American Crime and Incarceration

STA 199 Final Project

Oliver Greenwald, Karam Oubari, Steven Powell, Emely Gutierrez

Synopsis

This research project is aimed to further understand and analyze American prison incarceration and crime in the 50 states between 2001 and 2016. Analysis was done by focusing specifically on state prisoners (as opposed to federal) and the types of crimes that prisoners had been convicted of. The data set used observes 7 types of crimes that can be classified into violent crimes and property crimes.

Introduction and Data

(The introduction should introduce your general research question and your data (where it came from, how it was collected, what are the cases, what are the variables, etc.).)

The United States currently has the highest prison population in the world at around 2,094,000. This situation can be traced back to the 1970s, when politicians like Richard Nixon, Ronald Reagan, and more began the War on Drugs and their "tough on crime" agendas. Since then, rates of imprisonment have continued to increase under the guise of "public safety." However, the reality is that the US' crime rates are comparable to countries with similar economies, yet the rate of incarceration is significantly higher. The US has only 5% of the world's population, and yet it houses 25% of the world's incarcerated persons.

In this project we aim to understand the extent of the issue of mass incarceration, what sub-statistics can most explain the issue, how the numbers connect to public policy, and the future of the issue through basic predictive modeling.

Data Description

"The Bureau of Justice Statistics administers the National Prisoners Statistics Program (NPS), an annual data collection effort that began in response to a 1926 congressional mandate. The Uniform Crime Report (UCR) has served as the FBI's primary national data collection tool since a 1930 congressional mandate directed the Attorney General to 'acquire, collect, classify, and preserve identification, criminal identification, crime, and other records.' The FBI collects this information voluntarily submitted by local, state, and federal law enforcement agencies."

The user who updated this data set used the raw data from the NPS and wrangled it in Python to create the variables and observations. Within the data set there are 816 rows and 17 variables. The variables included in the set are US state, whether or not the data for that entry includes jails, the year, prisoner count on December 31st, whether or not the state in the data entry changed their system for reporting crime in comparison to previous years, whether or not crime totals are estimated, and numbers of: total state population, violent crimes, murders or manslaughter, rapes (using the old definition), rapes (using the new definition), robberies, aggravated assaults, property crime, burglaries, larceny, and vehicle theft.

Methodology

(The methodology section should include the variables used to address your research question, as well as any useful visualizations or summary statistics. As well, you should introduce and justify the statistical method(s) that you believe will be useful in answering your research question.)

The following variables, from the original data set, were used to address the research questions above:
• `jurisdiction`: a categorical variable for federal jurisdiction and the 50 states. Federal jurisdiction refers to the legal scope of the government's powers in the United States of America.
• `year`: a numerical variable for the year of each observation, 2001 through 2016.
• `prisoner_count`:a numerical variable for the number of prisoners in each state for the given year.
• `state_population`: a numerical variable for the population of each state in a given year.
• `violent_crime_total`: a numerical variable for the total number of violent crimes (equals the sum of `murder_manslaughter`, `rape_legacy`, `rape_revised`, `robbery`, and `agg_assault`). A violent crime is defined as a victim is harmed by or threatened with violence.
• `murder_manslaughter`: a numerical variable for the numbers of murders and manslaughters. Manslaughter is defined as an unlawful killing that doesn't involve malice aforethought—intent to seriously harm or kill, or extreme, reckless disregard for life.
• `rape_legacy`: a numerical variable for the number of rapes (using the old definition). The old definition of rape is: "The carnal knowledge of a female forcibly and against her will"
• `rape_revised`: a numerical variable for the number of rapes (using the new revised definition). The new definition of rape is: "Penetration no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of anotehr person, without the consent of the victim"
• `robbery`: a numerical variable for the number of robberies. Robbery is defined as the taking of something of value from another person using force or violence or the threat of force or violence
• `agg_assault`: a numerical variable for the number of aggravated assaults. Aggravated assault is defined as an attempt to cause serious bodily harm to an individual with disregard for human life. Factors that raise an assault to the aggravated level typically include the use of a weapon, the status of the victim, the intent of the perpetrator, and the degree of injury caused.
• `property_crime_total`: a numerical variable for the total number of property crimes (equals the sum of `burglary`, `larceny`, and `vehicle_theft`). In a property crime, a victim's property is stolen or destroyed, without the use or threat of force against the victim.
• `burglary`: a numerical variable for the number of burglaries. Burglary is defined as the unlawful entry into almost any structure (not just a home or business) with the intent to commit any crime inside (not just theft/larceny). No physical breaking and entering is required; the offender may simply trespass through an open door. Unlike robbery, which involves use of force or fear to obtain another person's property, there is usually no victim present during a burglary.
• `larceny`: a numerical variable for the number of larcenies. Larceny is defined as what most people think of as common theft - the taking of someone else's property without the use of force.
• `vehicle_theft`: a numerical variable for the number of vehicle thefts. Vehicle theft is defined as motor vehicle theft is defined as the theft or attempted theft of a motor vehicle. In the UCR Program, a motor vehicle is a self-propelled vehicle that runs on land surfaces and not on rails.

Statistical Methods

To answer the research questions we are exploring we utilized various statistical methods using R. More specifically, we used data visualization, and linear regression, to both see the current issue and extrapolate and make predictions outside the scope of the given data.

Summary Statistics and Visualizations

First, in this project we aim to understand the extent of the issue of mass incarceration through summary statistics and visualizations. In Figure 1, we visualize the total number of prisoners in the US from 2001 to 2016. As we can see the graph shows a steady incline from 2001 to 2008 and then a steady decline followed by a sharp decline thereafter. In Figure 2, we visualize the amount of crime in the US and notice that it

decreases quite steadily between 2001 and 2016. . In Figure 3.1, we examine the change in the proportion of the population that is imprisoned between 2001 and 2016. This shows us a steady increase between 2001 and 2007 before a steady decrease afterwards. Figure 3.2 shows the same information in terms of the percent of a states population that is incarcerated in a box plot rather than a line graph. The box plot shows that though the median percent of the population does not change significantly from year to year, certain years have a larger spread and large outliers. These outliers are further explained in Figure 4 where we mapped out the proportion of each state's population that was imprisoned in 2016 to get a clearer idea of which states contribute most the issue of mass incarceration. Most noticeably, Delaware, Oklahoma, and Arizona have have much higher imprisonment proportions than the rest of the country. These are likely the outliers in the boxplot from Figure 3.2.

