Report

December 10, 2019

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 length. 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

Here I will describe generally each data file and its contents, as well as give a small sample from each file

```
In [3]: incar = pd.read_csv('incarceration_trends.csv')
        pop = pd.read_csv('jail_population.csv')
        trend = pd.read_csv('jail_trends.csv')
        state = pd.read_csv('state_jail_data.csv')
        race = pd.read_csv('incarceration_by_race.csv')
        crime = pd.read_csv('crime_data.csv')
```

First we will examine `incarceration_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 [4]: print(incar[['Year','State prisons','Population']].sample(3))
```

```
    Year  State prisons  Population
14  1939         160088      130.88
62  1987         521289      242.29
38  1963         194155      189.24
```

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

```
   Pre-trial (unadjusted)  Convicted (unadjusted)
5                494200.0                291200.0
1                175669.0                166224.0
4                414800.0                269900.0
```

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

```
                Jail growth (1983-2013)  Percent pre-trial (2013)
Minnesota                          1.78                      0.60
Alabama                            2.21                      0.68
South Carolina                     3.11                      0.78
```

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

```
    CONFPOP  WHITE  BLACK  ASIAN  JUVMALE   MALE   FEM
39    24233  12271  10845     16       67  20692  3471
30    33341   9924  15926    275     1316  28991  2900
34     9585   5173   2321     57       47   8195  1341
```

`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 [8]: cols = ['Geography','Total','White','Black','White_rate','Black_rate']
        print(race[cols].sample(2))
```

```
        Geography  Total  White  Black  White_rate  Black_rate
50      Wisconsin  38102  21195  14518         432        4042
30  New Hampshire   4851   4288    337         347        2241
```

`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 [9]: 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
18  1978          497.8     9.0  31.0    195.8    262.1
26  1986          620.1     8.6  38.1    226.0    347.4
```

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. Each individual is recorded along with their age, gender, race, offense, and sentence length, among other things.

I also created a one-hot encoded this data set in order to run regressions on the data. Because of the size of the data, the regression data sets are only random subsets of the larger data set.

```
In [19]: inmates = pd.read_csv('individuals.csv')
         regr_df = pd.read_csv('small_regression_df.csv')
         race_regr_df = pd.read_csv('race_regression_df.csv')

In [11]: cols = ['LATEST ADMISSION DATE','AGE','RACE','SENTENCE DAYS']
         print(inmates[cols].sample(3))

