weighting Schema										
	PHT		YHT		ACC		TPF		TNF	
Cut-offs	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
0.1	1.00	0.78	1.00	0.86	1.00	0.80	1.00	0.82	1.00	0.79
0.2	1.00	0.86	1.00	0.89	1.00	0.87	1.00	0.88	1.00	0.87
0.3	1.00	0.89	1.00	0.91	1.00	0.90	1.00	0.91	1.00	0.90
0.4	1.00	0.91	0.95	0.93	0.98	0.92	0.93	0.93	0.99	0.92
0.5	0.96	0.93	0.87	0.95	0.88	0.94	0.84	0.95	0.90	0.94

Table 4 (cont'd)

0.6	0.86	0.95	0.79	0.96	0.78	0.96	0.66	0.97	0.83	0.95
0.7	0.76	0.96	0.72	0.97	0.61	0.98	0.34	0.99	0.71	0.97
0.8	0.60	0.98	0.54	0.98	0.29	0.99	0.00	1.00	0.46	0.99
0.9	0.30	0.99	0.33	0.99	0.00	0.97	0.00	0.97	0.20	1.00
	PHT		YHT		ACC		TPF		TNF	
Cut-offs	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
0.1	1.00	0.76	0.99	0.87	1.00	0.78	1.00	0.80	1.00	0.77
0.2	1.00	0.85	0.98	0.90	1.00	0.87	0.99	0.88	1.00	0.86
0.3	0.99	0.89	0.98	0.91	0.98	0.89	0.98	0.90	0.98	0.89
0.4	0.98	0.91	0.93	0.93	0.97	0.92	0.94	0.93	0.98	0.91
0.5	0.96	0.92	0.88	0.95	0.93	0.94	0.82	0.96	0.93	0.93
0.6	0.88	0.95	0.75	0.96	0.75	0.96	0.58	0.98	0.83	0.96
0.7	0.74	0.96	0.64	0.97	0.52	0.98	0.26	0.99	0.63	0.97
0.8	0.51	0.98	0.48	0.98	0.29	0.99	0.00	0.96	0.40	0.99
0.9	0.33	0.99	0.34	0.99	0.00	0.96	0.00	0.96	0.13	1.00

PHT – Unweighted average of expected probability
YHT – Unweighted average of hard predictions with a cut-off value of 0.333
ACC – Weighted by accuracy (used in main results)
TPF – Weighted by true positive fraction (sensitivity)
TNF – Weighted by true negative fraction (specificity)

4.2 Aim 2

4.2.1 Descriptive statistics

This study included 819,543 respondents from the NSDUH surveys between the years 2008 and 2019. The unweighted sample is 48% female, 60% White, 13% Black, 18% Hispanic, 2% Native American, 4% Asian, and 4% of more than one race or another race or ethnicity (Table 4). 11% used cannabis in the past month, and 3% qualified for past-year cannabis abuse or dependence. Table 5 provides the total unweighted sample characteristics as derived from the NSDUH Public Data Analysis System.

Table 5. Characteristics of the U.S. Population Under Study. Data from the U.S. National Surveys on Drug Use and Health.

Gender	%	n
Female	47.8%	322636
Male	52.2%	351885
Race		
White	59.9%	404314
Black	12.8%	86272
Native American	1.5%	10095
Native Hawaiian / Other Pacific Islander	0.5%	3380
Asian	4.1%	27907
More than one race	3.6%	24301
Hispanic	17.5%	118252
Age		
12-17 Years Old	28.1%	189789
18-25 Years Old	29.0%	195650
26-34 Years Old	12.7%	86000
35 or Older	30.1%	203082
Past-month cannabis use prevalence		
Did not use in the past month	88.7%	597984
Used within the past month	11.3%	76537
Past-year cannabis abuse or dependence		
No/Unknown	97.1%	654930
Yes	2.9%	19591
Unweighted Sample Total	100.0%	674521

Figures A.2 – A.6 show various combinations of the past-year cannabis use incidence for those aged 21 and older by state legal category. Upon visual inspection, the parallel lines assumption and assumption of no anticipation look to be met in every comparison of groups. The past-year cannabis use incidence ranges from as low as .25% for the illegal states in years

2008 and 2009 to over 2.5% in the states that legalized cannabis in 2014 (Oregon, Alaska, and the District of Columbia) in years 2018 and 2019.

4.2.2 Event Study Findings

Figures 10 and 11 show the primary findings for individuals aged 21 and older (Figure 10) and those between the ages of 12 and 20 (Figure 11). For those who were legally able to purchase cannabis (21 and older), the legalization of cannabis is estimated to have had no effect on newly incident cannabis use in the year that the legislation passed. However, between two and four years after legalization, RCLs were estimated to have increased incidence by 0.6% [95% Confidence Interval (CI) = 0.1, 1.0]. The corresponding estimate for the interval four to seven years after passage of the RCL is 1.3% [0.8, 1.8] (Figure 10). For the 12-to-20-year-olds, the estimated cannabis incidence does not vary appreciably in any period (Figure 11).

Figure 10. Estimated effect of time since cannabis legalization on cannabis incidence in the 21 and older age group with 95% confidence intervals.

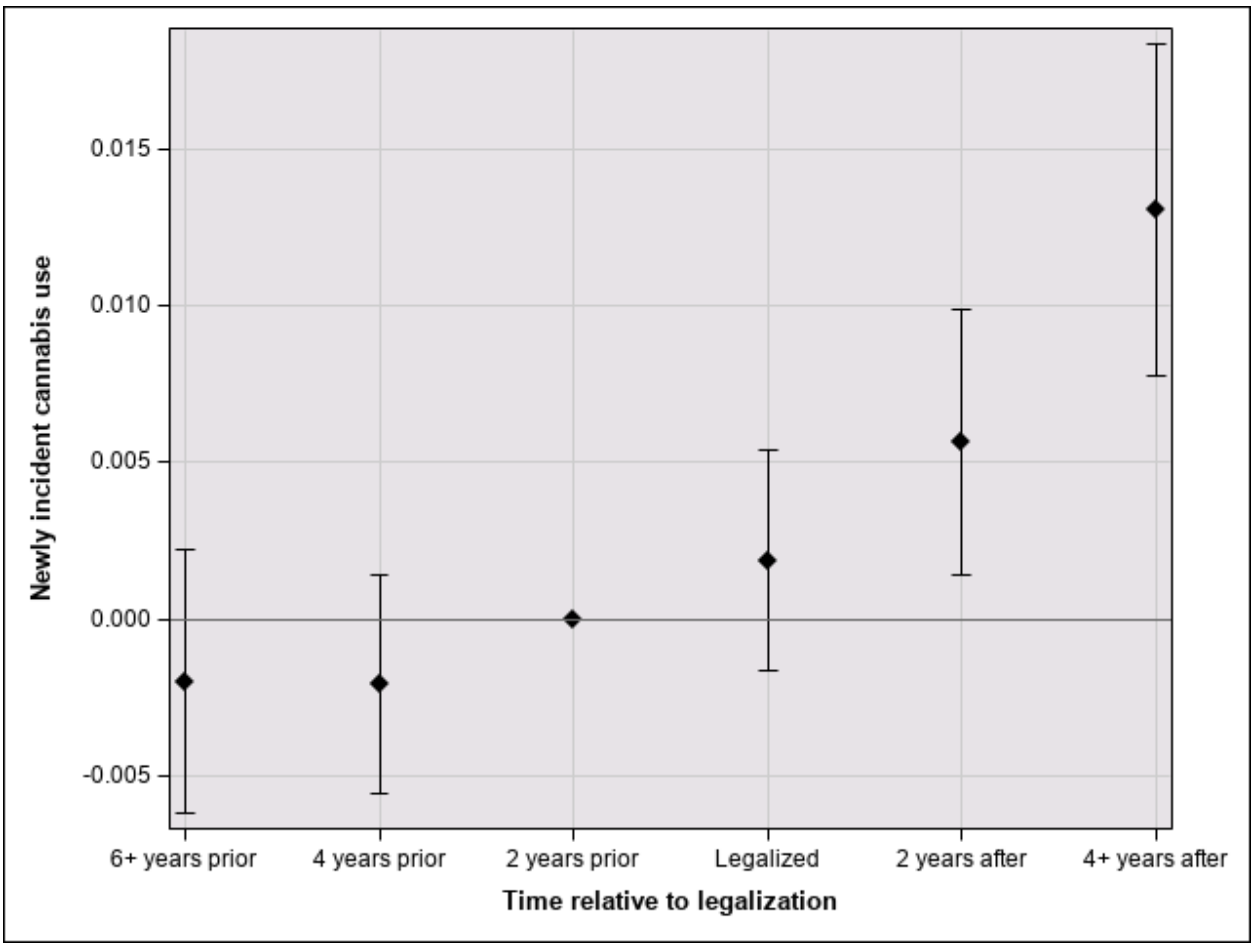
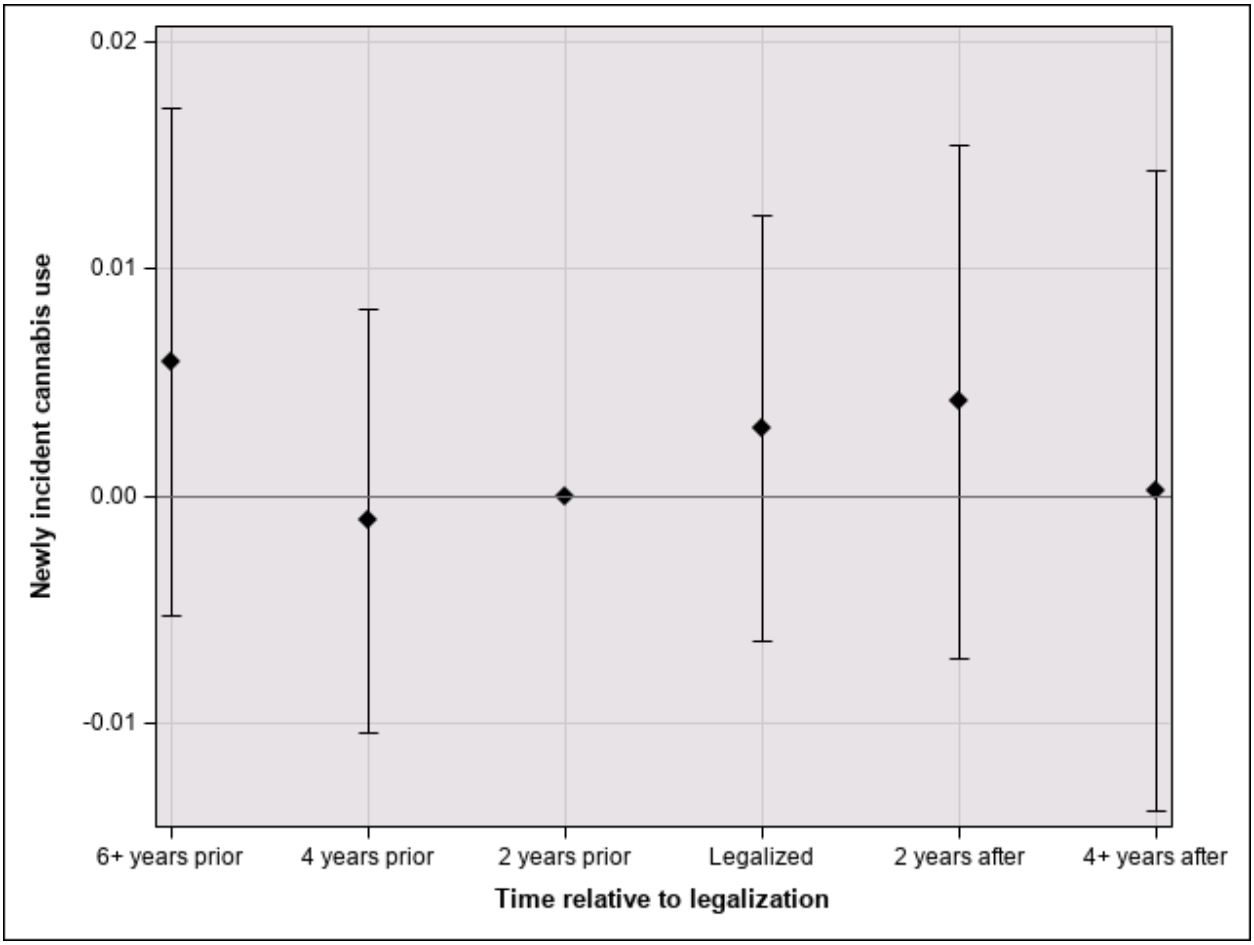


Figure 11. Estimated effect of time since legalization on incidence in those aged 12 to 20 with 95% confidence intervals.



4.2.3 DiD Findings

When including the total time post-legalization, the simple ATT estimate derived from the 2x2 DiD indicates no substantial differences in cannabis incidence before and after the laws were passed (p=0.12). However, since we expected no effect before cannabis sales became effective, we estimated a separate ATT for two years of legalization and later in the 21+ age group as 0.7% (p=0.003, [0.3, 1.1]). The estimated average treatment effects for those aged 12 to 20 years indicated no differences after the legalization date (p=0.27) or the effective date (p=.53).

4.2.4 Alternative Specifications and Robustness Checks

In our first alternate specification, we estimate that the effect of cannabis legalization increased the prevalence of cannabis use in the past month among those aged 21 and older by 3.2% between two and four years after legalization (p=0.0005, [1.6, 4.7]) (Figure S 7)**.** The corresponding estimate for the interval four to seven years after legalization is 4.3% (p=0.0002, [2.3, 6.2]) (Figure S 7). In the 12-to-20-year-old age group, no appreciable variation in estimated cannabis use prevalence is seen across these study intervals (P=0.39 and 0.33, respectively) (Figure S 8).

In the time placebo analysis based upon a randomized legalization date, the date of placebo legalization was set to the year 2011 for all the states that legalized cannabis through 2018. Figure S 9 shows an estimated coefficient that does increase slightly over time, yet the estimated effect of this 'placebo' policy change is null. Note especially that for the adolescents (<21 years), the coefficients are distributed more or less at random in relation to the zero value, with no appreciable differences or patterns (Figure S 10).

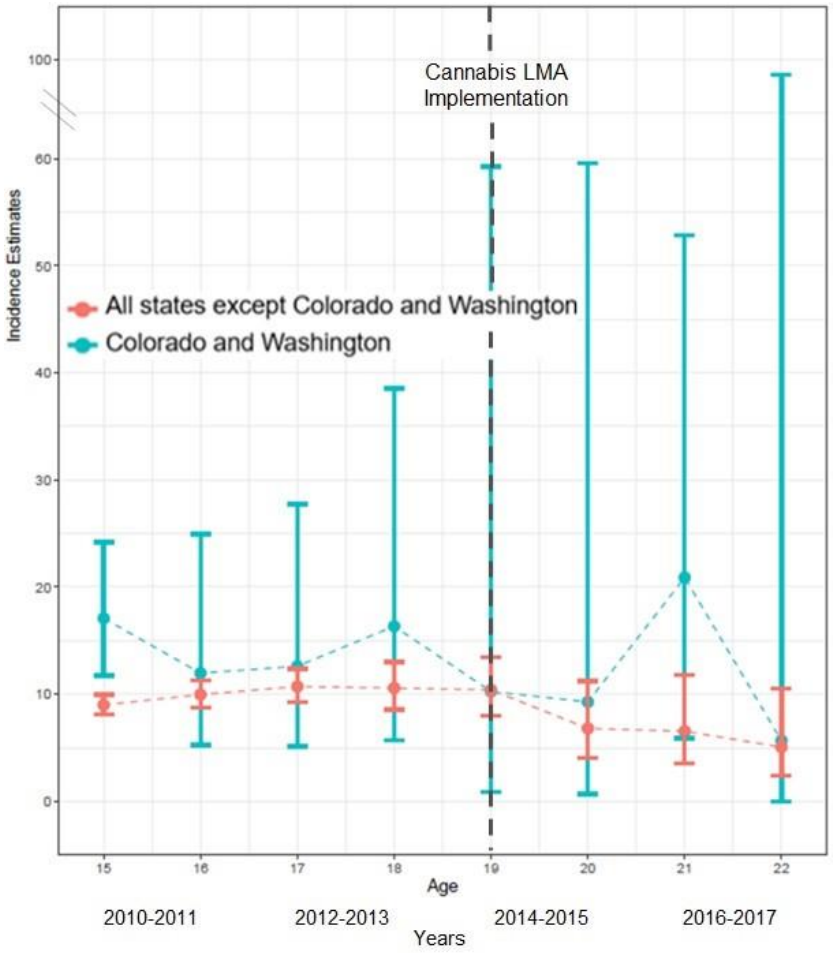
4.3 Aim 3

4.3.1 Panel study approach

Figure 12 shows cannabis incidence estimates based on the panel study approach restricted to the 1995-96 birth cohorts, with state contrasts based on RCL policies. The red lines show a relatively stable incidence by age in the 48 states where cannabis remains illegal for all ages with the highest incidence period occurring as expected between ages 16 and 18. As expected, incidence decreases in each successive year after this peak. In Colorado and Washington, incidence is higher at each age and also shows the expected peak of incidence in

adolescence. At age 21, we have the hypothesized increase in incidence at that age, however this sample does not have sufficient power and the confidence intervals overlap (Figure 12).

Figure 12. Trends in past-year cannabis incidence by age in Colorado and Washington vs. all other states in the US, 2010-2017.

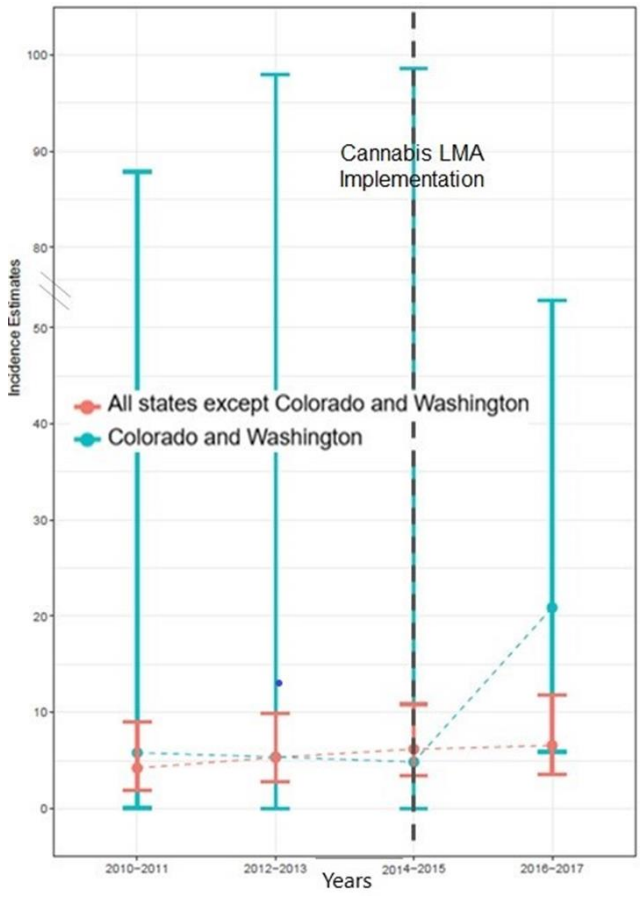


4.3.2 Stratification at age 21

Figure 13 shows RCL-stratified year-pair-specific estimated cannabis incidence, focusing on the NSDUH participants assessed at age 21. The mean cannabis incidence for non-RCL states (red) is relatively stable at about 5% becoming new users. For Colorado and Washington State, the corresponding estimate is close to the estimate for the other states until

after 2014-2015; the estimate for 2016-17 is just above 20%. Here, again, the statistical precision of estimates based on the R-DAS datasets is constrained, however the point estimate has increased fourfold (Figure 13).

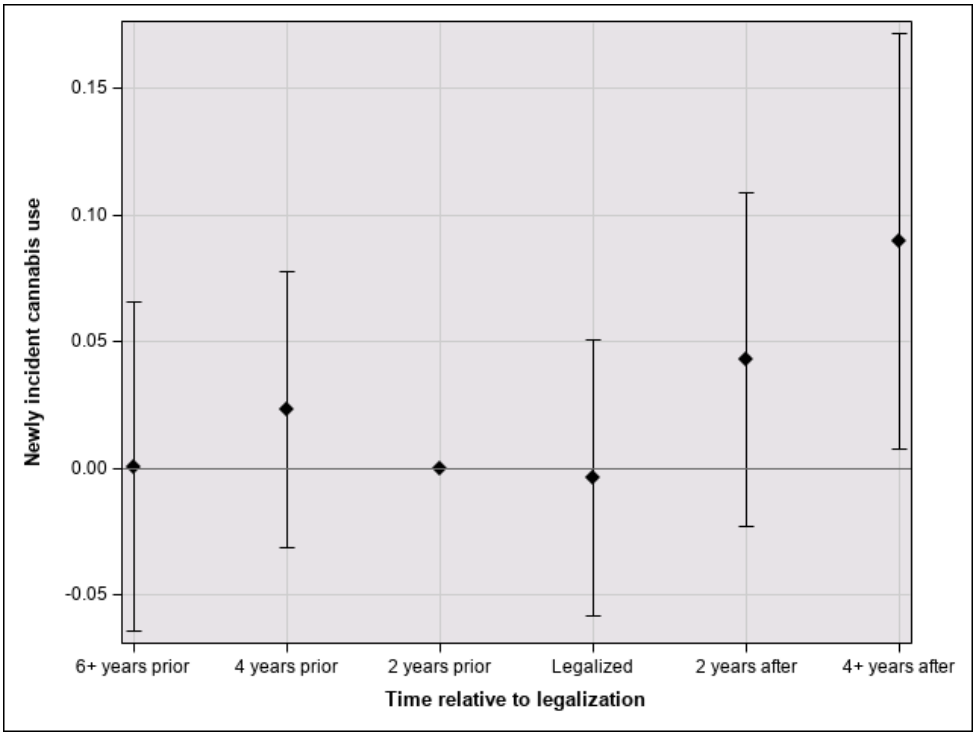
Figure 13. Trends in past-year cannabis incidence for 21 year-olds in Colorado and Washington vs. all other states in the US, 2010-2017.



The striking increase in the raw incidence estimate for 21 year-olds in Colorado and Washington justified further and more sophisticated analysis. Figure 14 shows the estimated effects when adding more states to the analysis, observing the outcome relative to each state's year of RCL, and doing so in the event study framework. The coefficients do not deviate from

zero prior to legalization which supports the assumption of parallel trends that we also see in figure 13. Two years after the passage of the RCLs, incidence increased 4.3% relative to the change in the non-RCL differences at baseline (figure 14). After four years, incidence increased by almost 9% in the RCL states relative to non-RCL state differences at baseline (figure 14). At age 21, the average incidence in non-RCL states was 5.4% and 6.7% in RCL states. Six years after the passage of the RCLs, incidence at age 21 in RCL states increased to 18.1% while it increased to only 6.9% in non-RCL states.

Figure 14. Estimated effect of time since cannabis legalization on cannabis incidence at age 21 with 95% confidence intervals.



5. Discussion

5.1 Aim 1

To my knowledge, this is the first predictive model of any policy change in the United States that uses publicly available data. With the use of similar methods and different machine learning algorithms, the results should be reproducible. While there are no accepted thresholds for policy prediction models, an AUC, sensitivity, specificity, and classification accuracy above 90% would be considered a very successful model in any prediction domain. Further, I find that the model is not sensitive to different outcome classifications as all the coding schemes produced similar results.

This research confirms the importance of prior cannabis use as the single most important predictor of cannabis policy liberalization. It extends this finding from the individual-level, where past cannabis use predicts support for cannabis liberalization, to the county-level, where a higher prevalence of cannabis use is predictive of future policy change. The findings that serious mental illness and cocaine use prevalences were largely predictive have not been reported before. Because I do not have a measure of cannabis use frequency in the model, I speculate that these associations may be indicative of heavier cannabis use in those areas. The significance of the county land area is likely an artifact of early cannabis reform occurring in the west where counties tend to be larger than the east, so is unlikely to have predictive utility in analyses of later time periods. If the definition of the outcome variable were expanded to include the more recent areas that have legalized cannabis in the Northeast, I expect this variable's importance would decrease. However, if it did not, it is not impossible to imagine that it could fit into the causal puzzle. Perhaps having more land facilitates growing cannabis, and this distal factor manifests in our legalization outcome.

Although sociodemographics such as age and race are included in the census principal components used as predictors, I cannot conclude what the individual effects of different age and race distributions may be on RCL in this sample. However, I can assume with some confidence that age and race differences by county are controlled for by including the census principal components. Lastly, I found that after controlling for the other covariates in the model, political affiliation was not as important as past drug use in predicting RCL. Furthermore, democratic affiliation was not even positively associated with RCL; rather, it was voting for a third-party candidate that was positively correlated with RCL.

While the prediction algorithm produced a fair number of false positives, savvy readers will note that many of the states in which the false positives appear have legalized cannabis since 2014. The model may be improved with an expanded definition of legalization which would include more exposed counties and produce a more balanced dataset for analysis.

This study has several limitations: 1) Reliance on self-reported mental health and drug use are subject to social desirability bias; however, there is no other nationally representative and publicly available dataset on drug use and mental health from which I could derive sub-state estimates. 2) I was not able to include all variables from the NSDUH that would be theoretically predictive of cannabis liberalization (e.g., perceived harmfulness of cannabis, past criminalization, etc.) as this data were unavailable in the NSDUH small area estimates. 3) Election data in Alaska and mental health and drug use estimates in Massachusetts and Connecticut are not county-specific. I attempted to mitigate the effect of this missing data by imputing state-level averages of these variables for all counties in their respective states. 4) I was unable to match some counties in the North Carolina NSDUH small area estimates with the other sources of data. Because this was a small number of counties, I decided it was best to leave these counties out

rather than impute state-level averages. 5) One small county in Hawaii (population <100) has no voter data. Because this was such a small county, I left it out of this analysis as well. 6) Lastly, county-level policies regarding recreational cannabis can change over time. Little documentation exists on changes in local cannabis policies, let alone on how they may change over time and why. A database of county-level cannabis policies that document changes over time would be of substantial value to this field.

Notwithstanding these limitations, this study shows that the legalization of recreational cannabis sales is a relatively predictable phenomenon using publicly available data that precede the policy change by two to four years. This conclusion was achieved as a result of the several strengths of this work which include: 1) The novelty of the design which could be used to predict other state or county-level policy changes; 2) using only publicly available data which allows the results to be replicated, and 3) having used nationally representative data, I were not limited to comparing counties within states, rather, I were able to predict county-level changes throughout the U.S.

Future research can build on these findings in several ways. Primarily, given the increasing interest in modelling the effects of policy, extensions of this method can be used to establish covariates for use in quasi-experimental methods. The causal interpretation of these quasi-experimental designs is only allowed when the covariate structures between areas can be considered balanced. Specific methods of interest include creating synthetic controls, propensity score matching, and difference-in-differences (Ben-Michael, Feller, & Rothstein, 2021; Imbens & Rubin, 2015; Roth, 2020; Wooldridge, 2021). Second, updating the model with the 2020 census data and more recent voter and NSDUH data may be used to predict the next wave of states and counties which legalize recreational cannabis. In this sample, the prediction

was 90% accurate for events that occurred between two and four years out. A larger sample may be more accurate with longer time horizons. Finally, decomposing the demographic information in the census may prove to be beneficial for the prediction model. It would allow for the replicability of earlier findings that race and age can be highly predictive of cannabis legalization.

5.2 Aim 2

These results show consistent evidence of an increase in past-year cannabis use incidence among those for whom cannabis became legal, but not for those for whom cannabis remained illegal because of their age. In the simple 2x2 DiD model, I estimate an average increase in past-year cannabis use incidence of 0.7 percentage points after recreational cannabis began being legally sold through the year 2019.

To understand the magnitude of these changes, I find it best to compare these changes in annual incidence to the raw incidence rates estimated by the NSDUH (figures S2-S6). Between 2008 and 2019, the overall estimate of past-year cannabis incidence in the 21 and older age group, independent of the state, was estimated to be just 0.5%. Thus, an increase in the incidence of 0.6% is more than double the proportion of new cannabis users in this age group.

This study advances our understanding of the effects of RCLs in several important ways. First, this is the first study of which I am aware that examines the effects of RCLs on cannabis use incidence. Prior studies on the associations between liberalizing cannabis policy and cannabis use epidemiology focused on past-month cannabis use prevalence (Gruber et al., 2016; Reed, 2016; Cerdá et al., 2017; Dilley et al., 2019; Kerr, Lui, & Ye, 2018; Everson et al., 2019; Martins et al., 2021; Reed, 2021; Paschall, García-Ramírez, & Grube, 2021), the prevalence

of daily or frequent users (Everson et al., 2019; Coley et al., 2021; Martins et al., 2021), and prevalence of CUD (Cerdá et al., 2020; Martins et al., 2021). The importance of understanding changes in cannabis use incidence in response to RCLs cannot be overstated. Prevalence of use and dependence syndromes and frequency of use are of great public health importance. Yet, they tell us nothing about whether new users are entering into the population of cannabis users. This study produces robust evidence that the legalization of cannabis increases the number of cannabis users entering the cannabis-using population where it becomes legal to sell cannabis.

Second, this is the first study of which I am aware that has examined the heterogeneity of treatment effects in the years following RCL. This event study DiD design allows for the estimation of effects by years relative to the passage of the RCL and the effective dates of implementation. This method has yielded three important pieces of evidence: 1.) That the effect of cannabis legalization estimated here is dynamic; 2.) that the trends of the estimates are different by age group, and 3.) the estimated effects of legalization have been an increase in the incidence and prevalence among those for whom cannabis has become legal.

Third, the use of a quasi-experimental DiD design allows for the causal interpretation of the estimated effects of RCL. Except for a few studies (Cerdá et al., 2017; Coley et al., 2021), the evidence produced in this literature has relied mostly on controlling for observed variables between the populations. The differences between states and populations that have legalized cannabis and those that have not are so vast that I question whether controlling for any number of observed variables is likely to produce valid estimates. The design I used allows for the control of unobserved variables if a few assumptions are met. Additionally, I have produced evidence and judged that the assumptions of no anticipation and parallel trends hold in this case.

Although the estimates of incidence cannot be compared to the findings of other studies as these are the first of their kind, I used the same method to estimate the effect of RCL on prevalence in the two age groups as has been done many times. The two studies that used quasi-experimental methods to estimate changes in cannabis use prevalence post-RCLs also reported null findings among adolescents (Cerdá et al., 2017; Coley et al., 2021). Although no quasi-experimental methods have been used to estimate changes in cannabis use prevalence in the adult population, the findings on prevalence are in line with those reported by Cerdá et al., Martins et al., and Reed's more recent findings (2020; 2021; 2021). Perhaps more importantly, the findings also support the seemingly conflicting earlier null findings in this age group (Reed, 2016; Kerr, Lui, & Ye, 2018; Smart and Pacula, 2019), and support a narrative that the increases in the use of cannabis (among both new users and existing) in the adult age group only began increasing after a few years when recreational cannabis shops began sales.

The strengths of this work are the robustness of the estimates, the novelty of the design in this space, and the interpretations that the design allow. The estimates of the effects of RCL on cannabis use incidence by age group were robust to both the check of face validity using the same method to estimate past-month prevalence, as well as the alternate specification using a time-placebo analysis. The use of a DiD event study design moves this field forward by allowing for a dynamic estimate of the causal effect of RCL on the outcome of choice. As I have demonstrated, it is not reasonable to assume that the effect of cannabis legalization is homogenous over time, especially not if the period includes the time before cannabis sales began. Therefore, future research on the effects of RCL should allow for effect heterogeneity. Although this is only one study, from which conclusions should not be drawn, this design allows for a visualization of the policy lag effect, about which much has been written (Cheng et al., 2019;

Hall & Weier, 2015). We see that the effect estimates are not linear and are beginning to take a sigmoid shape with the increase in cannabis use incidence and prevalence beginning to plateau, although more data is needed to confirm the trend.

Some limitations of this work include the self-report nature of the data, differing legal statuses of the drug under study, the sensitivity of the findings to different definitions of the study period, and an inability to control for sub-state level RCL. First-time cannabis use between one and twelve months ago was self-reported, leading to potential recall bias. However, the data collection method has been validated in previous methodological studies. The legality of cannabis use is, of course, different between the states in this sample which makes the outcome subject to some differential response bias; however, the assessment was conducted using confidential standardized audio computer-assisted self-interview modules which have been shown to reduce this bias.

The limitation regarding the definition of the study period is important, specifically to the estimate of the ATT. When including the two years immediately after legalization (before sales began) in the treatment period, the estimated effect is small. However, using a study design that allows for dynamic treatment effects and having estimates that are robust to alternate specifications may make this more of a feature of the study than a limitation. Given that the time-specific estimate of the causal effect of RCL in this two-year window is not appreciably different, these two data points combine to form a strong argument that the effect of cannabis legalization is driven by the opening of outlets where recreational cannabis is sold.

Another limitation of this work at the state level is that many counties and municipalities within states that have legalized recreational cannabis have chosen to ban the sale or cultivation

of cannabis within that sub-state area. For example, in Washington State, 15% of counties and 55% of municipalities have prohibited the sale of cannabis (MRSC, 2019) while in California, 69% of counties and 70% of cities prohibit the sale (Staggs et al., 2019). Similar to the null finding between the date of legalization and effective dates of cannabis sales, it is likely that the estimate of the effect of RCL at the state level is diminished by incorporating incidence for many individuals who reside in areas where recreational cannabis is effectively in this pre-implementation state. This sub-group heterogeneity is averaged out in the state-level estimates. While a county-specific analysis is beyond the scope of this study, future research should seek to replicate this analysis at the county level.

This study adds to the significant literature on the public health puzzle that is the effects of legalizing cannabis in the U.S. by introducing the first estimates of first-time cannabis use and using a quasi-experimental approach that allows for estimation of time-varying treatment effects. The strengths of the study include the large, nationally representative samples spanning eleven years, all U.S. states and Washington D.C., and both age groups of interest to the policy under study, a survey design that allows for state-specific estimates, and state of the art statistical methods from the recent econometrics literature on causal inference with staggered policy implementation. The legalization of cannabis has proven to be a dividing and contentious issue in the national political landscape with different risks and benefits to all paths forward. Thus, voters need the best available information on what the potential effects of legalizing cannabis might be to weigh the empirical evidence with their personal values to make the best decision for themselves and their neighbors. Likewise, policymakers and public health officials can use this evidence to plan for the changes that are likely to occur if cannabis is to be legalized in their districts.

5.3 Aim 3

In this aim, I built from Cheng and colleagues' recent alcohol LMA hypothesis and present evidence that a corresponding pattern can be seen for cannabis when it became legal at the same age (2018). Given the constraints on the statistical precision of the cannabis incidence estimates presented in Figures 12 and 13, the event study estimates that used all available data is a much more convincing portrayal of the hypothesis (Figure 14). Taken together, these analyses provide some early indication that legalizing a drug with an accompanying legal minimum age creates a predictable spike in incidence at the specified age. My hope is that, in turn, these epidemiological estimates of age-specific patterns can be used to guide the organization and deployment of public health tactics of early outreach and intervention, as well as prevention initiatives intended to reduce hazards of drug use onsets during adolescence and the transition to early adulthood.

Limitations of the research include reliance upon self-reports about age and timing of cannabis onsets as well as uncontrolled confounding between states in the unadjusted analyses. Though it does not explicitly control for any potentially confounding factors, the event study model does provide a more robust point of evidence. Given the large increase at age 21 between the two types of states, the question becomes, what else could possibly explain this increase? Indeed, given the specificity of the outcome and our hypothesis grounded in prior theory, a viable alternate explanation is hard to imagine.

One of my committee members suggested that the evidence from both aims 2 and 3 might be strengthened by additional statistical controls, such as a propensity scoring approach or adding known confounders to the event study models. This work would add even more

robustness to the evidence by taking into account known confounders between the policy conditions. Although the exploration of these potential alternative models was not completed as part of my dissertation, these elaborations constitute lines of potential research that can be undertaken in the future.

Notwithstanding these limitations, these analyses demonstrate the potential for a large shift in long-standing patterns in the age of first use for cannabis in the U.S. Additionally, the study findings are of interest because the hypothesis that LMA may be shaping age-specific drug use incidence has never been tested. The lag time for seeing such policy effects has been estimated to take five to ten years if cannabis follows the experience with alcohol legal minimum age in the U.S. (Cheng *et al*., 2019). Our event study results suggest that four may be sufficient.

If this pattern continues to develop, there are new public health considerations for this age group as well as the design and implementation of cannabis prevention campaigns. Targeted prevention campaigns for alcohol and tobacco use have been one of the larger successes of public health and prevention, partly due to age-specific and appropriately timed targeting (Haggerty & Mrazek, 1994; Nation *et al*., 2003). In a deviation from the traditional perspective that early adolescence is the optimal window for prevention, if incidence at age 21 remains high in the RCL states, public health campaigns that seek to reduce cannabis use may be optimized in separate approaches for the law-ignoring teens who first use cannabis illegally vs. the 21-year-olds who wait until cannabis use is legal for them. Although more research is needed to investigate the theorized policy-induced curve, if a sufficient number of states legalize recreational cannabis, or if a federal legalization occurs, I believe we will likely see age 21 become a common age for first trying cannabis.

6. Summary

This dissertation adds to the literature on the laws that govern recreational cannabis use and its epidemiology in several ways. First, by leveraging publicly available data and machine learning techniques, I successfully predicted which counties would legalize recreational cannabis sales in 2014 with high levels of discrimination and revealed the major driving forces of that local policy change. Second, I produced the first estimates of how recreational cannabis legalization affects first-time cannabis use among both the legal and under-aged sub-groups. Third, I have introduced a quasi-experimental approach that allows for estimation of time-varying treatment effects. Fourth, I established a need for it by showing that the effect coefficients varied over time. Finally, I proposed and tested the hypothesis that, after legalization, drug use incidence increases at the legal minimum age.

Taken together, the work and results of this dissertation suggest that:

1. Cannabis legalization is a predictable process driven mainly by prior cannabis use.

2. When implemented, recreational cannabis legalization does not affect adolescents' choice to use cannabis for the first time.

3. However, among those of legal age, the occurrence of newly incident cannabis use may increase by two to four times.

    3.a. I interpret this finding in my hypothesized framework; by legalizing cannabis, a barrier for individuals interested in trying it, but did not for fear of legal consequences, is removed. In epidemiologic terms, this is comparable to disease spreading among a novel population (i.e., the virgin soil hypothesis).

4. The removal of this barrier may lead to an epidemic of first-time cannabis users among those aged 21 and older in areas where recreational cannabis is legally sold.

5. As more states legalize cannabis, the age-specific cannabis use incidence curve might begin to resemble that of alcohol, with a peak in the adolescent years, but a larger peak at age 21.

APPENDICES

Appendix A: Supplemental Figures and Tables

Figure A.1. Percent of variance captured from over 1000 census variables in each principal component.

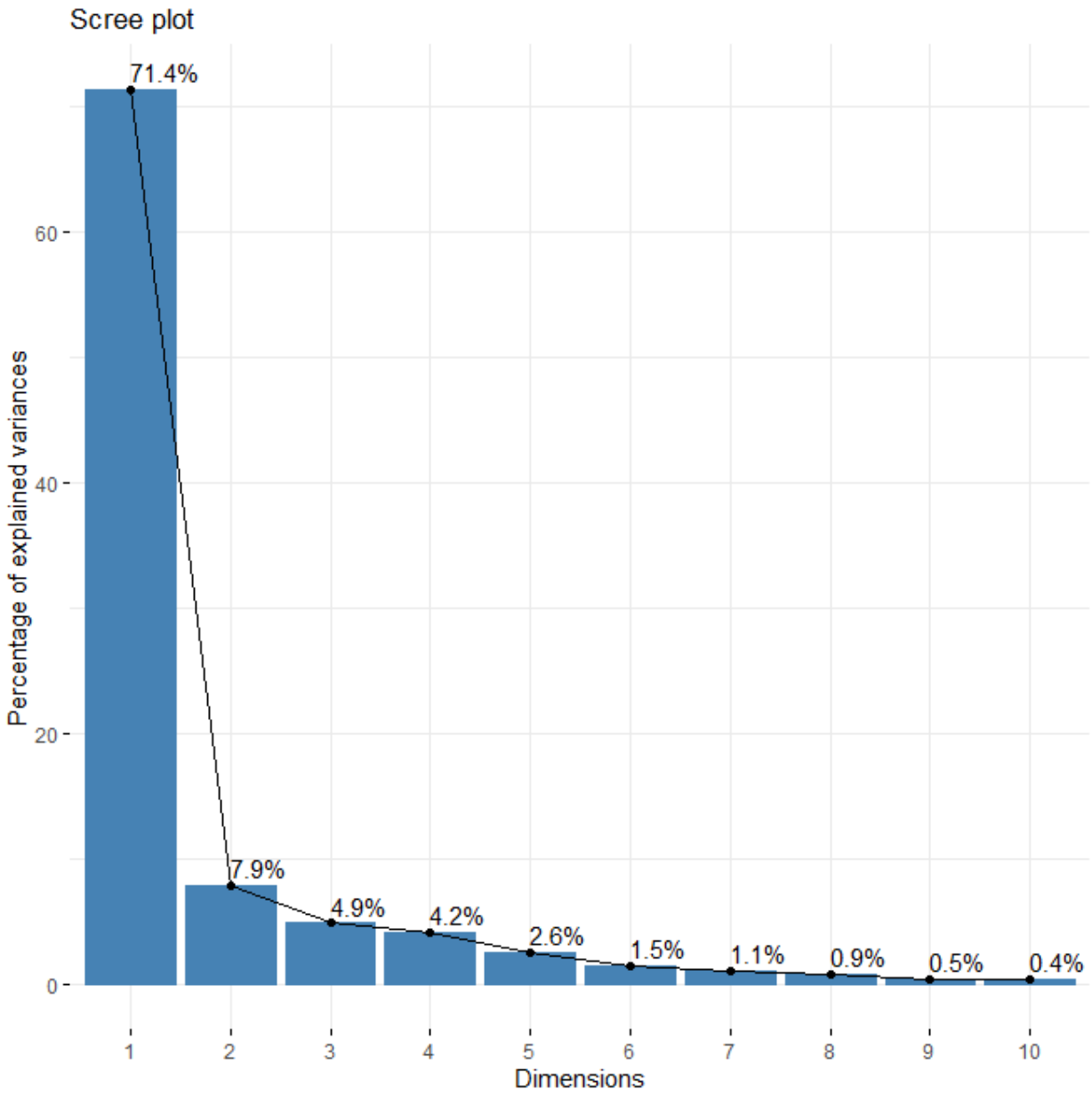


Figure A.2. Cannabis incidence in the 21 and older age group, first wave legalizing states vs untreated states.

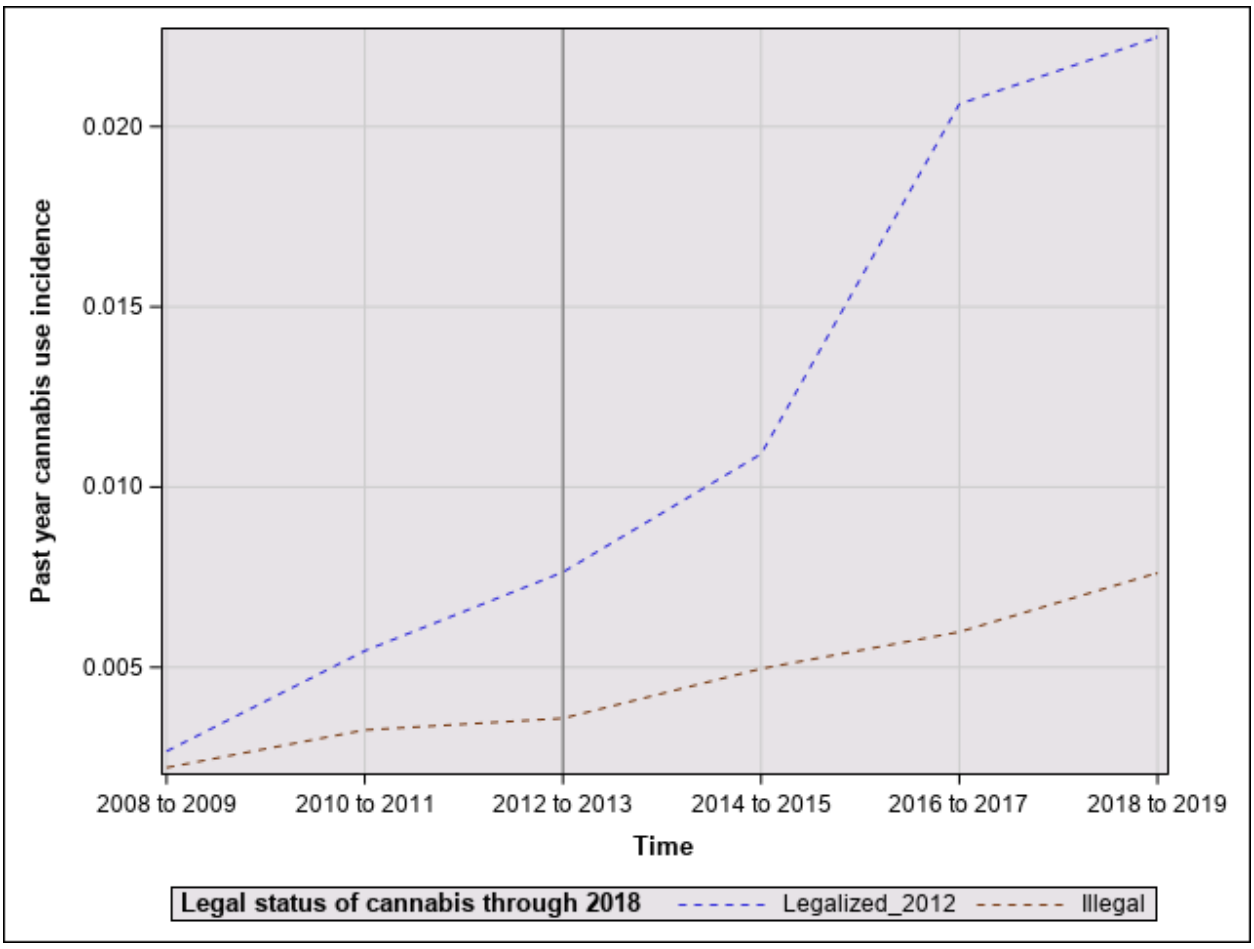


Figure A.3. Cannabis incidence in the 21 and older age group, second wave legalizing states vs untreated states.

