# FINAL REPORT - IMPACT OF LEGALIZATION OF MARIJUANA ON CRIME RATE

Rikarnob Bhattacharyya, Boyang Wei, Batool Fatima, Mainak Roy, Fazel Tabassum

Team - 4

## Executive summary

Marijuana is a psychoactive drug that has been widely used in the US. In 2012, Colorado became the first state to legalize the recreational use of marijuana. However, the impact of this policy remains difficult to evaluate, specifically its impact on crime rates. Our project aims to evaluate the causal relationship between marijuana legalization and crime rates through a difference in differences regression model. Based on the regression result, we hope to provide policy and social implications for policy-makers and the general public.

## Background

Marijuana is the most commonly used illicit drug in the United States. Over 94 million people have reported using it at least once, which accounts for nearly 30% of the total population. Marijuana was regarded as an illicit drug across the United States until 1973 when Oregon became the first state to decriminalize it. The policy indicated a relaxation, but not an exemption of criminal penalties associated with personal marijuana use. Almost 40 years later, in 2011, Colorado became the first state to legalize marijuana, which not only allows individual marijuana possession but in most cases, permits the legal production and sale of the drug.

The legalization of marijuana brought additional revenue for Colorado state. The state imposes a 15% excise tax from cultivator to retailer and a further 15% sales tax on the end customer. As a result, in 2018, the marijuana sale in Colorado topped $1.2 billion, with the state pulling in about $270 million in taxes.

Marijuana might be influencing violent crimes; however, what is the impact of the legalization of marijuana on violent crimes? Will there be more violent crimes owing to intoxication in public places? Or will there be fewer violent crimes because people stop illegally acquiring marijuana? In general, We hope to understand if there is a causal relationship between the legalization of marijuana and violent crimes. If so, what sign and the magnitude would the impact be.

## Threats to Causal Inference

Since we are trying to evaluate the impact of legalization of marijuana on crime rate, there are a few threats that we need to pay attention to when drawing this causal inference:

1. Measurement error
    a. Measurement error could exist in the dependent variable, which is the counts of monthly violent crimes in this case. Not all crimes may get reported, so it is possible that some crimes that occurred but not included in the dataset. Suppose the true number of violent crimes is y, but our measurement is $\hat{y} = y + e$, where e

is the measurement error. Therefore, we would not be estimating y = $\beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$; instead, we are estimating ŷ-e = $\beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$

2. Reverse causality
    a. Although the implementation of the policy may have an impact on crime rates, the crime rates may also have an impact on the implementation policy. One possible explanation is that the number of violent crimes caused by marijuana (ie: robbery, murder, etc) is increasing. To deal with this type of crime, the government decides to legalize marijuana. In this case, it is not viable for us to establish a causal relationship between legalization policy and crime rates since they are affecting each other.
3. Omitted variable bias
    a. There may be a lot of factors that are impacting crime rates other than one policy. For example, there is a correlation between temperature and crime rates, in which the number of crimes may increase due to hot weather. Besides, the crime rate may also increase due to an economic recession. More and more people become unemployed, and this population base could have a higher chance of conducting violent crimes. If these variables do exist and are ignored, the casual relationship will be incorrectly interpreted.

## Datasets / Measures

There are two parts of data being collected.

1. State Characteristics

a. In order to identify similar states as Colorado, we need to find attributes at a state level. These attributes/features include median age, total population, population of male, education level, employment rate and so on. The data is generated from the US Census Bureau and can be found here: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

b. The unit of observation is one state in the US

c. The data collected is in 2012

2. Crime Data

    a. After we identified similar states, we need to find monthly crime data to test the parallel trend. All of the data is generated directly and indirectly from the government source, which includes the Department of Public Safety and the Department of State Police. The feature that we collected is the number of violent crimes at a monthly level. The violent crime consistent of different crime types, including rape, aggravated assault, homicide and robbery. The link can be found below (Note that some data are manually collected from the annual report)

        i. https://coloradocrimestats.state.co.us/public/View/dispview.aspx?ReportId=32 (Colorado)