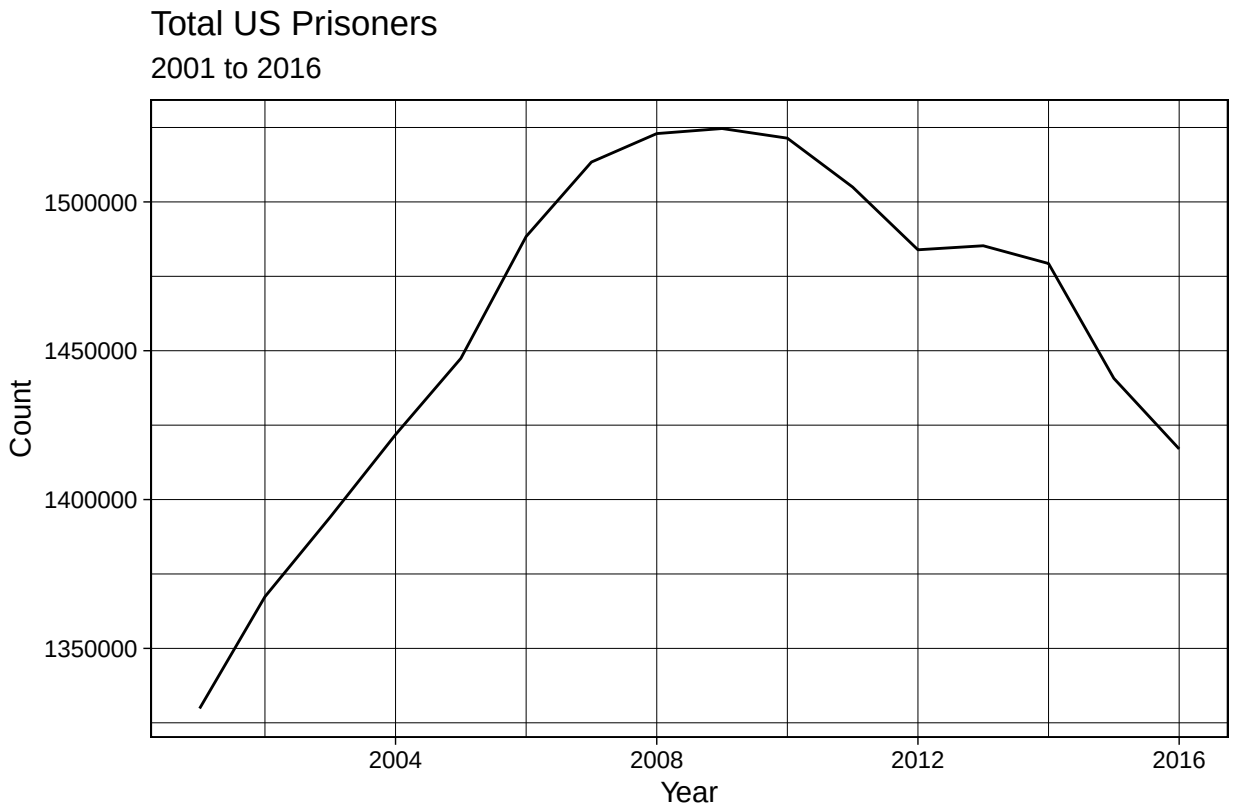


Figure 1

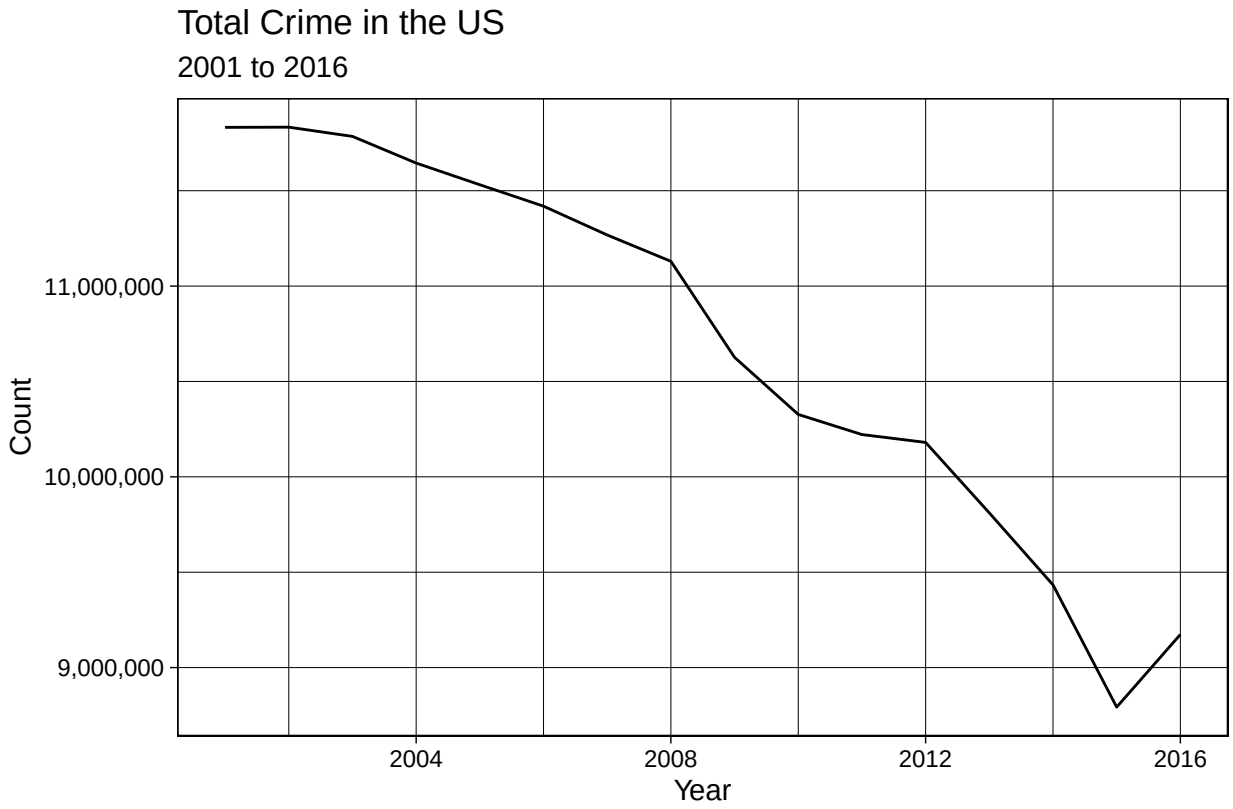


Figure 2

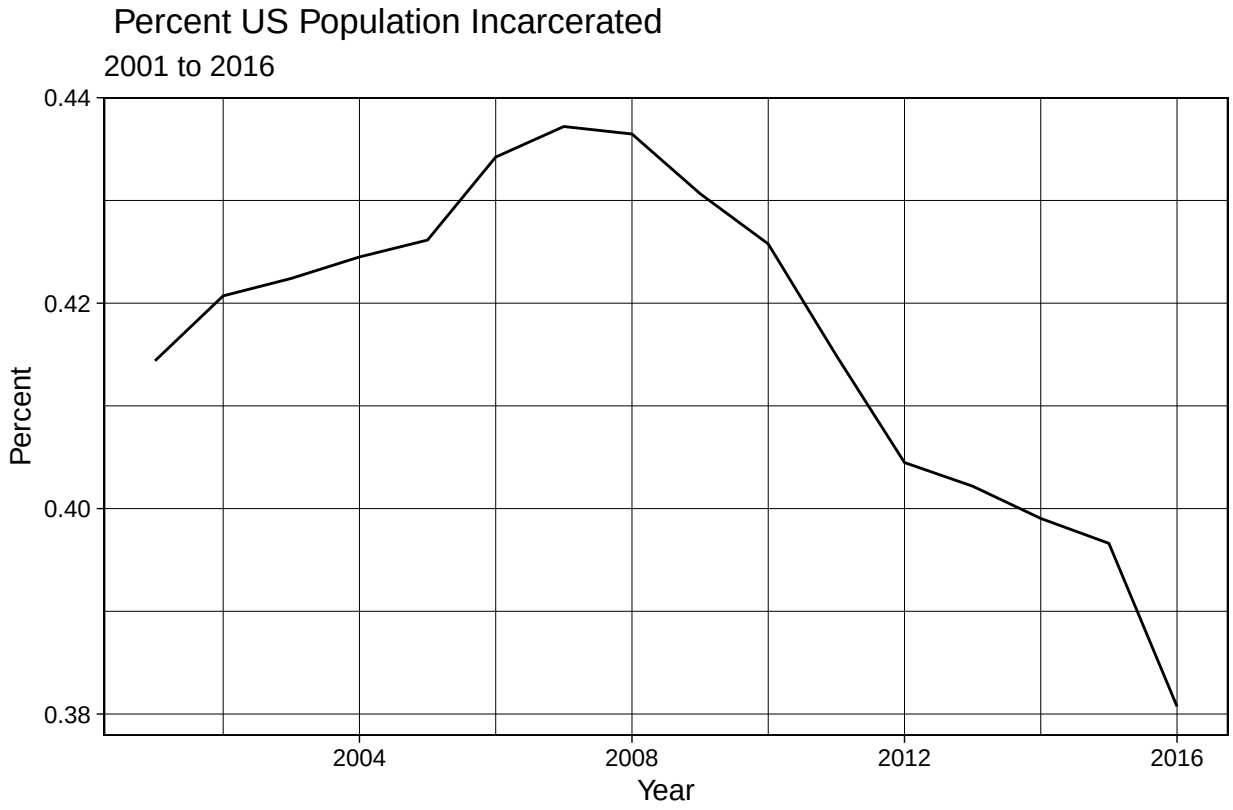