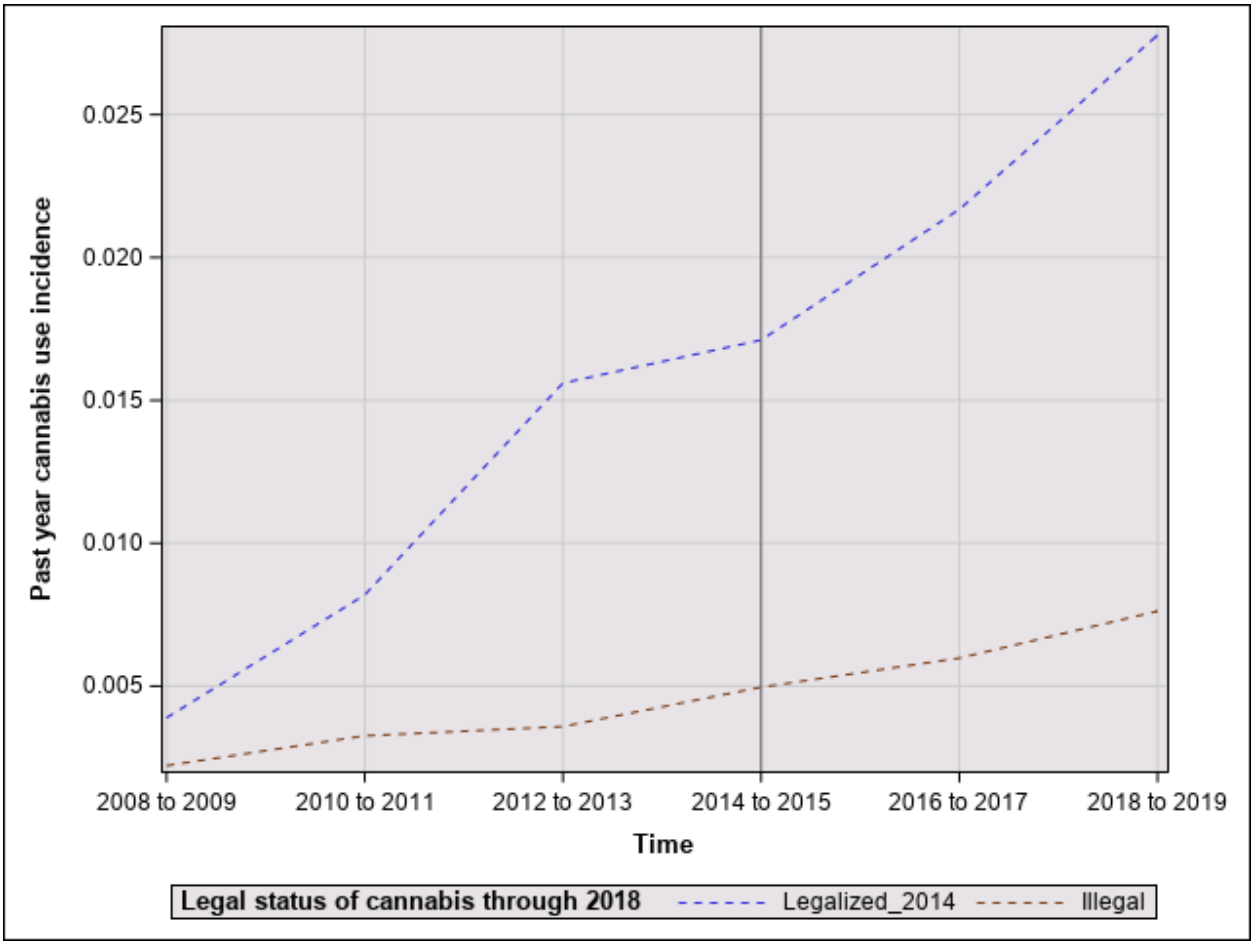


Figure A.4. Cannabis incidence in the 21 and older age group, third wave legalizing states vs untreated states.

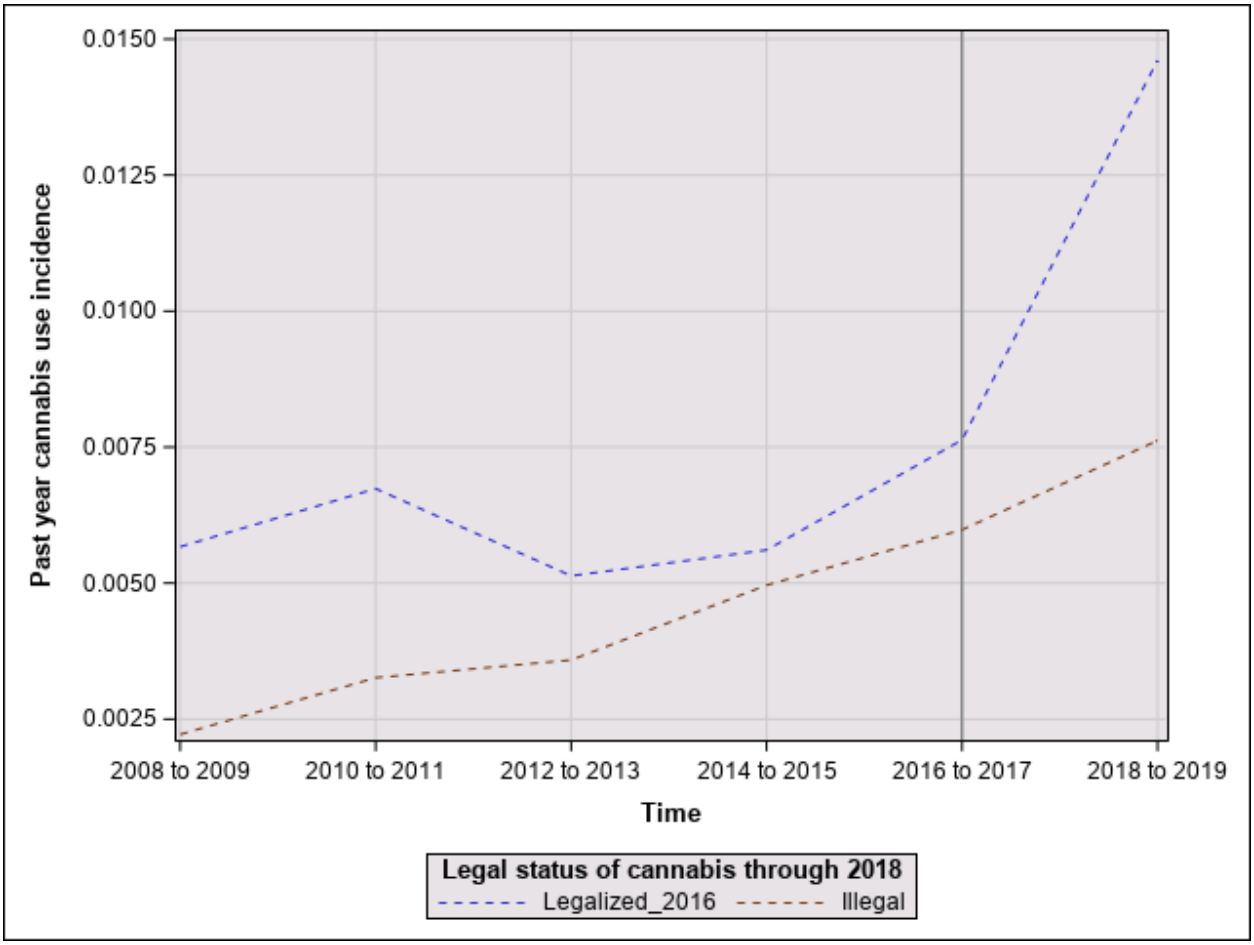


Figure A.5. Cannabis incidence in the 21 and older age group, first wave legalizing states vs third wave legalizing states.

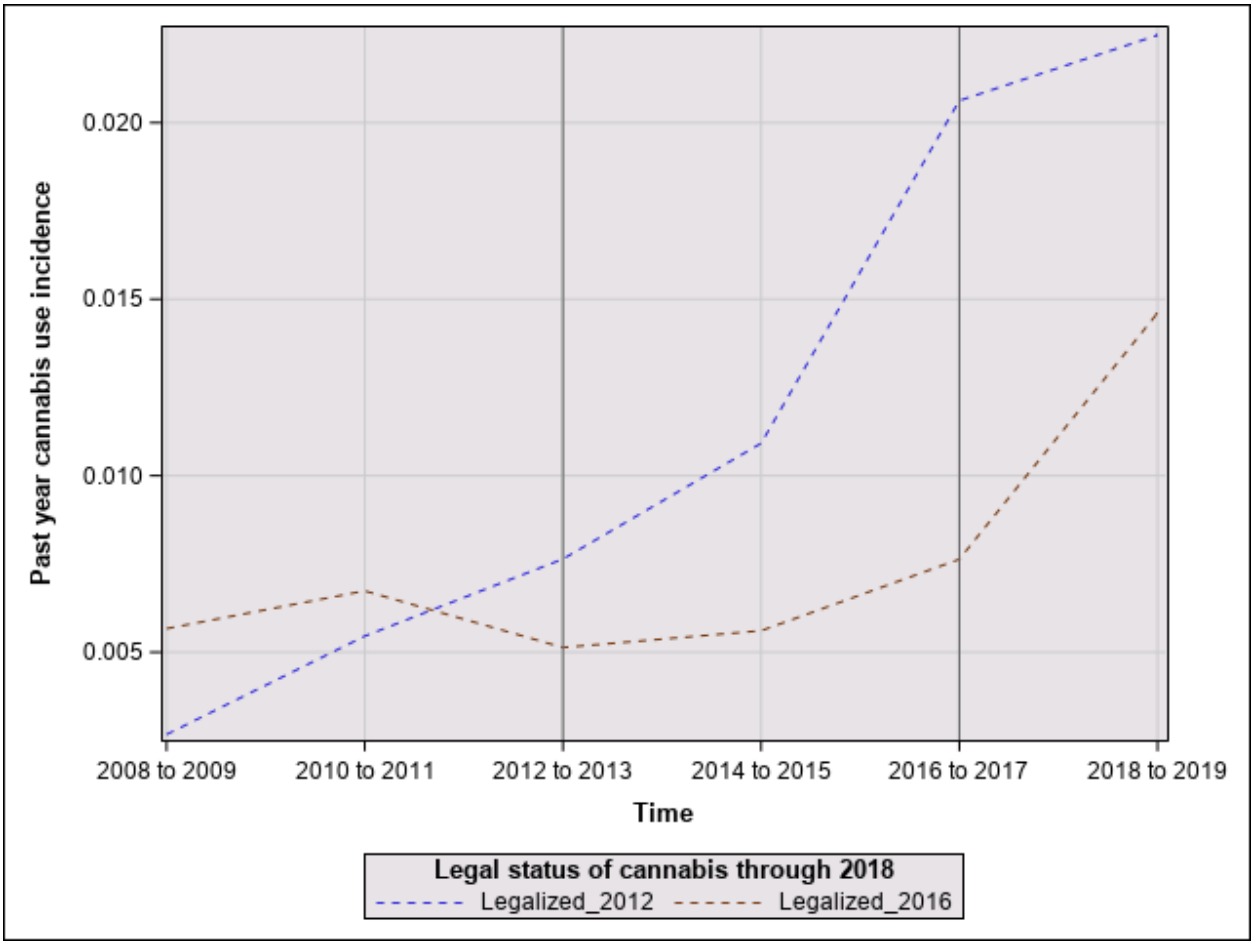


Figure A.6. Cannabis incidence in the 21 and older age group, second wave legalizing states vs third wave legalizing states.

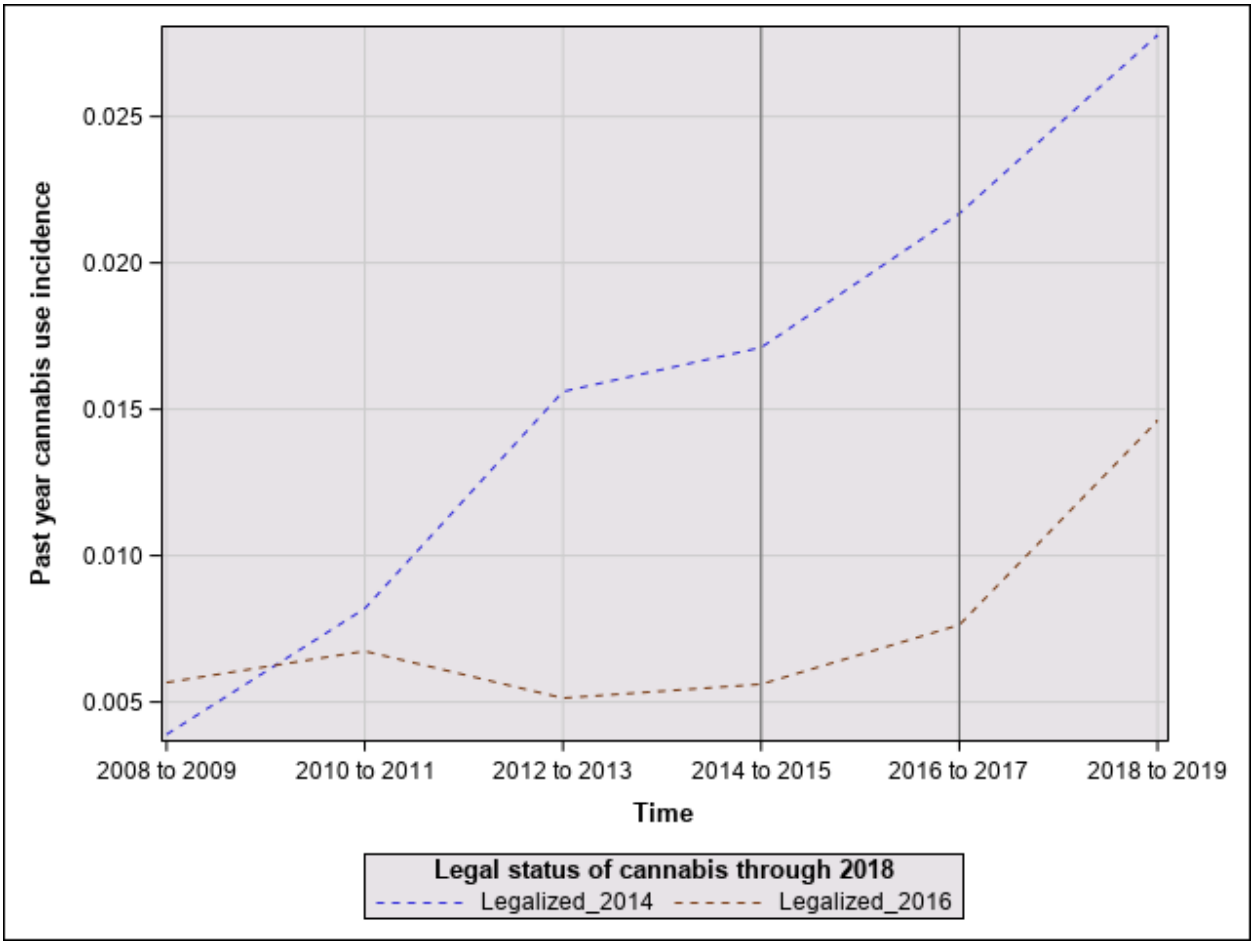


Figure A.7. Estimated effect of time since cannabis legalization on past-month cannabis prevalence in the 21 and older age group.

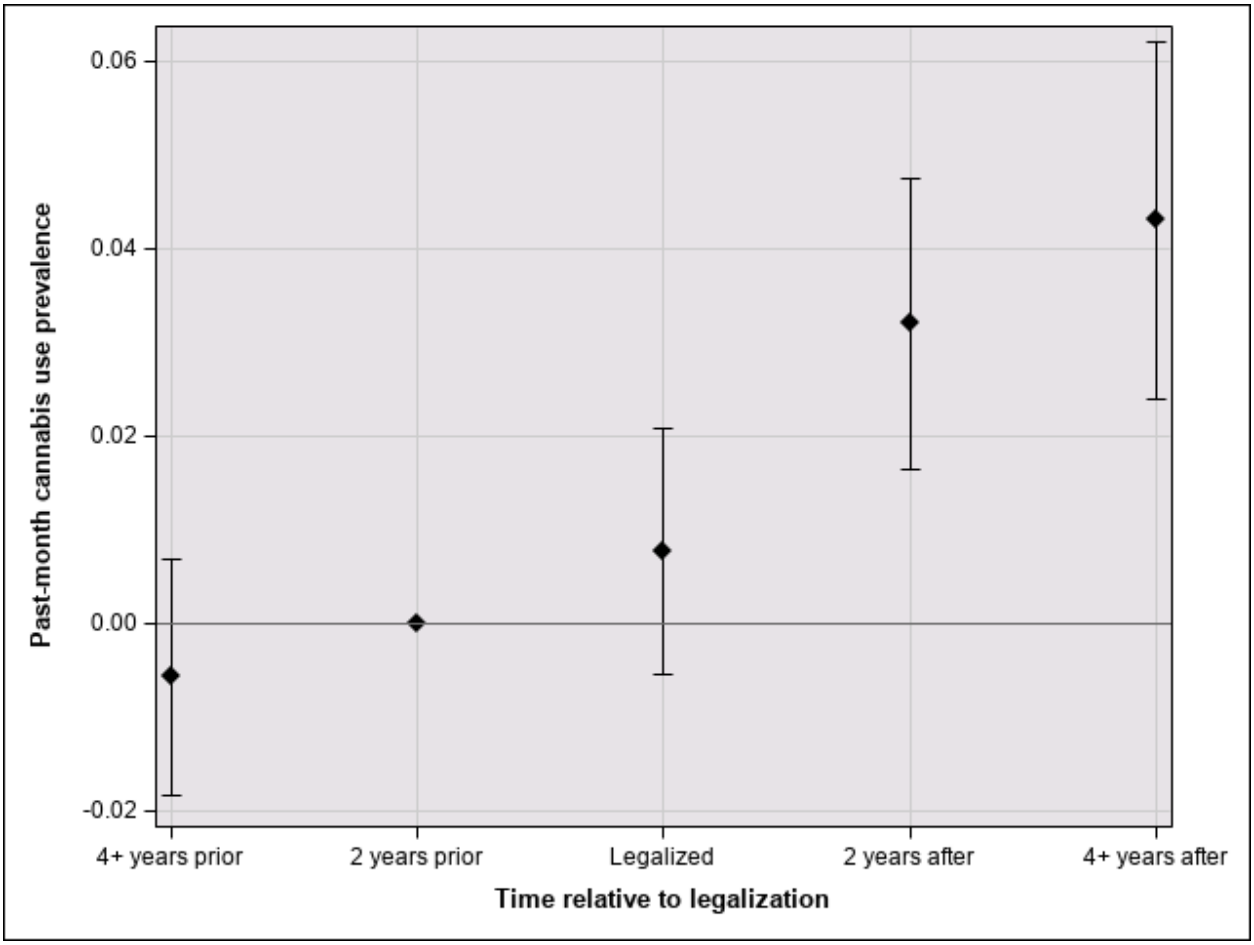


Figure A.8. Estimated effect of time since legalization on past-month cannabis prevalence in the 12 to 20 age group.

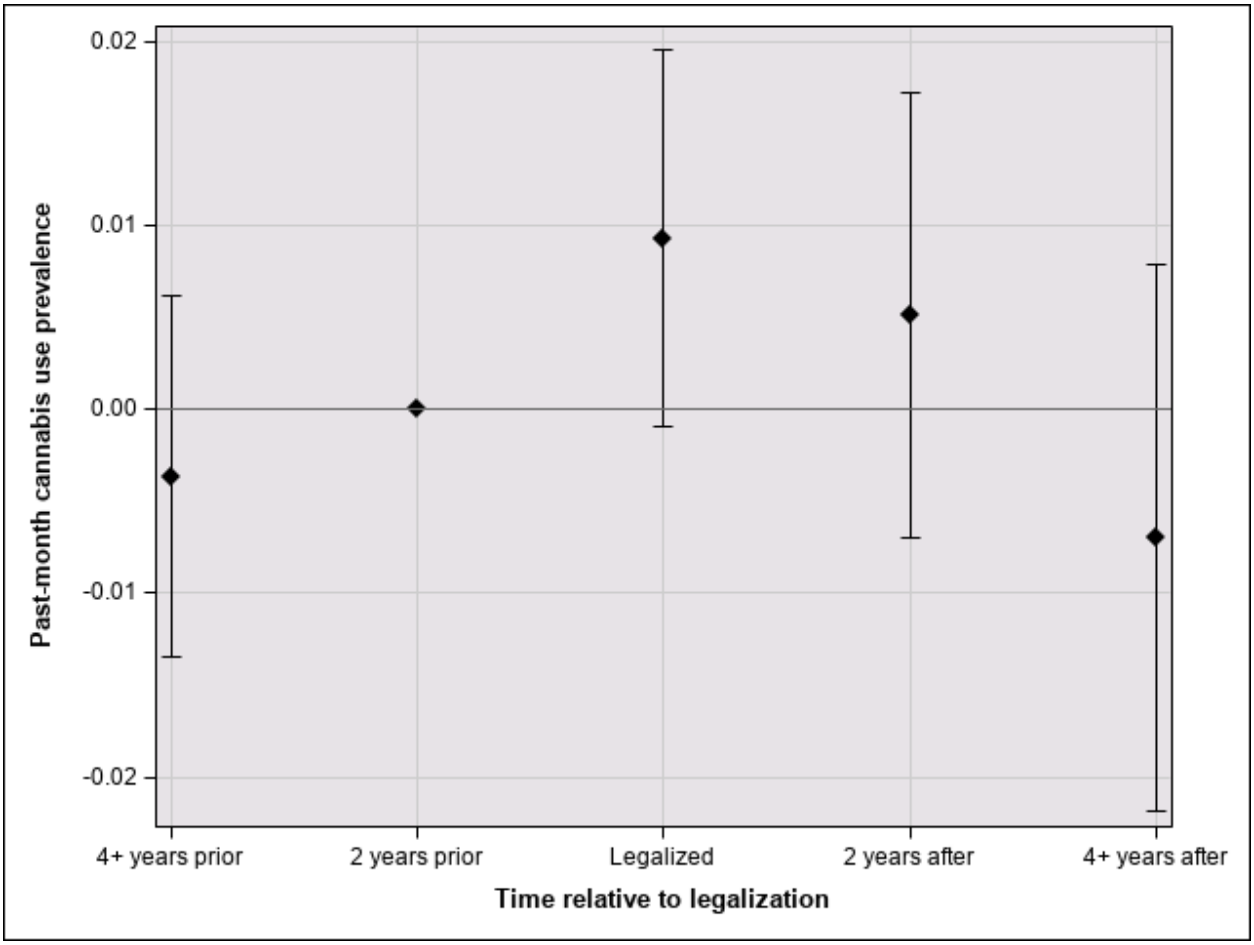


Figure A.9. Estimated placebo effect of time since cannabis legalization on past-year cannabis incidence in the 21 and older age group.

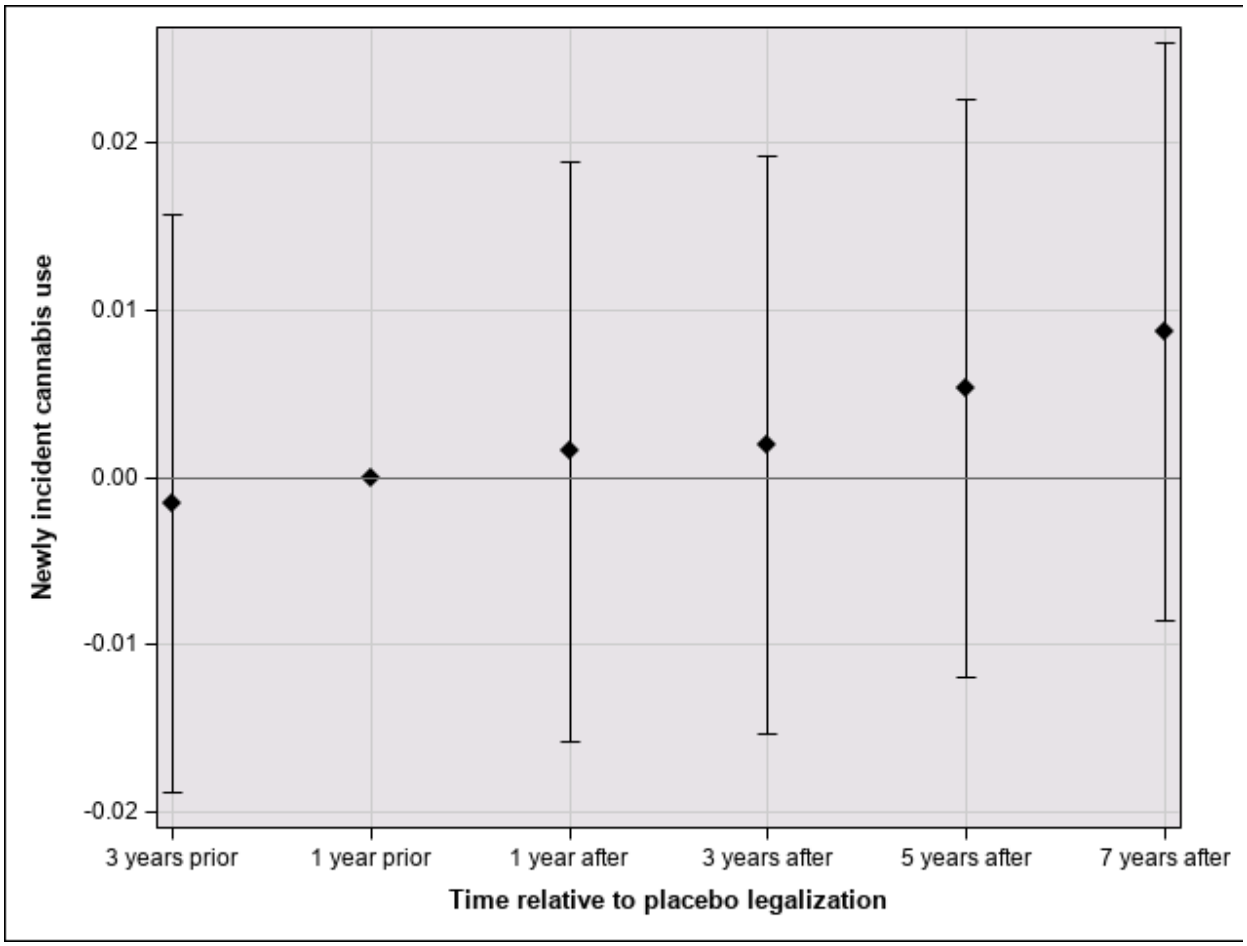


Figure A.10. Estimated placebo effect of time since cannabis legalization on past-year cannabis incidence in the 12 to 20 age group.

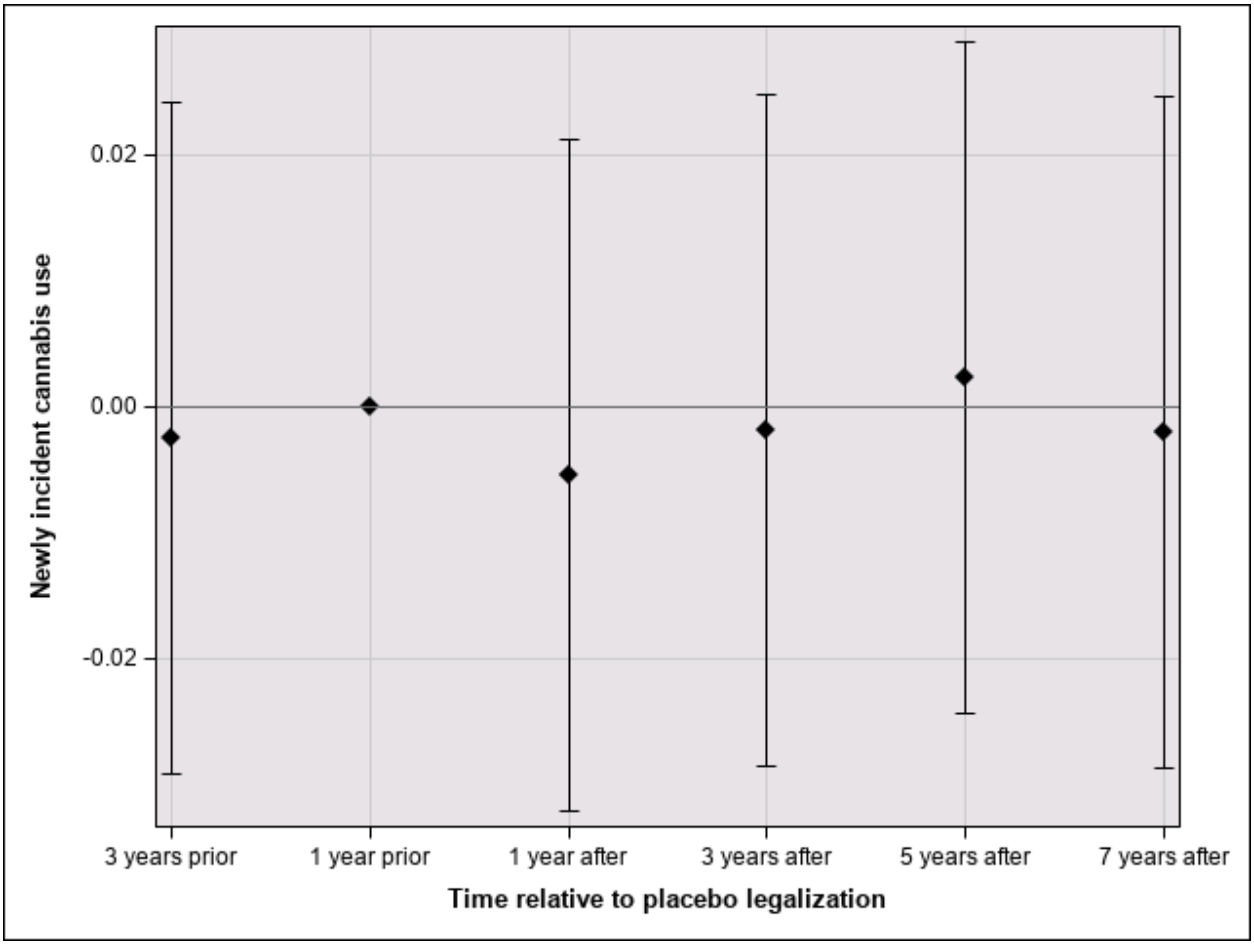


Table A.1. Predictors for legal cannabis sales in 2014 as represented by median z score over 1000 model iterations when proportions of voters are replaced by a binary indicating the party of the majority.

Variables	Median z score
Past month cannabis use prevalence [b]	3.83
Past year cocaine use prevalence [b]	1.34
Past year prevalence of serious mental illness [a]	-0.08
Land area	1.36
County voted majority democratic in 2012 presidential race	1.53
Past month cigarette use prevalence [b]	0.06
Census principal component 2	-0.25
Census principal component 1	-1.04
Past month alcohol use prevalence [b]	0.25
Past year alcohol use disorder prevalence [b]	0.80
Past year substance use disorder prevalence [b]	-0.50
Area water	0.11
Past year prevalence of suicidal thoughts [a]	-0.25
Footnotes [a] Prevalences of mental illnesses for individuals aged 18+ as sampled by the NSDUH [b] Prevalences of substance use for individuals aged 12+ as sampled by the NSDUH	

Appendix B: Program Code Used to Derive the Constructed Study

SAS

```
/*************************************************************************************************

* In Enterprise Guide, "Specify the page size for log and text output" under 'Results
General' must be *

* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *

*************************************************************************************************/

OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE validVarName=any;



title1'Dissertation';

title2'Aim 1: Legalization Prediction';



/*****************************************************************************

* The following macro variables are available to all users: *

*                                                                 *

* &Project     – the name of the project folder in which the .EGP file is stored    *

* &ProgName        – the name of the .egp file, without the extension                 *

* &ProgNode        - the name of the code node                                        *
```

```
* &ProgDir        – the path to the folder in which the .egp file is stored.          *

*******************************************************************************************/



/*************************************************************************

* The following macro variables are used in conjunction with the DOC_BLOCK   *

* macro to document programs and output.                                     *

*                                                                            *

* Do not use quotation marks when defining macro variables. If SAS syntax    *

* requires quotes, use double quotes when you reference the macro variable.  *

*************************************************************************/

** PROGRAMMER'S NAME ;

%LET PROGRAMMER      = Barrett Montgomery;



** LIST ALL SUBDIRECTORIES CALLED IN THE LIBNAME STATEMENT ;

** THESE CAN BE LEFT BLANK IF NOT NEEDED OR USED ;

%LET USEDIR1      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim

1\raw;
```

```
** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;

** USE &PROGNAME FOR SAVEFILE NAME ;

** LEAVE BLANK IF NO DATASET SAVED ;

%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\processed;


** THE NUMBERING SCHEME IS SAVEFILx_y WHERE X IS THE NUMBER OF THE
SAVEDIR AND ;

** Y IS THE NUMBER OF THE FILE WITHIN IT ;

** THIS SHOULD GENERALLY BE SET TO EITHER &PROGNAME OR &PROGNODE ;

%LET SAVEFIL1_1   = ;


** NAME FORMAT LIBRARY DIRECTORY ;

%LET FMTDIR       = ;


** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT
WORK ;
```

```
%LET PURPOSE1 =;



/************************************************************

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

************************************************************/



** INPUT FILES ;

LIBNAME USE "&USEDIR1";



** OUTPUT FILE DESTINATION ;

LIBNAME SAV  "&SAVEDIR1" ;



/*******************************

  ** Required Steps Outline **

1. Import data

2. Check data - note: nsduh county and nsduh tract files must be used in combination

3. Clean and merge data on sub-state area

4. Code sub-state region LRC implementation as of 2014
```

5. Model legalization at county level

**********/

title3"Step 1: Import Data";

proc import datafile= "C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim 1\Raw/countypres_2000_2016.csv" dbms=csv out=USE.election replace ;

GUESSINGROWS=**50000**;

run;

DATA USE.NSDUH_County;

set 'C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim 1\Raw/substate_county141516.sas7bdat';

run;

DATA USE.NSDUH_Tract;

set 'C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim 1\Raw/substate_tract141516.sas7bdat';

run;

```
proc import datafile= "C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Raw\County Merge/counties_geo_merge_2010-2012.xlsx" dbms=xlsx
out=USE.NSDUH_Crosswalk replace ;
run;


/*this macro imports all data in a folder with 2 options, the folder directory, and the type
of file
folder is saved in a macro variable called root and file type is csv.*/
/*%LET Root       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Raw\NSDUH;*/
/**/
/*%macro drive(dir,ext); */
/*   %local cnt filrf rc did memcnt name; */
/*   %let cnt=0;          */
/**/
/*   %let filrf=mydir;    */
/*   %let rc=%sysfunc(filename(filrf,&dir)); */
/*   %let did=%sysfunc(dopen(&filrf));*/
```

```
/*    %if &did ne 0 %then %do;   */

/*   %let memcnt=%sysfunc(dnum(&did));    */

/**/

/*    %do i=1 %to &memcnt;             */

/*                    */

/*     %let name=%qscan(%qsysfunc(dread(&did,&i)),-1,.);                  */

/*                  */

/*     %if %qupcase(%qsysfunc(dread(&did,&i))) ne %qupcase(&name) %then %do;*/

/*      %if %superq(ext) = %superq(name) %then %do;                        */

/*         %let cnt=%eval(&cnt+1);       */

/*         %put %qsysfunc(dread(&did,&i));  */

/*         proc import datafile="&dir\%qsysfunc(dread(&did,&i))" out=dsn&cnt */

/*          dbms=csv replace;*/

/*                              GUESSINGROWS=5000; */

/*         run;         */

/*       %end; */

/*     %end;  */

/**/
```

```
/*    %end;*/

/*      %end;*/

/*  %else %put &dir cannot be open.;*/

/*  %let rc=%sysfunc(dclose(&did));      */

/*            */

/* %mend drive;*/

/* */

/*%drive(&Root,csv) */

/**/

/*%macro combine;*/

/*data USE.NSDUH_SAE;*/

/*        length outcome $110 geography $250;*/

/*        set*/

/*                %do i = 1 %to 15;*/

/*                DSN&i*/

/*                %end;*/

/*        ;*/

/*run;*/
```

```
/*%mend combine;*/


/*%combine;


This data has been processed with standard errors addded, import that version*/


proc import datafile= "C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Raw\NSDUH/Appended_20102012_SAEs_SE.csv" dbms=csv out=USE.NSDUH_SAE
replace ;
        GUESSINGROWS=30000;
run;


title3"Step 2: Check Data";


proc contents data=USE.Census2010;
        title4'census data from 2010';
run;
proc print data=USE.Census2010 (obs=5);
run;
```

```
proc contents data=USE.election;
        title4'election data from 2000-2016';
run;
proc print data=USE.election (obs=5) ;
        where fips='5075';
run;
proc freq data=USE.election;
        table fips/list missing;
run;


proc contents data=USE.NSDUH_County;
        title4'County definitions from NSDUH 2014-2016';
run;
proc print data=USE.NSDUH_County (obs=5);
run;


proc contents data=USE.NSDUH_SAE;
```

```
        title4'NSDUH SAE 2010-2012';
run;
proc print data=USE.NSDUH_SAE (obs=5);
run;


proc freq data=USE.NSDUH_County;
        title4"Check to see if nsduh county names are same in both ds";
        table state*sbst16n / list missing;
        format state state.;
run;


proc freq data=USE.NSDUH_SAE;
        title4"Check to see if nsduh county names are same in both ds";
        table geography / list missing;
run;


/*merge steps
        make fips/county uniform for all datasets
```

Census data has Char var - leading zeroes for 4 digit fips

election data has num var with no leading zeroes for 4 digit fips

nsduh county file has two number vars for state and county

Need to concatenate in the form of SSCCC'

then need to concatenate state and sbst16n to merge with geography from NSDUH SAE

nsduh SAEs have only the region name as character variables

1. election fips needs to become char var with leading 0

2. Create working Crosswalk of NSDUH areas and counties FIPS Codes (FIPS_PCH)

5. Merge census, election, and nsduh crosswalk on FIPS_PCH

5. Merge in NSDUH SAEs on geography*/

```
proc print data=USE.Census2010 (obs=1);
    title4'census county variable';
    var fips;
run;
```

```
proc print data=USE.election (obs=1);
        title4'election county variable';
        var fips;
run;

proc freq data=USE.Census2010;
        title4'check values for state and make sure all align';
        table state/list missing;
run;

proc freq data=USE.election;
        title4'check values for state and make sure all align';
        table state/list missing;
run;

/* transform election fips from num var with no leading zeroes
        to a 5 character var with leading zeroes*/
title3"Step 3: Clean and Merge Data";
```

```
data election1;
        length  fips_pch $ 5 temp_fips 5;
   set USE.election;
        where year in (2008, 2012);
        if fips = "NA" then fips = ".";
   /* read fips as number, then format with leading zeroes */
   /* (pch = suffix for padded character) */
        temp_fips= input(fips, 5.);
   fips_pch = put(temp_fips, z5.);
        candidatevotesx=input(candidatevotes, 8.);


        label fips_pch = "county fips code, padded (leading zeroes) character variable";


        drop candidatevotes;
        rename candidatevotesx=candidatevotes;
run;
```

```
proc freq data=election1;
        title4'check new fips var in election data';
        table state*fips*temp_fips*fips_pch/list missing;
run;

proc contents data=election1;
        title4'Check new fips var in election data has correct format and length';
run;

data election2;
        set election1;
        drop fips temp_fips;

        /*recode kansas city, MO to be in Jackson County*/
        if fips_pch = "36000" then fips_pch = "29095";
run;

proc print data=election2;
```

```
        title4'what to do with Alaska';

        where year = 2008 and state="Alaska" and party='democrat';

run;



/*alaska population in 2010 is 710,231

something is wrong with the way alaska fips are coded



Official fips from Census.gov



AK,02,013,Aleutians East Borough,H1

AK,02,016,Aleutians West Census Area,H5

AK,02,020,Anchorage Municipality,H6

AK,02,050,Bethel Census Area,H5

AK,02,060,Bristol Bay Borough,H1

AK,02,068,Denali Borough,H1

AK,02,070,Dillingham Census Area,H5

AK,02,090,Fairbanks North Star Borough,H1

AK,02,100,Haines Borough,H1
```

