UNIVERSITY OF MINNESOTA
Driven to Discover℠

MSBA 6440 - Causal Inference via

Econometrics and Experimentation

March 1st, 2020

Impact of Cannabis Legalization on Traffic Accidents

Team 20

Chia-Hsuan Chou

Brady Engelke

Yidan Gao

Minya Na

William Wu

Table of Contents

Executive Summary

This study aims to provide substantive evidence to the disputed question: for US states that have legalized cannabis, has that decision casually affected fatal car accident rates?

The experiment design is composed of 2 steps: Propensity Score Matching (PSM) and Difference-in-Difference (DiD) to isolate the effect of cannabis legalization by identifying similar states in treatment (legalized) and control (illegal) groups. Legality was the outcome variable for PSM which was regressed on a set of matching variables collected at a yearly level from 01/14 - 01/17. These variables were averaged over three years to predict the probability a state would legalize cannabis at treatment time (01/17). Fatal crashes was the outcome variable for DiD and was regressed on state highway expenses, treatment flag, and the before-and-after treatment flag at a monthly level from 01/15 - 12/18. PSM found 2 matches: Maine & Minnesota, and Nevada & Texas. DiD on these matches failed to provide evidence to conclude a causal relationship between cannabis legalization and fatal crashes.

Potential threats to this study include disproportionate sample size, difficulty in predicting the likelihood of legalization on the treatment date, and the manifestation delay of cannabis legalization's effect on car accident rates is unknown. Total crash data, cannabis-related crash rates, and improved weather data would also improve this experiment. If these limitations are addressed, this study could serve as a reference for policymakers of non-legalized states, and a motivator for the development of effective cannabis-impairment assessment tools.

Situation Overview

The societal impact of cannabis legalization is not properly understood by legislative bodies within the U.S. Even researchers versed in Econometrics have struggled to find statistical evidence to answer the simple question: for states that have legalized cannabis, has that decision casually affected the rate of car accidents? Purnell and Howell (2018) concluded, “There are few convincing conclusions to be drawn concerning the risk of traffic accident fatalities from marijuana legalization [1].” Anderson, Hansen, and Rees (2013) asserted, “Because alternative mechanisms cannot be ruled out, the negative relationship between legalization and alcohol-related traffic fatalities does not necessarily imply that driving under the influence of marijuana is safer than driving under the influence of alcohol [2].” These findings and many other similarly inconclusive statements regarding the relationship of cannabis legalization and traffic accidents has left policymakers to rely upon intuition and anecdotal thinking.

In the past 20 years, many U.S. states have chosen to decriminalize and even legalize cannabis and many other states will likely follow suit - especially if the Democratic Party wins the 2020 Presidential Election. The true implications of these policy decisions made by progressive U.S. states must be rigorously studied to generate the necessary empirical evidence to inform the policy of states who have future legalization in mind as well as debates at the federal level. This study aims to provide such substantive evidence whether the legalization of cannabis is positively or negatively related to fatal car accidents to ensure these imminent discussions are supported by logic rooted in proper research.

Experimental Design

Our experiment utilized Propensity Score Matching (PSM) and Difference-in-Differences (DiD) to investigate any causal impact between cannabis legalization and fatal car accidents. We used Propensity Score Matching to recreate a randomized experiment since a state legalizing cannabis is inherently non-random. Legislators take in a variety of factors into consideration when making such a decision. By matching legalized states with similar illegal states on indicators that foreshadow whether a state will legalize or not, we can reconstruct randomized experimental conditions and isolate the effect of cannabis legalization.

In this experiment, we identified 4 states for the treatment group because they each legalized cannabis at the end of 2016. These states are California, Maine, Massachusetts, and Nevada. We used PSM to find matching states that still deem recreational cannabis as illegal. With matched states, we then leveraged a DiD regression to estimate the causal effect legalization had on fatal car accident rates. DiD cancels out the time-variant confounds common to both the control and treatment group as well as the time-invariant confounds unique to either group. With the confounds controlled, the causal effect of cannabis legalization and its relationship with fatal car accident rates could be isolated.