Figure 3.1

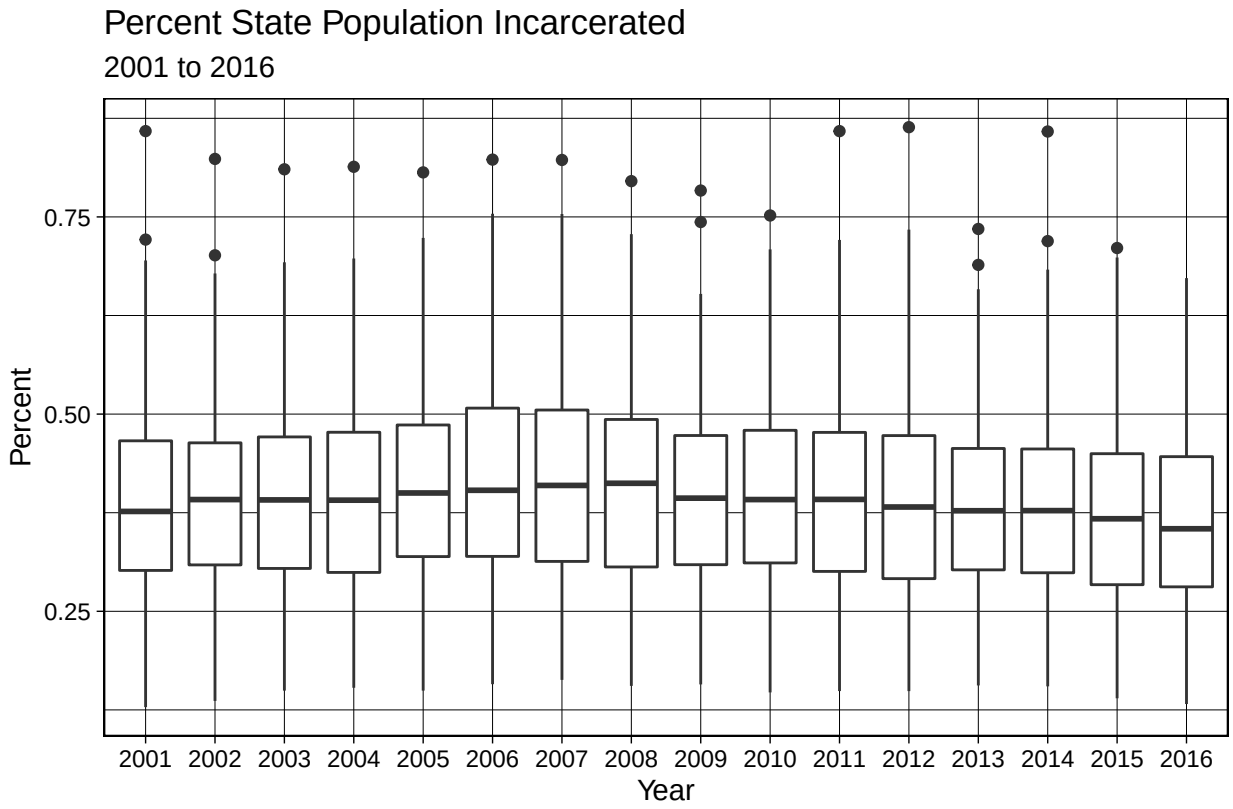


Figure 3.2

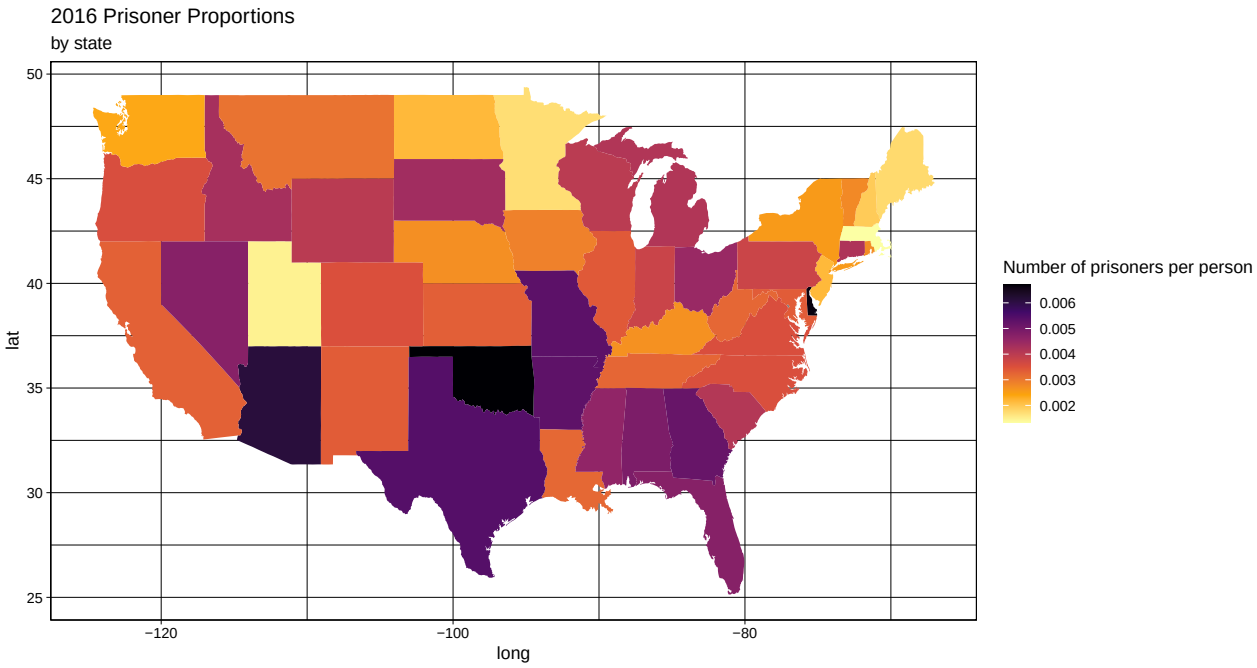


Figure 4