AK,02,105,Hoonah-Angoon Census Area,H5

AK,02,110,Juneau City and Borough,H6

AK,02,122,Kenai Peninsula Borough,H1

AK,02,130,Ketchikan Gateway Borough,H1

AK,02,150,Kodiak Island Borough,H1

AK,02,164,Lake and Peninsula Borough,H1

AK,02,170,Matanuska-Susitna Borough,H1

AK,02,180,Nome Census Area,H5

AK,02,185,North Slope Borough,H1

AK,02,188,Northwest Arctic Borough,H1

AK,02,195,Petersburg Census Area,H5

AK,02,198,Prince of Wales-Hyder Census Area,H5

AK,02,220,Sitka City and Borough,H6

AK,02,230,Skagway Municipality,H1

AK,02,240,Southeast Fairbanks Census Area,H5

AK,02,261,Valdez-Cordova Census Area,H5

AK,02,270,Wade Hampton Census Area,H5

AK,02,275,Wrangell City and Borough,H1

AK,02,282,Yakutat City and Borough,H1

AK,02,290,Yukon-Koyukuk Census Area,H5

impute means for alaskan voter data*/

```
proc print data=election2 ;
        where state='Alaska';
run;

proc sql;
        title4'calculate average alaskan voter data by year and party';
        create table voter_impute as
        select *, mean(candidatevotes) as vote_imp
        from election2
        where state = "Alaska"
        group by year, party;
run;
```

```
proc means data=election2 mean;
        var candidatevotes;
        class state year party;
        where state = "Alaska";
run;


data election3;
        set election2 voter_impute;
        if state = "Alaska" and vote_imp=. then delete;
        if state = "Alaska" then candidatevotes = vote_imp;
        drop vote_imp;
run;


proc print data=election3 noobs;
        where state = "Alaska";
run;


/*transform party votes data from long to wide*/
```

```
proc sql;
        title4'Create year-party-specific vote vars';
        create table partyvote1 as
        select fips_pch, sum(candidatevotes) as dem_2008_votes
        from election3
        where party="democrat" and year=2008
        group by fips_pch;
quit;


proc sql;
        create table partyvote2 as
        select fips_pch, sum(candidatevotes) as rep_2008_votes
        from election3
        where party="republican" and year=2008
        group by fips_pch;
quit;
```

```
proc sql;
        create table partyvote3 as
        select fips_pch, sum(candidatevotes) as other_2008_votes
        from election3
        where party="NA" and year=2008
        group by fips_pch;
quit;


proc sql;
        title4'Create year-party-specific vote vars';
        create table partyvote4 as
        select fips_pch, sum(candidatevotes) as dem_2012_votes
        from election3
        where party="democrat" and year=2012
        group by fips_pch;
quit;


proc sql;
```

```
        create table partyvote5 as
        select fips_pch, sum(candidatevotes) as rep_2012_votes
        from election3
        where party="republican" and year=2012
        group by fips_pch;
quit;

proc sql;
        create table partyvote6 as
        select fips_pch, sum(candidatevotes) as other_2012_votes
        from election3
        where party="NA" and year=2012
        group by fips_pch;
quit;

data partyyearvotes;
        merge partyvote1 partyvote2 partyvote3 partyvote4 partyvote5 partyvote6;
run;
```

```
proc surveyselect data=partyyearvotes n=5 out=SampleRep seed=2;;
        title4'Check summarization worked as intended';
run;
proc print data=partyyearvotes noobs;
        var fips_pch other: dem: rep:;
   where fips_pch in ("06087", "13173","26147", "37013", "45079");
run;


proc means data=election3 sum;
   class fips_pch party year;
   var candidatevotes;
   where fips_pch in ("06087", "13173","26147", "37013", "45079");
run;


data election4;
        set partyyearvotes;
```

```
        total_2008_votes = sum(other_2008_votes, dem_2008_votes, rep_2008_votes);
        total_2012_votes = sum(other_2012_votes, dem_2012_votes, rep_2012_votes) ;
        percentdem_2008_votes = dem_2008_votes/total_2008_votes;
        percentrep_2008_votes = rep_2008_votes/total_2008_votes;
        percentother_2008_votes = other_2008_votes/total_2008_votes;
        percentdem_2012_votes = dem_2012_votes/total_2012_votes;
        percentrep_2012_votes = rep_2012_votes/total_2012_votes;
        percentother_2012_votes = other_2012_votes/total_2012_votes;

        label total_2008_votes = 'total votes in district in 2008'
                total_2012_votes = 'total votes in district in 2012'
                percent_2008_votes ='percent of total votes in district for each party in
2008'
                percent_2012_votes ='percent of total votes in district for each party in
2012';
run;

proc print data=election4;
        title4'check percent voter and total votes variables';
```

```
   where fips_pch in ("06087", "13173","26147", "37013", "45079");
run;

/*successfully made all election data wide format, use election4 and merge by
fips_pch*/

data census1;
        length fips_pch $ 5;
        set      USE.Census2010;
        fips_pch = fips;
        rename state=state_fips;
run;

proc freq data=census1 (obs=50);
        title4'these need to be made numeric';
        table arealand * areawatr /list missing;
run;
```

```
proc freq data=census1;
        title4'drop these vars?';
        table sumlev division geocomp /list missing;
run;
/*keep division to use as a hierarchical variable*/


data census2;
        length state $20;
        set census1;


        if state_fips = '01' then state = "Alabama";
        if state_fips = '02' then state = "Alaska";
        if state_fips = '04' then state = "Arizona";
        if state_fips = '05' then state = "Arkansas";
        if state_fips = '06' then state = "California";
        if state_fips = '08' then state = "Colorado";
        if state_fips = '09' then state = "Connecticut";
        if state_fips = '10' then state = "Delaware";
```

```
if state_fips = '11' then state = "DC";

if state_fips = '12' then state = "Florida";

if state_fips = '13' then state = "Georgia";

if state_fips = '15' then state = "Hawaii";

if state_fips = '16' then state = "Idaho";

if state_fips = '17' then state = "Illinois";

if state_fips = '18' then state = "Indiana";

if state_fips = '19' then state = "Iowa";

if state_fips = '20' then state = "Kansas";

if state_fips = '21' then state = "Kentucky";

if state_fips = '22' then state = "Louisiana";

if state_fips = '23' then state = "Maine";

if state_fips = '24' then state = "Maryland";

if state_fips = '25' then state = "Massachusetts";

if state_fips = '26' then state = "Michigan";

if state_fips = '27' then state = "Minnesota";

if state_fips = '28' then state = "Mississippi";

if state_fips = '29' then state = "Missouri";
```

```
if state_fips = '30' then state = "Montana";

if state_fips = '31' then state = "Nebraska";

if state_fips = '32' then state = "Nevada";

if state_fips = '33' then state = "New Hampshire";

if state_fips = '34' then state = "New Jersey";

if state_fips = '35' then state = "New Mexico";

if state_fips = '36' then state = "New York";

if state_fips = '37' then state = "North Carolina";

if state_fips = '38' then state = "North Dakota";

if state_fips = '39' then state = "Ohio";

if state_fips = '40' then state = "Oklahoma";

if state_fips = '41' then state = "Oregon";

if state_fips = '42' then state = "Pennsylvania";

if state_fips = '44' then state = "Rhode Island";

if state_fips = '45' then state = "South Carolina";

if state_fips = '46' then state = "South Dakota";

if state_fips = '47' then state = "Tennessee";

if state_fips = '48' then state = "Texas";
```

```
if state_fips = '49' then state = "Utah";

if state_fips = '50' then state = "Vermont";

if state_fips = '51' then state = "Virginia";

if state_fips = '53' then state = "Washington";

if state_fips = '54' then state = "West Virginia";

if state_fips = '55' then state = "Wisconsin";

if state_fips = '56' then state = "Wyoming";

if state_fips = '60' then state = "American Samoa";

if state_fips = '66' then state = "Guam";

if state_fips = '69' then state = "Northern Mariana Islands";

if state_fips = '72' then state = "Puerto Rico";

if state_fips = '78' then state = "Virgin Islands";



area_water=input(areawatr, 15.);

area_land=input(arealand, 15.);

division_n =input(division, 8.);
```

```
        label area_land="the size, in square meters, of the land portions of geographic
entities for which the Census Bureau tabulates and disseminates data"
                area_water = "size, in square meters of inland, coastal, Great Lakes, and
territorial sea water";


        drop fips sumlev geocomp;
run;


proc freq data=census2;
        title4'check state fips and state variable creation';
        table state_fips * state/list missing;
run;


proc freq data=census2;
        title4'check state fips and state variable creation';
        table state_fips * county * fips_pch/list missing;
run;


proc print data=census2 (obs=50);
```

```
        title4'check area variable creation';

        var areawatr area_water arealand area_land division division_n;

run;



proc means data=census2;

        var area_water area_land;

run;



/*check duplication of county and state var for which one to keep*/



proc freq data=census2;

        title4'check duplication of county and state var for which one to keep';

        table state county/list missing;

run;



proc sort data=census2;

        by fips_pch;

run;
```

```
proc sort data=election4;
        by fips_pch;
run;


data test1;
        merge election4 (in=inelectionx) census2 (in=incensusx drop=arealand areawatr)
;
        by fips_pch;
        inelection=inelectionx;
        incensus=incensusx;

        state_fips=put(fips_pch,2.);
run;


proc contents data=test1;
        title4'Check merge of census and election data';
run;
```

```
proc freq data= test1;
        table incensus*inelection/list missing;
run;


proc freq data=test1;
        title4'Check fips and state fips';
        table fips_pch * state_fips/list missing;
run;


proc print data= test1 noobs n;
        title4'check discrepancies after merge: not in census data';
        where incensus=0;
run;
/*these are the connecticut/maine/rhode island non-counties that have voter data, can
be deleted


need to investigate why Alaska and Kansas City are specifcially not merging
```

```
kansas city exists in several counties, recoded to jackson county, where most of
popultion lies, in election 2 datastep*/


proc print data= test1 noobs n;
        title4'check discrepancies after merge: not in election data';
        var fips_pch county state:;
        where inelection=0;
run;
/*      alaska here probably the data not merging correctly
        look into county in hawaii
        Puerto rico can be deleted*/


proc print data= test1 noobs n;
        title4'Hawaii county not in election data';
        where inelection=0 and state="Hawaii";
run;
/*tiny village with less than 100 people, leave as is*/
```

```
data test2;
        set test1;
/*              delete the weird connecticut/maine/rhode island non-counties,  delete
Puerto Rico Census data*/
        if compress(fips_pch)= "." then delete;
        if state ="Puerto Rico" then delete;
run;


proc print data= test2;
        title4'Check test 2 fixes worked (should not print)';
        where state="Puerto Rico" or compress(fips_pch)= "." ;
run;


proc freq data= test2;
        title4'Check mismatches after test 2 fixes';
        table incensus*inelection/list missing;
run;
```

```
proc print data= test2;
        title4'Check mismatches after test 2 fixes';
        where incensus=0 or inelection=0;
run;


proc print data= test2;
        title4'Check Alsakan data uses averages';
        where state="Alaska";
run;


data test3;
        /* remove bad alaskan data but impute average voter data
                set in election for these to be 1*/
        set test2;
        where incensus=1;
        if state="Alaska" then do;
                dem_2008_votes                  = 3089.85;
                rep_2008_votes                  = 4846.03;
```

```
                other_2008_votes        = 219.05;
                dem_2012_votes          = 3066;
                rep_2012_votes          = 4116.9;
                other_2012_votes        = 329.475;
                total_2008_votes        = 8154.93;
                total_2012_votes        = 7512.38;
                percentdem_2008_votes   = 0.37889;
                percentrep_2008_votes   = 0.59425;
                percentother_2008_votes = 0.026861;
                percentdem_2012_votes   = 0.40813;
                percentrep_2012_votes   = 0.54802;
                percentother_2012_votes = 0.043858;
                inelection=1;
                end;
run;

proc print data= test3;
        title4'Check Alsakan data uses averages and correct fips';
```

```
        where state="Alaska";
run;

proc print data= test3;
        title4'last remianing issue';
        where inelection=0 or incensus=0;
run;

/*limitations of voter data: alsaka uses state averages, missing 1 hawaii county*/

/* compared crosswalk and my small area estimates with
https://www.samhsa.gov/data/report/2010-2012-nsduh-substate-region-definitions
changes to crosswalk made to match sae file based on above comparisons and
documentation*/

data nsduh_crosswalk2;
        set use.nsduh_crosswalk;
        where fips_pch ~='';
```

```
if geography = "Arizona Maricopa" then geography= "Arizona Central";

if geography = "Arizona Pima" then geography= "Arizona South A";

if geography = "Florida Circuit 17 (Broward)" then geography= "Florida
Broward (Circuit 17)";

if geography = "Florida Region F - Southern (Circuits 11 and 16)" then
geography= "Florida South (Circuits 11 and 16)";

if geography = "Illinois Region I (Cook)" then geography= "Illinois Region
1 (Cook)";

if geography = "Illinois Region II" then geography= "Illinois Region 2";

if geography = "Illinois Region III" then geography= "Illinois Region 3";

if geography = "Illinois Region IV" then geography= "Illinois Region 4";

if geography = "Illinois Region V" then geography= "Illinois Region 5";

if geography = "Kentucky Seven Counties" then geography= "Kentucky
Centerstone";

if geography = "Michigan Macomb" then geography= "Michigan Region
1";

if geography = "Michigan Oakland" then geography= "Michigan Region
8";

if geography = "Michigan Pathways and Western" then geography=
"Michigan Region 9";
```

```
run;



/* merge NSDUH crosswalk*/

proc sort data= nsduh_crosswalk2;

        by fips_pch;

run;



proc sort data= test3;

        by fips_pch;

run;



data test4;

        merge test3 nsduh_crosswalk2 (in=incrossx);

        by fips_pch;

        where fips_pch ~='';



        incross = incrossx;
```

```
        label    fips_pch = 'concatenated state_fips and county_fips in the form of
SSCCC, mergable with election and census data';
run;


proc freq data=test4;
        title4'nsduh merge check';
        table incross*incensus*inelection/list missing;
run;


proc freq data= test4;
        title4'Check mismatches after test 3 merge: not in election data';
        table qname/list missing;
        where inelection=0;
run;
/*same limitations of census data relating to alaska and hawaii*/


/*begin with nsduh SAEs, these also need to be transformed to one level...*/
```

```
proc sql;
        title4'prepare NSDUH SAEs for transpose and merge';
        select distinct(outcome)
        from  use.nsduh_sae;
quit;


data NSDUH_SAEs;
        set use.nsduh_sae ;
        length short_name $32 age_group2 $32;

/*remove aggregate areas and sub-county areas*/
        where geography not in ("District of Columbia Ward 1","District of Columbia
Ward 2",
                "District of Columbia Ward 3","District of Columbia Ward 4","District of
Columbia Ward 5",
                "District of Columbia Ward 6","District of Columbia Ward 7","District of
Columbia Ward 8",
                "California LA SPA 1 and 5", "California LA SPA 2", "California LA SPA
3",
```

"California LA SPA 4", "California LA SPA 6", "California LA SPA 7",
"California LA SPA 8",

"California Regions 13 and 19R",

"Florida Region A - Northwest", "Florida Region B - Northeast",

"Florida Region C - Central", "Florida Region D - Southeast"

"Florida Region E - Sun Coast", "Hawaii Kauai and Maui",

"Louisiana Regions 1 and 10", "Maine Aroostook/Downeast", "Missouri
Eastern"

"Missouri Northwest", "Nebraska Regions 1 and 2", "Nevada Region 3",

"New Hampshire Central", "New Hampshire Southern", "Texas Region
11",

"Texas Region 3", "Texas Region 6", "Texas Region 7", "Washington
Region 1",

"Washington Region 2", "Washington Region 3");

if outcome = "Alcohol Dependence in the Past Year" then short_name =
"PYAlcDepPrev";

if outcome = "Alcohol Use Disorder in the Past Year" then short_name =
"PYAUDPrev";

if outcome = "Alcohol Use in the Past Month" then short_name = "PMAlcPrev";

```
        if outcome = "Any Mental Illness in the Past Year" then short_name =
"PYAMIPrev";

        if outcome = "Average Annual Rate of First Use of Marijuana" then short_name =
"PYMJInc";

        if outcome = "Cigarette Use in the Past Month" then short_name =
"PMCigPrev";

        if outcome = "Cocaine Use in the Past Year" then short_name = "PYCocPrev";

        if outcome = "Past Year Suicidal Thoughts" then short_name = "PYSTPrev";

        if outcome = "Illicit Drug Dependence in the Past Year" then short_name =
"PYSUDPrev";

        if outcome = "Major Depressive Episode in the Past Year" then short_name =
"PYMDEPrev";

        if outcome = "Marijuana Use in the Past Month" then short_name =
"PMMJPrev";

        if outcome = "Marijuana Use in the Past Year" then short_name = "PYMJPrev";

        if outcome = "Serious Mental Illness in the Past Year" then short_name =
"PYSMIPrev";

        if outcome = "Tobacco Product Use in the Past Month" then short_name =
"PMTobPrev";
```

```
        if outcome = "Underage Alcohol Use in the Past Month" then short_name =
"PMUAlcPrev";



        if age_group = "12 or Older" then age_group2       = "Twelve_plus";

        if age_group = "12 to 17" then age_group2          = "Twelve_to_17" ;

        if age_group = "12 to 20" then age_group2          = "Twelve_to_20" ;

        if age_group = "18 or Older" then age_group2       = "Eighteen_plus" ;

        if age_group = "18 to 25" then age_group2          = "Eighteen_to_25" ;

        if age_group = "26 or Older" then age_group2       = "Twenty_six_plus" ;

run;



proc print data=NSDUH_SAEs;

        title3'interval check (should not print)';

        var geography estimate ci_lower ci_upper;

        where estimate > ci_upper or estimate < ci_lower;

run;



proc freq data=NSDUH_SAEs;
```

```
        title4'Chek to see if new vars worked';
        table outcome*short_name age_group*age_group2/list missing;
run;


proc sort data=NSDUH_SAEs;
        by geography;
run;


/*macro to transpose all variable values to make wide dataset*/
%macro multi_transp (var=, out=);
proc transpose data= NSDUH_SAEs delimiter=_ out=&out(drop=_NAME_) suffix=&var ;
        id age_group2 short_name ;
        by geography ;
        var &var   ;
run;


proc sort data= &out;
        by geography;
```

```
run;

%mend;



%multi_transp(var=estimate, out=NSDUH_SAEs1);

%multi_transp(var=ci_lower, out=NSDUH_SAEs2);

%multi_transp(var=ci_upper, out=NSDUH_SAEs3);

%multi_transp(var=L,                out=NSDUH_SAEs4);

%multi_transp(var=L_lower,          out=NSDUH_SAEs5);

%multi_transp(var=sel,              out=NSDUH_SAEs6);

%multi_transp(var=L_upper,          out=NSDUH_SAEs7);

%multi_transp(var=se,               out=NSDUH_SAEs8);



data sae_merge;

        merge NSDUH_SAEs1 NSDUH_SAEs2 NSDUH_SAEs3 NSDUH_SAEs4
NSDUH_SAEs5 NSDUH_SAEs6 NSDUH_SAEs7 NSDUH_SAEs8;

        by geography;



        label    fips_pch = 'concatenated state_fips and county_fips in the form of
SSCCC, mergable with election and census data';
```

```
run;

/*ensure this merge worked*/

data merge_check;
        set sae_merge;
        a1=Twelve_plus_PYAlcDepPrevestimate<Twelve_plus_PYAlcDepPrevci_lower;
        a2=Twelve_plus_PYAlcDepPrevestimate>Twelve_plus_PYAlcDepPrevci_upper;
        b1=Twelve_plus_PMAlcPrevestimate<Twelve_plus_PMAlcPrevci_lower;
        b2=Twelve_plus_PMAlcPrevestimate>Twelve_plus_PMAlcPrevci_upper;
        c1=Twelve_plus_PMCigPrevestimate<Twelve_plus_PMCigPrevci_lower;
        c2=Twelve_plus_PMCigPrevestimate>Twelve_plus_PMCigPrevci_upper;
        d1=Twelve_plus_PMMJPrevestimate<Twelve_plus_PMMJPrevci_lower;
        d2=Twelve_plus_PMMJPrevestimate>Twelve_plus_PMMJPrevci_upper;
        e1=Twelve_plus_PMTobPrevestimate<Twelve_plus_PMTobPrevci_lower;
        e2=Twelve_plus_PMTobPrevestimate>Twelve_plus_PMTobPrevci_upper;
        f1=Twelve_plus_PYAUDPrevestimate<Twelve_plus_PYAUDPrevci_lower;
        f2=Twelve_plus_PYAUDPrevestimate>Twelve_plus_PYAUDPrevci_upper;
```

```
        g1=Twelve_plus_PYAlcDepPrevestimate<Twelve_plus_PYAlcDepPrevci_lower;
        g2=Twelve_plus_PYAlcDepPrevestimate>Twelve_plus_PYAlcDepPrevci_upper;
        h1=Twelve_plus_PYCocPrevestimate<Twelve_plus_PYCocPrevci_lower;
        h2=Twelve_plus_PYCocPrevestimate>Twelve_plus_PYCocPrevci_upper;
        i1= Twelve_plus_PYMJIncestimate<Twelve_plus_PYMJIncci_lower;
        i2= Twelve_plus_PYMJIncestimate>Twelve_plus_PYMJIncci_upper;
        j1=Twelve_plus_PYMJPrevestimate<Twelve_plus_PYMJPrevci_lower;
        j2=Twelve_plus_PYMJPrevestimate>Twelve_plus_PYMJPrevci_upper;
        k1=Twelve_plus_PYSUDPrevestimate<Twelve_plus_PYSUDPrevci_lower;
        k2=Twelve_plus_PYSUDPrevestimate>Twelve_plus_PYSUDPrevci_upper;
run;

proc freq data=merge_check;
        title4'1 indicates where estimate is out of bounds';
        table a1 a2 b1 b2 c1 c2 d1 d2 e1 e2 f1 f2 g1 g2 h1 h2 i1 i2 j1 j2 k1 k2/list
missing;
run;
```

```
/*all estimates are within lower and upper confidence intervals

chack a random subset of variables for comparisons*/


proc surveyselect noprint data=nsduh_saes method=srs n=10 seed=2
out=merge_check1;
run;


proc print data=merge_check1 noobs;
        title4'random subset of long form observations';
        var geography outcome short_name age_group2 estimate ci_lower ci_upper;
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Eighteen_plus_PYMJPrevestimate Eighteen_plus_PYMJPrevci_lower
Eighteen_plus_PYMJPrevci_upper;
        where geography = "California Region 12R";
run;
```

```
proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Twelve_plus_PYAUDPrevestimate Twelve_plus_PYAUDPrevci_lower
Twelve_plus_PYAUDPrevci_upper;
        where geography = "Delaware Sussex";
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Eighteen_plus_PYAlcDepPrevestima Eighteen_plus_PYAlcDepPrevci_low
Eighteen_plus_PYAlcDepPrevci_upp;
        where geography = "Florida Circuit 12";
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var  Eighteen_to_25_PYCocPrevestimate Eighteen_to_25_PYCocPrevci_lower
Eighteen_to_25_PYCocPrevci_upper;
        where geography = "Florida Circuit 12";
```

```
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Twelve_plus_PYSUDPrevestimate Twelve_plus_PYSUDPrevci_lower
Twelve_plus_PYSUDPrevci_upper;
        where geography = "Massachusetts Central";
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Eighteen_plus_PYCocPrevestimate  Eighteen_plus_PYCocPrevci_lower
Eighteen_plus_PYCocPrevci_upper;
        where geography = "Michigan Kent";
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
```

```
        var Twenty_six_plus_PMMJPrevestimate Twenty_six_plus_PMMJPrevci_lower
Twenty_six_plus_PMMJPrevci_upper;
        where geography = "New Hampshire Central 1";
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Eighteen_to_25_PMTobPrevestimate Eighteen_to_25_PMTobPrevci_lower
Eighteen_to_25_PMTobPrevci_upper;
        where geography = "New York Region 2C: New York";
run;


proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Twelve_to_17_PMAlcPrevestimate Twelve_to_17_PMAlcPrevci_lower
Twelve_to_17_PMAlcPrevci_upper;
        where geography = "Oregon Region 4";
run;
```

```
proc print data= sae_merge noobs;
        title4'compare to wide form after merge';
        var Twelve_plus_PYAlcDepPrevestimate Twelve_plus_PYAlcDepPrevci_lower
Twelve_plus_PYAlcDepPrevci_upper;
        where geography = "West";
run;
/*all are exact matches*/


data sae_merge2;
        set sae_merge ;


/*create a state binary*/
        if geography
in("Alabama","Alaska","Arizona","Arkansas","California","Colorado","Connecticut",
        "Delaware","Florida","Georgia","Hawaii","Idaho","Illinois","Indiana","Iowa","Kan
sas","Kentucky",
        "Louisiana","Maine","Maryland","Massachusetts","Michigan","Minnesota","Miss
issippi","Missouri",
```

```
        "Montana","Nebraska","Nevada","New Hampshire","New Jersey","New
Mexico","New York","North Carolina",
        "North Dakota","Ohio",        "Oklahoma","Oregon","Pennsylvania","Rhode
Island","South Carolina",
        "South Dakota",        "Tennessee","Texas","Utah",
        "Vermont","Virginia","Washington","West Virginia",
        "Wisconsin","Wyoming") then do;
                agg_state=1;
                agg=1;
                agg_region=0;
        end;
        else if geography in ("Midwest", "Northeast", "South", "West", "United States")
then do
                agg_region=1;
                agg_state=0;
                agg=1;
         end;
         else do;
                agg_region=0;
```

```
                agg_state=0;

                agg=0;

        end;



                if geography = "New York Region 10" then geography = "New York
Region C";

                if geography = "New York Region 11" then geography = "New York
Region C";

                if geography = "New York Region 12" then geography = "New York
Region C";

                if geography = "New York Region 13" then geography = "New York
Region D";

                if geography = "New York Region 14" then geography = "New York
Region D";

                if geography = "New York Region 15" then geography = "New York
Region D";

                if geography = "New York Region 1: Long Island" then geography =
"New York Region B";

                if geography = "New York Region 2" then geography = "New York
Region A";
```

if geography = "New York Region 2: New York City" then geography = "New York Region A";

if geography = "New York Region 2A: Bronx" then geography = "New York Region A";

if geography = "New York Region 2C: New York" then geography = "New York Region A";

if geography = "New York Region 2D: Queens" then geography = "New York Region A";

if geography = "New York Region 6" then geography = "New York Region B";

if geography = "New York Region 7" then geography = "New York Region B";

if geography = "New York Region 8" then geography = "New York Region C";

if geography = "New York Region 9" then geography = "New York Region C";

if geography = "North Carolina Cardinal Innovations Healthcare Solutions 3" then geography = "North Carolina Cardinal Innovations";

if geography = "North Carolina CenterPoint Human Services" then geography = "North Carolina CenterPoint";

```
                if geography = "North Carolina Partners Behavioral Health Management"
then geography = "North Carolina Partners";

                if geography = "North Carolina Sandhills Center 1" then geography =
"North Carolina Sandhills";

                if geography = "North Carolina Sandhills Center 2" then geography =
"North Carolina Sandhills";

                if geography = "North Carolina Smoky Mountain Center 1" then
geography = "North Carolina Smoky Mountain";

                if geography = "North Carolina Smoky Mountain Center 2" then
geography = "North Carolina Smoky Mountain";

                if geography = "North Carolina Trillium Health Resources 1" then
geography = "North Carolina ECBH";

                if geography = "North Carolina Trillium Health Resources 2" then
geography = "North Carolina CoastalCare";



                if geography = "Michigan Detroit City" then fips_pch = "26163";

run;



proc freq data=sae_merge2;

        title4'check agg bins';
```

```
        table agg*agg_region*agg_state/list missing;
run;


proc sort data= sae_merge2;
        by geography  ;
run;


proc sort data= nsduh_crosswalk2;
        by geography  ;
run;


data nsduh_merge;
        merge sae_merge2 (in=insaex) nsduh_crosswalk2 (in=incrossx);
        by geography ;


        incross = incrossx;
        insae = insaex;
run;
```

```
proc freq data=nsduh_merge;
        title4'nsduh merge check';
        table incross*insae/list missing;
run;

proc freq data=nsduh_merge;
        title4'nsduh merge check where not in crosswalk data';
        table geography / list missing;
        where incross=0 and agg=0;
run;

proc print data=nsduh_merge noobs n;
        title4'why are so many county fips codes missing?';
        var name fips_pch geography;
        where fips_pch = "" and agg=0;
run;
```

```
proc freq data=nsduh_merge noprint;
        title4'find remaining fips code duplicates';
        table fips_pch / list missing out=fips_counts ;
run;


proc print data=fips_counts;
        where count>1;
run;


proc print data=nsduh_merge;
        var name fips_pch geography;
        where fips_pch in ("10003", "26163");
run;


proc print data=nsduh_merge;
        title4'Check empty geography';
        where geography="";
run;
```

```
data nsduh_merge2;
        length name $90 ;
        set nsduh_merge;
        /*remove wilmington city, detroit, and kalawao*/
        if geography in ("Delaware Wilmington City") then delete;
        if geography in ("Michigan Detroit City") then delete;
        if name in ("Kalawao County") then delete;
run;

proc freq data=nsduh_merge2  noprint;
        title4'Check to see if fips duplicate removal worked (should not print)';
        table fips_pch / list missing out=fips_counts ;
run;

proc print data=fips_counts n;
        where count>1 and fips_pch ~="";
run;
```

```
proc freq data=test4 noprint;
        title4'check census/election data for duplicate fips';
        table fips_pch/list missing out=countfips;
run;


proc print data=countfips;
        where count>1;
run;


proc print data=test4;
        var name fips_pch geography;
        where fips_pch in ("10003", "26163");
run;


proc sort data=nsduh_merge2;
        by fips_pch;
run;
```

```
proc sort data=test4;
        by fips_pch;
run;

data full_set;
        merge nsduh_merge2 test4;
        by fips_pch;
/*make sure these are not in final set*/
        if geography ="Delaware Wilmington City" then delete;
        if name ="Kalawao County" then delete;

        label agg = "Observation level is above county - aggregate"
                agg_region = "Observation level is US regional"
                agg_state = "Observation level is US state";
run;

proc freq data=full_set;
```

```
        title4'check final merge';
        table incensus*inelection*insae*incross/list missing;
run;

proc print data=full_set ;
        title4'check missing census data if not aggregate info';
        where incensus= . and agg ~=1;
run;

proc print data=full_set ;
        title4'check missing name values';
        var geography;
        where name= "" and agg ~=1;
run;
/*allmass/conn counties or aggregate areas*/

proc freq data=full_set;
        title4'check aggregate flag data';
```

```
        table agg*incensus*inelection*insae*incross/list missing;

run;



proc print data=full_set;

        var agg name geography incensus inelection insae incross;

        where agg=.;

run;



proc print data=full_set;

        title4'Check obs missing census/election data';

        var geography name county fips_pch incensus incross inelection insae agg;

        where (inelection=. or incensus=.) and agg ~=1;

run;



proc print data=full_set noobs;

        title4'why are there missing nsduh estimates?';

        var state geography fips_pch name incensus incross inelection insae;

        where  insae= 0 and agg ~=1 ;
```

```
run;

/*begin imputing SAE state averages for massachusetts and conn counties/

/* Mean imputation: Use PROC STDIZE to replace missing values with mean */
proc stdize data=full_set out=mass_imputes
        reponly             /* only replace; do not standardize */
        method=MEAN;        /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
        var Eighteen: Twelve: Twenty:;
        where geography= 'Massachusetts' or state='Massachusetts';
run;

proc stdize data=full_set out=conn_imputes
        reponly             /* only replace; do not standardize */
        method=MEAN;        /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
        var Eighteen: Twelve: Twenty:;
        where geography= 'Connecticut' or state='Connecticut';
run;
```

```
proc print data=full_set;
        title4'Imputed mass and conn SAEs';
        var  state name geography fips_pch Eighteen: Twelve: Twenty:;
        where geography in ('Massachusetts', 'Connecticut');
run;


proc print data=mass_imputes;
        var  state name geography fips_pch Twelve:;
run;


proc print data=conn_imputes;
        var  state name geography fips_pch Twelve:;
run;


proc print data=full_set;
        title4'drop these duplicates';
        var  state name geography fips_pch Twelve:;
```

```
        where state="Massachusetts" or state='Connecticut';

run;



data full_set2;

        set full_set (where=(state~= "Massachusetts" and state~= 'Connecticut'))

                mass_imputes conn_imputes;



        if agg=. then agg = 0;

        if inelection=. then inelection= 0;

        if incensus=. then incensus = 0;



/*placeholders for variables built in next data step*/

        RCL_2012 = 0;

        LAG_2014 = 0;

        sens_2014 = 0;

        sens2_2014 = 0;

        sens3_2014 = 0;
```

```
        if state in ("Massachusetts", 'Connecticut') then geography = catx(' ', state,
name);
run;


proc print data= full_set2;
        title4'check mass imputations';
        var state geography fips_pch Twelve:;
        where (state = 'Massachusetts' or geography= 'Massachusetts') and agg ~=1;
run;


proc print data= full_set2;
        title4'check Conn imputations';
        var state geography fips_pch Twelve:;
        where (state = 'Connecticut' or geography= 'Connecticut') and agg ~=1;
run;


proc freq data=full_set2;
        title4'check for any unexpected missingness or duplicates';
```

```
        table incensus*incross*insae*inelection / list missing;
run;


proc freq data=full_set2;
        title4'check for any unexpected missingness or duplicates';
        table rcl_2012 LAG_2014 / list missing;
run;


proc freq data=full_set2;
        title4'missing names, fips, census, election, and cross walk should all be
aggregate obs';
        table agg / list missing;
        where fips_pch='' or name='' or incensus=0 or inelection=0 or incross=0;
run;


proc print data=full_set2 noobs n;
        title4'what are these non-agg obs?';
        var name geography;
```

```
        where (fips_pch='' or name='' or incensus=0 or inelection=0 or incross=0) and
agg=0;
run;


proc print data=full_set2 noobs n;
        title4'check to see that we have all the data for massachusetts and connecticut
we need before dropping these observations';
        where state in ('Massachusetts','Connecticut');
run;


proc print data=full_set2 noobs n;
        title4'check Colorado counties for RCL coding';
        var state name county geography;
        where state in ("Colorado");
run;


proc print data=full_set2 noobs n;
        title4'check Washington counties for RCL coding';
        var state name county geography;
```

```
        where state in ("Washington");
run;


title3"Step 4: Code county level cannabis sales";


data full_set3;
        set full_set2;
        if geography in ("Connecticut Eastern", "Connecticut North Central",
                                "Connecticut Northwestern", "Connecticut South
Central",
                                "Connecticut Southwest", "Massachusetts
Boston",
                                "Massachusetts Central", "Massachusetts
Metrowest",
                                "Massachusetts Northeast", "Massachusetts
Southeast",
                                "Massachusetts Western") then delete;
```

/*Here we create five variables: RCL_2012, LAG_2014, sens_2014, sens2_2014, sens3_2014

RCL_2012: Counties where rec cannabis sales became legal in 2012 (not implemented until 2014, but voted on in 2012).

This incluides all Washington counties and 2/3 of Colorado counties

LAG_2014: counties that have local authority over cannabis sales and voted to allow in 2014.

This includes counties in Colorado, all of Alaska, and Oregon.

Washington must be coded as missing because cannabis is legal but no local authority was granted to counties

should not be counted in either exposure condition

sens_2014: Same as LAG_2014 but codes washington counties according to zoning laws outlined in the

I-502 Evaluation Plan and Preliminary Report on Implementation, exhibit 6

sens2_2014: same as sens_2014, but Alaska defined as missing because it has no variance

sens3_2014: same as LAG_2014, but Alaska defined as missing because it has no variance

/*64 counties in Colorado*/

if state = "Colorado" and name in ("Park County", "Conejos County", "Pitkin County", "Arapahoe County", "Adams County", "Douglas County",

"Eagle County", "Larimer County", "Weld County", "Gilpin County", "Boulder County", "Summit County", "Chaffee County", "Garfield County",

"Delta County", "Montezuma County", "Moffat County", "Gunnison County", "Saguache County", "Mesa County", "Denver County",

"Montrose County", "Dolores County", "La Plata County", "El Paso County", "Jefferson County", "Clear Creek County", "San Miguel County",

"Crowley County", "Grand County", "Routt County", "Huerfano County", "Las Animas County", "Lake County", "Morgan County", "Archuleta County",

"Pueblo County", "Ouray County", "Otero County", "Costilla County", "Sedgwick County", "San Juan County")

then do;

RCL_2012 = **1**;

LAG_2014 = **1**;

sens_2014 = **1**;

```
                sens2_2014 = 1;

                sens3_2014 = 1;

        end;

    else if state = "Colorado" and name not in ("Park County", "Conejos County",
"Pitkin County", "Arapahoe County", "Adams County", "Douglas County",

        "Eagle County", "Larimer County", "Weld County", "Gilpin County",
"Boulder County", "Summit County", "Chaffee County", "Garfield County",

        "Delta County", "Montezuma County", "Moffat County", "Gunnison
County", "Saguache County", "Mesa County", "Denver County",

        "Montrose County", "Dolores County", "La Plata County", "El Paso
County", "Jefferson County", "Clear Creek County", "San Miguel County",

        "Crowley County", "Grand County", "Routt County", "Huerfano County",
"Las Animas County", "Lake County", "Morgan County", "Archuleta County",

        "Pueblo County", "Ouray County", "Otero County", "Costilla County",
"Sedgwick County", "San Juan County")

    then do;

                RCL_2012 = 0;

                LAG_2014 = 0;

                sens_2014 = 0;

                sens2_2014 = 0;
```

```
                sens3_2014 = 0;
        end;

        /*29 counties in Alaska, some towns where cannabis is illegal but not many*/
        else if state = "Alaska" and name in ("Aleutians West Census Area", "Kodiak
Island Borough", "Bethel Census Area", "Aleutians East Borough",
                "Wade Hampton Census Area", "Dillingham Census Area", "Yukon-
Koyukuk Census Area", "Northwest Arctic Borough", "North Slope Borough",
                "Anchorage Municipality", "Denali Borough", "Hoonah-Angoon Census
Area", "Nome Census Area", "Lake and Peninsula Borough",
                "Valdez-Cordova Census Area", "Prince of Wales-Hyder Census Area",
"Southeast Fairbanks Census Area", "Fairbanks North Star Borough",
                "Kenai Peninsula Borough", "Matanuska-Susitna Borough", "Juneau City
and Borough", "Petersburg Census Area", "Ketchikan Gateway Borough",
                "Sitka City and Borough", "Skagway Municipality", "Wrangell City and
Borough", "Yakutat City and Borough", "Bristol Bay Borough", "Haines Borough" )
                then do;
                        RCL_2012 = 0;
                        LAG_2014 = 1;
                        sens_2014 = 1;
```

```
                sens2_2014 = .;

                sens3_2014 = .;

            end;


            /*36 counties in Oregon, 15 counties with banned cannabis, 21 w/o*/

        else if state = "Oregon" and name in ("Benton County", "Clackamas County",
"Clatsop County", "Columbia County", "Coos County",

            "Curry County", "Deschutes County", "Gilliam County", "Grant County",
"Hood River County", "Jackson County",

            "Josephine County", "Lane County", "Lincoln County", "Linn County",
"Multnomah County", "Polk County",

            "Tillamook County", "Wasco County", "Washington County", "Yamhill
County")

            then do;

                RCL_2012 = 0;

                LAG_2014 = 1;

                sens_2014 = 1;

                sens2_2014 = 1;

                sens3_2014 = 1;
```

```
                end;

        else if state = "Oregon" and name not in ("Benton County", "Clackamas
County", "Clatsop County", "Columbia County", "Coos County",

                "Curry County", "Deschutes County", "Gilliam County", "Grant County",
"Hood River County", "Jackson County",

                "Josephine County", "Lane County", "Lincoln County", "Linn County",
"Multnomah County", "Polk County",

                "Tillamook County", "Wasco County", "Washington County", "Yamhill
County")

                then do;

                        RCL_2012 = 0;

                        LAG_2014 = 0;

                        sens_2014 = 0;

                        sens2_2014 = 0;

                        sens3_2014 = 0;

                end;


        else if state = "Washington" and name in ("Clark County","Columbia
County","Franklin County","Garfield County",
```

```
                "Kittitas County","Lewis County","Pierce County","Wahkiakum
County","Walla Walla County","Yakima County")
                then do;
                        RCL_2012 = 1;
                        LAG_2014 = .;
                        sens_2014 = 0;
                        sens2_2014 = 0;
                        sens3_2014 = .;
                end;


        else if state = "Washington" and name not in
("Clark","Columbia","Franklin","Garfield",
                "Kittias","Lewis","Pierce","Wahkiakum","Walla Walla","Yakima")
                then do;
                        RCL_2012 = 1;
                        LAG_2014 = .;
                        sens_2014 = 1;
                        sens2_2014 = 1;
                        sens3_2014 = .;
```

```
        end;


        else if state = "Alaska"
        then do;
                RCL_2012 = 1;
                LAG_2014 = .;
                sens_2014 = 1;
                sens2_2014 = .;
                sens3_2014 = .;
        end;


else if agg = 1 then do;
        RCL_2012 = .;
        LAG_2014 = .;
        sens_2014 = .;
        sens2_2014 = .;
        sens3_2014 = .;
end;
```

label RCL_2012 = "County has at least one municipality that will allow recreational cannabis sales (year of vote)"

LAG_2014 = "County has at least one municipality where voters decided cannabis can be legally sold AND Local authority was granted to those voters (year of vote), washington coded missing"

sens_2014 = "County has at least one municipality where voters decided cannabis can be legally sold AND Local authority was granted to those voters (year of vote), washington coded using zoning bans"

sens2_2014 = "County has at least one municipality where voters decided cannabis can be legally sold AND Local authority was granted to those voters (year of vote), washington coded using zoning bans, alaska missing"

sens3_2014 ="County has at least one municipality where voters decided cannabis can be legally sold AND Local authority was granted to those voters (year of vote), washington and alaska missing";

run;

proc freq data=full_set3;

title4'check for any unexpected missingness or duplicates';