We pulled data from NHTSA on fatal car accidents, state revenue and expenditure information from the Census, incarceration stats from Kaggle, alcohol consumption from NIAAA, and demographic, weather, and political orientation data from multiple other publicly available data sources ranging from 01/14 - 12/18 and compiled them into two datasets. We specified the treatment date as 01/17 to allow for legalization measures to be fully implemented in each treatment state. The two data sets are described as follows:

- Dataset 1: Legality and associated matching variables collected at a yearly level, ranging from 01/14 - 01/17. This allowed for three years of data to be averaged so that the likelihood of legalization in 01/17 could be computed for all U.S states. Legality was a binary variable with 1 being fully legal and 0 indicating not fully legal. Note, states that legalized cannabis far before 01/17 (Alaska, Colorado, Washington, Oregon) were removed.
- Dataset 2: Fatal car accident data and associated time-variant control variables collected at a monthly level, ranging from 01/2015 - 12/18. This allowed for two years of monthly fatal car accident data before and after treatment to be compared.

A detailed delineation of the matching variables in dataset 1 and the time-variant control variables in dataset 2 can be found within their respective methodology sections, PSM and DiD.

Threats to Causal Inference

A major threat during PSM was the unbalanced sample which resulted in only 4 states being considered treatment and 42 states considered as control. This significantly shrunk the sample size leftover after matching to accurately estimate the DiD regressions coefficients. Furthermore, it is inherently difficult to predict whether a state is going to legalize cannabis at a certain time or not. It is a highly non-random event that likely has a multitude of confounds correlated with it. This expansive set of confounds made the issue of an unbalanced dataset even more troubling since the more confounds included in the PSM function, the more difficult it was to define a reasonably low caliper while finding as many matches as possible.

A major threat to the DiD analysis was the fact that it is difficult to measure how long it takes for the effect of cannabis legalization on fatal traffic accidents to manifest itself if it does at all. Additionally, the outcome variable of interest, fatal car accidents, is relatively small in magnitude and has a lower standard deviation compared to total car accidents. This made for less variability to feasibly be accounted for by the treatment variable as well as the time-variant control variables included in the regression.

Total car accident data would have been useful for measuring the impact of cannabis legalization on traffic accidents in totality, either indirect or direct effects could have been captured although there would have been more noise affecting the accuracy of the coefficient estimates of the regression. Another data limitation was that data regarding traffic accidents influenced by cannabis-impairment could not be found. Data on the accidents influenced by cannabis would be well-suited to measure the fine-tuned impact of cannabis legalization since noise would be minimized although indirect effects in accident rates would be unobservable.

The last note on difficulties encountered when sourcing data was that the time-variant confounds hypothesized to be important by the team could only be gathered for subsets of the experiment's full duration of 01/15 - 12/17. Alcohol consumption data was only found for 01/15 - 12/16 and weather data was found for only 07/16 - 06/17. Lastly, interference may have been at play since cannabis consumers that reside in control states are likely to drive to treatment states to use cannabis and/or bring back cannabis to their home states. This influx of traffic could very well lead to a higher rate of fatal car accidents.

Propensity Score Matching

Before implementing PSM, the team coerced the raw data scraped from the internet into a format that could be interpreted by a Logit or Probit regression. The data had to be in the format of one vector of matching variables per state so that the PSM function would produce one prediction of the likelihood of legalization for each of the 46 states included in the study.

Assumptions

To recreate a randomized trial, the team gathered all observable predictors that determine whether a state legalizes cannabis or not. This is inherently a difficult task to even hypothesize let alone actually collect all of this data over the intended duration of the study. Within the given time constraints, the team exhaustively researched for information relating to the decision to legalize cannabis which resulted in the following list of factors:

- Political party orientation
- Average age
- Proportion of population that is male
- Number of prisoners in jails, penitentiaries, etc.
- Proportion of population that is imprisoned
- Gallons of alcoholic beverages consumed per capita yearly
- The ratio of state taxes divided by total state revenue
- State expenses on public welfare, hospitals, health, police, correctional facilities
- Population density

Not all of these factors could be included in the PSM function since some of them were highly collinear and some were less related to cannabis legalization than others. Being aware of the tradeoff that the more matching variables included in the PSM function, the greater the difference would be between a treated state's likelihood of legalization and a control state's likelihood of legalization, the team sought to minimize the number of matching variables included while fully determining the likelihood of legalization. Without this effort, the caliper would have been raised to an unreasonably high level to obtain a sufficient number of matches for DiD. After dropping the highly correlated matching variables (0.75 or above) and variables that intuitively mattered less than others when predicting the likelihood of legalization, the team was left with the following set of matching variables:

- Average age
- Political party orientation
- Proportion of population that is male
- Number of prisoners in jails, penitentiaries, etc.
- The ratio of state taxes divided by total state revenue
- Gallons of alcoholic beverages consumed per capita yearly
- Population density

Model Specification

The "matchit" algorithm borrowed from the "MatchIt" R library was leveraged to find control states that matched to the treatment states. Below is a plot of all states and their associated propensity scores which sheds light as to the extent of how unbalanced the dataset is. California was omitted from the plot since its propensity score was so high it made it hard to visualize the rest of the data.