        LATEST ADMISSION DATE  AGE      RACE  SENTENCE DAYS
1272156            02/26/1991   56  HISPANIC          21915
762893             06/09/2015   34     ASIAN           2557
4322986            06/27/2013   34  HISPANIC           2192
```

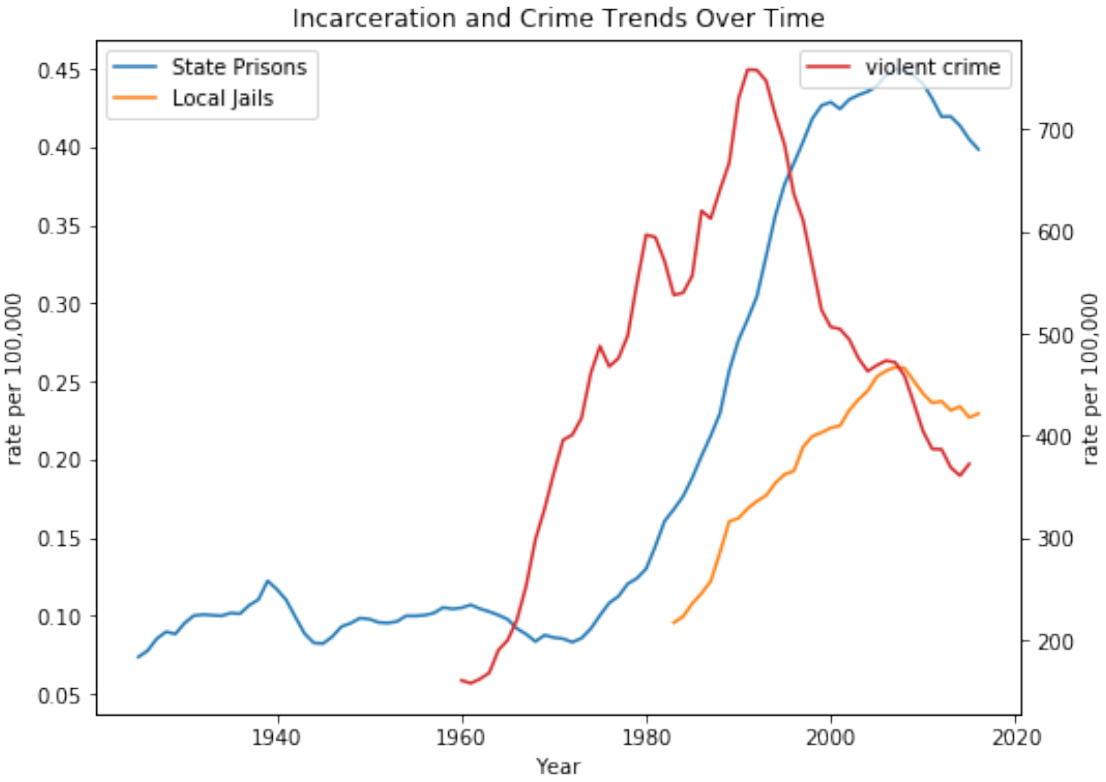
3 Analysis and Visualization

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.

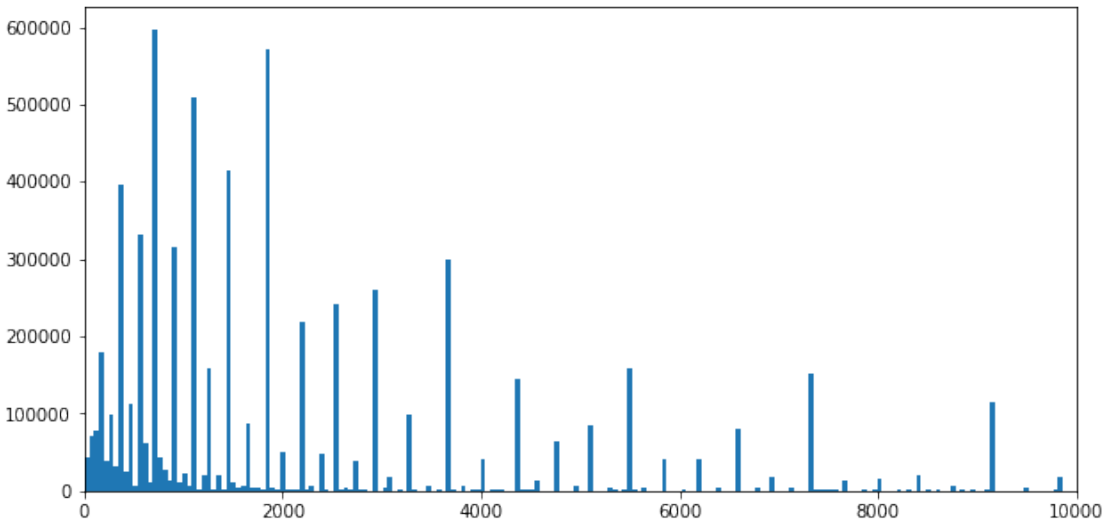
Here I have incarceration rates plotted against the violent crime rate in the United States. Near the beginning of the crime rate data