table incensus*incross*insae*inelection / list missing;

```
run;


proc freq data=full_set3;
        title4'check RCL 2014 coding';
        table state*rcl_2012*LAG_2014*sens_2014*sens2_2014*sens3_2014 / list
missing;
run;


proc freq data=full_set3;
        title4'missing names, fips, census, election, and cross walk should all be
aggregate obs';
        table agg*qname/ list missing;
        where incensus=0 or inelection=0 or incross=0 or insae=0;
run;


/*OK: aggregate areas, mass and conn counties*/


/*range check*/
data merge_check2;
```

```
set full_set3;

a1=Twelve_plus_PYAlcDepPrevestimate<Twelve_plus_PYAlcDepPrevci_lower;

a2=Twelve_plus_PYAlcDepPrevestimate>Twelve_plus_PYAlcDepPrevci_upper;

b1=Twelve_plus_PMAlcPrevestimate<Twelve_plus_PMAlcPrevci_lower;

b2=Twelve_plus_PMAlcPrevestimate>Twelve_plus_PMAlcPrevci_upper;

c1=Twelve_plus_PMCigPrevestimate<Twelve_plus_PMCigPrevci_lower;

c2=Twelve_plus_PMCigPrevestimate>Twelve_plus_PMCigPrevci_upper;

d1=Twelve_plus_PMMJPrevestimate<Twelve_plus_PMMJPrevci_lower;

d2=Twelve_plus_PMMJPrevestimate>Twelve_plus_PMMJPrevci_upper;

e1=Twelve_plus_PMTobPrevestimate<Twelve_plus_PMTobPrevci_lower;

e2=Twelve_plus_PMTobPrevestimate>Twelve_plus_PMTobPrevci_upper;

f1=Twelve_plus_PYAUDPrevestimate<Twelve_plus_PYAUDPrevci_lower;

f2=Twelve_plus_PYAUDPrevestimate>Twelve_plus_PYAUDPrevci_upper;

g1=Twelve_plus_PYAlcDepPrevestimate<Twelve_plus_PYAlcDepPrevci_lower;

g2=Twelve_plus_PYAlcDepPrevestimate>Twelve_plus_PYAlcDepPrevci_upper;

h1=Twelve_plus_PYCocPrevestimate<Twelve_plus_PYCocPrevci_lower;

h2=Twelve_plus_PYCocPrevestimate>Twelve_plus_PYCocPrevci_upper;

i1= Twelve_plus_PYMJIncestimate<Twelve_plus_PYMJIncci_lower;
```

```
        i2= Twelve_plus_PYMJIncestimate>Twelve_plus_PYMJIncci_upper;
        j1=Twelve_plus_PYMJPrevestimate<Twelve_plus_PYMJPrevci_lower;
        j2=Twelve_plus_PYMJPrevestimate>Twelve_plus_PYMJPrevci_upper;
        k1=Twelve_plus_PYSUDPrevestimate<Twelve_plus_PYSUDPrevci_lower;
        k2=Twelve_plus_PYSUDPrevestimate>Twelve_plus_PYSUDPrevci_upper;
run;

proc freq data=merge_check2;
        title4'1 indicates where estimate is out of bounds';
        table a1 a2 b1 b2 c1 c2 d1 d2 e1 e2 f1 f2 g1 g2 h1 h2 i1 i2 j1 j2 k1 k2/list
missing;
run;

proc means data=full_set3 nmiss;
        title4'Check Missingness in NSDUH data';
        var twelve: eighteen: twenty: ;
run;
/*smaller age groups are often suppressed, and overlapping age categories make no
sense to include in prediction
```

```
use only the twelve plus variables + 18+ mental health variables (MDE, SMI, AMI,
suicidal thoughts) because these variable aren't available for indiviudals under 18*/

/*save this version of dataset with local authority coding for easy access */
data sav.full_set_LAG;
        set full_set3;
        where agg ~=1;
        drop agg: incensus inelection incross insae;
run;

proc contents data=sav.full_set_LAG  ;
run;

proc sort data=sav.full_set_LAG;
        by state;
run;
proc print data= sav.full_set_lag;
        var name rcl_2012 lag_2014 sens_2014 sens2_2014 sens3_2014;
```

```
        where state in ("Alaska", "Colorado", "Oregon", "Washington");
        by state;
run;

/*Need a table 1*/

proc means data=sav.full_set_LAG;
        title3'table 1';
        class lag_2014;
        var T001_001 t003_002 t003_003 t011_002 t011_003 t011_004 t011_005
                T055_003 T055_004 T055_005 T055_006 T055_007 T055_008 T055_009
T055_010
                eighteen_plus_pysmiprevestimate eighteen_plus_pystprevestimate
percentdem_2012_votes percentrep_2012_votes
                twelve_plus_pmalcprevestimate twelve_plus_pmcigprevestimate
twelve_plus_pmmjprevestimate twelve_plus_pyaudprevestimate
                twelve_plus_pycocprevestimate twelve_plus_pysudprevestimate;
run;
```

```
/**/
/*proc sql noprint ;*/
/*      select name into :droplist separated by ' '*/
/*      from contents*/
/*      where name like '%upp%' or name like '%upper' */
/*              or name like '%se' or name like '%lower' */
/*              or name like '%estimate' escape '^';*/
/*quit;*/
/**/
/*%put=&droplist;*/
/**/
/*data trim_set ;*/
/*  set sav.full_set (drop=&droplist);*/
/*  where agg ~=1;*/
/*  drop agg: incensus inelection incross insae twelve_to_: twenty: eighteen_to:*/
/*              Eighteen_plus_PMAlcPrevL Eighteen_plus_PMAlcPrevsel*/
/*              Eighteen_plus_PMCigPrevL Eighteen_plus_PMCigPrevsel*/
/*              Eighteen_plus_PMMJPrevL Eighteen_plus_PMMJPrevsel*/
```

```
/*                Eighteen_plus_PMTobPrevL Eighteen_plus_PMTobPrevsel*/

/*                Eighteen_plus_PYAUDPrevL Eighteen_plus_PYAUDPrevsel*/

/*                Eighteen_plus_PYAlcDepPrevL Eighteen_plus_PYAlcDepPrevL_lowe*/

/*                Eighteen_plus_PYAlcDepPrevci_low Eighteen_plus_PYAlcDepPrevestima
Eighteen_plus_PYAlcDepPrevsel*/

/*                Eighteen_plus_PYCocPrevL Eighteen_plus_PYCocPrevsel*/

/*                Eighteen_plus_PYMJIncL Eighteen_plus_PYMJIncsel*/

/*                Eighteen_plus_PYMJPrevL Eighteen_plus_PYMJPrevsel */

/*                Eighteen_plus_PYSUDPrevL Eighteen_plus_PYSUDPrevsel ;*/

/*run;*/



/*save a trimmer set without excess nsduh variables (keep only L and SEL nsduh vars
for bootstrapping in R), aggregate areas, qc variables*/

/*data sav.trim_set;*/

/*        set trim_set;*/

/*run;*/

/**/

/*proc contents data=sav.full_set;*/

/*        title4'check final contents';*/
```

```
/*run;*/

/**/

/*proc contents data=sav.trim_set;*/

/*      title4'check final contents';*/

/*run;*/



/*make a version for appendix where voter variables are binaries*/



proc print data=sav.full_set_LAG;

        title3 'Do 3rd party voters outnumber dems or reps in any county?';

        var county;

        where (percentother_2012_votes > percentrep_2012_votes)

        or (percentother_2012_votes > percentdem_2012_votes);

run;



data voter_bins;

        set sav.full_set_LAG;

        dembin = 0;
```

```
        repbin = 0;


        if percentdem_2012_votes > percentrep_2012_votes then dembin=1;
        else if percentrep_2012_votes > percentdem_2012_votes then repbin = 1;
run;


proc freq data=voter_bins;
        table dembin*repbin/list missing;
run;




/*export for r*/


/*proc export data=sav.full_set_LAG*/
/*   outfile="C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Processed\County_RCL_LAG.csv"*/
/*   dbms=csv*/
/*   replace;*/
```

```
/*run;*/

/**/

/*proc export data=voter_bins*/

/*   outfile="C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Processed\voter_bins.csv"*/

/*   dbms=csv*/

/*   replace;*/

/*run;*/



proc sql;

        title4'Counts in full set';

        select count(distinct fips_pch) as uniq_counties, count (distinct geography) as
uniq_nsduh_sa

        from sav.full_set;

quit;



proc sql;

        title4'Counts in LAG';
```

```
        select count(distinct fips_pch) as uniq_counties, count (distinct geography) as
uniq_nsduh_sa
        from sav.full_set_LAG;
quit;


proc sql;
        title4'rcl counties per state: rcl_2012';
        select state, count(distinct(name))
        from sav.full_set_LAG
        where rcl_2012 in (1,.)
        group by state;
quit;


proc sql;
        title4'rcl counties per state: lag_2014';
        select state, count(distinct(name)), count(lag_2014)
        from sav.full_set_LAG
        where LAG_2014 in (1,.)
```

```
        group by state;

quit;



proc sql;

        title4'rcl counties per state: sens_2014';

        select state, count(distinct(name)), count(sens_2014)

        from sav.full_set_LAG

        where sens_2014 in (1,.)

        group by state;

quit;



proc sql;

        title4'rcl counties per state: sens2_2014';

        select state, count(distinct(name)), count(sens2_2014)

        from sav.full_set_LAG

        where sens2_2014 in (1,.)

        group by state;

quit;
```

```
proc sql;
        title4'rcl counties per state: sens3_2014';
        select state, count(distinct(name)), count(sens3_2014)
        from sav.full_set_LAG
        where sens3_2014 in (1,.)
        group by state;
quit;


proc sql;
        title4'check colorado rcl counties for mismatches in data step or missing
counties';
        select distinct name
        from sav.full_set
        where RCL_2014=1 and state="Colorado";
quit;
/*no text mismatches*/
```

```
proc print data=sav.full_set noobs n;
        title4'What does NC end up looking like?';
        var name geography incensus inelection incross insae Twelve:;
        where geography contains "North Carolina";
run;

title3"Step 5: Model legalization at county level";

proc contents data = sav.full_set out=varnames (keep=name) ;
run;

proc freq data=sav.full_set;
        title4'hierarchical variables, along with states';
        table  division division_n region / list missing;
run;

data predictorvars;
        set varnames;
```

```
        if name in ('COUNTY', 'QName', 'agg', 'agg_region', 'agg_state', 'fips_pch',
'geography',
                                    'incensus', 'incross', 'inelection', 'insae', 'name',
'state_fips') then delete;
run;


proc sql noprint;
        select trim(name) into : predictor_list separated by " "
        from predictorvars;
quit;


%put &predictor_list;


/*Too much missing data in these predictors for model convergence
        Missing data analysis
break down variables by source data to determine missing data patterns*/


proc sql noprint;
        select trim(name) into : censusvars separated by " "
```

```
        from predictorvars
        where index(name,"T0") > 0 or name in ('area_land', 'area_water');

        select trim(name) into : nsduhvars separated by " "
        from predictorvars
        where index(name,"Eighteen") > 0 or index(name,"Twelve")>0 or
index(name,"Twenty")>0 ;

        select trim(name) into : votervars separated by " "
        from predictorvars
        where index(name,"dem") > 0 or index(name,"rep")>0 or index(name,"other")>0
or index(name,"total")>0  ;
quit;

%put &censusvars;
%put &nsduhvars;
%put &votervars;

proc means data = sav.full_set n nmiss;
```

```
        title4'Check all numeric variables from census for missingness';
        var &censusvars;
        where agg=0;
run;
/*no census data missing, beautiful*/

proc means data = sav.full_set n nmiss;
        title4'Check all numeric variables from election data for missingness';
        var &votervars;
        where agg=0;
run;
/*no voter data missing with the Alaskan imputations, beautiful*/

proc means data = sav.full_set n nmiss;
        title4'Check all numeric variables from nsduh data for missingness';
        var &nsduhvars;
        where agg=0;
run;
```

```
/*NSDUH SAEs unsurprisingly the limiting factor due to data suppression issues.*/


data full_set2;
        set sav.full_set;
        if missing(Twenty_Six_or_O_PYMJInc) then miss=1;
        else miss=0;
run;


proc freq data=full_set2;
        table miss / list missing;
run;


proc means data=full_set2 n nmiss;
        title4'Check all numeric variables from nsduh data for missingness';
        var &nsduhvars;
        where agg=0 and miss=0;
run;
```

```
proc corr data=full_set2 plots=scatter;
        title4'Check to see if data is missing at random';
        var miss T001_001;
run;
/*Cannot argue this is missing at random
as expected, more missingness where there is less population*/

proc sql noprint;
        select trim(name) into : predictor_list separated by " "
        from predictorvars;
quit;

ods graphics on;

proc means data=sav.full_set;
        var Twelve_or_Older:;
run;
```

```
/*ALL MODELLING COMPLETED IN R*/

/*************************************************************************************************

* In Enterprise Guide, "Specify the page size for log and text output" under 'Results
General' must be *

* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *

*************************************************************************************************/

OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE validVarName=any;



title1'Dissertation';

title2'Aim 1: Legalization Prediction';



/*****************************************************************************

* The following macro variables are available to all users: *

*                                                                *

* &Project     – the name of the project folder in which the .EGP file is stored    *

* &ProgName        – the name of the .egp file, without the extension                 *

* &ProgNode        - the name of the code node                                        *
```

```
* &ProgDir        – the path to the folder in which the .egp file is stored.          *

*******************************************************************************/



/*******************************************************************************

* The following macro variables are used in conjunction with the DOC_BLOCK   *

* macro to document programs and output.                                      *

*                                                                             *

* Do not use quotation marks when defining macro variables. If SAS syntax    *

* requires quotes, use double quotes when you reference the macro variable.  *

*******************************************************************************/

** PROGRAMMER'S NAME ;

%LET PROGRAMMER     = Barrett Montgomery;



** LIST ALL SUBDIRECTORIES CALLED IN THE LIBNAME STATEMENT ;

** THESE CAN BE LEFT BLANK IF NOT NEEDED OR USED ;

%LET USEDIR1      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim

1\raw;
```

```
** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;

** USE &PROGNAME FOR SAVEFILE NAME ;

** LEAVE BLANK IF NO DATASET SAVED ;

%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\processed;


** THE NUMBERING SCHEME IS SAVEFILx_y WHERE X IS THE NUMBER OF THE
SAVEDIR AND ;

** Y IS THE NUMBER OF THE FILE WITHIN IT ;

** THIS SHOULD GENERALLY BE SET TO EITHER &PROGNAME OR &PROGNODE ;

%LET SAVEFIL1_1   = ;


** NAME FORMAT LIBRARY DIRECTORY ;

%LET FMTDIR       = ;


** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT
WORK ;
```

```
%LET PURPOSE1 =;



/************************************************************

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

************************************************************/



** INPUT FILES ;

LIBNAME USE "&USEDIR1";



** OUTPUT FILE DESTINATION ;

LIBNAME SAV  "&SAVEDIR1" ;



title3"Step 6: Data visualizations";



proc import datafile= "C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim

1/Processed/ens_lag.csv" dbms=csv out=fig replace ;

        GUESSINGROWS=3000;

run;
```

```
proc import datafile= "C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim
1/Processed/ens_sens1.csv" dbms=csv out=sens_fig replace ;
        GUESSINGROWS=3000;
run;

/*Prior analysis shows these two methods of coding outcomes makes prediction
harder*/
/*proc import datafile= "C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim
1/Processed/ens_sens2.csv" dbms=csv out=sens2_fig replace ;*/
/*        GUESSINGROWS=3000; */
/*run;*/
/**/
/*proc import datafile= "C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim
1/Processed/ens_sens3.csv" dbms=csv out=sens3_fig replace ;*/
/*        GUESSINGROWS=3000; */
/*run;*/

proc freq data=fig;
```

```
        title4'check missing';
        table county / list;
        where PHT = .;
run;

proc contents data=fig;
        title4'Check var names need changed';
run;

data fig1;
        set fig;
        length name $500;
        array thresh[9] thresh1-thresh9;
        thresh1=.1;
        thresh2=.2;
        thresh3=.3;
        thresh4=.4;
        thresh5=.5;
```

```
thresh6=.6;

thresh7=.7;

thresh8=.8;

thresh9=.9;

array PHT_Y [9] PHT_Y1-PHT_y9;

array YHT_Y [9] YHT_Y1-YHT_y9;

array TNF_Y [9] TNF_Y1-TNF_y9;

array TPF_Y [9] TPF_Y1-TPF_y9;

array FNF_Y [9] FNF_Y1-FNF_y9;

array acc_Y [9] acc_Y1-acc_y9;



do i=1 to 9;

        PHT_Y[i]=PHT>thresh[i];

        YHT_Y[i]=YHT>thresh[i];

        TNF_Y[i]=TNF>thresh[i];

        TPF_Y[i]=TPF>thresh[i];

        FNF_Y[i]=FNF>thresh[i];

        acc_Y[i]=acc>thresh[i];
```

```
end;


name = county;

prediction = (acc>.3333);


rename var1=county_num;


label Y                ="Actual county policy, legalize cannabis with local
control"
                N              = 'number of times county appeared in sims'
                county         = "County name"
                prediction= 'Binary prediction for the default model, accuracy weighted
probabilities with the prevalence based cutoff'
                YHT            = "Average predicted probability of county having legal
cannabis w/ local control weighted by hard-coded Y based on proportion of RCL
counties"
                PHT            = "Average predicted probability of county having legal
cannabis w/ local control"
                TNF            = "Average predicted probability of county having legal
cannabis w/ local control weighted by specificity of each simulation"
```

TPF = "Average predicted probability of county having legal cannabis w/ local control weighted by sensitivity of each simulation"

FNF = "Average predicted probability of county having legal cannabis w/ local control weighted by 1-specificity of each simulation"

acc = "Average predicted probability of county having legal cannabis w/ local control weighted by overall accuracy of each simulation";

drop i thresh: county;

run;

data sens_fig1;

length name $**500**;

set sens_fig;

array thresh[**9**] thresh1-thresh9;

thresh1=**.1**;

thresh2=**.2**;

thresh3=**.3**;

thresh4=**.4**;

thresh5=**.5**;

```
thresh6=.6;

thresh7=.7;

thresh8=.8;

thresh9=.9;

array sens_PHT_Y [9] sens_PHT_Y1-sens_PHT_y9;

array sens_YHT_Y [9] sens_YHT_Y1-sens_YHT_y9;

array sens_TNF_Y [9] sens_TNF_Y1-sens_TNF_y9;

array sens_TPF_Y [9] sens_TPF_Y1-sens_TPF_y9;

array sens_FNF_Y [9] sens_FNF_Y1-sens_FNF_y9;

array sens_acc_Y [9] sens_acc_Y1-sens_acc_y9;



do i=1 to 9;

        sens_PHT_Y[i]=PHT>thresh[i];

        sens_YHT_Y[i]=YHT>thresh[i];

        sens_TNF_Y[i]=TNF>thresh[i];

        sens_TPF_Y[i]=TPF>thresh[i];

        sens_FNF_Y[i]=FNF>thresh[i];

        sens_acc_Y[i]=acc>thresh[i];
```

```
end;


name = county;

sens_prediction = (acc>.3333);


rename var1=sens_county_num
                Y = sens_Y
                acc = sens_acc
                N = sens_N
                YHT= sens_YHT
                PHT=sens_PHT
                TNF=sens_TNF
                TPF=sens_TPF
                FNF=sens_FNF;


label   Y           ="Actual county policy, legalize cannabis with local control"
            N           = 'number of times county appeared in sims'
            county      = "County name"
```

```
                sens_prediction= 'Binary prediction for the default model, accuracy
weighted probabilities with the prevalence based cutoff'

                YHT          = "Average predicted probability of county having legal
cannabis w/ local control weighted by hard-coded Y based on proportion of RCL
counties"

                PHT          = "Average predicted probability of county having legal
cannabis w/ local control"

                TNF          = "Average predicted probability of county having legal
cannabis w/ local control weighted by specificity of each simulation"

                TPF          = "Average predicted probability of county having legal
cannabis w/ local control weighted by sensitivity of each simulation"

                FNF          = "Average predicted probability of county having legal
cannabis w/ local control weighted by 1-specificity of each simulation"

                acc          = "Average predicted probability of county having legal
cannabis w/ local control weighted by overall accuracy of each simulation";

              drop county  i thresh:;

run;



proc freq data=fig1;

      title4'Check binary variable creation works';
```

```
        table pht * PHT_Y1 * PHT_Y2 * PHT_Y3 * PHT_Y4 * PHT_Y5 * PHT_Y6 * PHT_Y7
* PHT_Y8 * PHT_Y9 / list missing;

        table acc*prediction/list missing;

        table prediction*Y/list missing senspec;

run;



proc freq data=sens_fig1;

        title4'Check binary variable creation works for sensitivty data';

        table sens_pht * sens_PHT_Y1 * sens_PHT_Y2 * sens_PHT_Y3 * sens_PHT_Y4 *
sens_PHT_Y5 * sens_PHT_Y6 * sens_PHT_Y7 * sens_PHT_Y8 * sens_PHT_Y9 / list
missing;

run;



proc contents data=fig1;

        title4"Contents of predictions data";

run;



proc freq data=fig1;

        title4'Prevalence of Actual County level RCLs';
```

```
        table Y /list missing;
run;


proc freq data=sens_fig1;
        title4'Prevalence of Actual County level RCLs in sens analysis';
        table sens_Y /list missing;
run;


/*check dona ana county new mexico in all sets*/


proc print data=fig1;
        var name county_num;
        where name contains "New Mexico";
run;
/*county_num = 785*/


proc print data=sens_fig1;
        var name sens_county_num;
```

```
        where name contains "New Mexico";
run;
/*sens_county_num = 793*/

data fig2;
        set fig1;
        if county_num = 785 then name="Dona Ana County, New Mexico";
run;

data sens_fig2;
        set sens_fig1;
        if sens_county_num = 793 then name="Dona Ana County, New Mexico";
run;

/*MAPPING*/

/*need to add fips codes to the county names
                add from my own data file
```

```
                check mismatches between names

rename dona ana to match in both files*/



data fips;

        length name $500;

        set sav.full_set_LAG (keep=qname fips_pch state rcl_2012 LAG_2014
sens_2014 sens2_2014 sens3_2014);



        fips_pch = strip(fips_pch);



        if fips_pch = "35013" then qname="Dona Ana County, New Mexico";



        name = qname;

        drop qname ;

run;



proc print data=fips;

        title3"Check Dona Ana in both sets for match";
```

```
        where state = "New Mexico";

        var _char_;

run;



proc print data=fig2;

        var name;

        where name contains ("New Mexico");

run;



proc print data=sens_fig2;

        var name;

        where name contains ("New Mexico");

run;



proc sort data=fig2;

        by name;

run;
```

```
proc sort data=sens_fig2;
        by name;
run;

proc sort data=fips;
        by name;
run;

data fig_wide3;
        merge fig2 (in=in1x) sens_fig2 (in=in2x)  fips (in=in3x) ;
        by name;

        in1=in1x;
        in2=in2x;
        in3=in3x;

run;
```

```
proc freq data=fig_wide3;
        title4'Look at merge observations';
        table in1*in2*in3  / list missing;
run;


proc freq data=fig_wide3 ;
        title4'Look at merge observations that do not appear in all';
        table in1*in2*in3*state /list missing ;
        where in1=0 or in2=0 ;
run;


proc sql;
        title3'States where False Positives Appear';
        select distinct state
        from fig_wide3
        where y=0 and prediction=1;
quit;
```

```
proc freq data=fig_wide3;
        title3'confusion matrix';
        table y *prediction /list missing;
        where y~=.;
run;
/*                          y
                            0         1
pred    0  2721 0
                  1  281  92


total 3094
*/


proc freq data=fig_wide3;
        title3'confusion matrix including WA';
        table sens_y *sens_prediction /list missing;
        where sens_y~=.;
run;
```

```
/*                          y

                            0         1

pred     0  2718 2

                    1  294  119



total 3133

*/



/*/*Create table for sensitivity analyses*/*/

/*ods excel file="C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Results\weights and cutpoints raw tables.xlsx";*/

/*proc freq data=fig_wide3;*/

/*          title3'Table of sensitivity and speciifcity for each weighted probability estimator
at each hard cut-point';*/

/*          table Y*pht_Y1 Y*pht_Y2 Y*pht_Y3 Y*pht_Y4 Y*pht_Y5 Y*pht_Y6 Y*pht_Y7
Y*pht_Y8 Y*pht_Y9/list missing ;*/

/*          table Y*yht_Y1 Y*yht_Y2 Y*yht_Y3 Y*yht_Y4 Y*yht_Y5 Y*yht_Y6 Y*yht_Y7
Y*yht_Y8 Y*yht_Y9/list missing ;*/
```

```
/*        table Y*acc_Y1 Y*acc_Y2 Y*acc_Y3 Y*acc_Y4 Y*acc_Y5 Y*acc_Y6 Y*acc_Y7
Y*acc_Y8 Y*acc_Y9/list missing ;*/

/*        table Y*tpf_Y1 Y*tpf_Y2 Y*tpf_Y3 Y*tpf_Y4 Y*tpf_Y5 Y*tpf_Y6 Y*tpf_Y7 Y*tpf_Y8
Y*tpf_Y9/list missing ;*/

/*        table Y*tnf_Y1 Y*tnf_Y2 Y*tnf_Y3 Y*tnf_Y4 Y*tnf_Y5 Y*tnf_Y6 Y*tnf_Y7 Y*tnf_Y8
Y*tnf_Y9/list missing ;*/

/*        table Y*fnf_Y1 Y*fnf_Y2 Y*fnf_Y3 Y*fnf_Y4 Y*fnf_Y5 Y*fnf_Y6 Y*fnf_Y7 Y*fnf_Y8
Y*fnf_Y9/list missing ;*/

/*        where Y ~=.;*/

/*run;*/

/*ods excel close;*/

/**/

/*ods excel file="C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
1\Results\weights and cutpoints raw tables sens.xlsx";*/

/*proc freq data=fig_wide3;*/

/*        title3'Table of sensitivity and speciifcity for each weighted probability estimator
at each hard cut-point';*/

/*        table sens_Y*sens_pht_Y1 sens_Y*sens_pht_Y2 sens_Y*sens_pht_Y3
sens_Y*sens_pht_Y4 sens_Y*sens_pht_Y5 sens_Y*sens_pht_Y6 sens_Y*sens_pht_Y7
sens_Y*sens_pht_Y8 sens_Y*sens_pht_Y9/list missing ;*/
```

```
/*        table sens_Y*sens_yht_Y1 sens_Y*sens_yht_Y2 sens_Y*sens_yht_Y3
sens_Y*sens_yht_Y4 sens_Y*sens_yht_Y5 sens_Y*sens_yht_Y6 sens_Y*sens_yht_Y7
sens_Y*sens_yht_Y8 sens_Y*sens_yht_Y9/list missing ;*/

/*        table sens_Y*sens_acc_Y1 sens_Y* sens_acc_Y2 sens_Y* sens_acc_Y3 sens_Y*
sens_acc_Y4 sens_Y* sens_acc_Y5 sens_Y* sens_acc_Y6 sens_Y*sens_acc_Y7
sens_Y* sens_acc_Y8 sens_Y*sens_acc_Y9/list missing ;*/

/*        table sens_Y*sens_tpf_Y1 sens_Y* sens_tpf_Y2 sens_Y* sens_tpf_Y3 sens_Y*
sens_tpf_Y4 sens_Y* sens_tpf_Y5 sens_Y* sens_tpf_Y6 sens_Y*sens_tpf_Y7
sens_Y*sens_tpf_Y8 sens_Y*sens_tpf_Y9/list missing ;*/

/*        table sens_Y*sens_tnf_Y1 sens_Y* sens_tnf_Y2 sens_Y* sens_tnf_Y3 sens_Y*
sens_tnf_Y4 sens_Y* sens_tnf_Y5 sens_Y* sens_tnf_Y6 sens_Y*sens_tnf_Y7
sens_Y*sens_tnf_Y8 sens_Y*sens_tnf_Y9/list missing ;*/

/*        table sens_Y*sens_fnf_Y1 sens_Y* sens_fnf_Y2 sens_Y* sens_fnf_Y3 sens_Y*
sens_fnf_Y4 sens_Y* sens_fnf_Y5 sens_Y* sens_fnf_Y6 sens_Y*sens_fnf_Y7
sens_Y*sens_fnf_Y8 sens_Y*sens_fnf_Y9/list missing ;*/

/*        where sens_Y ~=.;*/

/*run;*/

/*ods excel close;*/



/*maps can be made from this internal counties dataset*/;
```

```
data counties;

        set maps.uscounty;

        length  fips_pch $ 5 stfips_pch $ 2 cnfips_pch $3;



        where state <72; /*remove puerto rico*/



   stfips_pch = put(state, z2.);

        cnfips_pch = put(county, z3.);



        fips_pch = strip(stfips_pch)||strip(cnfips_pch);     /* (pch = suffix for padded
character) */



        label stfips_pch = "State Fips, padded character"

                cnfips_pch = "County Fips, padded character"

                fips_pch = "concatednated state and county fips, padded character";

run;



proc freq data=counties;
```

```
        title3'Check fips pch creation';

        table state*stfips_pch county*cnfips_pch stfips_pch*cnfips_pch*fips_pch/list
missing;

run;



proc print data=counties;

        title3'check dona ana county in new mexico';

        where fips_pch="35013";

run;



proc contents data=counties;

        title3'check to see if coordinates are projected and how';

run;

/*already projected*/



/*actions to make map work before merge



in counties
```

remove 15005 - Kalawao Hawaii previously removed

remove 30113 - Yellowstone National Park

Halifax county, Virginia (FIPS code 51083) includes the population of the former independent city South Boston city, Virginia (FIPS code 51780)

recode 51780 as 51083

in fig_wide

02105 and 02230 need to use shapefile data from 02232

replace those fips codes with 02232

02280 was split to create part of 02275 and all of 02195

02275 was created from part of 02280 and part of 02201

02198 was created from the remainder of the former 02201

02280 > 02275 02280 > 02195

02201 02198

duplicate 02275 and give 02280 02201

51560, formerly an independent city, merged with Alleghany county (FIPS code 51005)

recode 51560 as 51005

Broomfield County, Colorado (FIPS code 08014), formed on 11-15-2001. Population data is not available for Broomfield county, Colorado.

drop 08014

THESE CHANGES ARE ONLY TO MAKE MAPS FUNCTIONAL, OTHER ANALYSIS SHOULD BE COMPELTED USING PRIOR DATASETS

*/

data counties2;

set counties ;

where fips_pch not in('15005', '30113');

if fips_pch = '51780' then fips_pch = '51083';

if fips_pch = '51560' then fips_pch = '51005';

```
run;



data dupe;
        set fig_wide3 (where=(fips_pch='02275'));

        fips_pch='02280';
run;


data fig_wide4;
        set fig_wide3 dupe;
        where fips_pch not in('08014', '02195', '02198');

        if fips_pch in ('02105', '02230') then fips_pch = '02232';

        if fips_pch = '02275' then fips_pch = '02201';

  /*default outcomes for lag 2014*/
        map_lag = lag_2014;
```

```
if lag_2014=. then do;
        map_lag=2;
        prediction=2;
end;

if lag_2014 = 1 and prediction=1 then outcome="TP";
else if lag_2014 = 0 and prediction=0 then outcome="TN";
else if lag_2014 = 1 and prediction=0 then outcome="FN";
else if lag_2014 = 0 and prediction=1 then outcome="FP";

if prediction = . then prediction=2;

/* outcomes for sens 2014*/

map_sens = sens_2014;

if sens_2014=. then do;
```

```
        map_sens=2;

        sens_prediction=2;

    end;


    if sens_2014 = 1 and sens_prediction=1 then sens_outcome="TP";

    else if sens_2014 = 0 and sens_prediction=0 then sens_outcome="TN";

    else if sens_2014 = 1 and sens_prediction=0 then sens_outcome="FN";

    else if sens_2014 = 0 and sens_prediction=1 then sens_outcome="FP";


    if sens_prediction = . then sens_prediction=2;


    label lag_2014 = "County has at least one municipality where voters decided to
keep cannabis sales legal"

        sens_2014 = "County has at least one municipality where voters decided
to keep cannabis sales legal, including Washington"

        outcome = "Confusion matrix classification, unweighted probabilities
with cut-off=.4"

        sens_outcome = "Confusion matrix classification, unweighted
probabilities cut-off=.4, including Washington";
```

```
run;


proc freq data=fig_wide4;
        title3"check missing";
        table state;
        where sens_prediction = 2 or prediction =2;
run;


proc freq data=fig_wide4;
        title3"check missing";
        table y lag_2014/list missing;
run;


proc sort data=fig_wide3;
        by state fips_pch;
run;


proc print data= fig_wide3 noobs n;
```

```
        by state;
        var fips_pch;
        where state in ('Alaska', 'Colorado');
run;

proc sort data=fig_wide4;
        by state fips_pch;
run;

proc print data= fig_wide4 noobs n;
        by state;
        var fips_pch;
        where state in ('Alaska', 'Colorado');
run;

proc print data= fig_wide4;
        title3'confirm duping worked';
        where fips_pch in ('02275', '02201', '02280');
```

```
run;

proc freq data=fig_wide4;
        title3'Check new variable creation';
        table lag_2014 * map_lag outcome*lag_2014*prediction /list missing;
        table sens_2014 * map_sens sens_outcome*sens_2014*sens_prediction/list
missing;
run;

proc sort data=counties2;
        by fips_pch;
run;

proc sort data=fig_wide4;
        by fips_pch;
run;

data county_pred_map;
```

```
        merge counties2 (in=in1x ) fig_wide4(in=in2x drop=state);
        by fips_pch;
        in_shp = in1x;
        in_dat = in2x;
run;

/*check data for any problems*/
proc contents data=county_pred_map;
        title3'check merged data contents for any problems';
run;

proc freq data=county_pred_map;
        title3'check that all obs merged';
        table in_dat*in_shp/list missing;
run;

proc print data=county_pred_map;
        title3'check that all obs merged (should not print)';
```

```
        where in_shp = 0;
run;

proc print data=county_pred_map;
        title3'print data for the recoded fips records';
        var name x y pht ;
        where fips_pch in ('51780','51083','51005','02201', '02280', '02232', "35013");
run;

proc print data=county_pred_map;
        title3'fips codes missing from maps.uscounty (should not print)';
        var name fips_pch;
        where state=.;
run;

proc freq data=county_pred_map;
        title3'where are missaing values from confusion matrix';
        table state*lag_2014*pht/ list missing;
```

```
        where lag_2014 = . or pht=.;
        format state state.;
run;


proc contents data=county_pred_map;
run;


data sav.full_map_set;
        set county_pred_map;
        drop in_dat in_shp in1 in2 in3;
run;


/*make map layer for thicker state borders
Remove obs that do not apply to polygon borders*/


proc gremove data=maps.uscounty out=anno_outline;
        by state notsorted;
        id county;
```

```
run;


proc gmap map=anno_outline data=anno_outline;
        id state;
        choro segment / levels=1 stat=first nolegend coutline=grayaa;
run;


/* Create annotate dataset for diff state borders*/


data state_outline;
        set anno_outline;
        by state segment notsorted ;
        length function $8 color $8;
        color='gray33'; style='mempty'; when='a'; xsys='2';
        ysys='2';
        if first.segment then function='poly';
        else function='polycont';
run;
```

```
* graphics options to be used for all maps;
goptions reset=all ftext='calibri' htext=2;

/* Specify policy formats for mapping */
proc format ;
        value lag 0 = "Recreational cannabis illegal"
                  1 = "Locals allow recreational cannabis sales"
                  2 = "Cannabis legal, no local control";
run;
goptions colors=(cream darkgreen olive);

/*create map of actual policies*/

proc gmap map=maps.uscounty data=sav.full_map_set anno=state_outline;
        id state county;
        choro map_lag/ coutline=gray discrete stat=first;
        format map_lag lag.;
```

```
        label map_lag = "Cannabis policies by county, 2014";

run;

quit;



/*create map of predicted policies*/



proc format ;

        value con

                                0 = "Predicted illegal"

                                1 = "Predicted legal"

                                2 = "Data not used";

run;

goptions colors=(cream darkgreen CX142233);



proc gmap map=maps.uscounty data=sav.full_map_set anno=state_outline;

        id state county;

        choro prediction/ coutline=gray discrete stat=first;

        format prediction con.;
```

```
        label prediction = "Predicted recreational cannabis sale policy in 2014";
        legend1 label=(f=swissb j=c 'Cases') across=1 down=4 frame;
        title h=4 color=black 'Predicted recreational cannabis sales in 2014';
run;
quit;

/*create map of outcome classifiers
        just in Colorado (08)
        Oregon (41)
        Alaska (02)*/

proc format ;
        value $class
                                "TP" = "True Positive"
                                "TN" = "True Negative"
                                "FP" = "False Positive"
                                "FN" = "False Negative"
                                "  " = "Data not used";
```

```
run;

goptions colors=(red orange cyan darkgreen white);


proc gmap map=maps.uscounty (where=(state in (2,8,41))) data=sav.full_map_set
anno=state_outline;
        id state county;
        choro outcome/ coutline=gray discrete stat=first;
        format outcome $class.;
        title ;
        label outcome = "Diagnostic classification for each county, .4 cut-point";
run;
quit;


proc univariate data=fig_wide4;
        var acc;
        histogram acc;
run;
```

```
proc format;
        value p_hat
        0.435877 - 1 = 'Top 10%'
        0.0811583 - 0.43587699 = '75 - 90th percentile'
        0.0156888 - .081158299 = '50 - 75th percentile'
        0.00420628 - 0.015688799 = '25 - 50th percentile'
        0.0016763 - 0.0042062799 = '10 - 25th percentile'
        0 - 0.001676299 = 'Bottom 10%';
run;

goptions reset=all border colors= (VLIYG LIYG MOYG DAYG VDEYG VDAYG);
proc gmap map=maps.uscounty data=sav.full_map_set anno=state_outline all;
        id state county;
        choro acc/ coutline=gray cdefault=degb statistic=first discrete;
        label acc = "Accuracy weighted model probabilities";
        format acc p_hat.;
run;
quit;
```

```
proc freq data=sav.map_set_lag_2014;
        title4'states with FPs';
        table state;
        where fg = 0 and weighted_pred3 = 1;
        format state state.;
run;


proc sql;
        title3'Why do i have missing county averages?';
        select distinct(fips_pch) as miss_avg
        from sav.map_set
        where county_average = . and state ~=53;
run;


proc sql;
        title3'Why do i have missing county averages?';
        select distinct(fips_pch) as miss_consensus
```

```
        from sav.map_set
        where consensus = . and state ~=53;
run;

/*North carolina recodes from waaaay back, same counties i have no SAEs for...*/

proc print data=sav.full_set noobs;
        title3'NC data that I could not get SAEs for';
        var state  fips_pch geography incensus incross inelection insae;
        where state = "North Carolina" and insae=0;
run;

/*Maps for Sens analysis*/
goptions colors=(cream darkgreen olive);
proc gmap map=maps.uscounty data=sav.full_map_set anno=state_outline;
        id state county;
        choro map_sens/ coutline=gray discrete stat=first;
        format map_sens lag.;
```

```
        label map_sens = "Cannabis policies by county, 2014";
run;
quit;


goptions colors=(cream darkgreen CX142233);


proc gmap map=maps.uscounty data=sav.full_map_set anno=state_outline;
        id state county;
        choro sens_prediction/ coutline=gray discrete stat=first;
        format sens_prediction con.;
        label sens_prediction = "Predicted recreational cannabis sale policy in 2014";
        legend1 label=(f=swissb j=c 'Cases') across=1 down=4 frame;
        title h=4 color=black 'Predicted recreational cannabis sales in 2014';
run;
quit;


goptions colors=(red orange cyan darkgreen white);
```

```
proc gmap map=maps.uscounty (where=(state in (2,8,41, 53))) data=sav.full_map_set
anno=state_outline;
        id state county;
        choro sens_outcome_class3/ coutline=gray discrete stat=first;
        format sens_outcome_class3 $class.;
        title ;
        label sens_outcome_class3 = "Diagnostic classification for each county, .3 cut-
point";
run;
quit;


proc univariate data=fig_wide4;
        var sens_pht;
        histogram sens_pht;
run;


proc format;
        value sens_pht
        0.539441055 - 1 = 'Top 10%'
```

```
        0.113663725 - .53944105499 = '75 - 90th percentile'

        0.027264315 - 0.11366372499 = '50 - 75th percentile'

        0.009572451 - 0.02726431499 = '25 - 50th percentile'

        0.004093802 - 0.00957245099 = '10 - 25th percentile'

        0 - 0.00409380199 = 'Bottom 10%';

run;



goptions reset=all border colors= (VLIYG LIYG MOYG DAYG VDEYG VDAYG);

proc gmap map=maps.uscounty data=sav.full_map_set anno=state_outline all;

        id state county;

        choro sens_pht/ coutline=gray cdefault=degb statistic=first discrete;

        label sens_pht = "Accuracy weighted probabilities";

        format sens_pht p_hat.;

run;

quit;

/*************************************************************************************************

* In Enterprise Guide, "Specify the page size for log and text output" under 'Results

General' must be *
```

```
* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *
*****************************************************************************************************/
OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE ;


title1'Dissertation';
title2'Aim 2: Incidence after leglaization DiD';


/*****************************************************************************************
* The following macro variables are available to all users: *
*                                                                                        *
* &Project     – the name of the project folder in which the .EGP file is stored    *
* &ProgName        – the name of the .egp file, without the extension                  *
* &ProgNode        - the name of the code node                                         *
* &ProgDir         – the path to the folder in which the .egp file is stored.          *
*****************************************************************************************/


/*************************************************************************
```

```
* The following macro variables are used in conjunction with the DOC_BLOCK   *
* macro to document programs and output.                                      *
*                                                                             *
* Do not use quotation marks when defining macro variables. If SAS syntax    *
* requires quotes, use double quotes when you reference the macro variable.  *
*****************************************************************************/
** PROGRAMMER'S NAME ;
%LET PROGRAMMER      = Barrett Montgomery;


** DEFINE ALL NON-SAS FILES CALLED IN YOUR PROGRAM AS MACRO VARIABLES
by Drug;
** THESE CAN BE LEFT BLANK IF NOT NEEDED OR USED ;
%LET Elg       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Legal Timeline Categories\ELG;
%LET Rec       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Legal Timeline Categories\REC;
%LET ElgE      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Effective Date Categories\ELG;
```

```
%LET RecE      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Effective Date Categories\REC;

%LET PMMJ      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Alt Spec\PMMJ;


** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;

** USE &PROGNAME FOR SAVEFILE NAME ;

** LEAVE BLANK IF NO DATASET SAVED ;

%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Processed;


** NAME FORMAT LIBRARY DIRECTORY ;

%LET FMTDIR       =  ;


** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT
WORK ;

%LET PURPOSE1 = Difference in difference event study design to estimate the effect
of cannabis legalization on cannabis incidence;
```

```
/************************************************************

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

************************************************************/



** INPUT FILES ;



** OUTPUT FILE DESTINATION ;

LIBNAME SAV  "&SAVEDIR1" ;



/*this macro imports all data in a file with 2 options, the folder directory, and the type of

file

here, folder is saved in a macro variable named above and file type is csv.*/



title1"Download all data, append, and organize";

%macro drive(dir,ext);

  %local cnt filrf rc did memcnt name;

  %let cnt=0;
```

```
%let filrf=mydir;

%let rc=%sysfunc(filename(filrf,&dir));

%let did=%sysfunc(dopen(&filrf));

 %if &did ne 0 %then %do;

%let memcnt=%sysfunc(dnum(&did));



 %do i=1 %to &memcnt;



  %let name=%qscan(%qsysfunc(dread(&did,&i)),-1,.);



  %if %qupcase(%qsysfunc(dread(&did,&i))) ne %qupcase(&name) %then %do;

   %if %superq(ext) = %superq(name) %then %do;

     %let cnt=%eval(&cnt+1);

     %put %qsysfunc(dread(&did,&i));

     proc import datafile="&dir\%qsysfunc(dread(&did,&i))" out=dsn&cnt

      dbms=csv replace;

     run;
```

```
      %end;
     %end;


   %end;
    %end;
 %else %put &dir cannot be open.;
 %let rc=%sysfunc(dclose(&did));


%mend drive;


/*read in eligibility data for leglaization dates*/
 %drive(&Elg,csv)


data dsn1;
        set dsn1;
        years="2008 to 2009";
run;
```

```
data dsn2;
        set dsn2;
        years="2010 to 2011";
run;

data dsn3;
        set dsn3;
        years="2012 to 2013";
run;

data dsn4;
        set dsn4;
        years="2014 to 2015";
run;

data dsn5;
        set dsn5;
        years="2016 to 2017";
```

```
run;



data dsn6;
        set dsn6;
        years="2018 to 2019";
run;



data Elig;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
  eligible = catx('', 'eligible for past year initiatio'n, 'rc-eligible for past year initia'n);
run;



proc freq data=Elig;
        title3'check to see that new var works';
        table eligible*'eligible for past year initiatio'n* 'rc-eligible for past year initia'n/list
missing;
run;
```

```
/*read in for recmj data*/

%drive(&Rec,csv)



 data dsn1;

        set dsn1;

        years="2008 to 2009";

run;



data dsn2;

        set dsn2;

        years="2010 to 2011";

run;



data dsn3;

        set dsn3;

        years="2012 to 2013";

run;
```

```
data dsn4;
        set dsn4;
        years="2014 to 2015";
run;

data dsn5;
        set dsn5;
        years="2016 to 2017";
run;

data dsn6;
        set dsn6;
        years="2018 to 2019";
run;

data Rec;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
```

```
        recent_initiate = catx('','rc-past year initiate of marijua'n, 'rc-recent initiate of
marijuana'n, 'recent initiate of marijuana use'n);
run;


/*cleaning*/
/*prepare susbets of data where eligibility and rec = yes
to merge and create incidence estimates
 first for legal date then for effective dates*/
data elig1;
        set elig;
        where eligible = "1 - Yes";
        drop row: total: 'eligible for past year initiatio'n 'rc-eligible for past year initia'n
eligible 'Column % CI (lower)'n 'Column % CI (upper)'n;


        rename 'STATE NAME'n = legal_cat;
        rename 'final edited age'n = age;
        rename 'Column %'n = elig_phat;
        rename 'Column % SE'n = elig_phat_SE;
        rename 'Weighted Count'n = elig_count;
```

```
        rename 'Count SE'n = elig_count_se;


run;


data rec1;
        set rec;
        where recent_initiate = "1 - Yes";
        drop row: total: 'rc-past year initiate of marijua'n 'rc-recent initiate of marijuana'n
'recent initiate of marijuana use'n recent_initiate 'Column % CI (lower)'n 'Column % CI
(upper)'n;
        rename 'STATE NAME'n = legal_cat;
        rename 'final edited age'n = age;
        rename 'Column %'n = rec_phat;
        rename 'Column % SE'n = rec_phat_SE;
        rename 'Weighted Count'n = rec_count;
        rename 'Count SE'n = rec_count_se;
run;


proc sort data=elig1;
```

```
        by years legal_cat age ;
run;


proc sort data=rec1;
        by years legal_cat age ;
run;


/*estimate incidence by age group, state group, and years*/
data all_data;
        merge elig1 rec1;
        by years legal_cat age ;
        length time year_n 8.;


        if legal_cat="Not_Legalized_" then legal_cat="Illegal";
        incidence = rec_count/elig_count;


/*fixed effects for years as categorical may function differently
        if years was considered a continuous variable, year_n is to check this*/
```

```
        if years = '2008 to 2009' then year_n = 2008;

        else if years = '2010 to 2011' then year_n = 2010;

        else if years = '2012 to 2013' then year_n = 2012;

        else if years = '2014 to 2015' then year_n = 2014;

        else if years = '2016 to 2017' then year_n = 2016;

        else if years = '2018 to 2019' then year_n = 2018;


        /*need to calculate incidence SE or 95%CIs*/


        label incidence = "Percentage of past year initiates among persons at risk for
initiation"
                        legal_cat = "Legal status of cannabis through 2018"

                        year_n = "Numeric date of data, first year in year-pair";
run;


proc print data=all_data (obs=6);

        title3'check incidence calcs';
```

```
run;



proc freq data=all_data;
        title3'Check Time recode';
        table year_n*years/list missing;
run;



proc freq data=all_data;
        title3'combinations of legal cat and years to make relative time variable';
        table legal_cat* years/list;
        where legal_cat ~="Overall";
run;



proc means data=all_data;
        title3'average incidence in legal states';
        var incidence;
        where legal_cat not in("Overall", "Illegal");
run;
```

```
proc means data=all_data;
        title3'average incidence overall';
        var incidence;
        where legal_cat ="Overall" and age="Overall";
run;

proc means data=all_data;
        title3'average incidence 21+';
        var incidence;
        where legal_cat ="Overall" and age="21_Plus";
run;

data all_data2;
        set all_data;
        where legal_cat ~="Overall";

        /*create a time variable relative to year of legalization and year of data
```

time = how many years away froma states year of legalization is this data point?*/

if legal_cat = "Illegal" then time=**0**;

else if legal_cat = "Legalized_2012" and years = "2008 to 2009" then time = -**4**;

else if legal_cat = "Legalized_2014" and years = "2008 to 2009" then time = -**6**;

else if legal_cat = "Legalized_2016" and years = "2008 to 2009" then time = -**8**;

else if legal_cat = "Legalized_2018" and years = "2008 to 2009" then time = -**10**;

else if legal_cat = "Legalized_2012" and years = "2010 to 2011" then time = -**2**;

else if legal_cat = "Legalized_2014" and years = "2010 to 2011" then time = -**4**;

else if legal_cat = "Legalized_2016" and years = "2010 to 2011" then time = -**6**;

else if legal_cat = "Legalized_2018" and years = "2010 to 2011" then time = -**8**;

else if legal_cat = "Legalized_2012" and years = "2012 to 2013" then time = **0**;

else if legal_cat = "Legalized_2014" and years = "2012 to 2013" then time = -**2**;

else if legal_cat = "Legalized_2016" and years = "2012 to 2013" then time = -**4**;