        ii. https://www.azdps.gov/sites/default/files/media/Crime_In_Arizona_Report_2011.pdf
        https://www.azdps.gov/sites/default/files/media/Crime_In_Arizona_Report_2012.pdf

https://www.azdps.gov/sites/default/files/media/Crime_In_Arizona_Report_2013.pdf (Arizona)

iii. https://www.vsp.virginia.gov/downloads/Crime_in_Virginia/Crime_in_Virginia_2011.pdf

https://www.vsp.virginia.gov/downloads/Crime_in_Virginia/Crime_in_Virginia_2012.pdf

https://www.vsp.virginia.gov/downloads/Crime_in_Virginia/Crime_in_Virginia_2013.pdf (Virginia)

b. The unit of observation is the counts of violent crimes at a monthly level

c. The data collected is from November 2011 - December 2013, which is one year before and after marijuana was legalized in Colorado. The reason behind this was the assumption that there would be no significant changes within a state in a year prior and after

## Methodology: Selecting the similar states

1. The first step was to find states that are similar to Colorado in terms of "make-up"; i.e. states that have similar demographic composition in terms of the following variables –

   a. Difference from the rest of US in racial makeup (percentage figure)

   b. Share of population with bachelor's degree

   c. Median age

d. Median household income

e. Share of people who consider religion “very important”

f. Population

g. Population of males

h. Population under 18, and

i. Employment rate

2. We used Euclidean distance to find the closest neighbors to Colorado (since every variable is numerical). Given that there were variables in different scales, e.g. Age (in the order of 10s, and Income (in the order of $10^5$), we scaled the variables to a 0 – 1 scale

   a. Our assumption was, barring factors we couldn’t measure or observe (like other policy changes or other microeconomic indicators

3. With the distances calculated, we proceeded to take the top 5 states (i.e. the 5 closest neighbors in terms of Euclidean distance).

   a. Note: Since our methodology involved calculating difference in differences, we required data that had crime rates by month for the years under consideration (November 2011 – December 2013). We were able to get the data in the required panel structure only for Virginia and Arizona (in addition to Colorado, which was our treatment state) and we proceeded with these

4. Within November 2011 – December 2013, we plotted the crime trends in these states (Appendix A). The treatment period started on November 6, 2012, which was when legalization of marijuana occurred in Colorado. A quick eyeball test confirmed that Arizona and Colorado had more similarity in terms of a parallel trend than did Colorado and Virginia, prior to November 6.

    a. We carried out a dynamic difference in differences to find out which of the two states were more "similar" in terms of their trends (Appendix B). The results revealed that Arizona and Colorado were indeed more similar and so we proceeded with Colorado as treatment and Arizona as control

## Methodology: Difference in differences

We carried out a two-way panel regression to detect the difference in difference in crime rates between the two states. Our dependent variable was the log of crime count since we wanted to track the percentage change from before legalization to after legalization. The regression equation used was -

$$log(Crime\ count) = \beta_0 + \beta_1 * Legal + \beta_2 * After + \beta_3 * Legal * After$$

Where,

*Legal = Flag variable indicating whether marijuana is legal in a state, i.e. treatment - control flag*

*After = Flag variable indicating whether time period is before or after November 6, 2012*

$\beta_3$gives us the treatment effect, i.e. the difference in differences, that captures the percentage change in crime in going from before legalization to after legalization in Colorado.

We assumed that there would be unobservable confounds in between states and across time. To overcome this we considered a panel structure of State and Time and formulated the regression equation (Appendix C).

## Results & Conclusions

Using two-way panel regression we obtained a $\beta_3$ value of roughly -0.08 implying there is a 0.8% decrease in crimes. However, the corresponding p-value was 0.79 implying that we would obtain this result by chance (and not due to causation) 79% times.
The model was tweaked to introduce a placebo, i.e. an artificial treatment date. We pushed it forward by 3 months to August 2012. This was done to investigate changes in significance to $\beta_3$, if any. Even though we still did not obtain a significant p-value, it dropped to 0.43 and the $\beta_3$ coefficient was -0.024 indicating that there would be a 2.4% decrease in crimes.