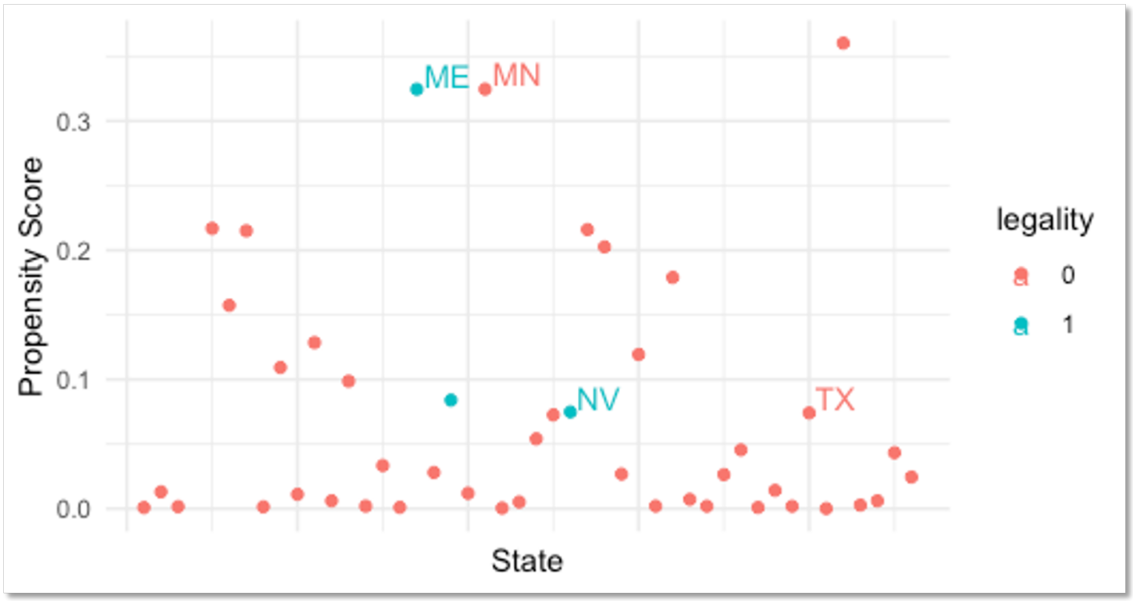
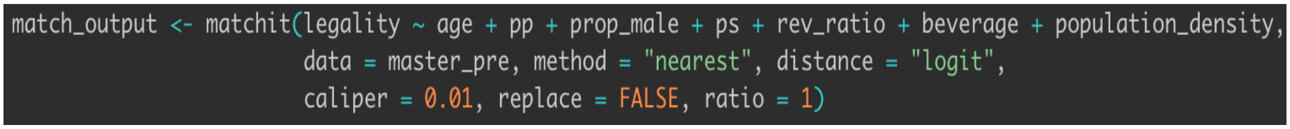


Figure 1: Propensity Scores for All States

It was interesting to find that California's propensity score was approximately 0.7, nearly twice as large as the next largest propensity score. This finding gave a rough confirmation that the PSM function was properly predicting the likelihood of a state to legalize cannabis since it aligns well with the understanding that California has always been progressive in decriminalizing cannabis and is likely to have a relatively high propensity score that reflects that.

The best parameter setting was found to be a caliper of 0.01, logit as the PSM function, and matching-without-replacement after standardization of the matching variables was used to ensure no one variable's scale would dominate the propensity score prediction. These parameter settings allowed the matchit algorithm to find two matches which was deemed sufficient for the ensuing DiD analysis. Below is the actual model specified in R:



```
match_output <- matchit(legality ~ age + pp + prop_male + ps + rev_ratio + beverage + population_density,
                        data = master_pre, method = "nearest", distance = "logit",
                        caliper = 0.01, replace = FALSE, ratio = 1)
```

Figure 2: PSM Function Specification

The Probit function was also tested as the PSM function to see if any significant changes in the results would be observed. No such changes were found, validating that the team had robust results.