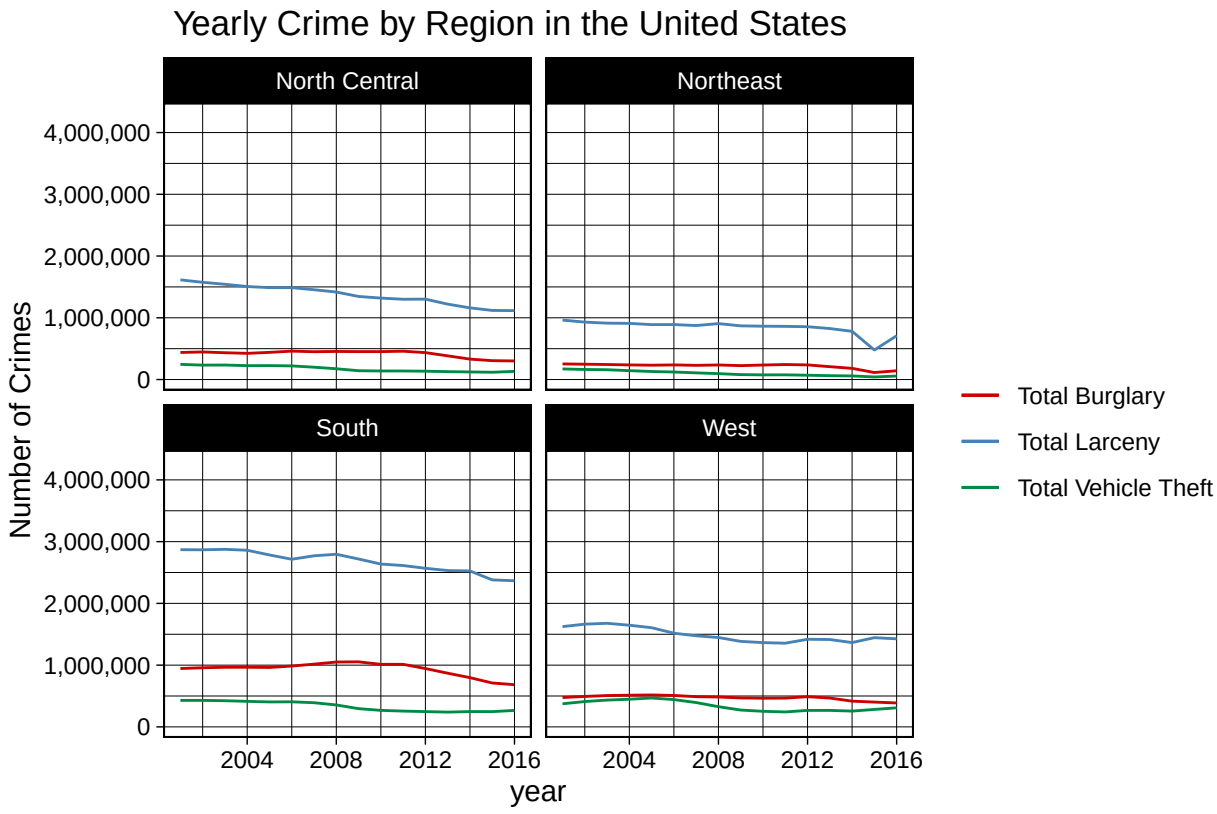
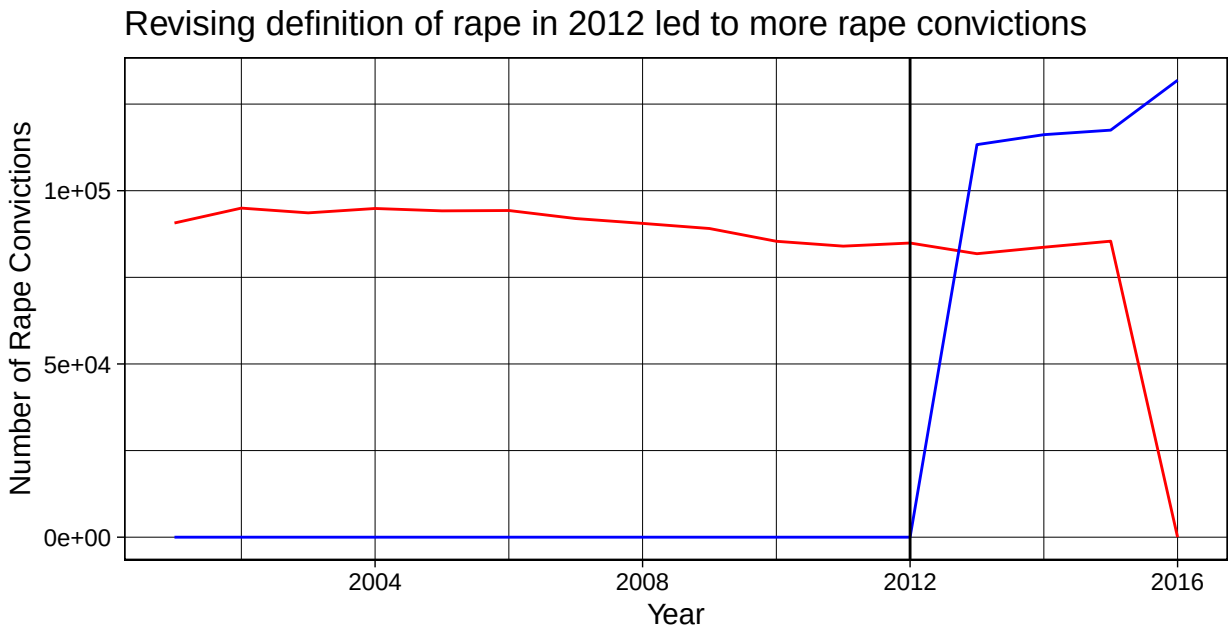


Figure 5



Red = Total Number of Prisoners convicted of rape by the Legacy Definition in state custody

Blue = Total Number of Prisoners convicted of rape by the Revised Definition in state custody