This led us to conclude the following -

1. Since results were statistically insignificant, there was no way we could conclude with certainty that the legalization of marijuana was a causal factor behind crime reduction. However, policymakers can view this result directionally instead of it being the gospel

truth. Lack of enough data points was a limitation and more data might lead to more conclusive (and encouraging) results

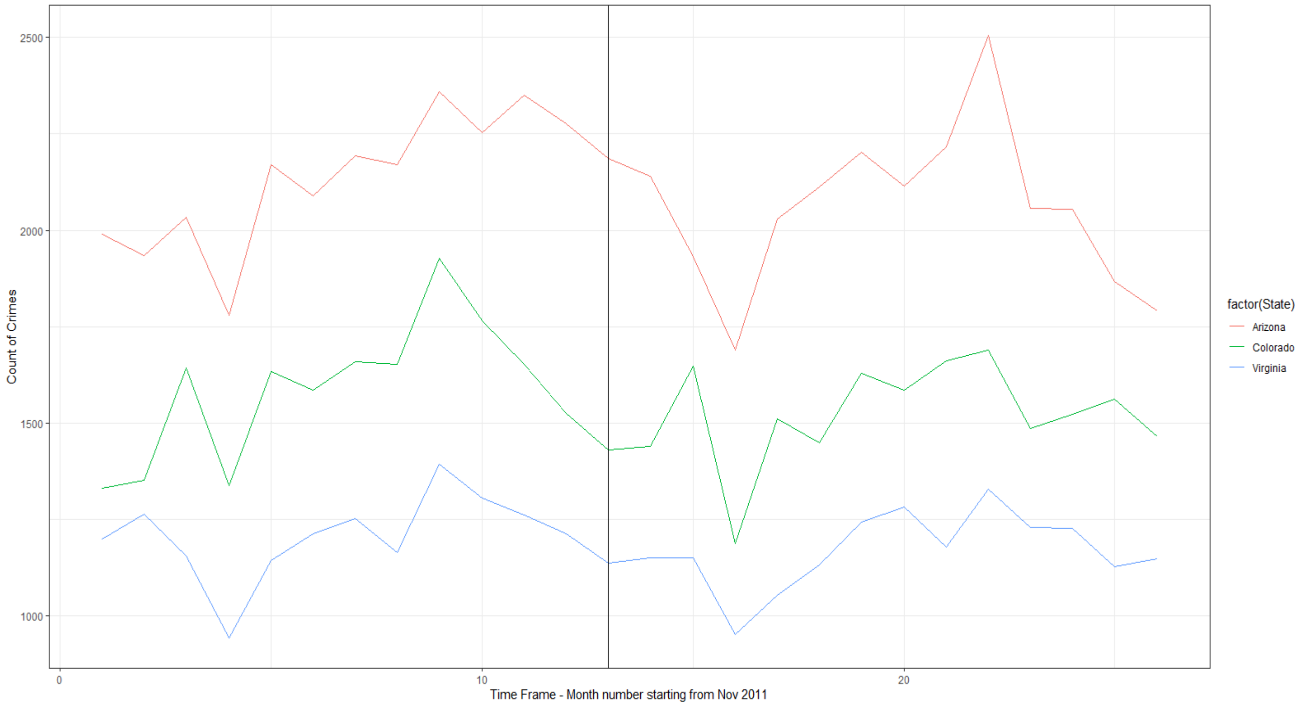
2. Secondly, the placebo test revealed something fairly interesting - the treatment effect improved and so did the significance level. This can be indicative of a certain "anticipation factor" coming into play prior to legalization. News of marijuana being legalized in the future might have caused people to behave differently and driven crime rates down prior to actual legalization. Nonetheless, again, drawing causal inferences should be avoided owing to lack of statistical significance

