Report

December 10, 2019

```
In [156]: import pandas as pd
          import numpy as np
          import scipy.linalg as la
          import scipy.stats as stats
          import statsmodels.api as sm
          from sklearn import linear_model, model_selection, metrics
          import sklearn
          import plotly.graph_objs as go
          import matplotlib.pyplot as plt
          import pprint
```

1 Introduction

The united states criminal justice system is large complicated machine that seeks to deliver justice when an offense has been committed. This system has been slowly evolving as our society and culture have changed. Many things that Americans take as natural in our criminal justice systems are quite abnormal among justice systems worldwide. Since the 1990s, America has seen a drastic increase in the incarcerated population. Many Americans believe that this drastic increase in incarceration is a result to increased rates of crime, and that this heightened rate is natural and just. To many it is unclear who is most affected by this drastic change in the application of justice in America. It is also unclear how they are so affected.

There is a lot of existing research exploring incarceration and the criminal justice system. There is a consensus that America incarcerates a larger proportion of its population than any other nation and that people of color are disproportionately affected by this high incarceration rate.

In this project I am interested in understanding more about the criminal justice system and the ways in which the law is being applied differently to people in America. I will explore different ways to quantify claims about mass incarceration and race. I will also be examining things that factor into sentence lengtjh. The factors I will be examining are: offense, age, race, gender, and admission date. There are many things that contribute to sentence length, however the scope of this project is limited to these factors.

2 Data

2.1 Source and Credibility

The data that I will be using in this analysis is gathered from primarily two sources. The first is the Bureau of Justice Statistics and the second is a link to a database hosted on Data.gov and

maintained by the State of Connecticut Department of Corrections. These are highly credible sources because they are primary sources for the data. These organizations are official government agencies which collect, maintain, and report on this data.

2.2 Gathering and Cleaning

All the data which I am using in this report are freely available to the public. Collection and cleaning was relatively simple as the source data was well maintained. The data that I collected from the Bureau of Justice Statistics (BJS) need to be formatted in a way that is easily read by the Python packages I will be using. This data was prepared in .xlsx files as to be easily human readable, however this is not generally easily ingested by programs. I extracted data that I found to be relevant into separate .csv files and kept the original files for reference. The files are

```
incarceration_counts.csv
jail_population.csv
jail_trends.csv
state_jail_data.csv
incarceration_by_race.csv
crime_data.csv.
```

The file that I obtained from the Connecticut Department of Corrections is a very well maintained database. The largest issue I had with this file was mild inconsistency with the way in which certain data was encoded (ex. race was encoded as both `WHITE` and `WHITE\t`). This was the data that I spent the most time working to engineer as it is the data set that I intend to use for different regression-related analyses. The files are

```
individuals.csv
regression_df.csv.
```

2.3 Contents

2.3.1 Bureau of Justice Statistics Data

First we will examine `incarceratio_trends.csv`. This data set records total jail and prison populations across the United States over time. This is useful in understanding general trends in the U.S. over time.

```
In [132]: print(incar[['Year','State prisons','Population']].sample(3))
```

```
    Year  State prisons  Population
50  1975         216462      215.97
12  1937         137432      128.82
28  1953         154216      160.18
```

Next is `jail_trends.csv` which has more information about the breakdown of the populations of United States prisons and jails. However the data is more infrequent that that of `incarceration_trends.csv`.

```
In [143]: cols = ['Pre-trial (unadjusted)','Convicted (unadjusted)']
          print(pop[cols].sample(3))
```

```
   Pre-trial (unadjusted)  Convicted (unadjusted)
5                494200.0                291200.0
6                453200.0                278000.0
1                175669.0                166224.0
```

`jail_trends.csv` contains data among the states collected in 2013, comparing different incarceration rate information. This data gives a general breakdown of why different people are being held at a state level

```
In [142]: cols = [
              'Jail growth (1983-2013)',
              'Percent pre-trial (2013)'
          ]
          print(trend[cols].sample(3))

         Jail growth (1983-2013)  Percent pre-trial (2013)
Montana                     4.26                      0.46
Texas                       1.84                      0.75
Maine                       1.77                      0.65
```

`state_jail_data.csv` contains race, gender, and age demographic data of state incarcerated populations. This data will help us understand the demographic breakdown of state incarcerated populations

```
In [141]: cols = ['CONFPOP','WHITE','BLACK','ASIAN','JUVMALE','MALE','FEM']
          print(state[cols].sample(3))

    CONFPOP  WHITE  BLACK  ASIAN  JUVMALE   MALE    FEM
4     84030  25818  17562   1705        6  73159  10865
7     65166  33351  27111    170      599  56032   8509
40    67418  24910  19619    147      322  58124   8931
```