else if legal_cat = "Legalized_2018" and years = "2012 to 2013" then time = -**6**;

```
        else if legal_cat = "Legalized_2012" and years = "2014 to 2015" then time = 2;

        else if legal_cat = "Legalized_2014" and years = "2014 to 2015" then time = 0;

        else if legal_cat = "Legalized_2016" and years = "2014 to 2015" then time = -2;

        else if legal_cat = "Legalized_2018" and years = "2014 to 2015" then time = -4;


        else if legal_cat = "Legalized_2012" and years = "2016 to 2017" then time = 4;

        else if legal_cat = "Legalized_2014" and years = "2016 to 2017" then time = 2;

        else if legal_cat = "Legalized_2016" and years = "2016 to 2017" then time = 0;

        else if legal_cat = "Legalized_2018" and years = "2016 to 2017" then time = -2;


        else if legal_cat = "Legalized_2012" and years = "2018 to 2019" then time = 6;

        else if legal_cat = "Legalized_2014" and years = "2018 to 2019" then time = 4;

        else if legal_cat = "Legalized_2016" and years = "2018 to 2019" then time = 2;

        else if legal_cat = "Legalized_2018" and years = "2018 to 2019" then time = 0;


/*create dummy variable for each relative time point*/

        if time=-10 then tminus10 =1; else tminus10=0;
```

```
        if time=-8 then tminus8 =1; else tminus8=0;

        if time=-6 then tminus6 =1; else tminus6=0;

        if time=-4 then tminus4 =1; else tminus4=0;

        if time=-2 then tminus2 =1; else tminus2=0;

        if time=0 then t0 =1; else t0=0;

        if time=2 then t2 =1; else t2=0;

        if time=4 then t4 =1; else t4=0;

        if time=6 then t6 =1; else t6=0;


        label    time = "Time relative to legalization based on beginning of year pair";

run;


proc contents data=all_data2;

run;


proc freq data=all_data2;

        title3'check new variable creation';

        table time*legal_cat*years
```

```
                time*tminus10*tminus8*tminus6*tminus4*tminus2*t0*t2*t4*t6/list missing;
        where legal_cat ~="Overall";
run;

proc freq data=all_data2;
        title3'determine where tails of legal date data distribution should be grouped
together';
        table time/list;
run;
/*<=-4, >=4*/

proc freq data=all_data2;
        title3'Check new time event dummies';
        table legal_cat/list missing;
run;

/* data step 3 creates new dummies that
can categorize all data before a certain time point
```

```
so that analysis on a single category is not done

 econs call it balancing leads and lags for short*/

data all_data3;

        set all_data2;



        if time<=-4 then tlt4=1; else tlt4=0;

        if time>=4 then tgt4=1; else tgt4=0;



        if time<=-6 then tlt6=1; else tlt6=0;

        if time>=6 then tgt6=1; else tgt6=0;



        if legal_cat = "Illegal" then legal = 0;

        else if time>=0 then legal = 1;

        else if time <0 then legal =0;



        if legal_cat = "Illegal" then effective = 0;

        else if time>=2 then effective = 1;

        else if time <2 then effective =0;
```

```
        if legal_cat = "Illegal" then legal_wave = 0;

        else if legal_cat = "Legalized_2012" then legal_wave = 1;

        else if legal_cat = "Legalized_2014" then legal_wave = 2;

        else if legal_cat = "Legalized_2016" then legal_wave = 3;

        else if legal_cat = "Legalized_2018" then legal_wave = 4;



        label legal = "Simple binary for RCL, 1 if year>=legalize date, 0 otherwise"

                        effective = "Simple binary for RCL effective, 1 if year>=effective
date, 0 otherwise";

run;



proc freq data=all_data3;

        title3'Check new time event dummies';

        table time*tlt4*tminus10*tminus8*tminus6*tminus4*tminus2*t0*t2*t4*t6*tgt4/list
missing;

run;



proc freq data=all_data3;
```

```
        title3'Check new legality dummy';
        table legal_cat*years*legal/list missing;
run;

proc freq data=all_data3;
        title3'Check new effective dummy';
        table legal_cat*years*effective/list missing;
run;

/*Save dataset*/
/*data sav.Legal_date;*/
/*      set all_data3;*/
/*run;*/
/**/
/*/*export to csv if needed*/*/
/*proc export  data= sav.Legal_date outfile='C:\Users\montg\Dropbox\Ph.D
Work\Dissertation\Data\Aim 2\Processed\Legal_date.csv'*/
/*      dbms=csv replace;*/
```

```
/*run;*/



/*/*/*/*/*/*/*/**/*/*/*/*/*/*/*/*/

/* REPEAT FOR EFFECTIVE DATE DATA */

/*/*/*/*/*/*/*/*/**/*/*/*/*/*/*/*/*/;



/*read in eligibility data for effective dates*/

%drive(&ElgE,csv)



data dsn1;

        set dsn1;

        years="2008 to 2009";

run;



data dsn2;

        set dsn2;

        years="2010 to 2011";

run;
```

```
data dsn3;
        set dsn3;
        years="2012 to 2013";
run;

data dsn4;
        set dsn4;
        years="2014 to 2015";
run;

data dsn5;
        set dsn5;
        years="2016 to 2017";
run;

data dsn6;
        set dsn6;
```

```
        years="2018 to 2019";
run;


data EligE;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
  eligible = catx('', 'eligible for past year initiatio'n, 'rc-eligible for past year initia'n);
run;


proc freq data=EligE;
        title3'check to see that new var works';
        table eligible*'eligible for past year initiatio'n* 'rc-eligible for past year initia'n/list
missing;
run;


%drive(&RecE,csv)


data dsn1;
        set dsn1;
```

```
        years="2008 to 2009";
run;


data dsn2;
        set dsn2;
        years="2010 to 2011";
run;


data dsn3;
        set dsn3;
        years="2012 to 2013";
run;


data dsn4;
        set dsn4;
        years="2014 to 2015";
run;
```

```
data dsn5;
        set dsn5;
        years="2016 to 2017";
run;

data dsn6;
        set dsn6;
        years="2018 to 2019";
run;

data RecE;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
        recent_initiate = catx('','rc-past year initiate of marijua'n, 'rc-recent initiate of
marijuana'n, 'recent initiate of marijuana use'n);
run;

/*cleaning*/
```

```
proc freq data=rece;
        title3'check to see that new var works';
        table recent_initiate*'rc-past year initiate of marijua'n *'rc-recent initiate of
marijuana'n *'recent initiate of marijuana use'n /list missing;
run;


proc print data=elige;
        title3'May not even need to use subtraction method, check for suppression
(should not print)';
        where 'Weighted Count'n = . or 'Count SE'n =.;
run;


proc print data=rece;
        title3'May not even need to use subtraction method, check for suppression
(should not print)';
        where 'Weighted Count'n = . or 'Count SE'n =.;
run;
/*no data from 2008-2009 period in ALaska, subtraction method doesnt even work*/
```

```
data elige1;
        set elige;
        where eligible = "1 - Yes";
        drop row: total: 'eligible for past year initiatio'n 'rc-eligible for past year initia'n
eligible 'Column % CI (lower)'n 'Column % CI (upper)'n;

        rename 'STATE NAME'n = legal_cat;
        rename 'final edited age'n = age;
        rename 'Column %'n = elig_phat;
        rename 'Column % SE'n = elig_phat_SE;
        rename 'Weighted Count'n = elig_count;
        rename 'Count SE'n = elig_count_se;

run;

data rece1;
        set rece;
        where recent_initiate = "1 - Yes";
```

```
        drop row: total: 'rc-past year initiate of marijua'n 'rc-recent initiate of marijuana'n
'recent initiate of marijuana use'n recent_initiate 'Column % CI (lower)'n 'Column % CI
(upper)'n;

        rename 'STATE NAME'n = legal_cat;

        rename 'final edited age'n = age;

        rename 'Column %'n = rec_phat;

        rename 'Column % SE'n = rec_phat_SE;

        rename 'Weighted Count'n = rec_count;

        rename 'Count SE'n = rec_count_se;

run;



proc sort data=elige1;

        by years legal_cat age ;

run;



proc sort data=rece1;

        by years legal_cat age ;

run;
```

```
/*estimate incidence by age group, state group, and years*/

data all_datae;

        merge elige1 rece1;

        by years legal_cat age ;

        length time 8.;



        incidence = rec_count/elig_count;



        /*need to calculate incidence SE or 95%CIs*/



        label incidence = "Percentage of past year initiates among persons at risk for
initiation"

                        legal_cat = "Legal status of cannabis through 2018";

run;



proc freq data=all_datae;

        title3'combinations of legal cat and years to make relative time variable';

        table years*legal_cat /list;
```

```
        where legal_cat ~="Overall";

run;



data all_datae2;

        set all_datae;

        where legal_cat ~="Overall";

        /*create a time variable relative to year of legalization and year of data

                time = how many years away froma states year of legalization is this data
point?*/



        if legal_cat = "Illegal" then time=0;

        else if legal_cat = "Effective_2014" and years = "2008 to 2009" then time = -6;

        else if legal_cat = "Effective_2015" and years = "2008 to 2009" then time = -7;

        else if legal_cat = "Effective_2016" and years = "2008 to 2009" then time = -8;

        else if legal_cat = "Effective_2017" and years = "2008 to 2009" then time = -9;

        else if legal_cat = "Effective_2018" and years = "2008 to 2009" then time = -10;



        else if legal_cat = "Effective_2014" and years = "2010 to 2011" then time = -4;
```

else if legal_cat = "Effective_2015" and years = "2010 to 2011" then time = -**5**;

else if legal_cat = "Effective_2016" and years = "2010 to 2011" then time = -**6**;

else if legal_cat = "Effective_2017" and years = "2010 to 2011" then time = -**7**;

else if legal_cat = "Effective_2018" and years = "2010 to 2011" then time = -**8**;

else if legal_cat = "Effective_2014" and years = "2012 to 2013" then time = -**2**;

else if legal_cat = "Effective_2015" and years = "2012 to 2013" then time = -**3**;

else if legal_cat = "Effective_2016" and years = "2012 to 2013" then time = -**4**;

else if legal_cat = "Effective_2017" and years = "2012 to 2013" then time = -**5**;

else if legal_cat = "Effective_2018" and years = "2012 to 2013" then time = -**6**;

else if legal_cat = "Effective_2014" and years = "2014 to 2015" then time = **0**;

else if legal_cat = "Effective_2015" and years = "2014 to 2015" then time = -**1**;

else if legal_cat = "Effective_2016" and years = "2014 to 2015" then time = -**2**;

else if legal_cat = "Effective_2017" and years = "2014 to 2015" then time = -**3**;

else if legal_cat = "Effective_2018" and years = "2014 to 2015" then time = -**4**;

else if legal_cat = "Effective_2014" and years = "2016 to 2017" then time = **2**;

```
        else if legal_cat = "Effective_2015" and years = "2016 to 2017" then time = 1;

        else if legal_cat = "Effective_2016" and years = "2016 to 2017" then time = 0;

        else if legal_cat = "Effective_2017" and years = "2016 to 2017" then time = -1;

        else if legal_cat = "Effective_2018" and years = "2016 to 2017" then time = -2;


        else if legal_cat = "Effective_2014" and years = "2018 to 2019" then time = 4;

        else if legal_cat = "Effective_2015" and years = "2018 to 2019" then time = 3;

        else if legal_cat = "Effective_2016" and years = "2018 to 2019" then time = 2;

        else if legal_cat = "Effective_2017" and years = "2018 to 2019" then time = 1;

        else if legal_cat = "Effective_2018" and years = "2018 to 2019" then time = 0;


/*create dummy variable for each relative time point*/

        if time=-10 then tminus10 =1; else tminus10=0;

        if time=-9 then tminus9 =1; else tminus9=0;

        if time=-8 then tminus8 =1; else tminus8=0;

        if time=-7 then tminus7 =1; else tminus7=0;

        if time=-6 then tminus6 =1; else tminus6=0;

        if time=-5 then tminus5 =1; else tminus5=0;
```

```
        if time=-4 then tminus4 =1; else tminus4=0;

        if time=-3 then tminus3 =1; else tminus3=0;

        if time=-2 then tminus2 =1; else tminus2=0;

        if time=-1 then tminus1 =1; else tminus1=0;

        if time=0 then t0 =1; else t0=0;

        if time=1 then t1 =1; else t1=0;

        if time=2 then t2 =1; else t2=0;

        if time=3 then t3 =1; else t3=0;

        if time=4 then t4 =1; else t4=0;


        label    time = "Time relative to legalization based on beginning of year pair";

run;


proc freq data=all_datae2;

        title3'check new variable creation';

        table time time*legal_cat*years


        time*tminus10*tminus9*tminus8*tminus7*tminus6*tminus5*tminus4*tminus3*tmi
nus2*tminus1*t0*t1*t2*t3*t4/list missing;
```

```
run;


proc sort data=all_datae2;
        by time;
run;


proc print data=all_datae2;
        title3'Look at data sorted by time to see how much info we are missing in this
coding scheme';
        var time legal_cat incidence;
        where legal_cat ~="Overall";
run;
/*Just the Alaskan data from 2008-2009 I already knew about*/


proc freq data=all_datae2;
        title3'determine where tails of effective date data distribution should be grouped
together';
        table time/list;
run;
```

```
/*<=-6, >=2*/

/* data step 3 creates new dummies that

can categorize all data before a certain time point

so that analysis on a single category is not done

 econs call it balancing leads and lags for short*/



/*the cut points are different for legalization dates

than for effective dates because of the way the data

is structured and the categorization by 1 year intervals

(more effective date 1 year interval categories)*/



data all_datae3;

        set all_datae2;



        if time<=-6 then tlt6=1; else tlt6=0;

        if time>=2 then tgt2=1; else tgt2=0;

run;
```

```
proc freq data=all_datae3;
        title3'Check new time event dummies';
        table
time*tlt6*tminus10*tminus9*tminus8*tminus7*tminus6*tminus5*tminus4*tminus3*tminus
2*tminus1*t0*t1*t2*t3*t4*tgt2/list missing;
run;

/*Save dataset*/

/*data sav.Effective_date;*/
/*      set all_datae3;*/
/*run;*/

/*export to csv if needed*/
/*proc export  data= sav.Effective_date outfile='C:\Users\montg\Dropbox\Ph.D
Work\Dissertation\Data\Aim 2\Processed\Effective_date.csv'*/
/*      dbms=csv replace;*/
/*run;*/
```

```
/*/*/*/*/*/*/*/*/*/**/*/*/*/*/*/*/*/*/*/*/

/* REPEAT FOR PREVALENCE LEGAL DATE DATA  */

/*/*/*/*/*/*/*/*/*/**/*/*/*/*/*/*/*/*/*/*/;



/*read in prevalence data for leglaization dates*/

 %drive(&PMMJ,csv)



 data dsn1;

          set dsn1;

          years="2008 to 2009";

run;



data dsn2;

          set dsn2;

          years="2010 to 2011";

run;
```

```
data dsn3;
        set dsn3;
        years="2012 to 2013";
run;

data dsn4;
        set dsn4;
        years="2014 to 2015";
run;

data dsn5;
        set dsn5;
        years="2016 to 2017";
run;

data dsn6;
        set dsn6;
        years="2018 to 2019";
```

```
run;



data prev;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
run;



/*read in for recmj data*/
%drive(&Rec,csv)



 data dsn1;
          set dsn1;
          years="2008 to 2009";
run;



data dsn2;
          set dsn2;
          years="2010 to 2011";
run;
```

```
data dsn3;
        set dsn3;
        years="2012 to 2013";
run;

data dsn4;
        set dsn4;
        years="2014 to 2015";
run;

data dsn5;
        set dsn5;
        years="2016 to 2017";
run;

data dsn6;
        set dsn6;
```

```
        years="2018 to 2019";
run;


data prev;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
        past_month_use = catx('','RC-MARIJUANA - PAST MONTH USE'n,
'MARIJUANA - PAST MONTH USE'n);
run;


proc freq data=prev;
        title3'check that prevalence indicator concatenated correctly';
        table past_month_use*'RC-MARIJUANA - PAST MONTH USE'n*'MARIJUANA -
PAST MONTH USE'n / list missing;
run;


/*cleaning*/
/*prepare susbets of data for past month users*/
data prev1;
        set prev;
```

```
        where past_month_use = "1 - Used within the past month";

        length time 8.;

        drop row: total: 'RC-MARIJUANA - PAST MONTH USE'n 'MARIJUANA - PAST
MONTH USE'n  'Column % CI (lower)'n 'Column % CI (upper)'n;



        /*need to calculate prevalence SE or 95%CIs?*/

        rename 'STATE NAME'n = legal_cat;

        rename 'final edited age'n = age;

        rename 'Column %'n = pmmj_phat;

        rename 'Column % SE'n = pmmj_SE;

        rename 'Weighted Count'n = pmmj_count;

        rename 'Count SE'n = pmmj_count_se;



        label pmmj_phat = "Percentage of population that used marijuana past month"

                        legal_cat = "Legal status of cannabis through 2018";



run;
```

```
proc means data=prev1 n mean stddev;
        title3'face validity check for prevalence estimates';
        var pmmj_phat;
        class legal_cat age;
run;


data prev2;
        set prev1;
        /*remove overall state category*/
        where legal_cat ~="Overall";

        /*create a time variable relative to year of legalization and year of data
                time = how many years away froma states year of legalization is this data
point?*/

        if legal_cat = "Illegal" then time=0;
        else if legal_cat = "Legal_2012" and years = "2008 to 2009" then time = -4;
        else if legal_cat = "Legal_2014" and years = "2008 to 2009" then time = -6;
```

else if legal_cat = "Legal_2016" and years = "2008 to 2009" then time = -**8**;

else if legal_cat = "Legal_2018" and years = "2008 to 2009" then time = -**10**;

else if legal_cat = "Legal_2012" and years = "2010 to 2011" then time = -**2**;

else if legal_cat = "Legal_2014" and years = "2010 to 2011" then time = -**4**;

else if legal_cat = "Legal_2016" and years = "2010 to 2011" then time = -**6**;

else if legal_cat = "Legal_2018" and years = "2010 to 2011" then time = -**8**;

else if legal_cat = "Legal_2012" and years = "2012 to 2013" then time = **0**;

else if legal_cat = "Legal_2014" and years = "2012 to 2013" then time = -**2**;

else if legal_cat = "Legal_2016" and years = "2012 to 2013" then time = -**4**;

else if legal_cat = "Legal_2018" and years = "2012 to 2013" then time = -**6**;

else if legal_cat = "Legal_2012" and years = "2014 to 2015" then time = **2**;

else if legal_cat = "Legal_2014" and years = "2014 to 2015" then time = **0**;

else if legal_cat = "Legal_2016" and years = "2014 to 2015" then time = -**2**;

else if legal_cat = "Legal_2018" and years = "2014 to 2015" then time = -**4**;

```
        else if legal_cat = "Legal_2012" and years = "2016 to 2017" then time = 4;

        else if legal_cat = "Legal_2014" and years = "2016 to 2017" then time = 2;

        else if legal_cat = "Legal_2016" and years = "2016 to 2017" then time = 0;

        else if legal_cat = "Legal_2018" and years = "2016 to 2017" then time = -2;


        else if legal_cat = "Legal_2012" and years = "2018 to 2019" then time = 6;

        else if legal_cat = "Legal_2014" and years = "2018 to 2019" then time = 4;

        else if legal_cat = "Legal_2016" and years = "2018 to 2019" then time = 2;

        else if legal_cat = "Legal_2018" and years = "2018 to 2019" then time = 0;


/*create dummy variable for each relative time point*/

        if time=-10 then tminus10 =1; else tminus10=0;

        if time=-8 then tminus8 =1; else tminus8=0;

        if time=-6 then tminus6 =1; else tminus6=0;

        if time=-4 then tminus4 =1; else tminus4=0;

        if time=-2 then tminus2 =1; else tminus2=0;

        if time=0 then t0 =1; else t0=0;

        if time=2 then t2 =1; else t2=0;
```

```
        if time=4 then t4 =1; else t4=0;

        if time=6 then t6 =1; else t6=0;



        label    time = "Time relative to legalization based on beginning of year pair";

run;



proc freq data=prev2;

        title3'combinations of legal cat and years to make relative time variable';

        table legal_cat* years/list;

run;



proc freq data=prev2;

        title3'check new variable creation';

        table time*legal_cat*years

                time*tminus10*tminus8*tminus6*tminus4*tminus2*t0*t2*t4*t6/list missing;

run;



proc freq data=prev2;
```

```
        title3'determine where tails of legal date data distribution should be grouped
together';
        table time/list;
run;
/*<=-4, >=4*/


/* data step 3 creates new dummies that
can categorize all data before a certain time point
so that analysis on a single category is not done
 econs call it balancing leads and lags for short*/
data prev3;
        set prev2;


        if time<=-4 then tlt4=1; else tlt4=0;
        if time>=4 then tgt4=1; else tgt4=0;
run;


proc freq data=prev3;
```

```
        title3'Check new time event dummies';

        table time*tlt4*tminus10*tminus8*tminus6*tminus4*tminus2*t0*t2*t4*t6*tgt4/list
missing;

run;



/*Save dataset*/

/*data sav.Prevalence;*/

/*        set prev3;*/

/*run;*/



/*export to csv if needed*/

/*proc export  data= sav.Prevalence outfile='C:\Users\montg\Dropbox\Ph.D
Work\Dissertation\Data\Aim 2\Processed\Prevalence.csv'*/

/*        dbms=csv replace;*/

/*run;*/



/*/*/*/*/*/*/*/**/*/*/*/*/*/*/*/

/*Placebo Analysis Dataset*/*/*/

/*/*/*/*/*/*/*/**/*/*/*/*/*/*/*/;
```

```
data time_placebo;
        set sav.Legal_date;

        if legal_cat = "Illegal" then legal=0;
        else if legal_cat = "Overall" then legal=.;
        else if time=0 and legal_cat ~= "Illegal" then legal=1;
        else if time<0 then legal=0;
        else if time>0 then legal=1;

        label    legal = "cannabis leagality binary";
run;

proc freq data= time_placebo;
        title3'check legal binary creation';
        table year_n*legal_cat*legal/list missing;
run;
```

```
/*randomly select a year to be the placebo time*/



/* pick a random number between 2010 and 2016 1000 times*/

data year_range;

        do i = 1 to 7;

          placebo_year = 2009 + i;

                id=1;

        output;

        end;



        keep id placebo_year;

run;



proc surveyselect data=year_range sampsize=1

        method=srs reps=1 out=placebos seed = 1124;

run;



data time_placebo2;
```

```
        set time_placebo;
        id=1;
run;

data time_placebo3;
        merge time_placebo2 placebos;
        by id;
        drop replicate id;
run;

proc print;
        var legal_cat year_n time ;
run;

data time_placebo4;
        merge time_placebo3;

        if legal_cat = "Illegal" then placebo_time=0;
```

```
        else if legal_cat ~= "Illegal" and years = "2008 to 2009" then placebo_time = -3;

        else if legal_cat ~= "Illegal" and years = "2010 to 2011" then placebo_time = -1;

        else if legal_cat ~= "Illegal" and years = "2012 to 2013" then placebo_time = 1;

        else if legal_cat ~= "Illegal" and years = "2014 to 2015" then placebo_time = 3;

        else if legal_cat ~= "Illegal" and years = "2016 to 2017" then placebo_time = 5;

        else if legal_cat ~= "Illegal" and years = "2018 to 2019" then placebo_time = 7;


/*create dummy variable for each relative time point*/


        if placebo_time=-3 then ptminus3 =1; else ptminus3=0;

        if placebo_time=-1 then ptminus1 =1; else ptminus1=0;

        if placebo_time=1 then pt1 =1; else pt1=0;

        if placebo_time=3 then pt3 =1; else pt3=0;

        if placebo_time=5 then pt5 =1; else pt5=0;

        if placebo_time=7 then pt7 =1; else pt7=0;

        if placebo_time=0 then pt0 =1; else pt0=0;


        if legal_cat= "Illegal" then placebo_legal = "Illegal";
```

```
        else  placebo_legal = "Legal";


        label placebo_time = "Time between observation and placebo year of cannabis
legalization"
                        placebo_legal = "2 groups for the plcebo trial, legalized in 2011
and illegal";
run;


proc freq data=time_placebo4;
        title3'check placebo time variable and placebo time binaries';
        table legal_cat*years*placebo_time
        placebo_time legal_cat*placebo_legal
        placebo_time*ptminus3*ptminus1*pt0*pt1*pt3*pt5*pt7/list missing;
run;


data sav.placebo;
        set time_placebo4;
run;
/*************************************************************************************************
```

```
* In Enterprise Guide, "Specify the page size for log and text output" under 'Results
General' must be *
* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *
*****************************************************************************************************/
OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE ;


title1'Dissertation';
title2'Aim 2: Table 1';


/*****************************************************************************************
* The following macro variables are available to all users: *
*                                                                *
* &Project     – the name of the project folder in which the .EGP file is stored    *
* &ProgName       – the name of the .egp file, without the extension                *
* &ProgNode       - the name of the code node                                       *
* &ProgDir        – the path to the folder in which the .egp file is stored.        *
*****************************************************************************************/
```

```
/*************************************************************************
* The following macro variables are used in conjunction with the DOC_BLOCK   *
* macro to document programs and output.                                     *
*                                                                            *
* Do not use quotation marks when defining macro variables. If SAS syntax    *
* requires quotes, use double quotes when you reference the macro variable.  *
*************************************************************************/
** PROGRAMMER'S NAME ;
%LET PROGRAMMER      = Barrett Montgomery;


** DEFINE ALL NON-SAS FILES CALLED IN YOUR PROGRAM AS MACRO VARIABLES
by Drug;
** THESE CAN BE LEFT BLANK IF NOT NEEDED OR USED ;
%LET dat     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Table 1;
%LET ABOD    = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Table 1\PDAS\ABODMRJ;
```

```
%LET AGE      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Table 1\PDAS\CATAGE;

%LET SEX      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Table 1\PDAS\IRSEX;

%LET MRJ      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Table 1\PDAS\MRJMON;

%LET RACE     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Table 1\PDAS\NEWRACE2;


** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;

** USE &PROGNAME FOR SAVEFILE NAME ;

** LEAVE BLANK IF NO DATASET SAVED ;

%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Processed;


** NAME FORMAT LIBRARY DIRECTORY ;

%LET FMTDIR       =  ;


** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;
```

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT WORK ;

%LET PURPOSE1 = Aim 2 sample demographics and whatnot;

/***

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

***/

** INPUT FILES ;

** OUTPUT FILE DESTINATION ;

LIBNAME SAV "&SAVEDIR1" ;

/*this macro imports all data in a file with 2 options, the folder directory, and the type of file

here, folder is saved in a macro variable named above and file type is csv.*/

title1"Download all data, append, and organize";

%macro drive(dir,ext);

```
%local cnt filrf rc did memcnt name;
%let cnt=0;


%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,&dir));
%let did=%sysfunc(dopen(&filrf));
 %if &did ne 0 %then %do;
%let memcnt=%sysfunc(dnum(&did));


 %do i=1 %to &memcnt;


  %let name=%qscan(%qsysfunc(dread(&did,&i)),-1,.);


  %if %qupcase(%qsysfunc(dread(&did,&i))) ne %qupcase(&name) %then %do;
   %if %superq(ext) = %superq(name) %then %do;
     %let cnt=%eval(&cnt+1);
     %put %qsysfunc(dread(&did,&i));
     proc import datafile="&dir\%qsysfunc(dread(&did,&i))" out=dsn&cnt
```

```
          dbms=csv replace;
         run;
       %end;
      %end;


    %end;
     %end;
  %else %put &dir cannot be open.;
  %let rc=%sysfunc(dclose(&did));


 %mend drive;


/*read in data*/
 %drive(&dat,csv)


data dsn1;
          set dsn1 ;
          years="2008 to 2009";
```

```
run;


data dsn2;
        set dsn2;
        years="2008 to 2009";
run;


data dsn3;
        set dsn3;
        years="2008 to 2009";
run;


data dsn4;
        set dsn4;
        years="2008 to 2009";
run;


data dsn5;
```

```
        set dsn5;
        years="2008 to 2009";
run;

data dsn6;
        set dsn6;
        years="2010 to 2011";
run;

data dsn7;
        set dsn7;
        years="2010 to 2011";
run;

data dsn8;
        set dsn8;
        years="2010 to 2011";
run;
```

```
data dsn9;
        set dsn9;
        years="2010 to 2011";
run;

data dsn10;
        set dsn10;
        years="2010 to 2011";
run;

data dsn11;
        set dsn11;
        years="2012 to 2013";
run;

data dsn12;
        set dsn12;
```

```
        years="2012 to 2013";
run;


data dsn13;
        set dsn13;
        years="2012 to 2013";
run;


data dsn14;
        set dsn14;
        years="2012 to 2013";
run;


data dsn15;
        set dsn15;
        years="2012 to 2013";
run;
```

```
data dsn16;
        set dsn16;
        years="2014 to 2015";
run;

data dsn17;
        set dsn17;
        years="2014 to 2015";
run;

data dsn18;
        set dsn18;
        years="2014 to 2015";
run;

data dsn19;
        set dsn19;
        years="2014 to 2015";
```

```
run;



data dsn20;
        set dsn20;
        years="2014 to 2015";
run;



data dsn21;
        set dsn21;
        years="2016 to 2017";
run;



data dsn22;
        set dsn22;
        years="2016 to 2017";
run;



data dsn23;
```

```
        set dsn23;
        years="2016 to 2017";
run;

data dsn24;
        set dsn24;
        years="2016 to 2017";
run;

data dsn25;
        set dsn25;
        years="2016 to 2017";
run;

data dsn26;
        set dsn26;
        years="2018 to 2019";
run;
```

```
data dsn27;
        set dsn27;
        years="2018 to 2019";
run;

data dsn28;
        set dsn28;
        years="2018 to 2019";
run;

data dsn29;
        set dsn29;
        years="2018 to 2019";
run;

data dsn30;
        set dsn30;
```

```
        years="2018 to 2019";

run;



data dat2008;

  set dsn1 dsn2 dsn3 dsn4 dsn5;

  keep years 'state name'n 'final edited age'n 'marijuana abuse or dependence -'n

                                'imputation revised gender'n 'marijuana - past month

use'n

                                'race recode (4 levels)'n 'Weighted Count'n 'Count SE'n

                                'Column %'n  'Column % SE'n;



;

  rename 'final edited age'n  = 'FINAL EDITED AGE'n

                            'marijuana abuse or dependence -'n = 'RC-MARIJUANA

DEPENDENCE OR ABUSE'n

                            'imputation revised gender'n = 'GENDER - IMPUTATION

REVISED'n

                             'marijuana - past month use'n = 'RC-MARIJUANA - PAST

MONTH USE'n
```

```
                        'race recode (4 levels)'n = 'RC-RACE RECODE (4 LEVELS)'n ;
run;


data dat2010;
  set dsn6 dsn7 dsn8 dsn9 dsn10;
  keep years 'state name'n 'final edited age'n 'marijuana abuse or dependence -'n
        'imputation revised gender'n 'marijuana - past month use'n
        'race recode (4 levels)'n 'Weighted Count'n 'Count SE'n
        'Column %'n  'Column % SE'n;
  rename 'final edited age'n  = 'FINAL EDITED AGE'n
                        'marijuana abuse or dependence -'n = 'RC-MARIJUANA
DEPENDENCE OR ABUSE'n
                        'imputation revised gender'n = 'GENDER - IMPUTATION
REVISED'n
                         'marijuana - past month use'n = 'RC-MARIJUANA - PAST
MONTH USE'n
                         'race recode (4 levels)'n = 'RC-RACE RECODE (4 LEVELS)'n ;
run;
```

```
data dat2012;
  set dsn11 dsn12 dsn13 dsn14 dsn15;
  keep years 'state name'n 'final edited age'n 'MARIJUANA ABUSE OR DEPENDENCE -
'n'imputation revised gender'n
        'marijuana - past month use'n 'race recode (4 levels)'n 'Weighted Count'n
'Count SE'n
        'Column %'n  'Column % SE'n;
  rename 'final edited age'n  = 'FINAL EDITED AGE'n
                        'marijuana abuse or dependence -'n = 'RC-MARIJUANA
DEPENDENCE OR ABUSE'n
                        'imputation revised gender'n = 'GENDER - IMPUTATION
REVISED'n
                         'marijuana - past month use'n = 'RC-MARIJUANA - PAST
MONTH USE'n
                         'race recode (4 levels)'n = 'RC-RACE RECODE (4 LEVELS)'n ;
run;


data dat2014;
  set dsn16 dsn17 dsn18 dsn19 dsn20;
```

keep years 'state name'n 'FINAL EDITED AGE'n 'RC-MARIJUANA DEPENDENCE OR ABUSE'n 'GENDER - IMPUTATION REVISED'n

'RC-MARIJUANA - PAST MONTH USE'n 'RC-RACE RECODE (4 LEVELS)'n 'Weighted Count'n 'Count SE'n 'Column %'n 'Column % SE'n;

run;

data dat2016;

set dsn21 dsn22 dsn23 dsn24 dsn25;

keep years 'state name'n 'FINAL EDITED AGE'n 'RC-MARIJUANA DEPENDENCE OR ABUSE'n 'GENDER - IMPUTATION REVISED'n

'RC-MARIJUANA - PAST MONTH USE'n 'RC-RACE RECODE (4 LEVELS)'n 'Weighted Count'n 'Count SE'n 'Column %'n 'Column % SE'n;

run;

data dat2018;

set dsn26 dsn27 dsn28 dsn29 dsn30;

keep years 'state name'n 'FINAL EDITED AGE'n 'RC-MARIJUANA DEPENDENCE OR ABUSE'n 'GENDER - IMPUTATION REVISED'n