## Limitations

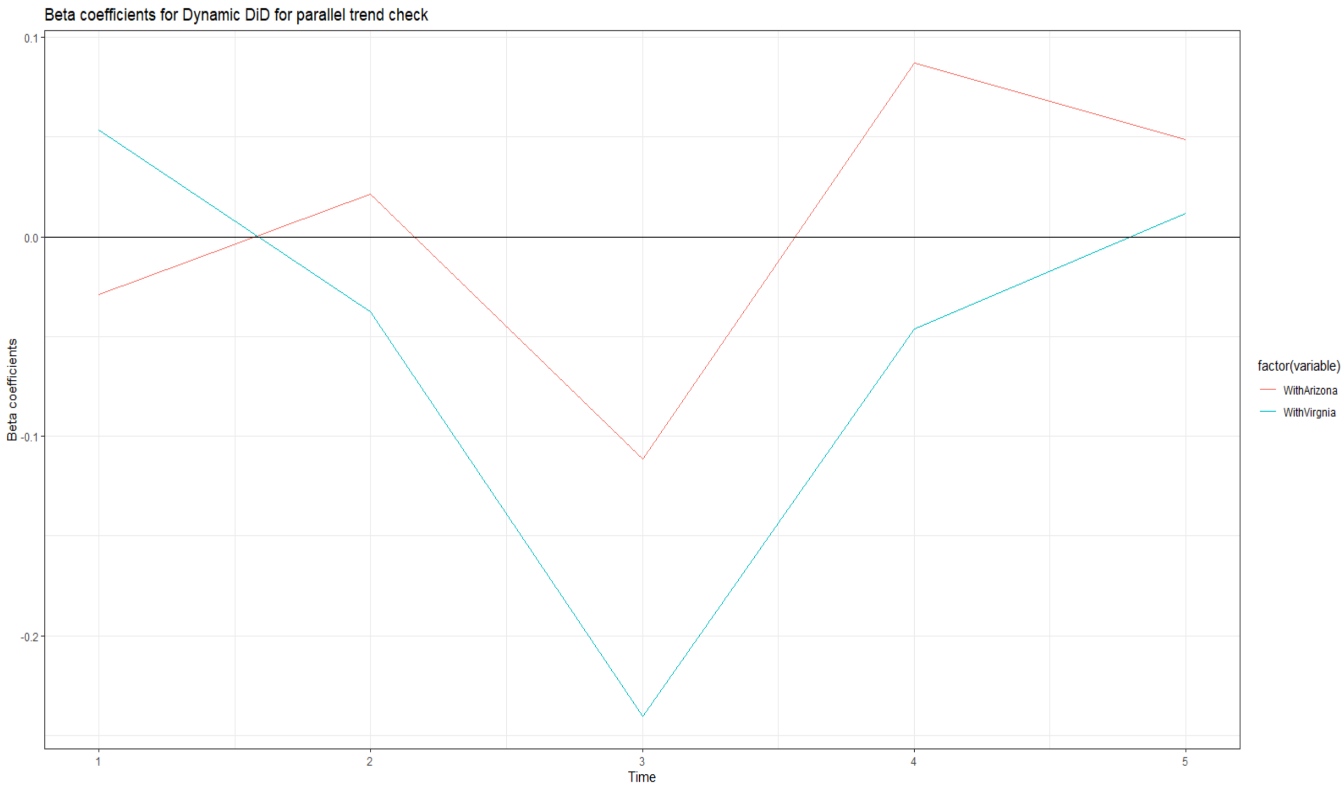
1. Given the results obtained are statistically insignificant (albeit, encouraging for lawmakers), causal inferences need to be drawn with caution;
2. Since we had only 26 time-points for each of the two states, the model suffered from a lack of data. A possible next step for policymakers might be to collect more data crime before and after the legalization event (especially after)
3. Being neighboring states, there might be interference between Arizona and Colorado that causes differences in behavior in people - e.g., there could be people traveling to and from Colorado at a rate that affects the demographics of the region
4. Since crime is dependent on a lot of factors outside of policy, causal attribution of decreasing crime rate to policy alone is difficult