Results

The following plot illustrates the effectiveness of the data treatment and parameter tuning made by the team to ensure useful propensity score matching results were found.

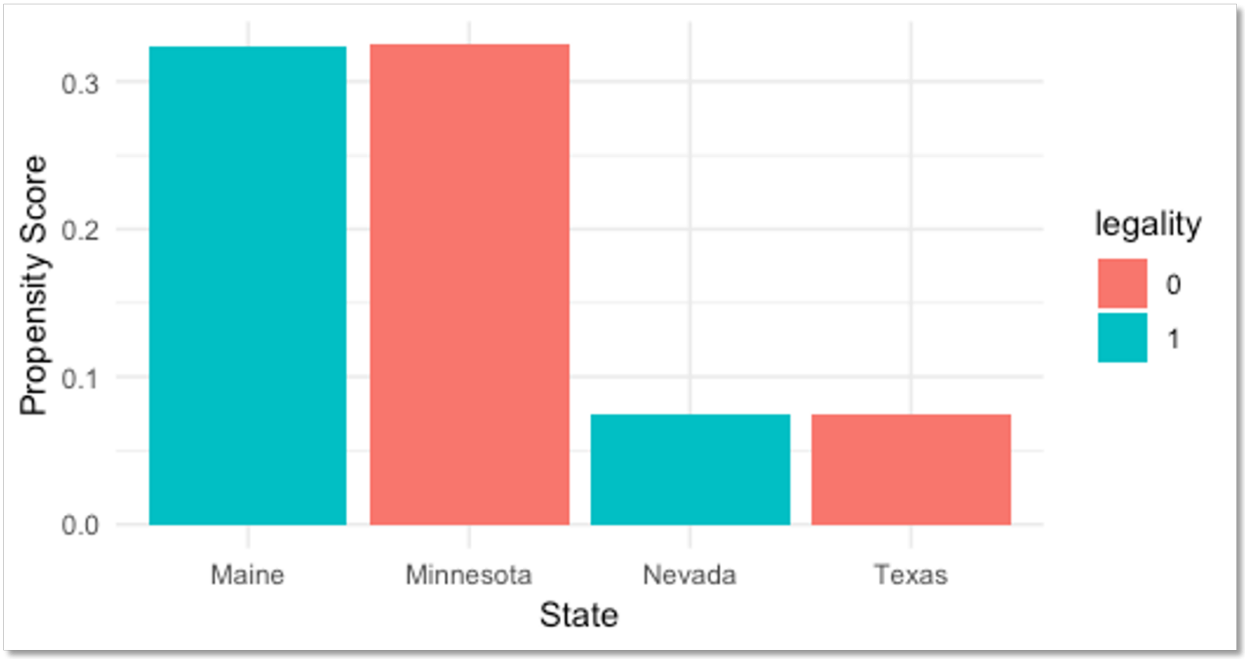


Figure 3: Propensity Scores of Matched States

With matches found for Maine and Nevada, the team now had a subset of the data that gave the best simulation of a randomized experiment.

Difference-in-Differences

After matching, the team isolated the causal effect legalization of cannabis had on fatal car crashes by utilizing DiD. Panel data was observed from 01/15 - 12/18 with an effective treatment date of 01/17.

Assumptions

Our assumptions when making inferences from the DiD results are as follows:

1. The number of fatal car crashes across the treatment and control groups follow a parallel trend
2. There are no unobserved, time-variant factors that are unique to either the control or treatment group not included as control variables in the regression
3. Illegal states are not influencing the legalized states

To verify the assumption of a parallel trend, a plot of fatal crashes between treatment and control groups was created.