'RC-MARIJUANA - PAST MONTH USE'n 'RC-RACE RECODE (4 LEVELS)'n 'Weighted Count'n 'Count SE'n 'Column %'n 'Column % SE'n;

```
run;


data all;
        set dat2008 dat2010 dat2012 dat2014 dat2016 dat2018;
run;


proc print data=all;
        where 'FINAL EDITED AGE'n ~='';
        var 'state name'n 'FINAL EDITED AGE'n years 'Weighted Count'n 'Count SE'n
'Column %'n  'Column % SE'n ;
run;


proc print data=all;
        where 'GENDER - IMPUTATION REVISED'n ~='';
        var 'state name'n years 'Weighted Count'n 'GENDER - IMPUTATION REVISED'n
'Count SE'n 'Column %'n     'Column % SE'n;
run;


proc print data=all;
```

```
        where 'RC-RACE RECODE (4 LEVELS)'n ~='';

        var 'state name'n years 'Weighted Count'n 'RC-RACE RECODE (4 LEVELS)'n
'Count SE'n 'Column %'n      'Column % SE'n;

run;



proc print data=all;

        where 'RC-MARIJUANA - PAST MONTH USE'n ~='';

        var 'state name'n years 'Weighted Count'n 'RC-MARIJUANA - PAST MONTH
USE'n 'Count SE'n 'Column %'n      'Column % SE'n;

run;



proc print data=all;

        where 'RC-MARIJUANA DEPENDENCE OR ABUSE'n ~='';

        var 'state name'n years 'Weighted Count'n 'RC-MARIJUANA DEPENDENCE OR
ABUSE'n 'Count SE'n 'Column %'n  'Column % SE'n;

run;



/*/*/*/* PDAS */*/*/*/
```

```
/*read in data*/;

%drive(&ABOD,csv)



data dsn1;

        set dsn1 ;

        year="2008";

run;



data dsn2;

        set dsn2 ;

        year="2009";

run;



data dsn3;

        set dsn3;

        year="2010";

run;
```

```
data dsn4;
        set dsn4;
        year="2011";
run;

data dsn5;
        set dsn5;
        year="2012";
run;

data dsn6;
        set dsn6;
        year="2013";
run;

data dsn7;
        set dsn7;
        year="2014";
```

```
run;



data dsn8;
        set dsn8;
        year="2015";
run;



data dsn9;
        set dsn9;
        year="2016";
run;



data dsn10;
        set dsn10;
        year="2017";
run;



data dsn11;
```

```
        set dsn11;

        year="2018";

run;



data dsn12;

        set dsn12;

        year="2019";

run;



data ABOD;

  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6 dsn7 dsn8 dsn9 dsn10 dsn11 dsn12;

  MJ_DEP = catx('','marijuana abuse or dependence -'n,'RC-MARIJUANA
DEPENDENCE OR ABUSE'n);

  keep year mj_dep 'Unweighted Count'n;

run;



%drive(&AGE,csv)
```

```
data dsn1;
        set dsn1 ;
        year="2008";
run;

data dsn2;
        set dsn2 ;
        year="2009";
run;

data dsn3;
        set dsn3;
        year="2010";
run;

data dsn4;
        set dsn4;
        year="2011";
```

```
run;


data dsn5;
        set dsn5;
        year="2012";
run;


data dsn6;
        set dsn6;
        year="2013";
run;


data dsn7;
        set dsn7;
        year="2014";
run;


data dsn8;
```

```
        set dsn8;
        year="2015";
run;

data dsn9;
        set dsn9;
        year="2016";
run;

data dsn10;
        set dsn10;
        year="2017";
run;

data dsn11;
        set dsn11;
        year="2018";
run;
```

```
data dsn12;
        set dsn12;
        year="2019";
run;


data age;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6 dsn7 dsn8 dsn9 dsn10 dsn11 dsn12;
  AGE = catx('','AGE CATEGORY'n,'RC-AGE CATEGORY'n);
  keep year AGE 'Unweighted Count'n;
run;


%drive(&SEX,csv)


data dsn1;
        set dsn1 ;
        year="2008";
run;
```

```
data dsn2;
        set dsn2 ;
        year="2009";
run;

data dsn3;
        set dsn3;
        year="2010";
run;

data dsn4;
        set dsn4;
        year="2011";
run;

data dsn5;
        set dsn5;
```

```
        year="2012";
run;

data dsn6;
        set dsn6;
        year="2013";
run;

data dsn7;
        set dsn7;
        year="2014";
run;

data dsn8;
        set dsn8;
        year="2015";
run;
```

```
data dsn9;
        set dsn9;
        year="2016";
run;

data dsn10;
        set dsn10;
        year="2017";
run;

data dsn11;
        set dsn11;
        year="2018";
run;

data dsn12;
        set dsn12;
        year="2019";
```

```
run;


data gender;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6 dsn7 dsn8 dsn9 dsn10 dsn11 dsn12;
  keep year 'IMPUTATION REVISED GENDER'n 'Unweighted Count'n;
run;


%drive(&MRJ,csv)


data dsn1;
        set dsn1 ;
        year="2008";
run;


data dsn2;
        set dsn2 ;
        year="2009";
run;
```

```
data dsn3;
        set dsn3;
        year="2010";
run;

data dsn4;
        set dsn4;
        year="2011";
run;

data dsn5;
        set dsn5;
        year="2012";
run;

data dsn6;
        set dsn6;
```

```
        year="2013";
run;


data dsn7;
        set dsn7;
        year="2014";
run;


data dsn8;
        set dsn8;
        year="2015";
run;


data dsn9;
        set dsn9;
        year="2016";
run;
```

```
data dsn10;
        set dsn10;
        year="2017";
run;

data dsn11;
        set dsn11;
        year="2018";
run;

data dsn12;
        set dsn12;
        year="2019";
run;

data mrj;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6 dsn7 dsn8 dsn9 dsn10 dsn11 dsn12;
  mrj = catx('','marijuana - past month use'n,'RC-MARIJUANA - PAST MONTH USE'n);
```

```
  keep year mrj'Unweighted Count'n;
run;


%drive(&RACE,csv)


data dsn1;
        set dsn1 ;
        year="2008";
run;


data dsn2;
        set dsn2 ;
        year="2009";
run;


data dsn3;
        set dsn3;
        year="2010";
```

```
run;


data dsn4;
        set dsn4;
        year="2011";
run;


data dsn5;
        set dsn5;
        year="2012";
run;


data dsn6;
        set dsn6;
        year="2013";
run;


data dsn7;
```

```
        set dsn7;
        year="2014";
run;

data dsn8;
        set dsn8;
        year="2015";
run;

data dsn9;
        set dsn9;
        year="2016";
run;

data dsn10;
        set dsn10;
        year="2017";
run;
```

```
data dsn11;
        set dsn11;
        year="2018";
run;


data dsn12;
        set dsn12;
        year="2019";
run;


data RACE;
  set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6 dsn7 dsn8 dsn9 dsn10 dsn11 dsn12;
  RACE = catx('','RC-RACE/HISPANICITY RECODE (7 LE'n,'RACE/HISPANICITY
RECODE (7 LEVEL'n);
  keep year RACE 'Unweighted Count'n;
run;
```

```
proc means data=abod sum;
        var 'Unweighted Count'n;
        class MJ_DEP;
run;

proc means data=age sum;
        var 'Unweighted Count'n;
        class AGE;
run;

proc means data=gender sum;
        var 'Unweighted Count'n;
        class 'IMPUTATION REVISED GENDER'n;
run;

proc means data=RACE sum;
        var 'Unweighted Count'n;
        class RACE;
```

```
run;


proc means data=mrj  sum;
        var 'Unweighted Count'n;
        class mrj;
run;
/*************************************************************************************************
* In Enterprise Guide, "Specify the page size for log and text output" under 'Results
General' must be *
* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *
*************************************************************************************************/
OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE ;


title1'Dissertation';
title2'Aim 2: Incidence after leglaization DiD';


/*****************************************************************************
```

```
* The following macro variables are available to all users: *
*                                                           *
* &Project     – the name of the project folder in which the .EGP file is stored    *
* &ProgName        – the name of the .egp file, without the extension                *
* &ProgNode        - the name of the code node                                       *
* &ProgDir         – the path to the folder in which the .egp file is stored.        *
*****************************************************************************************/


/*************************************************************************
* The following macro variables are used in conjunction with the DOC_BLOCK   *
* macro to document programs and output.                                     *
*                                                                            *
* Do not use quotation marks when defining macro variables. If SAS syntax    *
* requires quotes, use double quotes when you reference the macro variable.  *
*************************************************************************/
** PROGRAMMER'S NAME ;
%LET PROGRAMMER      = Barrett Montgomery;
```

```
** DEFINE ALL NON-SAS FILES CALLED IN YOUR PROGRAM AS MACRO VARIABLES
by Drug;

** THESE CAN BE LEFT BLANK IF NOT NEEDED OR USED ;

%LET Elg       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Legal Timeline Categories\ELG;

%LET Rec       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Legal Timeline Categories\REC;

%LET ElgE      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Effective Date Categories\ELG;

%LET RecE      = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Raw\Effective Date Categories\REC;


** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;

** USE &PROGNAME FOR SAVEFILE NAME ;

** LEAVE BLANK IF NO DATASET SAVED ;

%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
2\Processed;


** NAME FORMAT LIBRARY DIRECTORY ;
```

```
%LET FMTDIR      = ;


** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT
WORK ;

%LET PURPOSE1 = Differnce in differnce and event study design to estimate the effect
of cannabis legalization on cannabis incidence



/*************************************************************

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

*************************************************************/


** INPUT FILES ;


** OUTPUT FILE DESTINATION ;

LIBNAME SAV  "&SAVEDIR1" ;



/*This section recreates the various 2x2 plots that can be created with these data
```

```
meant to replicate figure 2 in  Goodman-Bacon's 2018 seminal working paper

        also easy visual check for parrallel trends assumption*/



proc freq data = sav.Legal_date;

        table legal_cat age / list missing;

run;



proc format;

        value $legal

         'Legalized_2012'='Legalized in 2012'

         'Legalized_2014'='Legalized in 2014'

         'Legalized_2016'='Legalized in 2016'

         'Legalized_2018'='Legalized in 2018';

         value $bin

          'Legalized_2012'='Legal'

         'Legalized_2014'='Legal'

         'Legalized_2016'='Legal'

         'Legalized_2018'='Legal';
```

```
run;


proc sgplot data=sav.Legal_date ;
        title3'Cannabis incidence in 21+ age group, first wave legalizing states vs
untreated states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legalized_2012','Illegal') and age = "21_Plus";
   refline "2012 to 2013" / axis=x label="First wave of cannabis legalization"
lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;


proc sgplot data=sav.Legal_date ;
        title3'Cannabis incidence in 21+ age group, second wave legalizing states vs
untreated states';
        series  x=years y=incidence / lineattrs=(pattern=2) group=legal_cat;
        where legal_cat in ('Legalized_2014','Illegal') and age = "21_Plus";
```

```
   refline "2014 to 2015" / axis=x label="Second wave of cannabis legalization"
lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;

proc sgplot data=sav.Legal_date ;
        title3'Cannabis incidence in 21+ age group, third wave legalizing states vs
untreated states';
        series  x=years y=incidence / lineattrs=(pattern=2) group=legal_cat;
        where legal_cat in ('Legalized_2016','Illegal') and age = "21_Plus";
   refline "2016 to 2017" / axis=x label="Third wave of cannabis legalization"
lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;
```

```
proc sgplot data=sav.Legal_date ;
        title3'Cannabis incidence in 21+ age group, first wave legalizing states vs
second wave legalizing states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legalized_2012','Legalized_2014') and age = "21_Plus";
   refline "2012 to 2013"  "2014 to 2015"/ axis=x lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;


proc sgplot data=sav.Legal_date ;
        title3'Cannabis incidence in 21+ age group, first wave legalizing states vs third
wave legalizing states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legalized_2012','Legalized_2016') and age = "21_Plus";
   refline "2012 to 2013"  "2016 to 2017"/ axis=x lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
```

```
        format legal_cat $legal.;
run;


proc sgplot data=sav.Legal_date ;
        title3'Cannabis incidence in 21+ age group, second wave legalizing states vs
third wave legalizing states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legalized_2014','Legalized_2016') and age = "21_Plus";
   refline "2014 to 2015"   "2016 to 2017"/ axis=x;
         xaxis grid label='Time';
   yaxis grid label='Past year cannabis use incidence' discreteorder=data;
run;


/***********************************************/
/*                    legal date analysis Plots                    */
/***********************************************/


ods graphics on;
```

```
title1 "Simple Incidence Plots";

title3 "All ages";

proc sgplot data=sav.Legal_date;

          where age="Overall";

   series x=years y=incidence / group=legal_cat;

run;



title3 "12-20";

proc sgplot data=sav.Legal_date;

          where age="12_20";

   series x=years y=incidence / group=legal_cat;

run;



title3 "21+";

proc sgplot data=sav.Legal_date;

          where age="21_Plus";

   series x=years y=incidence / group=legal_cat;

run;
```

```
title3 "21+";

proc sgplot data=sav.Legal_date;

        where age="21_Plus";

   series x=time y=incidence / group=legal_cat;

run;



proc freq data=sav.Legal_date;

        table legal_cat*years*incidence / list missing;

        where age = "21_Plus";

run;



/*Establish baseline*/



proc freq data=sav.Legal_date;

        table legal_cat*  / list missing;

        where age = "21_Plus";

run;
```

```
proc means data=sav.Legal_date mean;
        title3'Average incidence in 2 year period prior to legalization';
        var incidence;
        class age legal_cat ;
        where time=-2;
        format legal_cat $bin.;
run;

proc means data=sav.Legal_date mean;
        title3'Average incidence  where illegal';
        var incidence;
        class age legal_cat ;
        where legal_cat="Illegal";
        format legal_cat $bin.;
run;

title1 "Regression modelling";
```

```
title4 ;


proc sort data=sav.Legal_date;
        by years;
run;


ods graphics on;
proc glm data=sav.Legal_date ;
        title3 "Regression for panel event study, all ages";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus8(ref="0") tminus6(ref="0")
tminus4(ref="0") t0(ref="0") t2(ref="0") t4(ref="0") t6(ref="0");
        where age="Overall";
        model incidence = legal_cat tminus10 tminus8 tminus6 tminus4 t0 t2 t4 t6
/solution CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;


proc glm data=sav.Legal_date ;
```

```
        title3 "Regression for panel event study, 12 to 20 year olds";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus8(ref="0") tminus6(ref="0")
tminus4(ref="0") t0(ref="0") t2(ref="0") t4(ref="0") t6(ref="0");
        where age="12_20";
        model incidence = legal_cat tminus10 tminus8 tminus6 tminus4 t0 t2 t4 t6
/solution CLPARM ;
        ods output ParameterEstimates = ParamEstunderage;
run;


proc glm data=sav.Legal_date ;
        title3 "Regression for event study, ages 21 and up";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus8(ref="0") tminus6(ref="0")
tminus4(ref="0") t0(ref="0") t2(ref="0") t4(ref="0") t6(ref="0");
        where age="21_Plus";
        model incidence = legal_cat tminus10 tminus8 tminus6 tminus4 t0 t2 t4 t6
/solution CLPARM ;
        ods output ParameterEstimates = ParamEst21;
```

```
run;

title1"Manage Regression Output";

data ParamEsttotal1;
        set ParamEsttotal;
        where StdErr >. and Parameter in ("tminus10  1",
"tminus8   1",
"tminus6   1",
"tminus4   1",
"t0        1",
"t2        1",
"t4        1",
"t6        1");
run;

data ParamEstunderage1;
        set ParamEstunderage;
```

```
        where StdErr >. and Parameter in ("tminus10  1",
"tminus8   1",
"tminus6   1",
"tminus4   1",
"t0        1",
"t2        1",
"t4        1",
"t6        1");
run;

data ParamEst211;
        set ParamEst21;
        where StdErr >. and Parameter in ("tminus10  1",
"tminus8   1",
"tminus6   1",
"tminus4   1",
"t0        1",
"t2        1",
```

```
"t4        1",

"t6        1");

run;



title1 "Plot coefficients";



proc sgplot data=ParamEsttotal1 noautolegend;

        title3'Effect of time since legalization on incidence';

        title4'all ages, fixed effects for time and state categories';

        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL

        markerattrs=(symbol=diamondfilled);

   refline 0 / axis=y;

   xaxis grid;

   yaxis grid display=(nolabel) discreteorder=data;

run;



proc sgplot data=ParamEstunderage1 noautolegend;

        title3'Effect of time since legalization on incidence';
```

```
        title4'aged 12 to 20, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid;
    yaxis grid display=(nolabel) discreteorder=data;
run;

proc sgplot data=ParamEst211 noautolegend;
        title3'Effect of time since cannabis legalization on cannabis incidence';
        title4'ages 21 and up, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid;
    yaxis grid display=(nolabel) discreteorder=data;
run;
```

```
/*********************************************/
/*                    Simple DD estimate ATT for 21+     */
/*********************************************/

proc sort data=sav.Legal_date ;
        by years;
run;
proc glm data=sav.Legal_date ;
        title3 "Regression for event study, ages 21 and up";
        absorb years;
        class legal_cat (ref="Illegal") legal(ref="0");
        where age="21_Plus";
        model incidence = legal_cat legal /solution CLPARM ;
run;
/*small because includes data from before effective date...*/

proc glm data=sav.Legal_date ;
        title3 "Regression for event study, ages 21 and up";
```

```
        absorb years;
        class legal_cat (ref="Illegal") effective(ref="0");
        where age="21_Plus";
        model incidence = legal_cat effective /solution CLPARM ;
run;

proc glm data=sav.Legal_date ;
        title3 "Regression for event study, 20 and younger";
        absorb years;
        class legal_cat (ref="Illegal") legal(ref="0");
        where age="12_20";
        model incidence = legal_cat legal /solution CLPARM ;
run;

proc glm data=sav.Legal_date ;
        title3 "Regression for event study, 20 and younger";
        absorb years;
        class legal_cat (ref="Illegal") effective(ref="0");
```

```
        where age="12_20";
        model incidence = legal_cat effective /solution CLPARM ;
run;

/************************/
/*effective date analysis*/
/************************/

title1 "Simple Incidence Plots to check for parallel trends";
title3 "All ages";
proc sgplot data=sav.Effective_date;
        where age="Overall";
   series x=years y=incidence / group=legal_cat;
run;

title3 "12-20";
proc sgplot data=sav.Effective_date;
        where age="12_20";
```

```
   series x=years y=incidence / group=legal_cat;
run;


title3 "21+";
proc sgplot data=sav.Effective_date;
          where age="21_Plus";
   series x=years y=incidence / group=legal_cat;
run;


title1 "Regression modelling by effective date";
title4 ;


proc sort data=sav.Effective_date;
          by years;
run;


ods graphics on;
proc glm data=sav.Effective_date ;
```

```
        title3 "Regression for panel event study, all ages";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus9(ref="0") tminus8(ref="0")
tminus7(ref="0") tminus6(ref="0") tminus5(ref="0") tminus4(ref="0") tminus3(ref="0")
tminus2(ref="0") t0(ref="0") t1(ref="0") t2(ref="0") t3(ref="0") t4(ref="0") ;
        where age="Overall";
        model incidence = legal_cat tminus10 tminus9 tminus8 tminus7 tminus6
tminus5 tminus4 tminus3 tminus2 t0 t1 t2 t3 t4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;

proc glm data=sav.Effective_date ;
        title3 "Regression for panel event study, 12 to 20 year olds";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus9(ref="0") tminus8(ref="0")
tminus7(ref="0") tminus6(ref="0") tminus5(ref="0") tminus4(ref="0") tminus3(ref="0")
tminus2(ref="0") t0(ref="0") t1(ref="0") t2(ref="0") t3(ref="0") t4(ref="0") ;
        where age="12_20";
        model incidence = legal_cat tminus10 tminus9 tminus8 tminus7 tminus6
tminus5 tminus4 tminus3 tminus2 t0 t1 t2 t3 t4 /solution CLPARM ;
```

```
        ods output ParameterEstimates = ParamEstunderage;
run;


proc glm data=sav.Effective_date ;
        title3 "Regression for event study, ages 21 and up";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus9(ref="0") tminus8(ref="0")
tminus7(ref="0") tminus6(ref="0") tminus5(ref="0") tminus4(ref="0") tminus3(ref="0")
tminus2(ref="0") t0(ref="0") t1(ref="0") t2(ref="0") t3(ref="0") t4(ref="0") ;
        where age="21_Plus";
        model incidence = legal_cat tminus10 tminus9 tminus8 tminus7 tminus6
tminus5 tminus4 tminus3 tminus2 t0 t1 t2 t3 t4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEst21;
run;


title1"Manage Regression Output";


data ParamEsttotal1;
        set ParamEsttotal;
```

```
        where StdErr >. and Parameter in ("tminus10  1",
        "tminus9   1",
        "tminus8   1",
        "tminus7   1",
        "tminus6   1",
        "tminus5   1",
        "tminus4   1",
        "tminus3   1",
        "tminus2   1",
        "tminus1   1",
        "t0        1",
        "t1        1",
        "t2        1",
        "t3        1",
        "t4        1");
run;

data ParamEstunderage1;
```

```
        set ParamEstunderage;

        where StdErr >. and Parameter in ("tminus10  1",

        "tminus9   1",

        "tminus8   1",

        "tminus7   1",

        "tminus6   1",

        "tminus5   1",

        "tminus4   1",

        "tminus3   1",

        "tminus2   1",

        "tminus1   1",

        "t0        1",

        "t1        1",

        "t2        1",

        "t3        1",

        "t4        1");

run;
```

```
data ParamEst211;
        set ParamEst21;
        where StdErr >. and Parameter in ("tminus10  1",
        "tminus9   1",
        "tminus8   1",
        "tminus7   1",
        "tminus6   1",
        "tminus5   1",
        "tminus4   1",
        "tminus3   1",
        "tminus2   1",
        "tminus1   1",
        "t0        1",
        "t1        1",
        "t2        1",
        "t3        1",
        "t4        1");
run;
```

```
title1 "Plot coefficients";


proc sgplot data=ParamEsttotal1 noautolegend;
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence';
        title4'all ages, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid;
   yaxis grid display=(nolabel) discreteorder=data;
run;


proc sgplot data=ParamEstunderage1 noautolegend;
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence';
        title4'ages 12 to 20, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
```

```
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid;
    yaxis grid display=(nolabel) discreteorder=data;
run;


proc sgplot data=ParamEst211 noautolegend;
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence';
        title4'ages 21 and up, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid;
    yaxis grid display=(nolabel) discreteorder=data;
run;


/*****************************************/
```

```
/*analyze legal date and effective

  date with more balanced lags and leads */

/***************************************/



title1 "Regression modelling with balanced leads and lags";

title4 'Legal dates';



proc sort data=sav.Legal_date;

        by years;

run;



proc print data=sav.Legal_date; run;



ods graphics on;

proc glm data=sav.Legal_date ;

        title3 "Regression for panel event study, all ages";

        absorb years;
```

```
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus4(ref="0") t0(ref="0") t2(ref="0")
tgt4(ref="0");
        where age="Overall";
        model incidence = legal_cat tlt6 tminus4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;

proc glm data=sav.Legal_date ;
        title3 "Regression for panel event study, 12 to 20 year olds";
        absorb years;
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus4(ref="0") t0(ref="0") t2(ref="0")
tgt4(ref="0");
        where age="12_20";
        model incidence = legal_cat tlt6 tminus4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEstunderage;
run;

proc glm data=sav.Legal_date ;
        title3 "Regression for event study, ages 21 and up";
```

```
        absorb years;
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus4(ref="0") t0(ref="0") t2(ref="0")
tgt4(ref="0");
        where age="21_Plus";
        model incidence = legal_cat tlt6 tminus4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEst21;
run;


title1"Manage Regression Output";


data ParamEsttotal1;
        set ParamEsttotal;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus4   1",
"t0        1",
"t2        1",
"tgt4      1");
run;
```

```
data ParamEstunderage1;
        set ParamEstunderage end=eof;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus4   1",
        "t0        1",
        "t2        1",
        "tgt4      1");
        if parameter = "tlt6      1" then do;
                parameter = '6+ years prior' ;
                order=1;
        end;
        if parameter = "tminus4   1" then do;
                parameter = '4 years prior';
                order=2;
        end;
        if parameter = "t0        1" then do;
                parameter = 'Legalized';
```

```
        order=4;
end;
if parameter = "t2       1" then do;
        parameter = '2 years after';
        order=5;
end;
if parameter = "tgt4     1" then do;
        parameter = '4+ years after';
        order=6;
end;
if eof then do;
output;
        parameter = '2 years prior';
        estimate=0;
        stderr=0;
        tvalue=0;
        probt=0;
        lowercl=0;
```

```
                uppercl=0;
                order=3;
        end;
        output;
run;


proc sort data=ParamEstunderage1;
        by order;
run;


data ParamEst211;
        set ParamEst21 end=eof;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus4   1",
        "t0        1",
        "t2        1",
        "tgt4      1");
```

```
if parameter = "tlt6      1" then do;
        parameter = '6+ years prior' ;
        order=1;
end;
if parameter = "tminus4   1" then do;
        parameter = '4 years prior';
        order=2;
end;
if parameter = "t0        1" then do;
        parameter = 'Legalized';
        order=4;
end;
if parameter = "t2        1" then do;
        parameter = '2 years after';
        order=5;
end;
if parameter = "tgt4      1" then do;
        parameter = '4+ years after';
```

```
            order=6;
        end;
        if eof then do;
        output;
            parameter = '2 years prior';
            estimate=0;
            stderr=0;
            tvalue=0;
            probt=0;
            lowercl=0;
            uppercl=0;
            order=3;
        end;
        output;
run;

proc sort data=ParamEst211;
        by order;
```

```
run;

title1 "Plot coefficients";

proc sgplot data=ParamEsttotal1 noautolegend;
        title3'Effect of time since legalization on incidence';
        title4'all ages, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid;
   yaxis grid display=(nolabel) discreteorder=data;
run;

proc sgplot data=ParamEstunderage1 noautolegend;
        title3'Effect of time since legalization on incidence';
        title4'aged 12 to 20, balanced leads and lags, fixed effects for time and state
categories';
```

```
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid label='Time relative to legalization';
    yaxis grid label='Newly incident cannabis use' discreteorder=data;
run;

proc sgplot data=ParamEst211 noautolegend;
        title3'Effect of time since cannabis legalization on cannabis incidence';
        title4'ages 21 and up, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid label='Time relative to legalization';
    yaxis grid label='Newly incident cannabis use' discreteorder=data;
run;
```

```
title1 "Regression modelling with balanced leads and lags";
title4 'Effective dates';

proc sort data=sav.Effective_date;
        by years;
run;

ods graphics on;
proc glm data=sav.Effective_date ;
        title3 "Regression for panel event study, all ages";
        absorb years;
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus5(ref="0") tminus4(ref="0")
tminus3(ref="0") tminus2(ref="0") t0(ref="0") t1(ref="0") tgt2(ref="0") ;
        where age="Overall";
        model incidence = legal_cat tlt6 tminus5 tminus4 tminus3 tminus2 t0 t1 tgt2
/solution CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;
```

```
proc glm data=sav.Effective_date ;
        title3 "Regression for panel event study, 12 to 20 year olds";
        absorb years;
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus5(ref="0") tminus4(ref="0")
tminus3(ref="0") tminus2(ref="0") t0(ref="0") t1(ref="0") tgt2(ref="0") ;
        where age="12_20";
        model incidence = legal_cat tlt6 tminus5 tminus4 tminus3 tminus2 t0 t1 tgt2
/solution CLPARM ;
        ods output ParameterEstimates = ParamEstunderage;
run;


proc glm data=sav.Effective_date ;
        title3 "Regression for event study, ages 21 and up";
        absorb years;
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus5(ref="0") tminus4(ref="0")
tminus3(ref="0") tminus2(ref="0") t0(ref="0") t1(ref="0") tgt2(ref="0") ;
        where age="21_Plus";
        model incidence = legal_cat tlt6 tminus5 tminus4 tminus3 tminus2 t0 t1 tgt2
/solution CLPARM ;
```

```
        ods output ParameterEstimates = ParamEst21;
run;


title1"Manage Regression Output";


data ParamEsttotal1;
        set ParamEsttotal;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus5   1",
        "tminus4   1",
        "tminus3   1",
        "tminus2   1",
        "tminus1   1",
        "t0        1",
        "t1        1",
        "tgt2      1");
run;
```

```
data ParamEstunderage1;
        set ParamEstunderage;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus5   1",
        "tminus4   1",
        "tminus3   1",
        "tminus2   1",
        "tminus1   1",
        "t0        1",
        "t1        1",
        "tgt2      1");
run;


data ParamEst211;
        set ParamEst21;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus5   1",
        "tminus4   1",
```

```
        "tminus3   1",
        "tminus2   1",
        "tminus1   1",
        "t0        1",
        "t1        1",
        "tgt2      1");
run;

title1 "Plot coefficients";

proc sgplot data=ParamEsttotal1 noautolegend;
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence';
        title4'all ages, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid;
```

```
    yaxis grid display=(nolabel) discreteorder=data;
run;


proc sgplot data=ParamEstunderage1 noautolegend;
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence';
        title4'ages 12 to 20, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid;
    yaxis grid display=(nolabel) discreteorder=data;
run;


proc sgplot data=ParamEst211 noautolegend;
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence';
```

```
        title4'ages 21 and up, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid;
   yaxis grid display=(nolabel) discreteorder=data;
run;

/*****************************************/
/*/*/*/*Repeat procedure for prevalence to */*/*/
/*/*/*compare to prior results from Cerda */*/*/
/*/*/*/*/*/*/*/     and others  */*/*/*/*/*/*/*/*/
/*****************************************/;

title1 "Regression modelling with balanced leads and lags";
title4 'Prevalence by Legal dates';
```

```
proc sort data=sav.Prevalence;
        by years;
run;


ods graphics on;
proc glm data=sav.Prevalence ;
        title3 "Regression for panel event study, all ages";
        absorb years;
        class legal_cat (ref="Illegal") tlt4(ref="0") t0(ref="0") t2(ref="0") tgt4(ref="0");
        where age="Overall";
        model pmmj_phat = legal_cat tlt4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;


proc glm data=sav.Prevalence ;
        title3 "Regression for panel event study, 12 to 20 year olds";
        absorb years;
        class legal_cat (ref="Illegal") tlt4(ref="0") t0(ref="0") t2(ref="0") tgt4(ref="0");
```

```
        where age="12_20";
        model pmmj_phat = legal_cat tlt4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEstunderage;
run;

proc glm data=sav.Prevalence ;
        title3 "Regression for event study, ages 21 and up";
        absorb years;
        class legal_cat (ref="Illegal") tlt4(ref="0") t0(ref="0") t2(ref="0") tgt4(ref="0");
        where age="21_Plus";
        model pmmj_phat = legal_cat tlt4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEst21;
run;

title1"Manage Regression Output";

data ParamEsttotal1;
        set ParamEsttotal;
```

```
        where StdErr >. and Parameter in ("tlt4      1",
"t0       1",
"t2       1",
"tgt4     1");
run;


data ParamEstunderage1;
        set ParamEstunderage end=eof;
        where StdErr >. and Parameter in ("tlt4      1",
"t0       1",
"t2       1",
"tgt4     1");

        if parameter = "tlt4      1" then do;
                parameter = '4+ years prior';
                order=1;
        end;
        if parameter = "t0        1" then do;
```

```
        parameter = 'Legalized';
        order=3;
end;
if parameter = "t2       1" then do;
        parameter = '2 years after';
        order=4;
end;
if parameter = "tgt4     1" then do;
        parameter = '4+ years after';
        order=5;
end;
if eof then do;
output;
        parameter = '2 years prior';
        estimate=0;
        stderr=0;
        tvalue=0;
        probt=0;
```

```
                lowercl=0;

                uppercl=0;

                order=2;

        end;

        output;

run;



proc sort data= ParamEstunderage1;

        by order;

run;



data ParamEst211;

        set ParamEst21 end=eof;

        where StdErr >. and Parameter in ("tlt4      1",

"t0       1",

"t2       1",

"tgt4     1");
```

```
        if parameter = "tlt4      1" then do;
        parameter = '4+ years prior';
        order=1;
end;
if parameter = "t0        1" then do;
        parameter = 'Legalized';
        order=3;
end;
if parameter = "t2        1" then do;
        parameter = '2 years after';
        order=4;
end;
if parameter = "tgt4      1" then do;
        parameter = '4+ years after';
        order=5;
end;
if eof then do;
output;
```

```
                parameter = '2 years prior';
                estimate=0;
                stderr=0;
                tvalue=0;
                probt=0;
                lowercl=0;
                uppercl=0;
                order=2;
        end;
        output;
run;


proc sort data= ParamEst211;
        by order;
run;


title1 "Plot coefficients";
```

```
proc sgplot data=ParamEsttotal1 noautolegend;
        title3'Effect of time since legalization on past month cannabis prevalence';
        title4'all ages, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid label='Time relative to legalization';
   yaxis grid label='Past-month cannabis use prevalence' discreteorder=data;
run;


proc sgplot data=ParamEstunderage1 noautolegend;
        title3'Effect of time since legalization on past month cannabis prevalence';
        title4'aged 12 to 20, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid label='Time relative to legalization';
```

```
   yaxis grid label='Past-month cannabis use prevalence' discreteorder=data;
run;


proc sgplot data=ParamEst211 noautolegend;
        title3'Effect of time since cannabis legalization on past month cannabis
prevalence';
        title4'ages 21 and up, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid label='Time relative to legalization';
   yaxis grid label='Past-month cannabis use prevalence' discreteorder=data;
run;


/*Additional checks*/


proc means data=sav.Legal_date;
        title3'What is going on at t minus 6 in the legalization data?';
```

```
        var incidence ;

        class age tminus6;

run;



data r;

        set sav.Legal_date;



        /*alternate specification for replication in r*/



        if legal_cat = "Illegal" then legal=0;

        else if legal_cat = "Overall" then legal=.;

        else if time=0 and legal_cat ~= "Illegal" then legal=1;

        else if time<0 then legal=0;

        else if time>0 then legal=1;



        label    legal = "cannabis leagality binary";

run;
```

```
proc freq data=r;
        title3'check legal binary creation';
        table time*years*legal_cat*legal/list missing;
run;

proc glm data=r ;
        title3 "Regression for event study, altrernate specification, ages 21 and up";
        class legal (ref="0") time (ref="-2");
        where age="21_Plus";
        model incidence = legal*time /solution CLPARM ;
        ods output ParameterEstimates = ParamEst21;
run;

proc sort data=ParamEst21;
        by Parameter ;
run;

proc sgplot data=ParamEst21 noautolegend;
```

```
        title3'Effect of time since recreational cannabis dispensaries become operational
on cannabis incidence, 21+ alternate specification';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid;
   yaxis grid display=(nolabel) discreteorder=data;
run;

/*****************************************/
/*  analyze legal date with placebo RCL  */
/*****************************************/

proc sort data=sav.placebo;
        by years;
run;

ods graphics on;
```

```
proc glm data=sav.placebo ;
        title3 "PLACEBO Regression for panel event study, all ages";
        absorb years;
        class placebo_legal (ref="Illegal") ptminus3(ref="0")  pt0(ref="0") pt1(ref="0")
pt3(ref="0") pt5(ref="0") pt7(ref="0");
        where age="Overall";
        model incidence = placebo_legal ptminus3 pt0 pt1 pt3 pt5 pt7 /solution
CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;


proc glm data=sav.placebo ;
        title3 "PLACEBO Regression for panel event study, 12 to 20 year olds";
        absorb years;
        class placebo_legal (ref="Illegal") ptminus3(ref="0")  pt0(ref="0") pt1(ref="0")
pt3(ref="0") pt5(ref="0") pt7(ref="0");
        where age="12_20";
        model incidence = placebo_legal ptminus3 pt0 pt1 pt3 pt5 pt7 /solution
CLPARM ;
```

```
        ods output ParameterEstimates = ParamEstunderage;
run;


proc glm data=sav.placebo ;
        title3 "PLACEBO Regression for event study, ages 21 and up";
        absorb years;
        class placebo_legal (ref="Illegal") ptminus3(ref="0")  pt0(ref="0") pt1(ref="0")
pt3(ref="0") pt5(ref="0") pt7(ref="0");
        where age="21_Plus";
        model incidence = placebo_legal ptminus3 pt0 pt1 pt3 pt5 pt7 /solution
CLPARM ;
        ods output ParameterEstimates = ParamEst21;
run;


title1"Manage Regression Output";


data ParamEsttotal1;
        set ParamEsttotal;
        where StdErr >. and Parameter in ('ptminus3      1',
```

```
'ptminus1      1',
'pt0           1',
'pt1           1',
'pt3           1',
'pt5           1');
run;

data ParamEstunderage1;
        set ParamEstunderage end=eof;
        where StdErr >. and Parameter in ('ptminus3      1',
        'pt0           1',
        'pt1           1',
        'pt3           1',
        'pt5           1',
        'pt7           1');

        if parameter = 'ptminus3      1' then do;
                parameter = '3 years prior' ;
```

```
        order=1;
end;
if parameter = 'pt1          1' then do;
        parameter = '1 year after';
        order=3;
end;
if parameter = 'pt3          1' then do;
        parameter = '3 years after';
        order=4;
end;
if parameter = 'pt5          1' then do;
        parameter = '5 years after';
        order=5;
end;
if parameter = 'pt7          1' then do;
        parameter = '7 years after';
        order=6;
end;
```

```
        if eof then do;
        output;
                parameter = '1 year prior';
                estimate=0;
                stderr=0;
                tvalue=0;
                probt=0;
                lowercl=0;
                uppercl=0;
                order=2;
        end;
        output;
run;


proc sort data=ParamEstunderage1;
        by order;
run;
```

```
data ParamEst211;
        set ParamEst21 end=eof;
        where StdErr >. and Parameter in ('ptminus3      1',
        'pt0          1',
        'pt1          1',
        'pt3          1',
        'pt5          1',
        'pt7          1');

        if parameter = 'ptminus3      1' then do;
                parameter = '3 years prior' ;
                order=1;
        end;
        if parameter = 'pt1          1' then do;
                parameter = '1 year after';
                order=3;
        end;
        if parameter = 'pt3          1' then do;
```

```
        parameter = '3 years after';
        order=4;
end;
if parameter = 'pt5          1' then do;
        parameter = '5 years after';
        order=5;
end;
if parameter = 'pt7          1' then do;
        parameter = '7 years after';
        order=6;
end;
if eof then do;
output;
        parameter = '1 year prior';
        estimate=0;
        stderr=0;
        tvalue=0;
        probt=0;
```

```
                lowercl=0;

                uppercl=0;

                order=2;

        end;

        output;

run;



proc sort data=ParamEst211;

        by order;

run;




title1 "Plot coefficients";



proc sgplot data=ParamEsttotal1 noautolegend;

        title3'PLACEBO Effect of time since legalization on incidence';

        title4'all ages, fixed effects for time and state categories';

        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
```

```
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid label='Time relative to placebo legalization';
    yaxis grid label='Past year cannabis use incidence' discreteorder=data;
run;


proc sgplot data=ParamEstunderage1 noautolegend;
        title3'PLACEBO Effect of time since legalization on incidence';
        title4'aged 12 to 20, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid label='Time relative to placebo legalization';
    yaxis grid label='Newly incident cannabis use' discreteorder=data;
run;


proc sgplot data=ParamEst211 noautolegend;
```

```
        title3'PLACEBO Effect of time since cannabis legalization on cannabis
incidence';
        title4'ages 21 and up, fixed effects for time and state categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
   refline 0 / axis=y;
   xaxis grid label='Time relative to placebo legalization';
   yaxis grid label='Newly incident cannabis use'  discreteorder=data;
run;
/*************************************************************************************************
* In Enterprise Guide, "Specify the page size for log and text output" under 'Results
General' must be *
* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *
*************************************************************************************************/
OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE ;

title1'Dissertation';
```

```
title2'Aim 3: The LMA Effect';


/*****************************************************************************
* The following macro variables are available to all users: *
*                                                           *
* &Project     – the name of the project folder in which the .EGP file is stored    *
* &ProgName        – the name of the .egp file, without the extension                *
* &ProgNode        - the name of the code node                                       *
* &ProgDir         – the path to the folder in which the .egp file is stored.        *
*****************************************************************************/


/*************************************************************************
* The following macro variables are used in conjunction with the DOC_BLOCK   *
* macro to document programs and output.                                     *
*                                                                            *
* Do not use quotation marks when defining macro variables. If SAS syntax    *
* requires quotes, use double quotes when you reference the macro variable.  *
*************************************************************************/
```

```
** PROGRAMMER'S NAME ;

%LET PROGRAMMER      = Barrett Montgomery;


** DEFINE ALL NON-SAS FILES CALLED IN YOUR PROGRAM AS MACRO VARIABLES
by Drug;

** THESE CAN BE LEFT BLANK IF NOT NEEDED OR USED ;

%LET Elg       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
3\Raw\ELG;

%LET Rec       = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
3\Raw\REC;


** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;

** USE &PROGNAME FOR SAVEFILE NAME ;

** LEAVE BLANK IF NO DATASET SAVED ;

%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
3\Processed;


** NAME FORMAT LIBRARY DIRECTORY ;
```

```
%LET FMTDIR      =  ;


** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT
WORK ;

%LET PURPOSE1 = Difference in difference event study design to estimate the effect
of RCL on incidence for 21 year olds (LMA);



/*************************************************************

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

*************************************************************/


** INPUT FILES ;


** OUTPUT FILE DESTINATION ;

LIBNAME SAV  "&SAVEDIR1" ;


/*theroetical power calc*/
```

```
ods graphics on;

proc power;

        twosamplefreq alpha=.05 sides=2 test=pchi

        relativerisk= 1.5 to 3.0 by .05

        refproportion=.08

        groupweights=(1 2)

        ntotal= .

        power=.8;

        plot x=effect xopts=(crossref=yes ref=.70 .75 .80 .85 .90) ;

run;

ods graphics off;



/*this macro imports all data in a file with 2 options, the folder directory, and the type of

file

here, folder is saved in a macro variable named above and file type is csv.*/



title1"Download all data, append, and organize";

%macro drive(dir,ext);
```

```
%local cnt filrf rc did memcnt name;
%let cnt=0;


%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,&dir));
%let did=%sysfunc(dopen(&filrf));
 %if &did ne 0 %then %do;
%let memcnt=%sysfunc(dnum(&did));


 %do i=1 %to &memcnt;


  %let name=%qscan(%qsysfunc(dread(&did,&i)),-1,.);


  %if %qupcase(%qsysfunc(dread(&did,&i))) ne %qupcase(&name) %then %do;
   %if %superq(ext) = %superq(name) %then %do;
     %let cnt=%eval(&cnt+1);
     %put %qsysfunc(dread(&did,&i));
     proc import datafile="&dir\%qsysfunc(dread(&did,&i))" out=dsn&cnt
```

```
          dbms=csv replace;

         run;

       %end;

      %end;


    %end;

     %end;

  %else %put &dir cannot be open.;

  %let rc=%sysfunc(dclose(&did));


 %mend drive;


/*read in eligibility data for leglaization dates*/

 %drive(&Elg,csv)


data dsn1;

          set dsn1;

          years="2008 to 2009";
```

```
run;



data dsn2;
        set dsn2;
        years="2010 to 2011";
run;



data dsn3;
        set dsn3;
        years="2012 to 2013";
run;



data dsn4;
        set dsn4;
        years="2014 to 2015";
run;



data dsn5;
```

```
        set dsn5;
        years="2016 to 2017";
run;

data dsn6;
        set dsn6;
        years="2018 to 2019";
run;

data Elig;
        set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;
        eligible = catx('', 'eligible for past year initiatio'n, 'rc-eligible for past year initia'n);
run;

proc freq data=Elig;
        title3'check to see that new var works';
        table eligible*'eligible for past year initiatio'n* 'rc-eligible for past year initia'n/list
missing;
```

```
run;


/*read in for recmj data*/
%drive(&Rec,csv)


data dsn1;
        set dsn1;
        years="2008 to 2009";
run;


data dsn2;
        set dsn2;
        years="2010 to 2011";
run;


data dsn3;
        set dsn3;
        years="2012 to 2013";
```

```
run;



data dsn4;
        set dsn4;
        years="2014 to 2015";
run;



data dsn5;
        set dsn5;
        years="2016 to 2017";
run;



data dsn6;
        set dsn6;
        years="2018 to 2019";
run;



data Rec;
```

```
        set dsn1 dsn2 dsn3 dsn4 dsn5 dsn6;

        recent_initiate = catx('','rc-past year initiate of marijua'n, 'rc-recent initiate of
marijuana'n, 'recent initiate of marijuana use'n);

run;



proc freq data=rec;

        title3'check to see that new var works';

        table recent_initiate*'rc-past year initiate of marijua'n*'rc-recent initiate of
marijuana'n*'recent initiate of marijuana use'n/list missing;

run;



/*cleaning*/

/*prepare susbets of data where eligibility and rec = yes

to merge and create incidence estimates

 first for legal date then for effective dates*/

data elig1;

        set elig;

        where eligible = "1 - Yes" and 'final edited age'n = '21';
```

drop row: total: 'eligible for past year initiatio'n 'rc-eligible for past year initia'n eligible 'Column % CI (lower)'n 'Column % CI (upper)'n;

rename 'STATE NAME'n = legal_cat;

rename 'final edited age'n = age;

rename 'Column %'n = elig_phat;

rename 'Column % SE'n = elig_phat_SE;

rename 'Weighted Count'n = elig_count;

rename 'Count SE'n = elig_count_se;

run;

data rec1;

set rec;

where recent_initiate = "1 - Yes" and 'final edited age'n = '21';

drop row: total: 'rc-past year initiate of marijua'n 'rc-recent initiate of marijuana'n 'recent initiate of marijuana use'n recent_initiate 'Column % CI (lower)'n 'Column % CI (upper)'n;

```
        rename 'STATE NAME'n = legal_cat;
        rename 'final edited age'n = age;
        rename 'Column %'n = rec_phat;
        rename 'Column % SE'n = rec_phat_SE;
        rename 'Weighted Count'n = rec_count;
        rename 'Count SE'n = rec_count_se;
run;

proc sort data=elig1;
        by years legal_cat age ;
run;

proc sort data=rec1;
        by years legal_cat age ;
run;

/*estimate incidence by age group, state group, and years*/
```

```
data all_data;
        merge elig1 rec1;
        by years legal_cat age ;
        length time year_n 8.;

        if legal_cat="Not_Legalized_" then legal_cat="Illegal";
        incidence = rec_count/elig_count;

/*fixed effects for years as categorical may function differently
        if years was considered a continuous variable, year_n is to check this*/

        if years = '2008 to 2009' then year_n = 2008;
        else if years = '2010 to 2011' then year_n = 2010;
        else if years = '2012 to 2013' then year_n = 2012;
        else if years = '2014 to 2015' then year_n = 2014;
        else if years = '2016 to 2017' then year_n = 2016;
        else if years = '2018 to 2019' then year_n = 2018;
```

```
        /*need to calculate incidence SE or 95%CIs*/


        label incidence = "Percentage of past year initiates among persons at risk for
initiation"
                        legal_cat = "Legal status of cannabis through 2018"
                        year_n = "Numeric date of data, first year in year-pair";
run;


proc print data=all_data (obs=6);
        title3'check incidence calcs';
run;


proc freq data=all_data;
        title3'Check Time recode';
        table year_n*years/list missing;
run;


proc freq data=all_data;
```

```
        title3'combinations of legal cat and years to make relative time variable';
        table legal_cat* years/list;
        where legal_cat ~="Overall";
run;

proc means data=all_data;
        title3'average incidence at age 21 by category';
        var incidence;
        class legal_cat;
run;

proc freq data=all_data;
        title3'Check legal category and year combinations';
        table legal_cat*years/list missing;
run;

data all_data2;
        set all_data;
```

```
where legal_cat ~="Overall";


/*create a time variable relative to year of legalization and year of data
        time = how many years away froma states year of legalization is this data
point?*/


if legal_cat = "Illegal" then time=0;
else if legal_cat = "Legal_2012" and years = "2008 to 2009" then time = -4;
else if legal_cat = "Legal_2014" and years = "2008 to 2009" then time = -6;
else if legal_cat = "Legal_2016" and years = "2008 to 2009" then time = -8;
else if legal_cat = "Legal_2018" and years = "2008 to 2009" then time = -10;

else if legal_cat = "Legal_2012" and years = "2010 to 2011" then time = -2;
else if legal_cat = "Legal_2014" and years = "2010 to 2011" then time = -4;
else if legal_cat = "Legal_2016" and years = "2010 to 2011" then time = -6;
else if legal_cat = "Legal_2018" and years = "2010 to 2011" then time = -8;

else if legal_cat = "Legal_2012" and years = "2012 to 2013" then time = 0;
```

else if legal_cat = "Legal_2014" and years = "2012 to 2013" then time = -**2**;

else if legal_cat = "Legal_2016" and years = "2012 to 2013" then time = -**4**;

else if legal_cat = "Legal_2018" and years = "2012 to 2013" then time = -**6**;

else if legal_cat = "Legal_2012" and years = "2014 to 2015" then time = **2**;

else if legal_cat = "Legal_2014" and years = "2014 to 2015" then time = **0**;

else if legal_cat = "Legal_2016" and years = "2014 to 2015" then time = -**2**;

else if legal_cat = "Legal_2018" and years = "2014 to 2015" then time = -**4**;

else if legal_cat = "Legal_2012" and years = "2016 to 2017" then time = **4**;

else if legal_cat = "Legal_2014" and years = "2016 to 2017" then time = **2**;

else if legal_cat = "Legal_2016" and years = "2016 to 2017" then time = **0**;

else if legal_cat = "Legal_2018" and years = "2016 to 2017" then time = -**2**;

else if legal_cat = "Legal_2012" and years = "2018 to 2019" then time = **6**;

else if legal_cat = "Legal_2014" and years = "2018 to 2019" then time = **4**;

else if legal_cat = "Legal_2016" and years = "2018 to 2019" then time = **2**;

else if legal_cat = "Legal_2018" and years = "2018 to 2019" then time = **0**;

```
/*create dummy variable for each relative time point*/

        if time=-10 then tminus10 =1; else tminus10=0;

        if time=-8 then tminus8 =1; else tminus8=0;

        if time=-6 then tminus6 =1; else tminus6=0;

        if time=-4 then tminus4 =1; else tminus4=0;

        if time=-2 then tminus2 =1; else tminus2=0;

        if time=0 then t0 =1; else t0=0;

        if time=2 then t2 =1; else t2=0;

        if time=4 then t4 =1; else t4=0;

        if time=6 then t6 =1; else t6=0;


        label    time = "Time relative to legalization based on beginning of year pair";

run;


proc freq data=all_data2;

        title3'check new variable creation';

        table time*legal_cat*years
```

```
                time*tminus10*tminus8*tminus6*tminus4*tminus2*t0*t2*t4*t6/list missing;
run;

proc freq data=all_data2;
        title3'determine where tails of legal date data distribution should be grouped
together';
        table time/list;
run;
/*<=-4, >=4*/

/* data step 3 creates new dummies that
can categorize all data before a certain time point
so that analysis on a single category is not done

econs call it balancing leads and lags for short*/

data all_data3;
        set all_data2;
```

if time<=-**4** then tlt4=**1**; else tlt4=**0**;

if time>=**4** then tgt4=**1**; else tgt4=**0**;

if time<=-**6** then tlt6=**1**; else tlt6=**0**;

if time>=**6** then tgt6=**1**; else tgt6=**0**;

if legal_cat = "Illegal" then legal = **0**;

else if time>=**0** then legal = **1**;

else if time <**0** then legal =**0**;

if legal_cat = "Illegal" then effective = **0**;

else if time>=**2** then effective = **1**;

else if time <**2** then effective =**0**;

if legal_cat = "Illegal" then legal_wave = **0**;

else if legal_cat = "Legal_2012" then legal_wave = **1**;

else if legal_cat = "Legal_2014" then legal_wave = **2**;

```
        else if legal_cat = "Legal_2016" then legal_wave = 3;

        else if legal_cat = "Legal_2018" then legal_wave = 4;



        label legal = "Simple binary for RCL, 1 if year>=legalize date, 0 otherwise"

                        effective = "Simple binary for RCL effective, 1 if year>=effective
date, 0 otherwise";

run;



proc freq data=all_data3;

        title3'Check new time event dummies';

        table time*tlt4*tminus10*tminus8*tminus6*tminus4*tminus2*t0*t2*t4*t6*tgt4/list
missing;

run;



proc freq data=all_data3;

        title3'Check new legality dummy';

        table legal_cat*years*legal/list missing;

run;
```

```
proc freq data=all_data3;
        title3'Check new effective dummy';
        table legal_cat*years*effective/list missing;
run;


/*Save dataset*/
data sav.Legal_date_21;
        set all_data3;
run;


/*export to csv if needed*/
proc export    data= sav.Legal_date_21 outfile='C:\Users\montg\Dropbox\Ph.D
Work\Dissertation\Data\Aim 2\Processed\Legal_date_21.csv'
        dbms=csv replace;
run;


/*************************************************************************************************
* In Enterprise Guide, "Specify the page size for log and text output" under 'Results
General' must be *
```

```
* de-selected in order to be able to specify pagesize and linesize using an options
statement.        *
*************************************************************************************************/
OPTIONS PS=56 LS=160 NOCENTER NOFMTERR MPRINT ORIENTATION =
LANDSCAPE ;


title1'Dissertation';
title2'Aim 3: Incidence at age 21 after leglaization event study';


/*************************************************************************************
* The following macro variables are available to all users: *
*                                                                                    *
* &Project     – the name of the project folder in which the .EGP file is stored     *
* &ProgName        – the name of the .egp file, without the extension                *
* &ProgNode        - the name of the code node                                       *
* &ProgDir         – the path to the folder in which the .egp file is stored.        *
*************************************************************************************/


/*************************************************************************
```

```
* The following macro variables are used in conjunction with the DOC_BLOCK   *
* macro to document programs and output.                                      *
*                                                                             *
* Do not use quotation marks when defining macro variables. If SAS syntax    *
* requires quotes, use double quotes when you reference the macro variable.  *
*****************************************************************************/
** PROGRAMMER'S NAME ;
%LET PROGRAMMER      = Barrett Montgomery;


** DEFINE DIRECTORY AND FILE NAME OF ANY PERMANENT SAS DATASETS
SAVED IN THIS PROGRAM AS MACRO VARIABLES ;
** USE &PROGNAME FOR SAVEFILE NAME ;
** LEAVE BLANK IF NO DATASET SAVED ;
%LET SAVEDIR1     = C:\Users\montg\Dropbox\Ph.D Work\Dissertation\Data\Aim
3\Processed;


** NAME FORMAT LIBRARY DIRECTORY ;
%LET FMTDIR       =  ;
```

```
** GIVE A BRIEF DESCRIPTION OF OVERALL PURPOSE OF THIS PROGRAM ;

** YOU CAN USE SINGLE QUOTES OR NO QUOTES-- DOUBLE QUOTES WILL NOT
WORK ;

%LET PURPOSE1 = Difference in difference and event study design to estimate the
effect of cannabis legalization on cannabis incidence for 21 year olds



/*************************************************************

** CREATE LIBRARY REFERENCES TO DIRECTORIES SPECIFIED ABOVE **

*************************************************************/



** INPUT FILES ;



** OUTPUT FILE DESTINATION ;

LIBNAME SAV  "&SAVEDIR1" ;



/*This section recreates the various 2x2 plots that can be created with these data

meant to replicate figure 2 in  Goodman-Bacon's 2018 seminal working paper

        also easy visual check for parrallel trends assumption*/
```

```
proc freq data = sav.Legal_date_21;
        table legal_cat age / list missing;
run;


proc format;
        value $legal
         'Legalized_2012'='Legalized in 2012'
         'Legalized_2014'='Legalized in 2014'
         'Legalized_2016'='Legalized in 2016'
         'Legalized_2018'='Legalized in 2018';
         value $bin
          'Legalized_2012'='Legal'
         'Legalized_2014'='Legal'
         'Legalized_2016'='Legal'
         'Legalized_2018'='Legal';
run;
```

```
proc means data = sav.Legal_date_21;
        var incidence;
        class legal_cat ;
        where time=0;
        format legal_cat $bin.;
run;

proc sgplot data=sav.Legal_date_21 ;
        title3'Cannabis incidence in 21 year olds, first wave legalizing states vs
untreated states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legal_2012','Illegal');
   refline "2012 to 2013" / axis=x label="First wave of cannabis legalization"
lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
```

```
run;

proc sgplot data=sav.Legal_date_21 ;
        title3'Cannabis incidence in 21 year olds, second wave legalizing states vs
untreated states';
        series  x=years y=incidence / lineattrs=(pattern=2) group=legal_cat;
        where legal_cat in ('Legal_2014','Illegal');
   refline "2014 to 2015" / axis=x label="Second wave of cannabis legalization"
lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;

proc sgplot data=sav.Legal_date_21 ;
        title3'Cannabis incidence in 21 year olds, third wave legalizing states vs
untreated states';
        series  x=years y=incidence / lineattrs=(pattern=2) group=legal_cat;
        where legal_cat in ('Legal_2016','Illegal');
```

```
   refline "2016 to 2017" / axis=x label="Third wave of cannabis legalization"
lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;

proc sgplot data=sav.Legal_date_21 ;
        title3'Cannabis incidence in 21 year olds, first wave legalizing states vs second
wave legalizing states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legal_2012','Legal_2014');
   refline "2012 to 2013"   "2014 to 2015"/ axis=x lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;

proc sgplot data=sav.Legal_date_21 ;
```

```
        title3'Cannabis incidence in 21+ age group, first wave legalizing states vs third
wave legalizing states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legal_2012','Legal_2016');
   refline "2012 to 2013"  "2016 to 2017"/ axis=x lineattrs=(color=green);
        xaxis grid label='NSDUH year-pairs';
   yaxis grid label='Newly incident cannabis use' discreteorder=data;
        format legal_cat $legal.;
run;


proc sgplot data=sav.Legal_date_21 ;
        title3'Cannabis incidence in 21+ age group, second wave legalizing states vs
third wave legalizing states';
        series  x=years y=incidence /lineattrs=(pattern=2)  group=legal_cat;
        where legal_cat in ('Legal_2014','Legal_2016');
   refline "2014 to 2015"  "2016 to 2017"/ axis=x;
         xaxis grid label='Time';
   yaxis grid label='Past year cannabis use incidence' discreteorder=data;
run;
```

```
/*********************************************/
/*                    legal date analysis Plots                    */
/*********************************************/

ods graphics on;
title1 "Simple Incidence Plots";

proc sgplot data=sav.Legal_date_21;
   series x=years y=incidence / group=legal_cat;
run;

proc means data=sav.Legal_date_21 mean;
        title3'Average incidence in 2 year period prior to legalization';
        var incidence;
        class age legal_cat ;
        where time=-2;
        format legal_cat $bin.;
```

```
run;


proc means data=sav.Legal_date_21 mean;
        title3'Average incidence  where illegal';
        var incidence;
        class age legal_cat ;
        where legal_cat="Illegal";
        format legal_cat $bin.;
run;


title1 "Regression modelling";
title4 ;


proc sort data=sav.Legal_date_21;
        by years;
run;


ods graphics on;
```

```
proc glm data=sav.Legal_date_21 ;
        title3 "Regression for panel event study";
        absorb years;
        class legal_cat (ref="Illegal") tminus10(ref="0") tminus8(ref="0") tminus6(ref="0")
tminus4(ref="0") t0(ref="0") t2(ref="0") t4(ref="0") t6(ref="0");
        model incidence = legal_cat tminus10 tminus8 tminus6 tminus4 t0 t2 t4 t6
/solution CLPARM ;
        ods output ParameterEstimates = ParamEsttotal;
run;


title1"Manage Regression Output";


data ParamEsttotal1;
        set ParamEsttotal;
        where StdErr >. and Parameter in ("tminus10  1",
"tminus8   1",
"tminus6   1",
"tminus4   1",
"t0        1",
```

```
"t2        1",

"t4        1",

"t6        1");

run;



title1 "Plot coefficients";



proc sgplot data=ParamEsttotal1 noautolegend;

        title3'Effect of time since legalization on incidence';

        title4'all ages, fixed effects for time and state categories';

        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL

        markerattrs=(symbol=diamondfilled);

   refline 0 / axis=y;

   xaxis grid;

   yaxis grid display=(nolabel) discreteorder=data;

run;



proc sort data=sav.Legal_date_21 ;
```

```
        by years;
run;
proc glm data=sav.Legal_date_21 ;
        title3 "Regression for event study";
        absorb years;
        class legal_cat (ref="Illegal") legal(ref="0");
        model incidence = legal_cat legal /solution CLPARM ;
run;
/*small because includes data from before effective date...*/


proc glm data=sav.Legal_date_21 ;
        title3 "Regression for event study, 20 and younger";
        absorb years;
        class legal_cat (ref="Illegal") effective(ref="0");
        where age="12_20";
        model incidence = legal_cat effective /solution CLPARM ;
run;
```

```
/************************************************/
/* analyze legal date with balanced lags and leads */
/************************************************/

title1 "Regression modelling with balanced leads and lags";

proc sort data=sav.Legal_date_21;
        by years;
run;

ods graphics on;
proc glm data=sav.Legal_date_21 ;
        title3 "Regression for panel event study, all ages";
        absorb years;
        class legal_cat (ref="Illegal") tlt6(ref="0") tminus4(ref="0") t0(ref="0") t2(ref="0")
tgt4(ref="0");
        model incidence = legal_cat tlt6 tminus4 t0 t2 tgt4 /solution CLPARM ;
        ods output ParameterEstimates = ParamEst21;
```

```
run;

title1"Manage Regression Output";

data ParamEst211;
        set ParamEst21 end=eof;
        where StdErr >. and Parameter in ("tlt6      1",
        "tminus4   1",
        "t0        1",
        "t2        1",
        "tgt4      1");

        if parameter = "tlt6      1" then do;
                parameter = '6+ years prior' ;
                order=1;
        end;
        if parameter = "tminus4   1" then do;
                parameter = '4 years prior';
```

```
        order=2;
end;
if parameter = "t0       1" then do;
        parameter = 'Legalized';
        order=4;
end;
if parameter = "t2       1" then do;
        parameter = '2 years after';
        order=5;
end;
if parameter = "tgt4     1" then do;
        parameter = '4+ years after';
        order=6;
end;
if eof then do;
output;
        parameter = '2 years prior';
        estimate=0;
```

```
                stderr=0;
                tvalue=0;
                probt=0;
                lowercl=0;
                uppercl=0;
                order=3;
        end;
        output;
run;

proc sort data=ParamEst211;
        by order;
run;

title1 "Plot coefficients";

proc sgplot data=ParamEst211 noautolegend;
        title3'Effect of time since cannabis legalization on cannabis incidence';
```

```
        title4'ages 21 and up, balanced leads and lags, fixed effects for time and state
categories';
        scatter  x=parameter y=Estimate / yerrorlower=LowerCL yerrorupper=upperCL
        markerattrs=(symbol=diamondfilled);
    refline 0 / axis=y;
    xaxis grid label='Time relative to legalization';
    yaxis grid label='Newly incident cannabis use' discreteorder=data;
run;

/*Look at average incidence by group to determine % change*/



proc contents; run;


data dat;
        set sav.Legal_date_21;


        if legal_cat = "Illegal" then legal=0;
```

```
        else if legal_cat = "Overall" then legal=.;

        else if time=0 and legal_cat ~= "Illegal" then legal=1;

        else if time<0 then legal=0;

        else if time>0 then legal=1;



        label    legal = "cannabis leagality binary";

run;



proc means data=dat;

        var incidence;

        class legal time;

run;



proc means data=dat;

        var incidence;

        class legal years;

run;
```