## Appendix:

# Appendix A (Graph for parallel trend)




# Appendix B (Result of dynamic did)




# Appendix C (Full R Code)

```
## Reading in file with state demographic characteristics
state <- readxl::read_xlsx('D:/UMN - Materials/3. Spring/MSBA
6440/Project/State Similarity Index.xlsx')
View(state)

## Converting numeric columns from string to double
state_numeric <- data.frame(apply(state[,2:10], 2, as.double))

## Creating a normalized table for calculating Euclidean distance
xx <- normalize(state_numeric, method='standardize', range=c(0,1), margin=2)

## Calculating Euclidean distance
library(cluster)
gower.dist <- daisy(xx)
gower.dist <- as.matrix(xx)

## Taking 8th index - colorado state
View(sort(gower.dist[,8]))

## 12, 32, 15, 5, 21 - are the top 5 state indices
```

```
state %>% filter(rownames(state) %in% c('12', '32', '15', '5', '20'))

###
--------------------------------------------------------------------------------
---------------------------------------------

#### DiD Regression ####

library(plm)
library(dplyr)
library(ggplot2)

#### Load the data ####
data = readxl::read_xlsx("D:/UMN - Materials/3. Spring/MSBA
6440/Project/CrimeCountData.xlsx", sheet = 1)

data <- data %>% group_by(State) %>% mutate(time = 1:26) %>% ungroup()
colnames(data)[3] <- "CrimeCount"

## Fixing a placebo test date for the placebo test, i.e. an artifical start
time
data <- data %>% mutate(After1 = ifelse(time>=10,1,0))

## Making two datasets for later use in Dynamic DiD - one without Arizona, one
without Virginia
data_az_co <- data %>% filter(State != 'Virginia')
data_va_co <- data %>% filter(State != 'Arizona')


# As descriptive visualization, we take a look at the crime trends for each of
the 3 states
ggplot(data) + aes(x = time, y = CrimeCount, col=factor(State)) + geom_line() +
geom_vline(xintercept = 13) +
  xlab('Time Frame - Month number starting from Nov 2011') + ylab('Count of
Crimes') + theme_bw()

ggplot(data_az_co) + aes(x = time, y = CrimeCount, col=factor(State)) +
geom_line() + geom_vline(xintercept = 13)

#### Difference in Differences Regression ####
# Interpreting the treatment effect

data$Legal <- as.factor(data$Legal)
data$After <- as.factor(data$After)

## Two way panel regression to evaluate the difference in differences
did_fe <- plm(log(CrimeCount) ~ Legal + After + Legal*After, effect='twoway',
               index=c('State','time') ,model='within',data=data_az_co)
```

```
summary(did_fe)


## Two way panel regression to evaluate difference in differences with Placebo effect
did_fe_pl <- plm(log(CrimeCount) ~ Legal + After1 + Legal*After1, effect='twoway',
              index=c('State','time') ,model='within',data=data_az_co)

summary(did_fe_pl)


# Dynamic DiD to figure out the parallel trend - AZ and CO; VA and CO
did_dyn1 <- lm(log(CrimeCount) ~ Legal + factor(time) + Legal*factor(time), data=data_az_co)
summary(did_dyn1)

did_dyn2 <- lm(log(CrimeCount) ~ Legal + factor(time) + Legal*factor(time), data=data_va_co)
summary(did_dyn2)


## Making a dataframe of coefficients of the dynamic DiD test
coeff_data<- data.frame(did_dyn1$coefficients[3:14], did_dyn2$coefficients[3:14])
colnames(coeff_data) <- c('WithArizona','WithVirgnia')
coeff_data_melted <- melt(coeff_data)
coeff_data_melted <- coeff_data_melted %>% group_by(variable) %>% mutate(time = c(1:12))


## plotting dynamic DiD results
ggplot(coeff_data_melted[coeff_data_melted$time<6,]) + aes(x = time, y = value, col=factor(variable)) + geom_line() +
   geom_hline(yintercept = 0) +
   theme_bw() + xlab('Time') + ylab('Beta coefficients') +
   ggtitle('Beta coefficients for Dynamic DiD for parallel trend check')
```