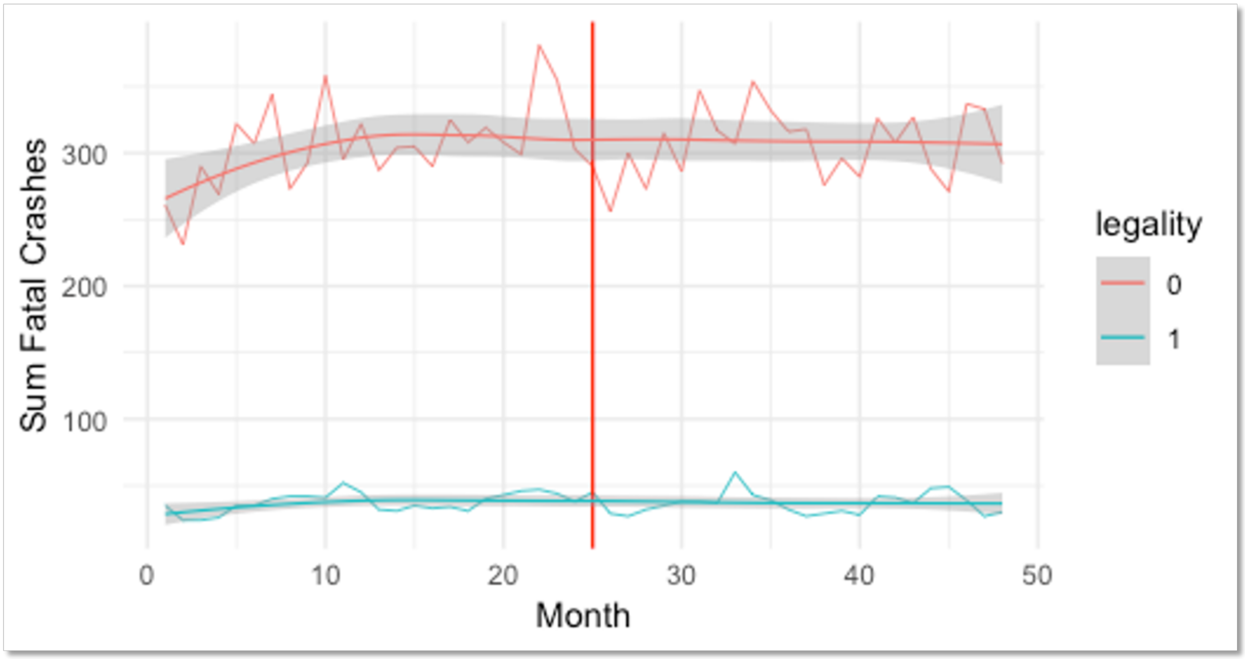


Figure 4: Sum of Fatal Car Crashes Over Time for Control (0) vs Treatment (1)

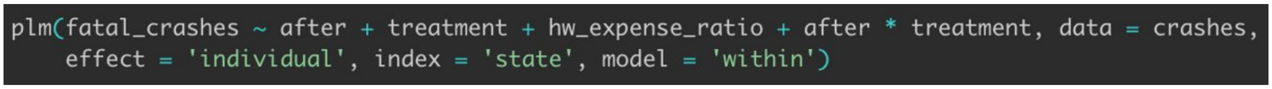
From figure 4 it was noticed that there were substantial variations in the control group's fatal crash rate, mainly resulting from the high rate of fatal car crashes in Texas. The aggregated sum of fatal crashes and the 95% confidence interval for both control and treatment are generally parallel and smooth throughout the study which served as a rough check that the parallel trend did hold. However, if significant results are found a Dynamic DiD analysis would be conducted to rigorously confirm that there is indeed a parallel trend.

Model Specification

The dependent variable of interest, before and after treatment flag, and time-variant confounds that are likely to be unique to each state are as follows:

- DV: Fatal car crashes per month
- Treatment: legality (1 = fully legalized, 0 = illegal)
- Before-after Flag: before 01/2017 = 0 and after 01/2017 = 1
- Time-variant confound: state highway expense divided by total area

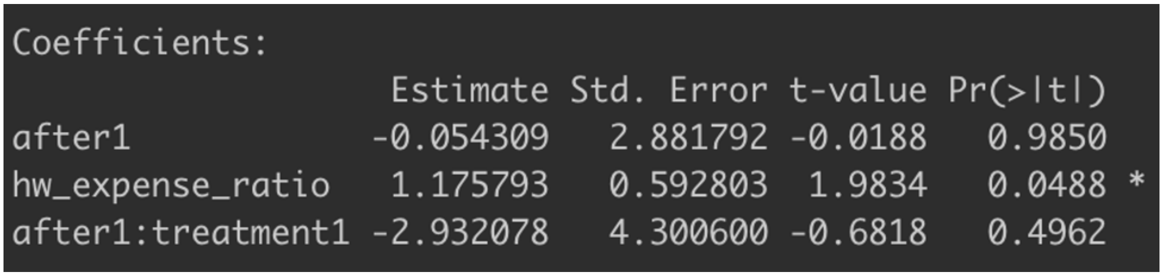
Population density was also included in the regression, but it made the results worse, so it was omitted from the final regression. After the decisions were made on which variables to include, the plm function (panel data linear regression) was borrowed from the plm library in R as the model for causal inference which can be found below.



```
plm(fatal_crashes ~ after + treatment + hw_expense_ratio + after * treatment, data = crashes,
    effect = 'individual', index = 'state', model = 'within')
```