Figure 6

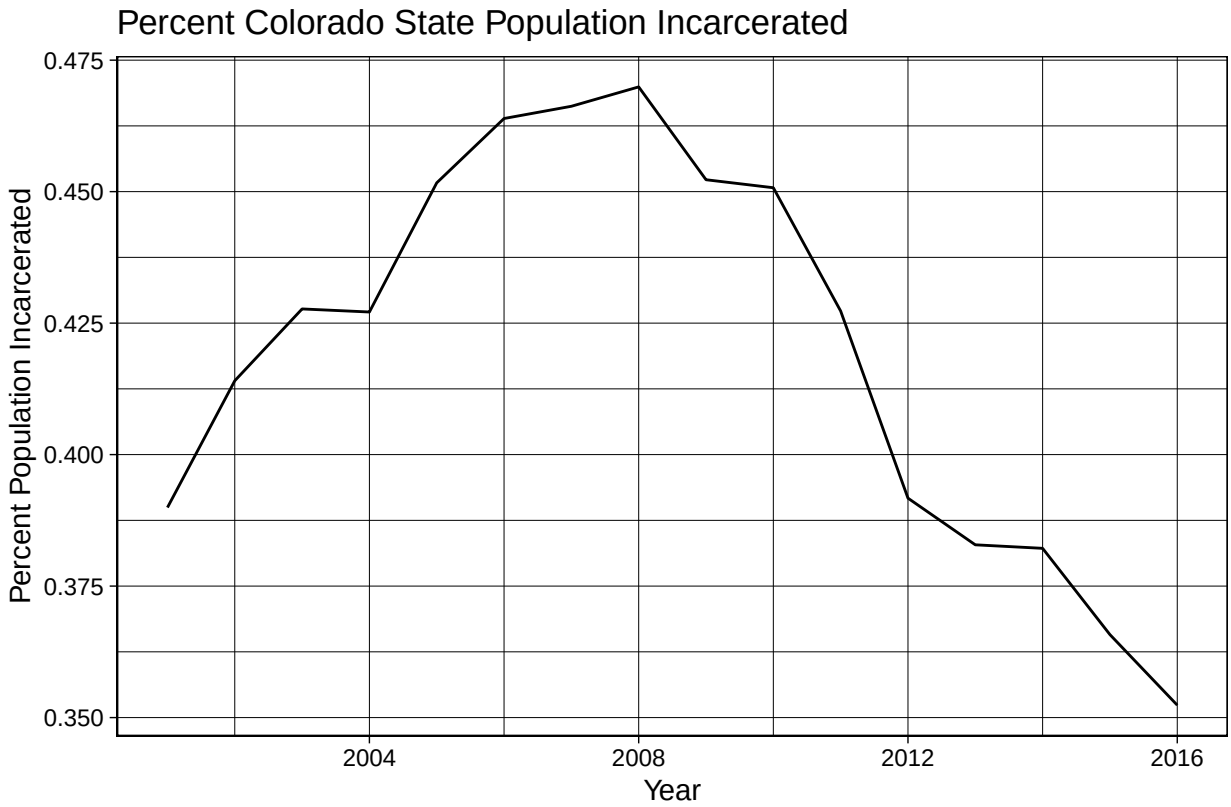


Figure 7.1

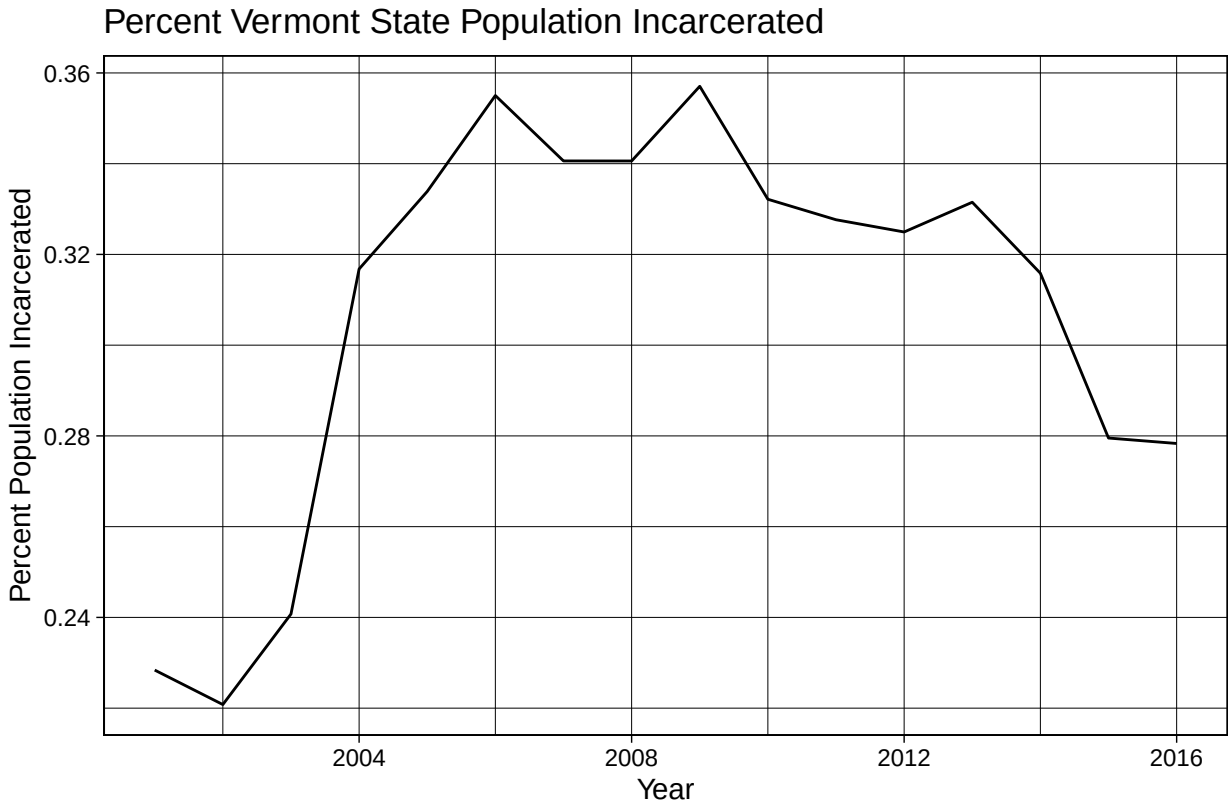


Figure 7.2

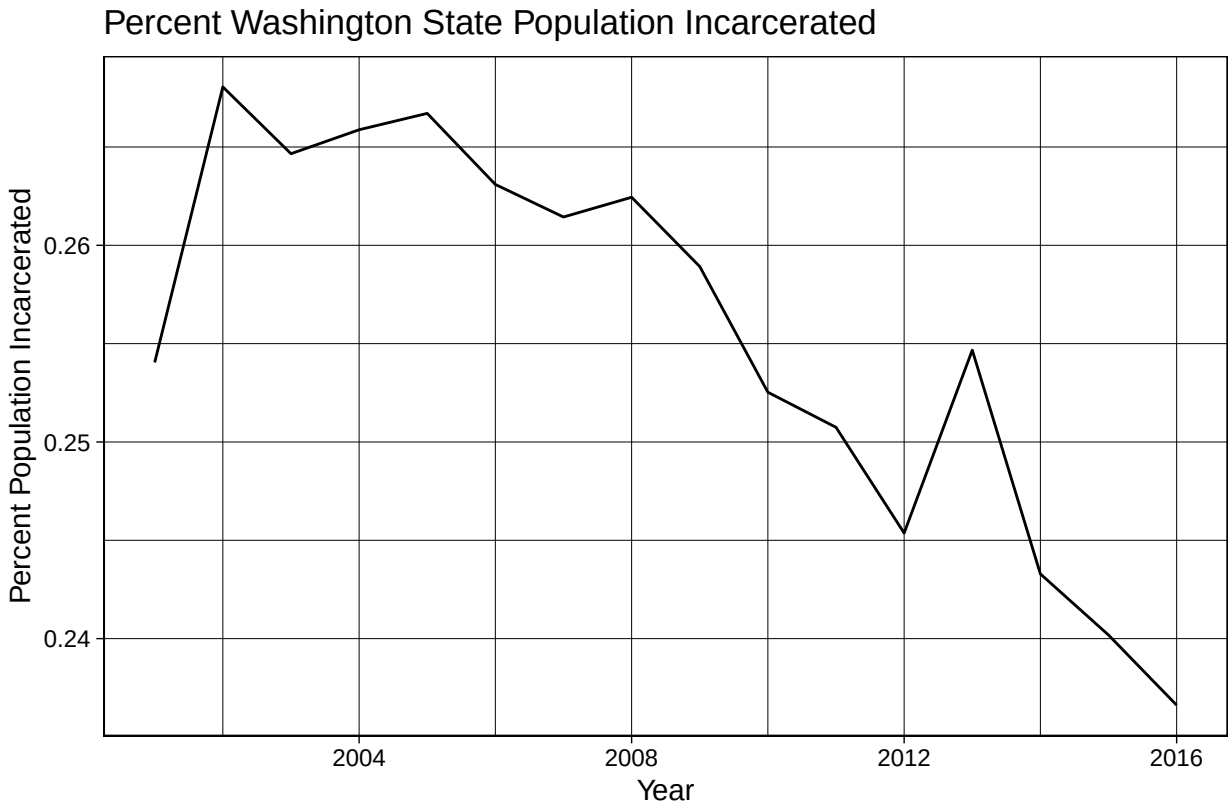


Figure 7.3

Results

(Showcase how you arrived at answers to your question using any techniques we have learned in this class (and some beyond, if you're feeling adventurous). Provide the main results from your analysis. The goal is not to do an exhaustive data analysis (i.e., do not calculate every statistic and procedure you have learned for every variable), but rather let me know that you are proficient at asking meaningful questions and answering them with results of data analysis, that you are proficient in using R, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research questions.)

Understanding the extent of the issue of mass incarceration:

Political rhetoric that has come to the forefront in recent decades is the idea of law and order, being tough on crime, and ensuring punishment to the fullest extent of the law. We went into this project knowing that the number of incarcerated individuals in the US is the highest of any other developed country with a similar economy and governmental system. Knowing this, we would assume that if the US has a strong belief that we need to be tough on crime, the high rates of incarceration are due to high crime rates, violent or otherwise. With this in mind, we wanted to find if the data would support this hypothesis.

Figure 1 and Figure 2:
What is immediately evident is that even when crime of all types was sharply increasing, the number of incarcerated individuals was still steadily decreasing. Based on this alone, though we cannot definitively draw this conclusion, it would appear that rates of incarceration are actually not due to how much or how little crime is being committed and is instead due to other factors which we will explore later in our analyses. Possible explanations might be policy changes, new sentencing guidelines, amount of policing, and a plethora of other reasons.

Figure 3.1, Figure 3.2, and Figure 4:
After drawing the previous conclusion, we thought there may be another explanation, even if crime was going down it could be that US population was increasing, when we graphed it above, it would appear that the percent of the US population which is imprisoned shows very strong similarities to the Total US Prison Population graph. However, that being said, even if the population is increasing, if the crime rate is staying the same, or decreasing as we observed previously, there is still no valid reason to incarcerate more people.

Sub-statistics that can better explain the issue:

How does crime differ for different regions of the US?

Figure 5:
To see how crime differs by region, we made a variable called `region` which splits up the United States into 4 distinct regions (North Central, Northeast, South, and West). We decided to look at only property crime (burglary, larceny, and vehicle theft) and so we plotted the total buglary, total larceny, total vehicle theft, and overall total property crime from 2000-2016 on each regional graph to see what the trends looked like.

One positive note with these graphs is that each region has a overall general decline in the amount of property crime throughout 2000-2016. This is only a speculation, but this would suggest that trends of property crime does not differ throughout each region but it does for each crime.

The south overwhelmingly has the total highest property crime total burglary, and total larceny.

An outlier seems to be in the northeast graph in 2015 when there was sharp decrease with larceny and burglary, and thus total property crime. There must have been policy that affected the Northeast only during that time period that lessened crime

Because of the high number of crimes in the south and the downward spime for the northwest graph that the north central and west regions are most similar in numbers of crimes committed over this 16 year span.