R

```
library(ROCR)

library(reshape2)

library(dplyr)

library(plyr)

library(tidyr)

library(FactoMineR)

library(factoextra)

library(caret)

library(InformationValue)

library(ISLR)

library(verification)

library(Epi)

library(pROC)



## ## ----------------------------------------------------
```

```
## ## read-in dataset

## ## ------------------------------------------------------

remove(list=ls()); gc()

dat <-read.csv("C:/Users/montg/Dropbox/Ph.D
Work/Dissertation/Data/Aim 1/Processed/County_RCL_LAG.csv", header
= T,
              na.strings=c("",".","NA", "NULL",  "N/A"))

  dim(dat)

#  [1] 3142 1598



## make all names lower cases

names(dat) <- tolower(names(dat))



## ------------------------------------------

## get rid of census data and replace with pca

## ------------------------------------------
```

```
kp <- c(grep("t0", names(dat), value = TRUE))

 ##  ,("area", names(dat), value = TRUE)) ## adds land area to PCA grep



pop.dat <- dat[, kp]

dat.pca <- prcomp(pop.dat, scale = TRUE)



## percent of variance explained

## first 2 PCS explain 79.3% of variabiility

fviz_eig(dat.pca, addlabels = TRUE)



## individual-level information

ind <- get_pca_ind(dat.pca)



## add PC information

dat$PCA1 <- ind$coord[,1]

dat$PCA2 <- ind$coord[,2]
```

```
dat$PCA3 <- ind$coord[,3]

dat$PCA4 <- ind$coord[,4]

dat$PCA5 <- ind$coord[,5]

dat$PCA6 <- ind$coord[,6]

dat$PCA7 <- ind$coord[,7]

dat$PCA8 <- ind$coord[,8]

dat$PCA9 <- ind$coord[,9]

dat$PCA10 <- ind$coord[,10]



dat <- dat %>% dplyr::select(-kp)

rm(pop.dat); rm(dat.pca); rm(ind); rm(kp)



#check PCs added correctly

glimpse(dat)



## Save
```

```
saveRDS(dat, file = "C:/Users/montg/Dropbox/Ph.D
Work/Dissertation/Data/Aim 1/Processed/clean_data.rds")

library(ROCR)

library(reshape2)

library(dplyr)

library(plyr)

library(tidyr)

library(FactoMineR)

library(factoextra)

library(caret)

library(InformationValue)

library(ISLR)

library(verification)

library(Epi)

library(pROC)
```

```
## ## ------------------------------------------------------

## ## read-in dataset

## ## ------------------------------------------------------

remove(list=ls()); gc()

dat <- readRDS(file = "C:\\Users\\montg\\Dropbox\\Ph.D

Work\\Dissertation\\Data\\Aim 1\\Processed\\clean_data.rds")



  dim(dat)



## make all names lower cases

names(dat) <- tolower(names(dat))



## ------------------------------------------------------

## remove duplicated variables

## ------------------------------------------------------

rmv <- c("name",
```

"geography",

"fips_pch",

"dem_2008_votes",

"rep_2008_votes",

"other_2008_votes",

"dem_2012_votes",

"rep_2012_votes",

"other_2012_votes",

"total_2008_votes",

"total_2012_votes",

"percentdem_2008_votes",

"percentrep_2008_votes",

"percentother_2008_votes",

"state",

"county",

"region",

```
        "division",
        "state_fips",
        "division_n"
        )
dat <- dat %>% dplyr::select(-rmv)
rm(rmv)


dim(dat)


## -----------------------------------------------------
## remove duplicate NSDUH variables
## -----------------------------------------------------
## keep only 12+ and 18+ estimates
rmv <- c(grep("twelve_to", names(dat), value = TRUE),
        grep("eighteen_to", names(dat), value = TRUE),
        grep("twenty_six", names(dat), value = TRUE))
```

```
dat <- dat %>% dplyr::select(-rmv)

rm(rmv)



# now, remove eighteen plus variables for which we have twelve plus
estimates

rmv <- c(grep("eighteen_plus_pymjinc", names(dat), value = TRUE),
         grep("eighteen_plus_pmalc", names(dat), value = TRUE),
         grep("eighteen_plus_pmcig", names(dat), value = TRUE),
         grep("eighteen_plus_pmmj", names(dat), value = TRUE),
         grep("eighteen_plus_pmtob", names(dat), value = TRUE),
         grep("eighteen_plus_pyalc", names(dat), value = TRUE),
         grep("eighteen_plus_pyaud", names(dat), value = TRUE),
         grep("eighteen_plus_pycoc", names(dat), value = TRUE),
         grep("eighteen_plus_pysud", names(dat), value = TRUE),
         grep("eighteen_plus_pycoc", names(dat), value = TRUE),
         grep("eighteen_plus_pymj", names(dat), value = TRUE)
```

```
        )

dat <- dat %>% dplyr::select(-rmv)

rm(rmv)



# all duplicate age groups removed



## ---------------------

## Variable Selection

## ---------------------

# # Excluded Eighteen_plus_PYAMIPrevestimate and
Eighteen_plus_PYMDEPrevestimate

# because mental health variables are about 70-80% correlated with each
other,

# SMI and ST are least correlated with each other (69%) and somewhat
correlated with policy.

# # Excluded Twelve_plus_PMTobPrevestimate since tobacco use is 96%
correlated with
```

```
# cigarettes and cigarettes are the better predictor for policy.

# # Excluded Twelve_plus_PYAlcDepPrevestimate because Alcohol
Dependence is

# the older (DSM-IV) way of coding AUD, it does not appear in all years
and is

# 86% correlated with AUD.

# # Excluded Twelve_plus_PYMJPrevestimate and
Twelve_plus_PYMJIncestimate.

# Among the marijuana variables, all are 80-90% correlated with each
other

# and past month prevalence is the best predictor of policy of the three.



rmv <- c(grep("eighteen_plus_pyamiprev", names(dat), value = TRUE),

       grep("eighteen_plus_pymdeprev", names(dat), value = TRUE),

        grep("twelve_plus_pmtobprev", names(dat), value = TRUE),

        grep("twelve_plus_pyalcdepprev", names(dat), value = TRUE),

        grep("twelve_plus_pymjprev", names(dat), value = TRUE),
```

```
        grep("twelve_plus_pymjinc", names(dat), value = TRUE)

        )

dat <- dat %>% dplyr::select(-rmv)

rm(rmv)



 dim(dat)



## Save

saveRDS(dat, file = "C:/Users/montg/Dropbox/Ph.D

Work/Dissertation/Data/Aim 1/Processed//final_data.rds")

library(ggplot2)

library(pROC)



set.seed(1000)

## -----------------------------------------------------

## read-in dataset
```

```
## ------------------------------------------------------

remove(list=ls()); gc()



dat <- readRDS("C:\\Users\\montg\\Dropbox\\Ph.D

Work\\Dissertation\\Data\\Aim 1\\Processed/final_data.rds")



## remove "Other" votes and alternate specifications for the outcomes

dat <- subset(dat, se=-c(percentdem_2012_votes, rcl_2012, sens_2014,

sens2_2014, sens3_2014))



## make all names lower cases

names(dat) <- tolower(names(dat))



## drop PCs if desired

dat <- subset(dat, se=-c(pca3, pca4, pca5, pca6, pca7, pca8, pca9,

pca10))
```

```
## drop missing outcomes from analysis

dat <- subset(dat, !is.na(lag_2014))

## dat <- within(dat,                  # xt: standardize large numbers

## {

##     area_water <- as.vector(scale(area_water))

##     area_land  <- as.vector(scale(area_land))

## })

dat <- na.omit(dat)                 # xt: 3103 to 3094



## ------------------------------------------------------

## Create macro variables for L and seL

## ------------------------------------------------------

seL <- grep("sel$", names(dat), value = TRUE)             # xt: sd of log

muL <- setdiff(grep("l$", names(dat), value = TRUE), seL)  # xt: mu of log

vnL <- sub('l$', '', muL)                                 # xt: var names

P <- length(vnL)
```

```
## ------------------------------------------------------

## define train/test sample and case control ratio

## ------------------------------------------------------

## ratio of train-to-test data split

trn <- 0.7; tst <- 0.3

## ratio of legalized-to-nonlegalized to sample

zrs <- 1.6; ons <- .8



## ------------------------------------------------------

## define objects to store results

## ------------------------------------------------------

var.lst <- list() # data to store variable importance

prd.lst <- list() # data to store county-level predictions

roc.lst <- list() # data to store AUC stuff

acc.lst <- list() # data to store accurary assessment
```

```
## ------------------------------------------------------

## start simulations

## ------------------------------------------------------

B <- 1e3

for(i in 1:B)

{

   ## bootstrap, training and testing DO NOT share counties.

   tmp <- with(dat,

   {

      i0 <- which(lag_2014 == 0); n0 <- length(i0) # 0s

      i1 <- which(lag_2014 == 1); n1 <- length(i1) # 1s

      ## divide (a) training and (b) testing, then permute

      a0 <- sample(i0, n0 * trn); b0 <- sample(setdiff(i0, a0))

      a1 <- sample(i1, n1 * trn); b1 <- sample(setdiff(i1, a1))

      ## bootstrap, number of 1s as the basic number
```

```
    a <- c(rep(a0, len=n1 * zrs * trn), rep(a1, len=n1 * ons * trn))
    b <- c(rep(b0, len=n1 * zrs * tst), rep(b1, len=n1 * ons * tst))
    list(a=sort(a), b=sort(b))
  })
  tmp <- rbind(cbind(dat='a', dat[tmp$a, ]), cbind(dat='b', dat[tmp$b, ]))
  N <- nrow(tmp)
  ## check
  with(tmp, intersect(qname[dat=='a'], qname[dat=='b'])) # must be
empty
  with(tmp, table(dat, lag_2014)) # row should be trn:tst, col should be
zro:ons


  ## ------------------------------------------------
  ## introduce variability to NSDUH RDAS estimates
  ## ------------------------------------------------
  ## xt: expand standard deviation to 3 times
```

```
    .x <- rnorm(N * P, unlist(tmp[, muL]), unlist(tmp[, seL]))

    .x <- matrix(1 / (1 + exp(-.x)), N, P, dimnames=list(NULL, vnL))

    tmp <- cbind(tmp[, !grepl("^(twelve|eighteen)", names(tmp))], .x)



    ## --------------------------

    ## Perform training testing split

    ## --------------------------

    A <- subset(tmp, dat == 'a', -dat)

    B <- subset(tmp, dat == 'b', -dat)



    ## --------------------------

    ## fit a GLM model, predict testing data

    ## --------------------------

    mld <- glm(lag_2014 ~ ., data    = subset(A, se=-qname), family =
'binomial')
```

```
    pht <- predict(mld,      newdata = subset(B, se=-qname), type   =
'response')


    ## ------------------------------------
    ## store county-level predictions
    ## ------------------------------------
    ## xt: fraction of 0s as threshold
    res <- cbind(sim=i, B, pht=pht, yht=0 + (pht > mean(B$lag_2014 < 1)))
    prd.lst[[i]] <- res


    ## ------------------------------------
    ## store predictor statistics
    ## ------------------------------------
    var.lst[[i]] <- data.frame(sim=i, variables=names(coef(mld))[-1],
Z=summary(mld)$coef[-1, 3], row.names=NULL)
```

```
    ## ------------------------------------

    ## store prediction accuracy results

    ## ------------------------------------

    tbl <- with(res, table(lag_2014, yht))

    acc.lst[[i]] <- data.frame(

       sim=i,

       TP  = tbl[2, 2], TN = tbl[1, 1], FP = tbl[1, 2], FN = tbl[2, 1],

       TPF = tbl[2, 2] / sum(tbl[2, ]), TNF = tbl[1, 1] / sum(tbl[1, ]),

       FPF = tbl[1, 2] / sum(tbl[1, ]), FNF = tbl[2, 1] / sum(tbl[2, ]),

       acc = sum(diag(tbl)) / sum(tbl))

    ## -------------------------

    ## create a ROC curve

    ## -------------------------

    ROC <- with(res, roc(lag_2014, pht, quiet = TRUE))

    roc.lst[[i]] <- with(ROC, data.frame(sim=i, TPR=rev(sensitivities),
FPR=rev(1 - specificities)))
```

```
}

var2 <- do.call(rbind, var.lst)    # predictor z-scores

acc2 <- do.call(rbind, acc.lst)    # accurary performance

roc2 <- do.call(rbind, roc.lst)    # roc

prd2 <- do.call(rbind, prd.lst)    # truth and prediction, per county, per sim



## ROC for all simulations combined (overall)

rocA <- with(prd2, roc(lag_2014, pht, quiet = TRUE))

aucA <- round(auc(rocA), 3)

rocA <- with(rocA, data.frame(TPR=rev(sensitivities), FPR=rev(1-

specificities)))



## ------------------------------------

## plot all ROC curves with the average

## ------------------------------------
```

```
setwd("C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim
1/results/")

g <- ggplot() + theme_bw()

g <- g + geom_abline(intercept = 0, slope = 1, size = 0.5, linetype =
"dashed") # ref

g <- g + geom_line(aes(x=FPR, y=TPR, group=sim), roc2, alpha=0.3) # per
sim

g <- g + geom_line(aes(x=FPR, y=TPR), rocA, color="red", size=2)    #
overall

g <- g + geom_text(aes(x = 0.75, y=0.4, label=paste0("AUC: ", aucA)),
color="red", size=10)

g <- g + scale_x_continuous(limits = c(0, 1)) + scale_y_continuous(limits =
c(0, 1))

ggsave("roc_glm.png", g)



## ----------------------------------------

## summarize predictive power of variables
```

```
## ------------------------------------------

round(do.call(rbind, with(var2, by(Z, variables, summary))))[, 2:5], 3)



#                          1st Qu. Median       Mean 3rd Qu.

# area_land                0.733  1.346  320390.08   1.932

# area_water               0.111  0.467  228688.61   0.829

# eighteen_plus_pysmiprev   0.527  1.351  441991.16   1.986

# eighteen_plus_pystprev   -1.107 -0.326 -194993.85   0.332

# pca1                     -1.649 -0.715 -195687.63   0.555

# pca2                     -1.588 -0.802 -216793.23  -0.007

# percentdem_2012_votes    -1.619 -0.912 -598545.37  -0.202

# percentrep_2012_votes    -1.782 -1.100 -612526.57  -0.392

# twelve_plus_pmalcprev    -0.004  0.618  268324.83   1.314

# twelve_plus_pmcigprev    -1.567 -0.829 -362955.85  -0.010

# twelve_plus_pmmjprev      2.203  2.871 1202101.03   3.326

# twelve_plus_pyaudprev    -0.085  0.536  392370.13   1.279
```

twelve_plus_pycocprev 1.636 2.239 333135.16 2.685

twelve_plus_pysudprev -1.320 -0.481 35547.05 0.260

Summarize Classification Accuracies

----

summary(acc2)

sim TP TN FP FN TPF

Min. : 1.0 Min. : 7.00 Min. :33.00 Min. : 0.000 Min. : 0.000 Min. :0.3182

1st Qu.: 250.8 1st Qu.:16.00 1st Qu.:40.00 1st Qu.: 2.000 1st Qu.: 3.000 1st Qu.:0.7273

Median : 500.5 Median :17.00 Median :41.00 Median : 3.000 Median : 5.000 Median :0.7727

Mean : 500.5 Mean :17.05 Mean :40.97 Mean : 3.026 Mean : 4.945 Mean :0.7752

3rd Qu.: 750.2 3rd Qu.:19.00 3rd Qu.:42.00 3rd Qu.: 4.000 3rd Qu.: 6.000 3rd Qu.:0.8636

Max. :1000.0 Max. :22.00 Max. :44.00 Max. :11.000 Max. :15.000 Max. :1.0000

TNF FPF FNF acc

Min. :0.7500 Min. :0.00000 Min. :0.0000 Min. :0.7576

1st Qu.:0.9091 1st Qu.:0.04545 1st Qu.:0.1364 1st Qu.:0.8485

Median :0.9318 Median :0.06818 Median :0.2273 Median :0.8788

Mean :0.9312 Mean :0.06877 Mean :0.2248 Mean :0.8792

3rd Qu.:0.9545 3rd Qu.:0.09091 3rd Qu.:0.2727 3rd Qu.:0.9091

Max. :1.0000 Max. :0.25000 Max. :0.6818 Max. :0.9848

#slightly lower accuracy but much better balance of sens and spec

---

ensemble predictions per county using different weights

```
## ------------------------------------------------

WTS <- c('TPF', 'TNF', 'FNF', 'acc')

mrg <- merge(prd2, acc2, by="sim")[, c(names(prd2), WTS)]

ens <- by(mrg, mrg$qname, function(x)

{

   wts <- t(colMeans(x[, 'pht'] * x[, WTS])) # esemble voted probability

   data.frame(

      county = x[1, 'qname'], N = nrow(x), Y = x[1, 'lag_2014'],

      YHT = mean(x[, 'yht']), # mean of per-simulation hard-call

      PHT = mean(x[, 'pht']), # mean of per-simulation preditive probability

      wts)

})

ens <- data.frame(do.call(rbind, ens), row.names=NULL)

## check performance

CRI <- c("YHT", "PHT", WTS) # criteria

round(cor(ens[, 'Y'], ens[, CRI]), 3)
```

```
## make hard-calls

THD <- zrs / (zrs + ons) # xt: fraction of 0s as threshold

hdc <- cbind(ens[, !names(ens) %in% CRI], (ens[, CRI] > THD) + 0)



with(hdc, table(Y, TNF))

##    TNF

## Y      0    1

##   0 2751  251

##   1    3   89



with(hdc, table(Y, FNF))

##    FNF

## Y      0

##   0 3002

##   1   92
```

```
with(hdc, table(Y, acc))
##    acc
## Y      0    1
##   0 2746  256
##   1    3   89

with(hdc, table(Y, PHT))
##    PHT
## Y      0    1
##   0 2732  270
##   1    1   91

with(hdc, table(Y, YHT))
##    YHT
## Y      0    1
```

```
##   0 2734  268

##   1    1   91



## Counties with false positive predictions

subset(ens, Y == 0 & TNF>THD)



## Output for SAS mapping

write.csv(ens, file='C:\\Users\\montg\\Dropbox\\Ph.D

Work\\Dissertation\\Data\\Aim 1\\Processed/ens_lag.csv')

library(ggplot2)

library(pROC)



set.seed(1000)

## ---------------------------------------------------

## read-in dataset

## ---------------------------------------------------
```

```
remove(list=ls()); gc()


dat <- readRDS("C:/Users/montg/Dropbox/Ph.D
Work/Dissertation/Data/Aim 1/Processed/final_data.rds")


## remove "Other" votes and alternate specifications for the outcomes
dat <- subset(dat, se=-c(percentother_2012_votes, rcl_2012, lag_2014,
sens2_2014, sens3_2014))


## make all names lower cases
names(dat) <- tolower(names(dat))


## drop PCs if desired
dat <- subset(dat, se=-c(pca3, pca4, pca5, pca6, pca7, pca8, pca9,
pca10))


## drop missing outcomes from analysis
```

```
dat <- subset(dat, !is.na(sens_2014))

## dat <- within(dat,                  # xt: standardize large numbers

## {

##     area_water <- as.vector(scale(area_water))

##     area_land  <- as.vector(scale(area_land))

## })

dat <- na.omit(dat)                 # 3142 to 3133



## ------------------------------------------------------

## Create macro variables for L and seL

## ------------------------------------------------------

seL <- grep("sel$", names(dat), value = TRUE)             # xt: sd of log

muL <- setdiff(grep("l$", names(dat), value = TRUE), seL)  # xt: mu of log

vnL <- sub('l$', '', muL)                                 # xt: var names

P <- length(vnL)
```

```
## ------------------------------------------------------

## define data slit sample samples

## ------------------------------------------------------

## ratio of train-to-test data split

trn <- 0.7; tst <- 0.3

## ratio of legalized-to-nonlegalized to sample

zrs <- 1.6; ons <- .8



## ------------------------------------------------------

## define objects to store results

## ------------------------------------------------------

var.lst <- list() # data to store variable importance

prd.lst <- list() # data to store county-level predictions

roc.lst <- list() # data to store AUC stuff

acc.lst <- list() # data to store accurary assessment
```

```
## ------------------------------------------------------

## start simulations

## ------------------------------------------------------

B <- 1e3

for(i in 1:B)

{

   ## bootstrap, training and testing DO NOT share counties.

   tmp <- with(dat,

   {

      i0 <- which(sens_2014 == 0); n0 <- length(i0) # 0s

      i1 <- which(sens_2014 == 1); n1 <- length(i1) # 1s

      ## divide (a) training and (b) testing, then permute

      a0 <- sample(i0, n0 * trn); b0 <- sample(setdiff(i0, a0))

      a1 <- sample(i1, n1 * trn); b1 <- sample(setdiff(i1, a1))

      ## bootstrap, number of 1s as the basic number

      a <- c(rep(a0, len=n1 * zrs * trn), rep(a1, len=n1 * ons * trn))
```

```
    b <- c(rep(b0, len=n1 * zrs * tst), rep(b1, len=n1 * ons * tst))
    list(a=sort(a), b=sort(b))
  })
  tmp <- rbind(cbind(dat='a', dat[tmp$a, ]), cbind(dat='b', dat[tmp$b, ]))
  N <- nrow(tmp)
  ## check
  with(tmp, intersect(qname[dat=='a'], qname[dat=='b'])) # must be
empty
  with(tmp, table(dat, sens_2014)) # row should be trn:tst, col should be
zro:ons


  ## ------------------------------------------------
  ## introduce variability to NSDUH RDAS estimates
  ## ------------------------------------------------
  ## xt: expand standard deviation to 3 times
  .x <- rnorm(N * P, unlist(tmp[, muL]), unlist(tmp[, seL]))
```

```
    .x <- matrix(1 / (1 + exp(-.x)), N, P, dimnames=list(NULL, vnL))

    tmp <- cbind(tmp[, !grepl("^(twelve|eighteen)", names(tmp))], .x)



    ## --------------------------

    ## Perform training testing split

    ## --------------------------

    A <- subset(tmp, dat == 'a', -dat)

    B <- subset(tmp, dat == 'b', -dat)



    ## --------------------------

    ## fit a GLM model, predict testing data

    ## --------------------------

    mld <- glm(sens_2014 ~ ., data    = subset(A, se=-qname), family =
'binomial')

    pht <- predict(mld,      newdata = subset(B, se=-qname), type   =
'response')
```

```
    ## ------------------------------------

    ## store county-level predictions

    ## ------------------------------------

    ## xt: fraction of 0s as threshold

    res <- cbind(sim=i, B, pht=pht, yht=0 + (pht > mean(B$sens_2014 < 1)))

    prd.lst[[i]] <- res



    ## ------------------------------------

    ## store predictor statistics

    ## ------------------------------------

    var.lst[[i]] <- data.frame(sim=i, variables=names(coef(mld))[-1],
Z=summary(mld)$coef[-1, 3], row.names=NULL)



    ## ------------------------------------

    ## store prediction accuracy results
```

```
    ## ------------------------------------

    tbl <- with(res, table(sens_2014, yht))

    acc.lst[[i]] <- data.frame(

       sim=i,

       TP  = tbl[2, 2], TN = tbl[1, 1], FP = tbl[1, 2], FN = tbl[2, 1],

       TPF = tbl[2, 2] / sum(tbl[2, ]), TNF = tbl[1, 1] / sum(tbl[1, ]),

       FPF = tbl[1, 2] / sum(tbl[1, ]), FNF = tbl[2, 1] / sum(tbl[2, ]),

       acc = sum(diag(tbl)) / sum(tbl))

    ## -------------------------

    ## create a ROC curve

    ## -------------------------

    ROC <- with(res, roc(sens_2014, pht, quiet = TRUE))

    roc.lst[[i]] <- with(ROC, data.frame(sim=i, TPR=rev(sensitivities),
FPR=rev(1 - specificities)))
}

var2 <- do.call(rbind, var.lst)    # predictor z-scores
```

```
acc2 <- do.call(rbind, acc.lst)    # accurary performance

roc2 <- do.call(rbind, roc.lst)    # roc

prd2 <- do.call(rbind, prd.lst)    # truth and prediction, per county, per sim



## ROC for all simulations combined (overall)

rocA <- with(prd2, roc(sens_2014, pht, quiet = TRUE))

aucA <- round(auc(rocA), 3)

rocA <- with(rocA, data.frame(TPR=rev(sensitivities), FPR=rev(1-

specificities)))



## ---------------------------------------

## plot all ROC curves with the average

## ---------------------------------------

setwd("C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim

1/results/")

g <- ggplot() + theme_bw()
```

```
g <- g + geom_abline(intercept = 0, slope = 1, size = 0.5, linetype =
"dashed") # ref

g <- g + geom_line(aes(x=FPR, y=TPR, group=sim), roc2, alpha=0.3) # per
sim

g <- g + geom_line(aes(x=FPR, y=TPR), rocA, color="red", size=2)    #
overall

g <- g + geom_text(aes(x = 0.75, y=0.4, label=paste0("AUC: ", aucA)),
color="red", size=10)

g <- g + scale_x_continuous(limits = c(0, 1)) + scale_y_continuous(limits =
c(0, 1))

ggsave("roc_glm_sens.png", g)



## ----------------------------------------

## summarize predictive power of variables

## ----------------------------------------

round(do.call(rbind, with(var2, by(Z, variables, summary)))[, 2:5], 3)
```

```
# area_land                0.963  1.559  1.583   2.176

# area_water               0.327  0.656  0.645   0.967

# eighteen_plus_pysmiprev   2.386  2.884  2.815   3.343

# eighteen_plus_pystprev   -0.377  0.367  0.338   1.084

# pca1                    -1.692 -0.907 -0.614   0.298

# pca2                    -1.669 -0.898 -0.831  -0.135

# percentdem_2012_votes    -1.805 -1.048 -1.108  -0.418

# percentrep_2012_votes    -1.945 -1.215 -1.282  -0.665

# twelve_plus_pmalcprev    -0.153  0.521  0.504   1.251

# twelve_plus_pmcigprev    -2.107 -1.465 -1.413  -0.798

# twelve_plus_pmmjprev      2.976  3.552  3.406   4.028

# twelve_plus_pyaudprev     0.195  0.879  0.855   1.578

# twelve_plus_pycocprev     1.681  2.288  2.175   2.814

# twelve_plus_pysudprev    -1.240 -0.570 -0.521   0.167


## ----------------------------------------
```

Summarize Classification Accuracies

----

summary(acc2)

sim TP TN FP FN TPF

Min. : 1.0 Min. :10.00 Min. :47.00 Min. : 0.000 Min. : 0.000 Min. :0.3448

1st Qu.: 250.8 1st Qu.:20.00 1st Qu.:53.00 1st Qu.: 2.000 1st Qu.: 5.000 1st Qu.:0.6897

Median : 500.5 Median :22.00 Median :54.00 Median : 4.000 Median : 7.000 Median :0.7586

Mean : 500.5 Mean :21.89 Mean :54.12 Mean : 3.882 Mean : 7.114 Mean :0.7547

3rd Qu.: 750.2 3rd Qu.:24.00 3rd Qu.:56.00 3rd Qu.: 5.000 3rd Qu.: 9.000 3rd Qu.:0.8276

Max. :1000.0 Max. :29.00 Max. :58.00 Max. :11.000 Max. :19.000 Max. :1.0000

TNF FPF FNF acc

```
#  Min.   :0.8103   Min.   :0.00000   Min.   :0.0000   Min.   :0.7241

#  1st Qu.:0.9138   1st Qu.:0.03448   1st Qu.:0.1724   1st Qu.:0.8506

#  Median :0.9310   Median :0.06897   Median :0.2414   Median :0.8736

#  Mean   :0.9331   Mean   :0.06693   Mean   :0.2453   Mean   :0.8736

#  3rd Qu.:0.9655   3rd Qu.:0.08621   3rd Qu.:0.3103   3rd Qu.:0.8966

#  Max.   :1.0000   Max.   :0.18966   Max.   :0.6552   Max.   :0.9655



#slightly lower accuracy but much better balance of sens and spec



## ------------------------------------------------

## ensemble predictions per county using different weights

## ------------------------------------------------

WTS <- c('TPF', 'TNF', 'FNF', 'acc')

mrg <- merge(prd2, acc2, by="sim")[, c(names(prd2), WTS)]

ens <- by(mrg, mrg$qname, function(x)

{
```

```
    wts <- t(colMeans(x[, 'pht'] * x[, WTS])) # esemble voted probability
    data.frame(
        county = x[1, 'qname'], N = nrow(x), Y = x[1, 'sens_2014'],
        YHT = mean(x[, 'yht']), # mean of per-simulation hard-call
        PHT = mean(x[, 'pht']), # mean of per-simulation preditive probability
        wts)
})
ens <- data.frame(do.call(rbind, ens), row.names=NULL)
## check performance
CRI <- c("YHT", "PHT", WTS) # criteria
round(cor(ens[, 'Y'], ens[, CRI]), 3)



## make hard-calls
THD <- zrs / (zrs + ons) # xt: fraction of 0s as threshold
hdc <- cbind(ens[, !names(ens) %in% CRI], (ens[, CRI] > THD) + 0)
```

```
with(hdc, table(Y, TNF))

with(hdc, table(Y, FNF))

with(hdc, table(Y, acc))

with(hdc, table(Y, PHT))

with(hdc, table(Y, YHT))

## Output for SAS mapping

write.csv(ens, file='C:\\Users\\montg\\Dropbox\\Ph.D
Work\\Dissertation\\Data\\Aim 1\\Processed/ens_sens1.csv')

library(ggplot2)

library(pROC)
```

```
set.seed(1000)

## ---------------------------------------------------------

## read-in dataset

## ---------------------------------------------------------

remove(list=ls()); gc()



dat <- readRDS("C:/Users/montg/Dropbox/Ph.D
Work/Dissertation/Data/Aim 1/Processed/final_data.rds")



## remove "Other" votes and alternate specifications for the outcomes

dat <- subset(dat, se=-c(percentother_2012_votes, rcl_2012, lag_2014,
sens_2014, sens3_2014))



## make all names lower cases

names(dat) <- tolower(names(dat))
```

```
## drop PCs if desired

dat <- subset(dat, se=-c(pca3, pca4, pca5, pca6, pca7, pca8, pca9,

pca10))



## drop missing outcomes from analysis

dat <- subset(dat, !is.na(sens2_2014))

## dat <- within(dat,                  # xt: standardize large numbers

## {

##     area_water <- as.vector(scale(area_water))

##     area_land  <- as.vector(scale(area_land))

## })

dat <- na.omit(dat)                 # 3142 to 3133



## ------------------------------------------------------

## Create macro variables for L and seL

## ------------------------------------------------------
```

```
seL <- grep("sel$", names(dat), value = TRUE)              # xt: sd of log

muL <- setdiff(grep("l$", names(dat), value = TRUE), seL)  # xt: mu of log

vnL <- sub('l$', '', muL)                                  # xt: var names

P <- length(vnL)



## ------------------------------------------------------

## define data slit sample samples

## ------------------------------------------------------

## ratio of train-to-test data split

trn <- 0.7; tst <- 0.3

## ratio of legalized-to-nonlegalized to sample

zrs <- 1.6; ons <- .8



## ------------------------------------------------------

## define objects to store results

## ------------------------------------------------------
```

```
var.lst <- list() # data to store variable importance

prd.lst <- list() # data to store county-level predictions

roc.lst <- list() # data to store AUC stuff

acc.lst <- list() # data to store accurary assessment



## ------------------------------------------------------

## start simulations

## ------------------------------------------------------

B <- 1e3

for(i in 1:B)

{

   ## bootstrap, training and testing DO NOT share counties.

   tmp <- with(dat,

   {

      i0 <- which(sens2_2014 == 0); n0 <- length(i0) # 0s

      i1 <- which(sens2_2014 == 1); n1 <- length(i1) # 1s
```

```
    ## divide (a) training and (b) testing, then permute
    a0 <- sample(i0, n0 * trn); b0 <- sample(setdiff(i0, a0))
    a1 <- sample(i1, n1 * trn); b1 <- sample(setdiff(i1, a1))
    ## bootstrap, number of 1s as the basic number
    a <- c(rep(a0, len=n1 * zrs * trn), rep(a1, len=n1 * ons * trn))
    b <- c(rep(b0, len=n1 * zrs * tst), rep(b1, len=n1 * ons * tst))
    list(a=sort(a), b=sort(b))
  })
  tmp <- rbind(cbind(dat='a', dat[tmp$a, ]), cbind(dat='b', dat[tmp$b, ]))
  N <- nrow(tmp)
  ## check
  with(tmp, intersect(qname[dat=='a'], qname[dat=='b'])) # must be
empty
  with(tmp, table(dat, sens2_2014)) # row should be trn:tst, col should be
zro:ons
```

```
## ------------------------------------------------

## introduce variability to NSDUH RDAS estimates

## ------------------------------------------------

## xt: expand standard deviation to 3 times

.x <- rnorm(N * P, unlist(tmp[, muL]), unlist(tmp[, seL]))

.x <- matrix(1 / (1 + exp(-.x)), N, P, dimnames=list(NULL, vnL))

tmp <- cbind(tmp[, !grepl("^(twelve|eighteen)", names(tmp))], .x)



## ---------------------------

## Perform training testing split

## ---------------------------

A <- subset(tmp, dat == 'a', -dat)

B <- subset(tmp, dat == 'b', -dat)



## ---------------------------

## fit a GLM model, predict testing data
```

```
    ## ---------------------------

    mld <- glm(sens2_2014 ~ ., data    = subset(A, se=-qname), family =
'binomial')

    pht <- predict(mld,      newdata = subset(B, se=-qname), type   =
'response')



    ## --------------------------------------

    ## store county-level predictions

    ## --------------------------------------

    ## xt: fraction of 0s as threshold

    res <- cbind(sim=i, B, pht=pht, yht=0 + (pht > mean(B$sens2_2014 <
1)))

    prd.lst[[i]] <- res



    ## ------------------------------------

    ## store predictor statistics

    ## ------------------------------------
```

```
    var.lst[[i]] <- data.frame(sim=i, variables=names(coef(mld))[-1],
Z=summary(mld)$coef[-1, 3], row.names=NULL)



    ## ------------------------------------

    ## store prediction accuracy results

    ## ------------------------------------

    tbl <- with(res, table(sens2_2014, yht))

    acc.lst[[i]] <- data.frame(

        sim=i,

        TP  = tbl[2, 2], TN = tbl[1, 1], FP = tbl[1, 2], FN = tbl[2, 1],

        TPF = tbl[2, 2] / sum(tbl[2, ]), TNF = tbl[1, 1] / sum(tbl[1, ]),

        FPF = tbl[1, 2] / sum(tbl[1, ]), FNF = tbl[2, 1] / sum(tbl[2, ]),

        acc = sum(diag(tbl)) / sum(tbl))

    ## -------------------------

    ## create a ROC curve

    ## -------------------------
```

```
   ROC <- with(res, roc(sens2_2014, pht, quiet = TRUE))

   roc.lst[[i]] <- with(ROC, data.frame(sim=i, TPR=rev(sensitivities),
FPR=rev(1 - specificities)))

}

var2 <- do.call(rbind, var.lst)    # predictor z-scores

acc2 <- do.call(rbind, acc.lst)    # accurary performance

roc2 <- do.call(rbind, roc.lst)    # roc

prd2 <- do.call(rbind, prd.lst)    # truth and prediction, per county, per sim



## ROC for all simulations combined (overall)

rocA <- with(prd2, roc(sens2_2014, pht, quiet = TRUE))

aucA <- round(auc(rocA), 3)

rocA <- with(rocA, data.frame(TPR=rev(sensitivities), FPR=rev(1-
specificities)))



## ----------------------------------------
```

```
## plot all ROC curves with the average

## ------------------------------------

setwd("C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim
1/results/")

g <- ggplot() + theme_bw()

g <- g + geom_abline(intercept = 0, slope = 1, size = 0.5, linetype =
"dashed") # ref

g <- g + geom_line(aes(x=FPR, y=TPR, group=sim), roc2, alpha=0.3) # per
sim

g <- g + geom_line(aes(x=FPR, y=TPR), rocA, color="red", size=2)    #
overall

g <- g + geom_text(aes(x = 0.75, y=0.4, label=paste0("AUC: ", aucA)),
color="red", size=10)

g <- g + scale_x_continuous(limits = c(0, 1)) + scale_y_continuous(limits =
c(0, 1))

ggsave("roc_glm_sens2.png", g)
```

```
## ----------------------------------------

## summarize predictive power of variables

## ----------------------------------------

round(do.call(rbind, with(var2, by(Z, variables, summary))))[, 2:5], 3)




## ----------------------------------------

## Summarize Classification Accuracies

## ----------------------------------------

summary(acc2)




## ------------------------------------------------

## ensemble predictions per county using different weights

## ------------------------------------------------
```

```
WTS <- c('TPF', 'TNF', 'FNF', 'acc')

mrg <- merge(prd2, acc2, by="sim")[, c(names(prd2), WTS)]

ens <- by(mrg, mrg$qname, function(x)

{

    wts <- t(colMeans(x[, 'pht'] * x[, WTS])) # esemble voted probability

    data.frame(

        county = x[1, 'qname'], N = nrow(x), Y = x[1, 'sens2_2014'],

        YHT = mean(x[, 'yht']), # mean of per-simulation hard-call

        PHT = mean(x[, 'pht']), # mean of per-simulation preditive probability

        wts)

})

ens <- data.frame(do.call(rbind, ens), row.names=NULL)

## check performance

CRI <- c("YHT", "PHT", WTS) # criteria

round(cor(ens[, 'Y'], ens[, CRI]), 3)
```

```
## make hard-calls

THD <- zrs / (zrs + ons) # xt: fraction of 0s as threshold

hdc <- cbind(ens[, !names(ens) %in% CRI], (ens[, CRI] > THD) + 0)



with(hdc, table(Y, TNF))



with(hdc, table(Y, FNF))



with(hdc, table(Y, acc))



with(hdc, table(Y, PHT))



with(hdc, table(Y, YHT))



## Output for SAS mapping
```

```
write.csv(ens, file='C:\\Users\\montg\\Dropbox\\Ph.D
Work\\Dissertation\\Data\\Aim 1\\Processed/ens_sens2.csv')
library(ggplot2)
library(pROC)


set.seed(1000)
## ----------------------------------------------------------
## read-in dataset
## ----------------------------------------------------------
remove(list=ls()); gc()


dat <- readRDS("C:/Users/montg/Dropbox/Ph.D
Work/Dissertation/Data/Aim 1/Processed/final_data.rds")


## remove "Other" votes and alternate specifications for the outcomes
dat <- subset(dat, se=-c(percentother_2012_votes, rcl_2012, lag_2014,
sens_2014, sens2_2014))
```

```
## make all names lower cases

names(dat) <- tolower(names(dat))



## drop PCs if desired

dat <- subset(dat, se=-c(pca3, pca4, pca5, pca6, pca7, pca8, pca9,

pca10))



## drop missing outcomes from analysis

dat <- subset(dat, !is.na(sens3_2014))

## dat <- within(dat,                   # xt: standardize large numbers

## {

##     area_water <- as.vector(scale(area_water))

##     area_land  <- as.vector(scale(area_land))

## })

dat <- na.omit(dat)                  # 3142 to 3133
```

```
## ------------------------------------------------------

## Create macro variables for L and seL

## ------------------------------------------------------

seL <- grep("sel$", names(dat), value = TRUE)              # xt: sd of log

muL <- setdiff(grep("l$", names(dat), value = TRUE), seL)  # xt: mu of log

vnL <- sub('l$', '', muL)                                  # xt: var names

P <- length(vnL)



## ------------------------------------------------------

## define data slit sample samples

## ------------------------------------------------------

## ratio of train-to-test data split

trn <- 0.7; tst <- 0.3

## ratio of legalized-to-nonlegalized to sample

zrs <- 1.6; ons <- .8
```

```
## ------------------------------------------------------

## define objects to store results

## ------------------------------------------------------

var.lst <- list() # data to store variable importance

prd.lst <- list() # data to store county-level predictions

roc.lst <- list() # data to store AUC stuff

acc.lst <- list() # data to store accurary assessment



## ------------------------------------------------------------

## start simulations

## ------------------------------------------------------------

B <- 1e3

for(i in 1:B)

{

   ## bootstrap, training and testing DO NOT share counties.
```

```
tmp <- with(dat,

{

   i0 <- which(sens3_2014 == 0); n0 <- length(i0) # 0s

   i1 <- which(sens3_2014 == 1); n1 <- length(i1) # 1s

   ## divide (a) training and (b) testing, then permute

   a0 <- sample(i0, n0 * trn); b0 <- sample(setdiff(i0, a0))

   a1 <- sample(i1, n1 * trn); b1 <- sample(setdiff(i1, a1))

   ## bootstrap, number of 1s as the basic number

   a <- c(rep(a0, len=n1 * zrs * trn), rep(a1, len=n1 * ons * trn))

   b <- c(rep(b0, len=n1 * zrs * tst), rep(b1, len=n1 * ons * tst))

   list(a=sort(a), b=sort(b))

})

tmp <- rbind(cbind(dat='a', dat[tmp$a, ]), cbind(dat='b', dat[tmp$b, ]))

N <- nrow(tmp)

## check
```

```
    with(tmp, intersect(qname[dat=='a'], qname[dat=='b'])) # must be
empty

    with(tmp, table(dat, sens3_2014)) # row should be trn:tst, col should be
zro:ons



    ## ------------------------------------------------

    ## introduce variability to NSDUH RDAS estimates

    ## ------------------------------------------------

    ## xt: expand standard deviation to 3 times

    .x <- rnorm(N * P, unlist(tmp[, muL]), unlist(tmp[, seL]))

    .x <- matrix(1 / (1 + exp(-.x)), N, P, dimnames=list(NULL, vnL))

    tmp <- cbind(tmp[, !grepl("^(twelve|eighteen)", names(tmp))], .x)



    ## ---------------------------

    ## Perform training testing split

    ## ---------------------------
```

```
    A <- subset(tmp, dat == 'a', -dat)

    B <- subset(tmp, dat == 'b', -dat)



    ## ---------------------------

    ## fit a GLM model, predict testing data

    ## ---------------------------

    mld <- glm(sens3_2014 ~ ., data    = subset(A, se=-qname), family =
'binomial')

    pht <- predict(mld,      newdata = subset(B, se=-qname), type   =
'response')



    ## ---------------------------------------

    ## store county-level predictions

    ## ---------------------------------------

    ## xt: fraction of 0s as threshold

    res <- cbind(sim=i, B, pht=pht, yht=0 + (pht > mean(B$sens3_2014 <
1)))
```

```
    prd.lst[[i]] <- res



    ## ------------------------------------

    ## store predictor statistics

    ## ------------------------------------

    var.lst[[i]] <- data.frame(sim=i, variables=names(coef(mld))[-1],
Z=summary(mld)$coef[-1, 3], row.names=NULL)



    ## ------------------------------------

    ## store prediction accuracy results

    ## ------------------------------------

    tbl <- with(res, table(sens3_2014, yht))

    acc.lst[[i]] <- data.frame(

       sim=i,

       TP  = tbl[2, 2], TN = tbl[1, 1], FP = tbl[1, 2], FN = tbl[2, 1],

       TPF = tbl[2, 2] / sum(tbl[2, ]), TNF = tbl[1, 1] / sum(tbl[1, ]),
```

```
      FPF = tbl[1, 2] / sum(tbl[1, ]), FNF = tbl[2, 1] / sum(tbl[2, ]),

      acc = sum(diag(tbl)) / sum(tbl))

   ## -------------------------

   ## create a ROC curve

   ## -------------------------

   ROC <- with(res, roc(sens3_2014, pht, quiet = TRUE))

   roc.lst[[i]] <- with(ROC, data.frame(sim=i, TPR=rev(sensitivities),
FPR=rev(1 - specificities)))

}

var2 <- do.call(rbind, var.lst)    # predictor z-scores

acc2 <- do.call(rbind, acc.lst)    # accurary performance

roc2 <- do.call(rbind, roc.lst)    # roc

prd2 <- do.call(rbind, prd.lst)    # truth and prediction, per county, per sim



## ROC for all simulations combined (overall)

rocA <- with(prd2, roc(sens3_2014, pht, quiet = TRUE))
```

```
aucA <- round(auc(rocA), 3)

rocA <- with(rocA, data.frame(TPR=rev(sensitivities), FPR=rev(1-
specificities)))



## ----------------------------------------

## plot all ROC curves with the average

## ----------------------------------------

setwd("C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim
1/results/")

g <- ggplot() + theme_bw()

g <- g + geom_abline(intercept = 0, slope = 1, size = 0.5, linetype =
"dashed") # ref

g <- g + geom_line(aes(x=FPR, y=TPR, group=sim), roc2, alpha=0.3) # per
sim

g <- g + geom_line(aes(x=FPR, y=TPR), rocA, color="red", size=2)    #
overall
```

```
g <- g + geom_text(aes(x = 0.75, y=0.4, label=paste0("AUC: ", aucA)),
color="red", size=10)
g <- g + scale_x_continuous(limits = c(0, 1)) + scale_y_continuous(limits =
c(0, 1))
ggsave("roc_glm_sens3.png", g)



## ----------------------------------------
## summarize predictive power of variables
## ----------------------------------------
round(do.call(rbind, with(var2, by(Z, variables, summary)))[, 2:5], 3)




## ----------------------------------------
## Summarize Classification Accuracies
## ----------------------------------------
```

```
summary(acc2)




## ------------------------------------------------

## ensemble predictions per county using different weights

## ------------------------------------------------

WTS <- c('TPF', 'TNF', 'FNF', 'acc')

mrg <- merge(prd2, acc2, by="sim")[, c(names(prd2), WTS)]

ens <- by(mrg, mrg$qname, function(x)

{

    wts <- t(colMeans(x[, 'pht'] * x[, WTS])) # esemble voted probability

    data.frame(

        county = x[1, 'qname'], N = nrow(x), Y = x[1, 'sens3_2014'],

        YHT = mean(x[, 'yht']), # mean of per-simulation hard-call

        PHT = mean(x[, 'pht']), # mean of per-simulation preditive probability

        wts)
```

```
})

ens <- data.frame(do.call(rbind, ens), row.names=NULL)

## check performance

CRI <- c("YHT", "PHT", WTS) # criteria

round(cor(ens[, 'Y'], ens[, CRI]), 3)



## make hard-calls

THD <- zrs / (zrs + ons) # xt: fraction of 0s as threshold

hdc <- cbind(ens[, !names(ens) %in% CRI], (ens[, CRI] > THD) + 0)



with(hdc, table(Y, TNF))



with(hdc, table(Y, FNF))



with(hdc, table(Y, acc))
```

```
with(hdc, table(Y, PHT))



with(hdc, table(Y, YHT))



## Output for SAS mapping

write.csv(ens, file='C:\\Users\\montg\\Dropbox\\Ph.D

Work\\Dissertation\\Data\\Aim 1\\Processed/ens_sens3.csv')

library(ggplot2)

library(pROC)



set.seed(1000)

## ---------------------------------------------------------

## read-in dataset

## ---------------------------------------------------------

remove(list=ls()); gc()
```

```
dat <- readRDS("C:\\Users\\montg\\Dropbox\\Ph.D
Work\\Dissertation\\Data\\Aim 1\\Processed/final_data_voterbins.rds")


## remove "Other" votes and alternate specifications for the outcomes
dat <- subset(dat, se=-c(percentdem_2012_votes, rcl_2012, sens_2014,
sens2_2014, sens3_2014))


## make all names lower cases
names(dat) <- tolower(names(dat))


## drop PCs if desired
dat <- subset(dat, se=-c(pca3, pca4, pca5, pca6, pca7, pca8, pca9,
pca10))


## drop missing outcomes from analysis
dat <- subset(dat, !is.na(lag_2014))
## dat <- within(dat,                  # xt: standardize large numbers
```

```
## {
##     area_water <- as.vector(scale(area_water))
##     area_land  <- as.vector(scale(area_land))
## })
dat <- na.omit(dat)                 # xt: 3103 to 3094


## ----------------------------------------------------
## Create macro variables for L and seL
## ----------------------------------------------------
seL <- grep("sel$", names(dat), value = TRUE)             # xt: sd of log
muL <- setdiff(grep("l$", names(dat), value = TRUE), seL)  # xt: mu of log
vnL <- sub('l$', '', muL)                                  # xt: var names
P <- length(vnL)


## ----------------------------------------------------
## define train/test sample and case control ratio
```

```
## ------------------------------------------------------

## ratio of train-to-test data split

trn <- 0.7; tst <- 0.3

## ratio of legalized-to-nonlegalized to sample

zrs <- 1.6; ons <- .8



## ------------------------------------------------------

## define objects to store results

## ------------------------------------------------------

var.lst <- list() # data to store variable importance

prd.lst <- list() # data to store county-level predictions

roc.lst <- list() # data to store AUC stuff

acc.lst <- list() # data to store accurary assessment



## ------------------------------------------------------

## start simulations
```

```
## ------------------------------------------------------

B <- 1e3

for(i in 1:B)

{

   ## bootstrap, training and testing DO NOT share counties.

   tmp <- with(dat,

   {

      i0 <- which(lag_2014 == 0); n0 <- length(i0) # 0s

      i1 <- which(lag_2014 == 1); n1 <- length(i1) # 1s

      ## divide (a) training and (b) testing, then permute

      a0 <- sample(i0, n0 * trn); b0 <- sample(setdiff(i0, a0))

      a1 <- sample(i1, n1 * trn); b1 <- sample(setdiff(i1, a1))

      ## bootstrap, number of 1s as the basic number

      a <- c(rep(a0, len=n1 * zrs * trn), rep(a1, len=n1 * ons * trn))

      b <- c(rep(b0, len=n1 * zrs * tst), rep(b1, len=n1 * ons * tst))

      list(a=sort(a), b=sort(b))
```

```
    })

    tmp <- rbind(cbind(dat='a', dat[tmp$a, ]), cbind(dat='b', dat[tmp$b, ]))

    N <- nrow(tmp)

    ## check

    with(tmp, intersect(qname[dat=='a'], qname[dat=='b'])) # must be
empty

    with(tmp, table(dat, lag_2014)) # row should be trn:tst, col should be
zro:ons



    ## ------------------------------------------------

    ## introduce variability to NSDUH RDAS estimates

    ## ------------------------------------------------

    ## xt: expand standard deviation to 3 times

    .x <- rnorm(N * P, unlist(tmp[, muL]), unlist(tmp[, seL]))

    .x <- matrix(1 / (1 + exp(-.x)), N, P, dimnames=list(NULL, vnL))

    tmp <- cbind(tmp[, !grepl("^(twelve|eighteen)", names(tmp))], .x)
```

```
    ## --------------------------

    ## Perform training testing split

    ## --------------------------

    A <- subset(tmp, dat == 'a', -dat)

    B <- subset(tmp, dat == 'b', -dat)



    ## --------------------------

    ## fit a GLM model, predict testing data

    ## --------------------------

    mld <- glm(lag_2014 ~ ., data    = subset(A, se=-qname), family =
'binomial')

    pht <- predict(mld,      newdata = subset(B, se=-qname), type   =
'response')



    ## ------------------------------------
```

```
    ## store county-level predictions

    ## ------------------------------------

    ## xt: fraction of 0s as threshold

    res <- cbind(sim=i, B, pht=pht, yht=0 + (pht > mean(B$lag_2014 < 1)))

    prd.lst[[i]] <- res



    ## ------------------------------------

    ## store predictor statistics

    ## ------------------------------------

    var.lst[[i]] <- data.frame(sim=i, variables=names(coef(mld))[-1],
Z=summary(mld)$coef[-1, 3], row.names=NULL)



    ## ------------------------------------

    ## store prediction accuracy results

    ## ------------------------------------

    tbl <- with(res, table(lag_2014, yht))
```

```
    acc.lst[[i]] <- data.frame(
        sim=i,
        TP  = tbl[2, 2], TN = tbl[1, 1], FP = tbl[1, 2], FN = tbl[2, 1],
        TPF = tbl[2, 2] / sum(tbl[2, ]), TNF = tbl[1, 1] / sum(tbl[1, ]),
        FPF = tbl[1, 2] / sum(tbl[1, ]), FNF = tbl[2, 1] / sum(tbl[2, ]),
        acc = sum(diag(tbl)) / sum(tbl))
    ## --------------------------
    ## create a ROC curve
    ## --------------------------
    ROC <- with(res, roc(lag_2014, pht, quiet = TRUE))
    roc.lst[[i]] <- with(ROC, data.frame(sim=i, TPR=rev(sensitivities),
FPR=rev(1 - specificities)))
}
var2 <- do.call(rbind, var.lst)    # predictor z-scores
acc2 <- do.call(rbind, acc.lst)    # accurary performance
roc2 <- do.call(rbind, roc.lst)    # roc
```

```
prd2 <- do.call(rbind, prd.lst)    # truth and prediction, per county, per sim



## ROC for all simulations combined (overall)

rocA <- with(prd2, roc(lag_2014, pht, quiet = TRUE))

aucA <- round(auc(rocA), 3)

rocA <- with(rocA, data.frame(TPR=rev(sensitivities), FPR=rev(1-

specificities)))



## ---------------------------------------

## plot all ROC curves with the average

## ---------------------------------------

setwd("C:/Users/montg/Dropbox/Ph.D Work/Dissertation/Data/Aim

1/results/")

g <- ggplot() + theme_bw()

g <- g + geom_abline(intercept = 0, slope = 1, size = 0.5, linetype =

"dashed") # ref
```

```
g <- g + geom_line(aes(x=FPR, y=TPR, group=sim), roc2, alpha=0.3) # per
sim

g <- g + geom_line(aes(x=FPR, y=TPR), rocA, color="red", size=2)    #
overall

g <- g + geom_text(aes(x = 0.75, y=0.4, label=paste0("AUC: ", aucA)),
color="red", size=10)

g <- g + scale_x_continuous(limits = c(0, 1)) + scale_y_continuous(limits =
c(0, 1))

ggsave("roc_glm.png", g)



## ----------------------------------------

## summarize predictive power of variables

## ----------------------------------------

round(do.call(rbind, with(var2, by(Z, variables, summary)))[, 2:5], 3)



#                    1st Qu. Median       Mean 3rd Qu.

# area_land               0.733  1.346  320390.08   1.932
```