Figure 5 - DiD Specification

Results



```
Coefficients:
                   Estimate Std. Error t-value Pr(>|t|)
after1            -0.054309   2.881792 -0.0188   0.9850
hw_expense_ratio   1.175793   0.592803  1.9834   0.0488 *
after1:treatment1 -2.932078   4.300600 -0.6818   0.4962
```

Figure 6: DiD Regression Results

The p-value for the interaction term was approximately 0.4962, meaning that there is a 50% chance that results at least as extreme as what was observed would be computed in future studies if there is no causal relationship between cannabis legalization and fatal crashes. In other words, insufficient evidence was obtained to conclude that there is a relationship between cannabis legalization and fatal crashes. Oddly, there is a very high chance the amount a state spends on its highways divided by its total area has a positive relationship with fatal crash rates.

Three different approaches to manipulating the model in figure 5 were tested to see if significant results could be found. First, weather data was included in the model, an external factor that varies across states and is likely correlated with fatal crashes. The weather data obtained only covered 03/16 - 12/18, which shrunk the analysis duration from 4 years to 1 year. The incomplete weather data crippled the team in analyzing the long-term impact of cannabis legalization on car accidents and significantly weakened the support for the coefficient estimates. Secondly, the target variable was manipulated. Fatal crashes per month was converted to a percentage increase from the previous month via a log transformation. It was hypothesized that this transformation would make the target variable more comparable across states. Nonetheless, no significant changes in the results were obtained. Lastly, since Texas had a much higher magnitude and variability in fatal crash rate than the other three states, Texas, and the corresponding treatment state, Nevada, were excluded from the analysis. Still, there were no significant changes in the results found.

Limitations & Improvements

If a smaller variation in the treatment effect could have been detected via more support or an improved target variable, significant results may have been attained. To make such an improvement, the current target variable, fatal car crash rate, should be substituted by the total car crash rate. Furthermore, the time-variant confounds included in the DiD regression could be improved. Instead of using population density, the number of vehicles registered through the state divided by the total area of that state could be substituted for population density in the DiD regression. This confound is likely more related to a state's car crash rate than population density since it is a better approximation of how dense traffic is.

As to the weather data, this data likely suffered from selection bias since it was collected only when there was a car crash and then aggregated to a monthly level. The weather data collected should be independent of car crashes to avoid this threat to causal inference in future studies. If such improvements are addressed and significant results are found, that study could motivate the development of effective cannabis assessment tests similar to the alcohol breathalyzer to better understand cannabis-impairment levels and crash rates and inform policy in states pondering future cannabis legalization.

Appendix

References

[1] Spencer Purnell & Allie Howell, "Does Marijuana Legalization Increase Traffic Accidents," Reason Foundation Public Policy Research, [Online], September 2018. Available: https://reason.org/wp-content/uploads/2018/08/evaluating-research-marijuana-legalization-traffic-accidents.pdf [Accessed February 5, 2020]

[2] Mark Anderson, Benjamin Hansen, & Daniel Rees, "Medical Marijuana Laws, Traffic Fatalities, and Alcohol Consumption," Journal of Law and Economics, vol. 56, [Online], May 2013. Available: https://www.jstor.org/stable/10.1086/668812?seq=1 [Accessed February 5, 2020]

Code

All project code can be found @ https://github.umn.edu/engel746/msba6440

Included in the Github Repository:

- *(1)_munging1.R* - Initial data cleaning and merging of data sources
- *(2)_munging2.R* - Final data cleaning and creation of dataframes for PSM and DiD
- *(3)_psm.R* - Propensity Score Matching analysis with exploratory plots
- *(4)_DiD.R* - Difference-in-Differences analysis with exploratory plots