`incarceration_by_race.csv` contains race demographic data for incarcerated populations by institution. This will allow us to understand distributions of state incarcerated populations

```
In [9]: print(race.sample(2))

    Unnamed: 0        GEOID  GEOID2 Geography  Total  White  Black  Indian  \
47          47  0400000US51    51.0  Virginia  65240  26216  37518     191
3            3  0400000US04     4.0   Arizona  67767  36160   8246    6723

    Asian   NPI  ...  RATE  White_rate  Black_rate  Indian_rate  Asian_rate  \
47    285    48  ...   815         478        2418          654          65
3    1136  1029  ...  1060         775        3184         2267         643

    NPI_rate  Other_rate  Two_rate  Hisp_rate  White_not_hisp_rate
```

```
47       803          261       137        482                  466
3       8136         1690       732       1453                  633

[2 rows x 34 columns]
```

`crime_data.csv` records crime rates over time in this U.S. This data set will help us understand how crime relates to incarceration.

```
In [144]: cols = ['Year','Violent crime','Murder','Rape','Robbery','Assault']
          print(crime[cols].sample(3))
```

```
    Year  Violent crime  Murder  Rape  Robbery  Assault
11  1971          396.0     8.6  20.5    188.0    178.8
43  2003          475.8     5.7  32.3    142.5    295.4
35  1995          684.5     8.2  37.1    220.9    418.3
```

2.3.2 Connecticut Department of Corrections Data

This data set contains individual information for 7.77 million people that have been processed by the justice system and recorded by the Connecticut Department of Corrections.

```
In [11]: inmates = pd.read_csv('individuals.csv')
```

```
In [154]: cols = ['LATEST ADMISSION DATE','AGE','RACE','SENTENCE DAYS']
          print(inmates[cols].sample(3))
```

```
        LATEST ADMISSION DATE  AGE   RACE  SENTENCE DAYS
3597359            06/28/2018   34  WHITE            365
16475              01/29/2016   26  BLACK           2557
2282618            07/19/2016   52  BLACK            914
```

Here is a sample from the altered dataframe that has race, gender, and offence one-hot encoded.

3 Analysis and Visualization

3.1 Increased rates

Here we can see the drastic increase of the amount of incarceration in the U.S. Something that is interesting to note is that Jails are defined as places for people who have a sentence less than 1 year, or who are awaiting trial. So we see that at its peak in 2008, there were more people awaiting trial than there were being held in federal prison in 1991.

Here I have incarceration rates plotted against the violent crime rate in the United States. Near the beginning of the crime rate data we might assume some amount of inaccuracy, since the crime rate seems to be less than the incarceration rate, however there is an indisputable spike in crime rates in the 1980s and 1990s. Something to note is that the incarceration rate seems to lag behind

about 20 years. Another thing to note is that there has been a strict decrease in violent crime (and in all crime) since the 90s, however we do not see the same decrease in the incarceration rate.

Some would argue that we see this decrease in crime because of the increase of incarceration. I would disagree. Consider the scales of the different curves we see on the chart. Both are in terms of the rate per 100,000, but the crime rate is more that 200 times higher than the incarceration rate in state prisons. Unless 0.5% of people who committed crime in the 90s were committing more than half of all crime, I would argue that there is some other cause for the decrease in crime rates.

3.2 Artifacts of Prison Policy

One might expect sentence lengths to be distributed somewhat smoothly. However there are standard sentence lengths and mandatory minimum sentences that influence the distribution of sentence lenghts, making the distribution not smooth.

3.3 Racial Disparities

Here we will explore how the criminal justice system affects people of different races.

3.3.1 Sentence Length

The first chart that we will explore is the distribution of sentence lengths among people of different races. Here we are using the data of more the 7.7 million individuals processed by the criminal justice system. The medians are compariable on the scale at which sentence length is given. Asians and American Indians have the longest median sentence length, however the standard deviation in their sentence lengths is much less than what is seen in the other racial categories. I would attribute this to there being many more Hispanic, Black, and White people that receive extreme sentence lenghts and this is seen in the max sentence lengths.

This boxplot is where we begin to see disparity in the way people of different races are sentenced. While the first two quartiles of each racial group's sentence lengths are similar, we see that the top two quartiles vary widely.

The following histograms will assist us in understanding these distributions.

While it may appear that these sentences are distributed fairly equivilently, we can examine the kurtosis of the distribution to understand how much of the weight of the distribution is found in the extremities. Groups with high kurtosis have a higher probability of receiving a sentence that is far from the mean.

As we can see, Blacks and Hispanics have the greatest kurtosis. We can expect the very small kurtosis in American Indians and Asians because they had such a small standard deviation.

3.4 Age Discrepency

Here we can see that there is slight bias against young people, as they seem to get slightly longer sentences. Mostly we see that young people are more likely to be arrested and convicted of a crime than older people are.

3.5 Standard Sentence Length

Something interesting that emerges from this hexbin plot is a depiction of standard sentence lengths. These became more common as the idea of mandatory minimum sentences has become

more popular. This is seen in the plot as straight bright horizontal lines. Sentences are given at these standard lengths, often times because it is not legal to give a shorter sentence.