area_water 0.111 0.467 228688.61 0.829

eighteen_plus_pysmiprev 0.527 1.351 441991.16 1.986

eighteen_plus_pystprev -1.107 -0.326 -194993.85 0.332

pca1 -1.649 -0.715 -195687.63 0.555

pca2 -1.588 -0.802 -216793.23 -0.007

percentdem_2012_votes -1.619 -0.912 -598545.37 -0.202

percentrep_2012_votes -1.782 -1.100 -612526.57 -0.392

twelve_plus_pmalcprev -0.004 0.618 268324.83 1.314

twelve_plus_pmcigprev -1.567 -0.829 -362955.85 -0.010

twelve_plus_pmmjprev 2.203 2.871 1202101.03 3.326

twelve_plus_pyaudprev -0.085 0.536 392370.13 1.279

twelve_plus_pycocprev 1.636 2.239 333135.16 2.685

twelve_plus_pysudprev -1.320 -0.481 35547.05 0.260

Summarize Classification Accuracies

--

summary(acc2)

sim TP TN FP FN TPF

Min. : 1.0 Min. : 7.00 Min. :33.00 Min. : 0.000 Min. : 0.000 Min. :0.3182

1st Qu.: 250.8 1st Qu.:16.00 1st Qu.:40.00 1st Qu.: 2.000 1st Qu.: 3.000 1st Qu.:0.7273

Median : 500.5 Median :17.00 Median :41.00 Median : 3.000 Median : 5.000 Median :0.7727

Mean : 500.5 Mean :17.05 Mean :40.97 Mean : 3.026 Mean : 4.945 Mean :0.7752

3rd Qu.: 750.2 3rd Qu.:19.00 3rd Qu.:42.00 3rd Qu.: 4.000 3rd Qu.: 6.000 3rd Qu.:0.8636

Max. :1000.0 Max. :22.00 Max. :44.00 Max. :11.000 Max. :15.000 Max. :1.0000

TNF FPF FNF acc

Min. :0.7500 Min. :0.00000 Min. :0.0000 Min. :0.7576

```
#  1st Qu.:0.9091   1st Qu.:0.04545   1st Qu.:0.1364   1st Qu.:0.8485

#  Median :0.9318   Median :0.06818   Median :0.2273   Median :0.8788

#  Mean   :0.9312   Mean   :0.06877   Mean   :0.2248   Mean   :0.8792

#  3rd Qu.:0.9545   3rd Qu.:0.09091   3rd Qu.:0.2727   3rd Qu.:0.9091

#  Max.   :1.0000   Max.   :0.25000   Max.   :0.6818   Max.   :0.9848



#slightly lower accuracy but much better balance of sens and spec



## ------------------------------------------------

## ensemble predictions per county using different weights

## ------------------------------------------------

WTS <- c('TPF', 'TNF', 'FNF', 'acc')

mrg <- merge(prd2, acc2, by="sim")[, c(names(prd2), WTS)]

ens <- by(mrg, mrg$qname, function(x)

{

   wts <- t(colMeans(x[, 'pht'] * x[, WTS])) # esemble voted probability
```

```
    data.frame(
        county = x[1, 'qname'], N = nrow(x), Y = x[1, 'lag_2014'],
        YHT = mean(x[, 'yht']), # mean of per-simulation hard-call
        PHT = mean(x[, 'pht']), # mean of per-simulation preditive probability
        wts)
})
ens <- data.frame(do.call(rbind, ens), row.names=NULL)
## check performance
CRI <- c("YHT", "PHT", WTS) # criteria
round(cor(ens[, 'Y'], ens[, CRI]), 3)


## make hard-calls
THD <- zrs / (zrs + ons) # xt: fraction of 0s as threshold
hdc <- cbind(ens[, !names(ens) %in% CRI], (ens[, CRI] > THD) + 0)


with(hdc, table(Y, TNF))
```

```
##    TNF
## Y      0    1
##   0 2751  251
##   1    3   89
```

```
with(hdc, table(Y, FNF))
```

```
##    FNF
## Y      0
##   0 3002
##   1   92
```

```
with(hdc, table(Y, acc))
```

```
##    acc
## Y      0    1
##   0 2746  256
##   1    3   89
```

```
with(hdc, table(Y, PHT))
##    PHT
## Y      0    1
##   0 2732  270
##   1    1   91

with(hdc, table(Y, YHT))
##    YHT
## Y      0    1
##   0 2734  268
##   1    1   91

## Counties with false positive predictions
subset(ens, Y == 0 & TNF>THD)
```

```
## Output for SAS mapping

write.csv(ens, file='C:\\Users\\montg\\Dropbox\\Ph.D

Work\\Dissertation\\Data\\Aim 1\\Processed/ens_lag.csv')
```

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Alaska State Legislature. (2014). House joint resolution no. 14. Retrieved from https://www.akleg.gov/basis/Bill/Text/32?Hsid=HJR014A

2. Andersen, H. (2007). History and philosophy of modern epidemiology.

3. Angrist, J. D. and Pischke, J.-S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

4. Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton university press.

5. Athey, S., & Imbens, G. W. (2021). Design-based analysis in difference-in-differences settings with staggered adoption. Journal of Econometrics.

6. Baicker, K., & Svoronos, T. (2019). Testing the validity of the single interrupted time series design (No. w26080). National Bureau of Economic Research.

7. Baker, K. M. (1975). Condorcet, from natural philosophy to social mathematics.

8. Barber, Elizabeth Wayland. (1992). Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean. Princeton University Press.

9. Beeching, J. (1977). The Chinese Opium Wars. United Kingdom: Harcourt Brace Jovanovich.

10. Beltz, L., Mosher, C., & Schwartz, J. (2020). County-Level Differences in Support for Recreational Cannabis on the Ballot. *Contemporary Drug Problems*, *47*(2), 149-164.

11. Ben-Michael, E., Feller, A., & Rothstein, J. (2021). Synthetic Controls with Staggered Adoption (No. w28886). *National Bureau of Economic Research.*

12. Ben-Michael, E., Feller, A., & Stuart, E. A. (2021). A trial emulation approach for policy evaluations with group-level longitudinal data. Epidemiology (Cambridge, Mass.), 32(4), 533.

13. Bernal, J. L., Cummins, S. and Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. International journal of epidemiology 46 348{355.

14. Blocker, J. (2006). Did Prohibition Really Work? Am J Pub Health, 96(2).

15. Boland, P. J. (1989). Majority systems and the Condorcet jury theorem. Journal of the Royal Statistical Society: Series D (The Statistician), 38(3), 181-189.

16. Bonnie, R. J., & Whitebread, C. H. (1970). The forbidden fruit and the tree of knowledge: an inquiry into the legal history of American marijuana prohibition. Virginia Law Review, 971-1203.

17. Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L. et al. (2015). Inferring causal impact using bayesian structural time-series models. The Annals of Applied Statistics 9 247{274.

18. C. C. Bakels. (2003). The contents of ceramic vessels in the Bactria-Margiana Archaeological Complex, Turkmenistan. Electron. J. Vedic Stud. 9.

19. C.J. van Boxtel, B. Santoso and I.R. Edwards (2008). Drug Benefits and Risks: International Textbook of Clinical Pharmacology, revised 2nd edition. IOS Press and Uppsala Monitoring Centre.

20. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. Journal of Econometrics, 225(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

21. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

22. Campbell, D. T. and Cook, T. D. (1979). Quasi-experimentation: Design & analysis issues for field settings. Rand McNally College Publishing Company Chicago.

23. Caulkins, J. P., Goyeneche, L. A., Guo, L., Lenart, K., & Rath, M. (2021). Outcomes associated with scheduling or up-scheduling controlled substances. International Journal of Drug Policy, 91, 103110.

24. Cerdá, M., Mauro, C., Hamilton, A., Levy, N. S., Santaella-Tenorio, J., Hasin, D., ... & Martins, S. S. (2020). Association between recreational marijuana legalization in the United States and changes in marijuana use and cannabis use disorder from 2008 to 2016. JAMA Psychiatry, 77(2), 165-171.

25. Cerdá, M., Mauro, C., Hamilton, A., Levy, N. S., Santaella-Tenorio, J., Hasin, D., ... & Martins, S. S. (2020). Association between recreational marijuana legalization in the United States and changes in marijuana use and cannabis use disorder from 2008 to 2016. *JAMA Psychiatry*, *77*(2), 165-171.

26. Cerdá, M., Wall, M., Feng, T., Keyes, K. M., Sarvet, A., Schulenberg, J., ... & Hasin, D. S. (2017). Association of state recreational marijuana laws with adolescent marijuana use. JAMA pediatrics, 171(2), 142-149.

27. Cerdá, M., Wall, M., Feng, T., Keyes, K. M., Sarvet, A., Schulenberg, J., ... & Hasin, D. S. (2017). Association of state recreational marijuana laws with adolescent marijuana use. *JAMA pediatrics*, *171*(2), 142-149.

28. Chan, J. T., & Zhong, W. (2019). Reading China: Predicting policy change with machine learning.

29. Chang, H. (1970). Commissioner Lin and the Opium War. United States: Norton.

30. Chen, C. Y., Dormitzer, C. M., Gutiérrez, U., Vittetoe, K., González, G. B., & Anthony, J. C. (2004). The adolescent behavioral repertoire as a context for drug exposure: behavioral autarcesis at play. *Addiction*, *99*(7), 897-906.

31. Cheng, H. G., Augustin, D., Glass, E. H., & Anthony, J. C. (2019). Nation-scale primary prevention to reduce newly incident adolescent drug use: the issue of lag time. PeerJ, 7, e6356. https://doi.org/10.7717/peerj.6356

32. Cheng, H. G., Augustin, D., Glass, E. H., & Anthony, J. C. (2019). Nation-scale primary prevention to reduce newly incident adolescent drug use: the issue of lag time. *PeerJ*, *7*, e6356. https://doi.org/10.7717/peerj.6356

33. Cheng, H. G., Cantave, M. D., & Anthony, J. C. (2016). Alcohol experiences viewed mutoscopically: newly incident drinking of twelve-to twenty-five-year-olds in the United States, 2002–2013. *Journal of studies on alcohol and drugs*, *77*(3), 405-412.

34. Cheng, H. G., Lopez-Quintero, C., & Anthony, J. C. (2018). Age of onset or age at assessment-that is the question: Estimating newly incident alcohol drinking and rapid transition to heavy drinking in the United States, 2002-2014. *International Journal of Methods in Psychiatric Research*, *27*(1), e1587.

35. Cheng, H. G., Lopez-Quintero, C., & Anthony, J. C. (2018). Age of onset or age at assessment—that is the question: Estimating newly incident alcohol drinking and rapid transition to heavy drinking in the United States, 2002–2014. *International journal of methods in psychiatric research*, *27*(1), e1587.

36. Coley, R. L., Kruzik, C., Ghiani, M., Carey, N., Hawkins, S. S., & Baum, C. F. (2021). Recreational Marijuana Legalization and Adolescent Use of Marijuana, Tobacco, and Alcohol. Journal of Adolescent Health, 69(1), 41–49. https://doi.org/10.1016/j.jadohealth.2020.10.019

37. Coley, R. L., Kruzik, C., Ghiani, M., Carey, N., Hawkins, S. S., & Baum, C. F. (2021). Recreational Marijuana Legalization and Adolescent Use of Marijuana, Tobacco, and Alcohol. *Journal of Adolescent Health*, *69*(1), 41–49. https://doi.org/10.1016/j.jadohealth.2020.10.019

38. Colorado Constitution, Amendment 64. (2012).

39. Colorado Municipal League (2019) Municipal Retail Marijuana Status. Retrieved from https://www.cml.org/home/topics-key-issues/municipal-retail-marijuana-laws

40. Cunningham, S. (2020). Causal Inference. *The Mixtape*, *1*.

41. Daniller, A. (2019). Two-thirds of Americans support marijuana legalization. Pew Research Center, 14.

42. Darnell, A. (2015). *I-502 Evaluation plan and preliminary report on implementation*. Washington State Institute for Public Policy.

43. Dawson, D. A., Goldstein, R. B., Patricia Chou, S., June Ruan, W., & Grant, B. F. (2008). Age at First Drink and the First Incidence of Adult-Onset DSM-IV Alcohol Use Disorders. Alcoholism: Clinical and Experimental Research, 32(12), 2149–2160. https://doi.org/10.1111/j.1530-0277.2008.00806.x

44. Degenhardt, Louisa et al. (2008). Toward a global view of alcohol, tobacco, cannabis, and cocaine use: findings from the WHO World Mental Health Surveys. *PLoS medicine* vol. 5,7: e141. doi:10.1371/journal.pmed.0050141

45. Department of Justice. Title 21 United States Code (USC) Controlled Substances Act (1970).

46. Dilley, J. A., Hitchcock, L., McGroder, N., Greto, L. A., & Richardson, S. M. (2017). Community-level policy responses to state marijuana legalization in Washington State. *International Journal of Drug Policy*, *42*, 102-108.

47. Dilley, J. A., Richardson, S. M., Kilmer, B., Pacula, R. L., Segawa, M. B., & Cerdá, M. (2019). Prevalence of cannabis use in youths after legalization in Washington state. *JAMA pediatrics*, 173(2), 192-193.

48. Durkheim, E. (2005). Suicide: A study in sociology. Routledge.

49. Ebrey, P. B. (1999). The Cambridge Illustrated History of China. Kiribati: Cambridge University Press.

50. Edward M. Brecher, et al. (1972). The Consumers Union Report on Licit and Illicit Drugs.

51. Everson EM, Dilley JA, Maher JE et al. Post-legalization opening of retail cannabis stores and adult cannabis use in Washington State, 2009-2016. Am J Public Health 2019;109:1294-301.

52. Everson EM, Dilley JA, Maher JE et al. Post-legalization opening of retail cannabis stores and adult cannabis use in Washington State, 2009-2016. (2019) *Am J Public Health*; 109:1294-301.

53. Farr, W. (2000). Vital statistics: memorial volume of selections from the reports and writings. 1885. Bulletin of the World Health Organization, 78(1), 88.

54. Feige, C., & Miron, J. A. (2008). The opium wars, opium legalization and opium consumption in China. Applied Economics Letters, 15(12), 911-913.

55. Fontes, M. A., Bolla, K. I., Cunha, P. J., Almeida, P. P., Jungerman, F., Laranjeira, R. R., ... & Lacerda, A. L. (2011). Cannabis use before age 15 and subsequent executive functioning. *The British Journal of Psychiatry*, *198*(6), 442-447.

56. Gallup Social and Policy Issues (2016). Gallup Polls [Gallup Poll Social Series]. Retrieved from http:// https://news.gallup.com/poll/196568/americans-views-shift-toughness-justice-system.aspx

57. Gallup Social and Policy Issues (2016). *Gallup Polls* [Gallup Poll Social Series]. Retrieved July 10, 2020, available at https://news.gallup.com/poll/196568/americans-views-shift-toughness-justice-system.aspx

58. Gallup. (2020). *Support for Legal Marijuana Inches Up to New High of 68%*. Gallup.Com. Retrieved December 29, 2021, from https://news.gallup.com/poll/323582/support-legal-marijuana-inches-new-high.aspx

59. Galston, W. A., & Dionne Jr, E. J. (2013). The new politics of marijuana legalization: Why opinion is changing. *Governance Studies at Brookings*, 1-17.

60. Galton, Francis (1877). Typical Laws of Heredity . Nature 15, 492–495 https://doi.org/10.1038/015492a0

61. Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

62. Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., & Kruger, L. (1990). The empire of chance: How probability changed science and everyday life (No. 12). Cambridge University Press.

63. Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics.

64. Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

65. Graunt, J. (1939). Natural and political observations made upon the bills of mortality (No. 2). Johns Hopkins Press.

66. Gruber, K., Anderson, A., Calanan, R., VanDyke, M., Barker, L., Burris, D., & Tolliver, R. (2016). Marijuana Use Among Adolescents in Colorado: Results from the 2013 Healthy Kids Colorado Survey. 10.

67. Gruber, K., Anderson, A., Calanan, R., VanDyke, M., Barker, L., Burris, D., & Tolliver, R. (2016). Marijuana Use Among Adolescents in Colorado: Results from the 2013 Healthy Kids Colorado Survey. 10.

68. Haggerty, R. J., & Mrazek, P. J. (1994). Reducing risks for mental disorders: Frontiers for preventive intervention research. National Academies Press.

69. Hald, A. (1998). A History of Mathematical Statistics from 1750 to 1930 (Vol. 314). Wiley-Interscience.

70. Hall, W., & Weier, M. (2015). Assessing the public health impacts of legalizing recreational cannabis use in the USA. *Clinical pharmacology & therapeutics*, *97*(6), 607-615.

71. Helmer, J., & Vietorisz, T. (1974). Drug use, the labor market and class conflict (Vol. 43, No. 8). Drug Abuse Council.

72. Herodotus, A. D. Godley. (1920). Trans. The Histories. Harvard Univ. Press.

73. Hill, A. B. (1951) The Clinical Trial. BMJ , 278-282.

74. Hill, A. B. (1952) The Clinical Trial. New England Journal of Medicine 247, 113-119.

75. Hill, A. B. (1953) Observations and Experiment. New England Journal of Medicine 248, 995-1001.

76. Hill, A. B. (1965) The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine 58, 295-300.

77. Hill, A. B. and Doll, R. (1956). Lung Cancer and Tobacco. The BMJ's Questions Answered by Professor A. Bradford Hill and R. Doll. BMJ 1(4976), 1160-1163.

78. Horwood, L. J., Fergusson, D. M., Hayatbakhsh, M. R., Najman, J. M., Coffey, C., Patton, G. C., ... & Hutchinson, D. M. (2010). Cannabis use and educational achievement: findings from three Australasian cohort studies. *Drug and alcohol dependence*, *110*(3), 247-253.
https://news.gallup.com/poll/323582/support-legal-marijuana-inches-new-high.aspx

79. Humphreys, K., Edwards, G., Caulkins, J. P., Babor, T., Foxcroft, D. R., Rehm, J., Fischer, B., Obot, I. S., Babor, T. F., Reuter, P. (2010). Drug Policy and the Public Good. United Kingdom: OUP Oxford.

80. Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

81. Jonnes, J. (1999). Hep-cats, narcs, and pipe dreams: A history of America's romance with illegal drugs. JHU Press.

82. Kandel, D. B., & Logan, J. A. (1984). Patterns of Drug Use from Adolescence to Young Adulthood: 1. Periods of Risk for Initiation, Continued Use, and Discontinuation.

83. Kendall, M. G. (1956). Studies in the history of probability and statistics: II. The beginnings of a probability calculus. Biometrika, 43(1/2), 1-14.

84. Kerr WC, Lui C, Ye Y. Trends and age, period and cohort effects for marijuana use prevalence in the 1984-2015 US national alcohol surveys. Addiction. 2018;113:473–81. doi:10.1111/add.14031.

85. Kerr WC, Lui C, Ye Y. Trends and age, period and cohort effects for marijuana use prevalence in the 1984-2015 US national alcohol surveys. Addiction. 2018;113:473–81. doi:10.1111/add.14031.

86. Kolb, L., & Du Mez, A. G. (1924). The prevalence and trend of drug addiction in the United States and factors influencing it. Public Health Reports (1896-1970), 1179-1204.

87. Kramer, M. (1957). A discussion of the concepts of incidence and prevalence as related to epidemiologic studies of mental disorders. American Journal of Public Health and the Nations Health, 47(7), 826-840.

88. Kramer, M. (1957). A discussion of the concepts of incidence and prevalence as related to epidemiologic studies of mental disorders. American Journal of Public Health and the Nations Health, 47(7), 826-840.

89. L.W. King. (1915). The Code of Hammurabi. Trans. Paulo J. S. Pereira.

90. Labouvie, E., Bates, M. E., & Pandina, R. J. (1997). Age of first use: its reliability and predictive utility. Journal of Studies on Alcohol, 58(6), 638–643. https://doi.org/10.15288/jsa.1997.58.638

91. Lapouse, R. (1967). Problems in studying the prevalence of psychiatric disorder. American Journal of Public Health and the Nations Health, 57(6), 947-954.

92. Lapouse, R. (1967). Problems in studying the prevalence of psychiatric disorder. American Journal of Public Health and the Nations Health, 57(6), 947-954.

93. *Legal Recreational Marijuana States and DC - Recreational Marijuana—ProCon.org* (2021). Recreational Marijuana. Retrieved November 29, 2021, available at https://marijuana.procon.org/legal-recreational-marijuana-states-and-dc/

94. MacMahon, B., T. F. Pugh, and J. Ipsen. (1960). Epidemiologic Methods. Boston: Little, Brown & Co.

95. Maistrov, L. E. (2014). Probability theory: A historical sketch. Academic Press.

96. Martins, S. S., Segura, L. E., Levy, N. S., Mauro, P. M., Mauro, C. M., Philbin, M. M., & Hasin, D. S. (2021). Racial and Ethnic Differences in Cannabis Use Following Legalization in US States With Medical Cannabis Laws. JAMA Network Open, 4(9), e2127002. https://doi.org/10.1001/jamanetworkopen.2021.27002

97. Martins, S. S., Segura, L. E., Levy, N. S., Mauro, P. M., Mauro, C. M., Philbin, M. M., & Hasin, D. S. (2021). Racial and Ethnic Differences in Cannabis Use Following Legalization in US States With Medical Cannabis Laws. *JAMA Network Open*, *4*(9), e2127002. https://doi.org/10.1001/jamanetworkopen.2021.27002

98. McWilliams, J. C. (1990). The Protectors: Harry J. Anslinger and the Federal Bureau of Narcotics, 1930-1962. United Kingdom: University of Delaware Press.

99. Melchior, M., Nakamura, A., Bolze, C., Hausfater, F., El Khoury, F., Mary-Krause, M., & Da Silva, M. A. (2019). Does liberalisation of cannabis policy influence levels of use in adolescents and young adults? A systematic review and meta-analysis. BMJ Open, 9(7), e025880.

100. Melchior, M., Nakamura, A., Bolze, C., Hausfater, F., El Khoury, F., Mary-Krause, M., & Da Silva, M. A. (2019). Does liberalisation of cannabis policy influence levels of use in adolescents and young adults? A systematic review and meta-analysis. *BMJ Open*, *9*(7), e025880.

101. Merlin, M. D. (2003). Archaeological evidence for the tradition of psychoactive plant use in the old world. Economic Botany, 57(3), 295-323.

102. Midgette, G., & Reuter, P. (2020). Has Cannabis Use Among Youth Increased After Changes in Its Legal Status? A Commentary on Use of Monitoring the Future for Analyses of Changes in State Cannabis Laws. Prevention Science, 21(1), 137–145. https://doi.org/10.1007/s11121-019-01068-4

103. MIT Election Data and Science Lab, 2018. County Presidential Election Returns 2000-2020, Retrieved from https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V9, UNF:6:qSwUYo7FKxI6vd/3Xev2Ng== [fileUNF]'

104. Montgomery, B. W., Anthony, J. C., & Vsevolozhskaya, O. (2021). An Epidemiological Hypothesis of Policy-Shaped Drug Use Onset Curves. Biomedical Journal of Scientific & Technical Research, 38(1), 29994-29998.

105. Mosher, C. J., & Akins, S. (2019). In the weeds: Demonization, legalization, and the evolution of us marijuana policy. Temple University Press.

106. MRSC. (2019). Marijuana Regulation in Washington State. Retrieved from http://mrsc.org/Home/Explore-Topics/Legal/Regulation/Marijuana-Regulation-in-Washington-State.aspx

107. MRSC. (2019). Marijuana Regulation in Washington State. Retrieved December 4, 2019, from http://mrsc.org/Home/Explore-Topics/Legal/Regulation/Marijuana-Regulation-in-Washington-State.aspx

108. Musto, D. F. (1999). The American disease: Origins of narcotic control. Oxford University Press.

109. Nation, M., Crusto, C., Wandersman, A., Kumpfer, K. L., Seybolt, D., Morrissey-Kane, E., & Davino, K. (2003). What works in prevention: Principles of effective prevention programs. American psychologist, 58(6-7), 449.

110. National Conference of State Legislatures [NCSL]. (2019). State medical marijuana Laws. Retrieved June 25, 2019. From http://www.ncsl.org/research/health/state-medical-marijuana-laws.aspx. Accessed 26 Aug 2019.

111. National Surveys on Drug Use and Health, 2014. 2010-2012 NSDUH Substate Region Definitions. Retrieved from https://www.samhsa.gov/data/report/2010-2012-nsduh-substate-region-definitions

112. New York Times. (1914) via Schaeffer's Drug Library. http://www.druglibrary.org/schaffer/History/Negro_cocaine_fiends.htm accessed November 11th, 2020.

113. Nobel Prize Outreach AB. (2022). The Nobel Prize in Physiology or Medicine 1929. NobelPrize.org. Retrieved April 5, 2022 from https://www.nobelprize.org/prizes/medicine/1929/summary/

114. Oregon Legislature. (2014). Chapter 475B — Cannabis Regulation. Retrieved December 30, 2021, from https://www.oregonlegislature.gov/bills_laws/ors/ors475B.html

115. Pacula RL, Kilmer B, Wagenaar AC et al. Developing public health regulations for marijuana: lessons from alcohol and tobacco. Am J Public Health 2014;104:1021-8.

116. Pacula RL, Kilmer B, Wagenaar AC et al. Developing public health regulations for marijuana: lessons from alcohol and tobacco. Am J Public Health 2014;104:1021-8.

117. Parker MA, Anthony JC. (2015). Epidemiological evidence on extra-medical use of prescription pain relievers: transitions from newly incident use to dependence among 12–21 year olds in the United States using meta-analysis, 2002-13. PeerJ 3:e1340 https://doi.org/10.7717/peerj.1340

118. Paschall, M. J., García-Ramírez, G., & Grube, J. W. (2021). Recreational marijuana legalization and use among California adolescents: findings from a statewide survey. Journal of studies on alcohol and drugs, 82(1), 103-111.

119. Paschall, M. J., García-Ramírez, G., & Grube, J. W. (2021). Recreational marijuana legalization and use among California adolescents: findings from a statewide survey. *Journal of studies on alcohol and drugs*, *82*(1), 103-111.

120. Payán, D. D., Brown, P., & Song, A. V. (2021). County-Level Recreational Marijuana Policies and Local Policy Changes in Colorado and Washington State (2012-2019). *The Milbank Quarterly*.

121. Pew. (2013). Majority now supports legalizing Marijuana. *Pew Research Center*. Retrieved December 30, 2021, from http://www.people-press.org/2013/04/04/majority-now-supports-legalizing-marijuana

122. Pew. (2015). 63% of Republican Millennials favor marijuana legalization. *Pew Research Center*. Retrieved December 30, 2021, from https://www.pewresearch.org/fact-tank/2015/02/27/63-of-republican-millennials-favor-marijuana-legalization/

123. Pew. (2016). Support for marijuana legalization continues to rise. *Pew Research Center*. Retrieved December 30, 2021, from https://www.pewresearch.org/fact-tank/2016/10/12/support-for-marijuana-legalization-continues-to-rise/

124. Pew. (2019). Two-thirds of Americans support marijuana legalization. *Pew Research Center*. Retrieved December 30, 2021, from https://www.pewresearch.org/fact-tank/2019/11/14/americans-support-marijuana-legalization/

125. Pew. (2021). Americans overwhelmingly say marijuana should be legal for recreational or medical use. *Pew Research Center*. Retrieved December 29, 2021, from https://www.pewresearch.org/fact-tank/2021/04/16/americans-overwhelmingly-say-marijuana-should-be-legal-for-recreational-or-medical-use/

126. Pillard, R. C. (1970). Marihuana. New England Journal of Medicine, 283(6), 294-303.

127. Prescott, C. A., & Kendler, K. S. (1999). Age at First Drink and Risk for Alcoholism: A Noncausal Association. Alcoholism: Clinical and Experimental Research, 23(1), 101–107. https://doi.org/10.1111/j.1530-0277.1999.tb04029.x

128. Public Law No. 223, 63rd Cong., approved December 17, 1914.

129. R. C. Clarke, M. D. Merlin. (2013) Cannabis: Evolution and Ethnobotany. University of California Press.

130. Reed, J. (2016). Marijuana Legalization in Colorado: Early Findings: A Report Pursuant to Senate Bill 13-283 (March 2016). 147.

131. Reed, J. (2016). *Marijuana Legalization in Colorado: Early Findings: A Report Pursuant to Senate Bill 13-283 (March 2016)*. 147.

132. Reed, J. (2021). Impacts of marijuana legalization in Colorado: A Report Pursuant to C.R.S. 24-33.4-516. Retrieved December 10, 2021, available at https://cdpsdocs.state.co.us/ors/docs/reports/2021-SB13-283_Rpt.pdf

133. Ren, M., Tang, Z., Wu, X., Spengler, R., Jiang, H., Yang, Y., & Boivin, N. (2019). The origins of cannabis smoking: Chemical residue evidence from the first millennium BCE in the Pamirs. Science advances, 5(6), eaaw1391.

134. Roth, J. (2020), "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," working paper, Brown University Department of Economics.

135. Shimonovich, M., Pearce, A., Thomson, H., Keyes, K., & Katikireddi, S. V. (2020). Assessing causality in epidemiology: revisiting Bradford Hill to incorporate developments in causal thinking. European Journal of Epidemiology, 1-15.

136. Smart, R., & Pacula, R. L. (2019). Early evidence of the impact of cannabis legalization on cannabis use, CUD, and the use of other substances: findings from state policy evaluations. The American journal of drug and alcohol abuse, 45(6), 644-663.

137. Smith L. (2018) How a racist hate-monger masterminded America's War on Drugs. Medium. Accessed April 5, 2022, available at https://timeline.com/harry-anslinger-racist-war-on-drugs-prison-industrial-complex-fb5cbc281189

138. Socia, K. M., & Brown, E. K. (2017). Up in smoke: The passage of medical Marijuana legislation and enactment of dispensary moratoriums in Massachusetts. *Crime & Delinquency*, *63*(5), 569-591.

139. Spillane, J. F. (2004). Debating the controlled substances act. Drug and Alcohol Dependence, 76(1), 17-29.

140. Staggs, B., Wheeler, I., Aitken, D., & Lawrence, P. (2019). What are the marijuana laws in your California city? Explore our database of local cannabis policies. Retrieved from https://www.ocregister.com/2018/01/03/what-are-the-marijuana-laws-in-your-california-city-explore-our-database-of-local-cannabis-policies-2/.

141. Staggs, B., Wheeler, I., Aitken, D., & Lawrence, P. (2019). What are the marijuana laws in your California city? Explore our database of local cannabis policies. Retrieved December 3, 2019, from https://www.ocregister.com/2018/01/03/what-are-the-marijuana-laws-in-your-california-city-explore-our-database-of-local-cannabis-policies-2/.

142. STATSAMERCIA (2021). Indiana Business Research Center. Indiana University Kelley School of Business. Retrieved from http://www.statsamerica.org/CityCountyFinder/Default.aspx

143. Substance Abuse and Mental Health Services Administration (2014). National survey on drug use and health.

144. Substance Abuse and Mental Health Services Administration (2021). CBHSQ Data. Retrieved September 11, 2021, from https://www.samhsa.gov/data/

145. Substance Abuse and Mental Health Services Administration. (2019). Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No. PEP19-5068, NSDUH Series H-54). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/

146. Thornton, M. Alcohol Prohibition Was a Failure (1991). Pol. Analysis No. 157. Washington: Cato.

147. United Nations Office of Drugs and Crime. (2020). https://dataunodc.un.org/data/drugs/Prevalence-general accessed December 15th, 2020

148. United States Census Bureau. (2012). 2010 Census. Retrieved from https://data.census.gov/cedsci/

149. United States. Surgeon General's Advisory Committee on Smoking. (1964). Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service (No. 1103). US Department of Health, Education, and Welfare, Public Health Service.

150. Varberg, D. (1963). The development of modern statistics. The Mathematics Teacher. 56 (4), 252-257. Retrieved September 7, 2021, from http://www.jstor.org/stable/27956805

151. Volkow, N. D., Baler, R. D., Compton, W. M., & Weiss, S. R. (2014). Adverse health effects of marijuana use. New England Journal of Medicine, 370(23), 2219-2227.

152. Volkow, N. D., Baler, R. D., Compton, W. M., & Weiss, S. R. (2014). Adverse health effects of marijuana use. New England Journal of Medicine, 370(23), 2219-2227.

153. Wagner, F. (2002). From First Drug Use to Drug Dependence Developmental Periods of Risk for Dependence upon Marijuana, Cocaine, and Alcohol. Neuropsychopharmacology, 26(4), 479–488. https://doi.org/10.1016/S0893-133X(01)00367-0

154. Wagner, F. (2002). From First Drug Use to Drug Dependence Developmental Periods of Risk for Dependence upon Marijuana, Cocaine, and Alcohol. Neuropsychopharmacology, 26(4), 479–488. https://doi.org/10.1016/S0893-133X(01)00367-0

155. Warner, Jessica; Her, Minghao; Gmel, Gerhard; Rehm, Jürgen (2001). Can Legislation Prevent Debauchery? Mother Gin and Public Health in 18th-Century England. American Journal of Public Health. 91: 375–84. doi:10.2105/ajph.91.3.375. PMC 1446560. PMID 11236401.

156. Washington State Liquor control Board, Initiative 502. (2012).

157. Wilkins, C., Tremewan, J., Rychert, M., Atkinson, Q., Fischer, K., & Forsyth, G. L. (2022). Predictors of voter support for the legalization of recreational cannabis use and supply via a national referendum. International Journal of Drug Policy, 99, 103442.

158. Wooldridge, Jeffrey M., Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators (August 17, 2021). Available at SSRN: https://ssrn.com/abstract=3906345 or http://dx.doi.org/10.2139/ssrn.3906345

159. Wright, Hamilton. (1910). Report on the International Opium Commission and on the Opium Problem as Seen Within the United States and Its Possessions. U.S. Senate, 61st Congress, 2nd Session, Document #377.

160. Wu, L.-T., Korper, S. P., Marsden, M. E., Lewis, C., & Bray, R. M. (2003). Use of Incidence and Prevalence in the Substance Use Literature: A Review. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

161. Wu, L.-T., Korper, S. P., Marsden, M. E., Lewis, C., & Bray, R. M. (2003). Use of Incidence and Prevalence in the Substance Use Literature: A Review. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

162. Zhang, A. X., & Counts, S. (2015, April). Modeling ideology and predicting policy change with social media: Case of same-sex marriage. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2603-2612).