It is also interesting to see that the types of crimes themselves are seemingly related. But this we mean that for each graph, from the highest number to the lowest number, the crimes are larceny, burglary, then vehicle theft.

If you were looking at a region to live in the united states and buy a house or car, it seems that the best place to do that would be in the North East, followed closely by the north central and west region. Thus it would be advised to not choose the south, although they are on a decline in property crimes committed.

How the results and the issue connect to non-statistical public policy

How does policy impact incarceration? This will be explored through the lens of changing definition of rape as well as through three states which decriminalized marijuana within the time frame of the data set.

Figure 6:
In 2012 the FBI changed their definition of rape from: "The old definition was 'The carnal knowledge of a female forcibly and against her will.' Many agencies interpreted this definition as excluding a long list of sex offenses that are criminal in most jurisdictions, such as offenses involving oral or anal penetration, penetration with objects, and rapes of males. The new Summary definition of Rape is: 'Penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim.' "

We wanted to see how the change in definition affected the number of prisoners convicted for rape. Based on the graph, it seems like after the new definition was implemented, the number of prisoners who are incarcerated for rape increased. We will perform tests to find the p-value and try to determine if this is a correlation relationship or a causal relationship.

This one change in policy had obvious impacts on the number of people incarcerated for the crime. While it is only a small example, it supports theories on incarceration that emphasize the role of laws and policies on the number of people incarcerated rather than the actual number of crimes committed.

This is assuming a causal relationship between the policy change and the number of people incarcerated. The creation and enforcement of policy has a greater impact on who is incarcerated and how many people are incarcerated for each crime than the actual rate of crime. Despite public perception assuming the opposite relationship. Many times the policies can be used to adjust a narrative to support a political movement or idea at the time. Crime rates can increase when policy is adjusted to include more crime as seen with the change in defintion for rape. This is not to say that this change contributed significantly to mass incarceration or that the change in definition was unjust. Rather, this example in our data exemplifies the impact of policies on incarceration rather than crime rate.

Figure 7.1, Figure 7.2, and Figure 7.3:
Colorado and Washington were the first US states to decriminalize the use of marijuana for recreational purposes in 2012 and Vermont followed suit in 2013. We wanted to see if there would be any notable decline in incarceration in these three states after recreational legalization. It is unfortunate that drug crimes were not included in the data set as the War on Drugs is responsible for 30-50% of the incarceration that we see today. Nonetheless, we thought we could seek to find if there was a notable decline in incarceration after 2012 in Colorado and Washington and after 2013 in Vermont. Figures 7.2 and 7.3 for Vermont and Washington state, respectively,show pretty strong evidence for a sharp decline in incarceration after recreational legalization. Figure 7.1 for Colorado is less convincing as the incarceration rate leading up to 2012 was already on the decline, however, the incarceration most certainly does not increase with recreational legalization, which is definitely not enough to call it a causal relationship, however, it is pretty compelling evidence.

Future of the issue through basic predictive modeling:

To explore the future of prison incarceration in the United States, we chose to create a linear model that we could use to extrapolate and make broad predictions about years past 2016. In addition, we though it would be interesting to compare these predicted numbers to the actual statistics from 2017, 2018, and 2019. This way we could evaluate the strength of our linear model and make a general assessment as to whether it is possible to predict the future of prison incarceration using past data, or whether these statistics are better determined by other factors such as public policy changes.

```
## # A tibble: 2 x 2
##   term        estimate
```

```
##   <chr>          <dbl>
## 1 (Intercept)  5.03
## 2 year        -0.00230
```

pred. % incarcerated = 5.03374618 - 0.00229865(year)

- This linear model implies that in the year 0, on average, the predicted percentage incarcerated in the United States was about 5% (a huge and unreasonable extrapolation). It also implies that all else held constant with each passing year, the percentage incarcerated in the United States is expected to decrease by about .002%.

pred. % incarcerated in 2017 = 5.03374618 - 0.00229865(2017)
pred. % incarcerated in 2017 = 0.3973691
actual % incarcerated in 2017 = 0.4570901

pred. % incarcerated in 2018 = 5.03374618 - 0.00229865(2018)
pred. % incarcerated in 2018 = 0.3950705
actual % incarcerated in 2018 = 0.4496333

pred. % incarcerated in 2019 = 5.03374618 - 0.00229865(2019)
pred. % incarcerated in 2019 = 0.3927718
actual % incarcerated in 2019 = 0.4373857

When compared to the actual data it is clear that this data slightly underestimates the percent incarcerated in the United States when extrapolated. In addition, it is likely inaccurate to extrapolate further than just a few years, since it is unlikely that incarceration rates will continue to drop forever.

Discussion

(This section is a conclusion and discussion. This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. Also, critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data and appropriateness of the statistical analysis should also be discussed here. A paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project should also be included.)

• If we were to continue with this project we would improve our analysis by looking for more political and event-based research to aid our graphs. For example, in figure 5 there is an obvious dip in every crime for the northeast region in ~2015. We spent a good amount of time trying to research this and see if there was any policy implemented that would cause a significant dip or an event that would cause crime to go down, but we were unsuccessful. Part of the issue was that it may have been one state that caused most of the crime and they themselves had something happen in 2015 that caused the entire northeast region to dip. But because we put states into regions, it would be tedious, for the amount of time we have to finish this project, to calculate the crime for every state in it's own graph and then see if one or a few states caused the dip. So we believe that if we had more time, we would eventually be able to figure this out to add more concrete observations/knowledge to the discussions of our graphs.

• We could test more of the literature on mass incarceration that revolves around drug incarcerations if we could find a comparable data set to combine with ours. The "War on Drugs" is viewed as one of the drivers of mass incarceration in America but data on incarcerations for drug charges are not included in this data set. There is also a relationship between drug incarcerations and violent crimes that would be interesting to investigate further.

• Additionally, having access to demographic data among the prison populations in the various states would allow us to investigate the racial and socioeconomic disparities of mass incarceration. This would be important to understand the implications of the population data especially if the data we found was in line with most literature that states with "one out of every three Black boys born today can expect to be sentenced to prison, compared 1 out 6 Latino boys; one out of 17 white boys" (NAACP).

• Considering that this is true population data instead of a sample population, it was not possible to perform hypothesis testing or bootstrap sampling. It could be interesting to find individual statistics of each conviction over a period of time to provide more context to our base rates. However, there is a benefit to having the true population data because our results will be reliable. In addition, our use of base rates mean we don't really have a clear picture of the individuals and specific cases that are impacted by the issue.

• The data set also has a few discrepancies in how the prison populations are reported. Some states included jails in their incarcerated population data while others only included prisons. This information is provided in the data under the "include_jails" column but we did not factor this into our analysis as there is no way to separate the jail and prison populations in the data set. Additionally, throughout the data set there are four changes in how crime is reported in Alabama and Montana in 2011, New York in 2012, and Georgia in 2015. The actual changes in the methods of reporting are not included in the data set but recognizing that the way crimes were reported would affect the data. Lastly, in Kentucky and Illinois, there are multiple years where the crime totals are estimates rather than exact data. This could also affect our data and our analysis and should be taken into consideration.

• In terms of the linear regression models, it is also important to note that these models might be better represented by another type of regression model such as a logistic one. This was not evaluated in the scope of the project. It is also relevant to note that the extrapolations for predicting future crime are extremely rough and aren't likely to accurately predict further than a few years into the future because previous crime rates are likely less indicative of future crime than a factor such as policy.

References

C., Chris. "Crime and Incarceration in the United States: 21st Century State Crime and Prison Custody Statistics." Kaggle, 2018, www.kaggle.com/ christophercorrea/prisoners-and-crime-in-united-states?select=crime_and_ incarceration_by_state.csv.

Chettiar, Inimai. Why Mass Incarceration Really Is the New Jim Crow. American Civil Liberties Union, 8 Mar. 2012, www.aclu.org/blog/racial- justice/why-mass-incarceration-really-new-jim-crow.

"Frequently Asked Questions about the Change in the UCR Definition of Rape." Federal Bureau of Investigations , 11 Dec. 2014.

"Highest to Lowest - Prison Population Total." World Prison Brief, Institute for Crime and Justice Policy Research, www.prisonstudies.org/ highest-to-lowest/prison-population-total?field_region_taxonomy_tid=All.

Kang-Brown, Jacob, et al. People in Prison in 2019. Vera Institute for Justice, May 2020, www.vera.org/publications/people-in-prison-in-2019.

"Mass Incarceration." American Civil Liberties Union, www.aclu.org/issues/ smart-justice/mass-incarceration.

"Prison Population Counts." Bureau of Justice Statistics (BJS), Office of Justice Programs, 2019, www.bjs.gov/index.cfm?ty=tp.

US Census Bureau. "US Population By Zip Code for Both 2000 and 2010." Kaggle, 2017, www.kaggle.com/census/us-population-by-zip-code.