we might assume some amount of inaccuracy, since the crime rate seems to be less than the incarceration rate, however there is an indisputable spike in crime rates in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

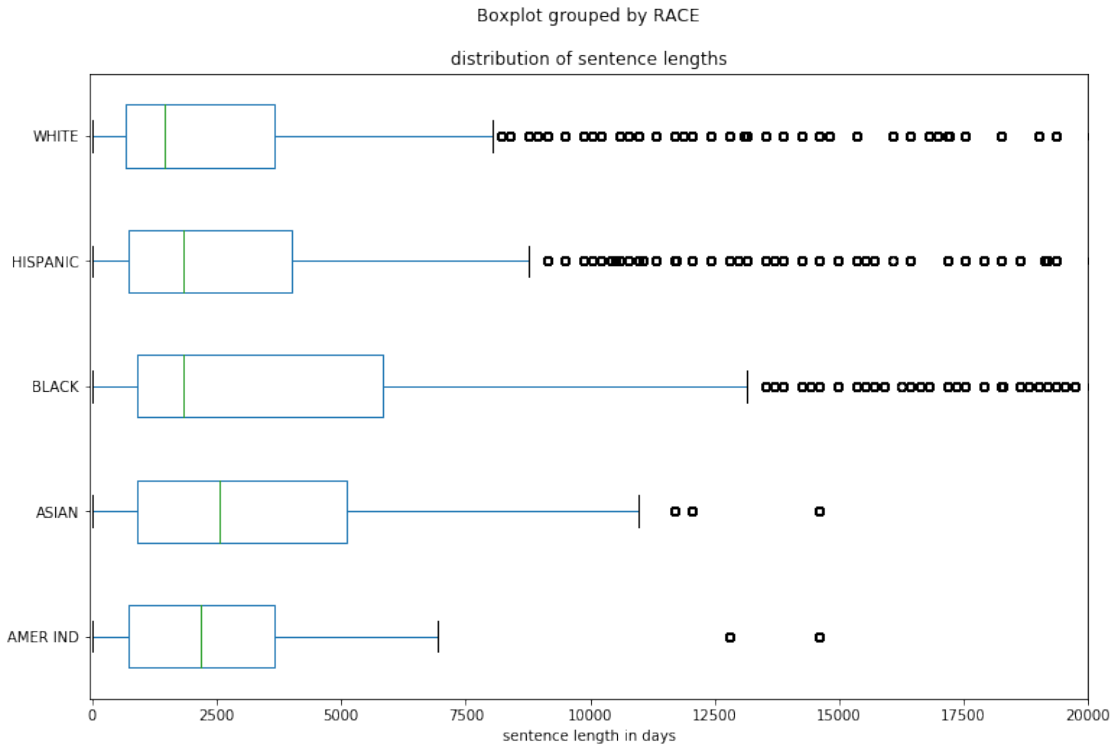
3.3.1 Sentence Length

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.

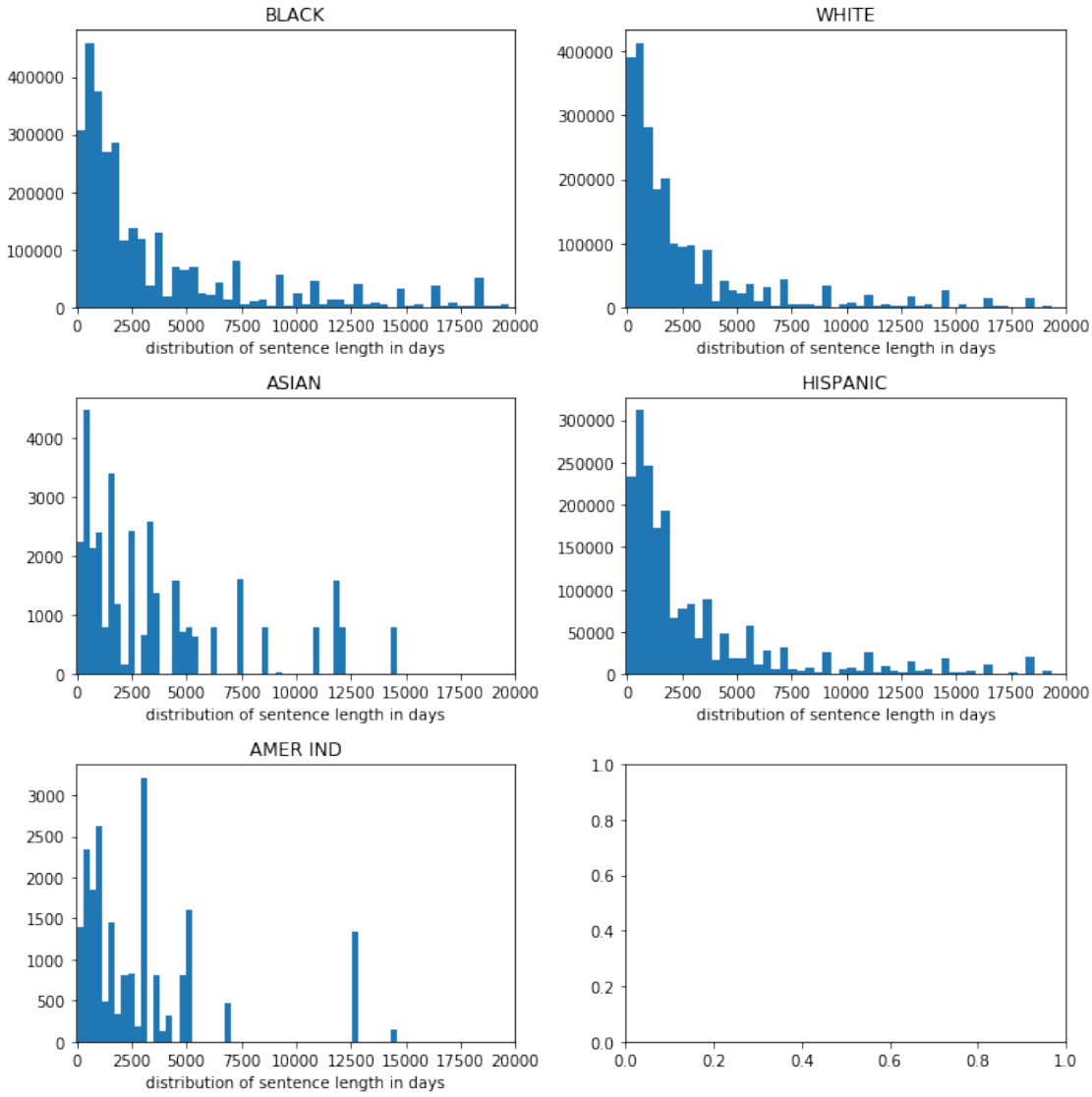
```
Median sentence lengths:
 RACE
AMER IND    2192
ASIAN       2557
BLACK       1826
HISPANIC    1826
WHITE       1461
Name: SENTENCE DAYS, dtype: int64


Standard deviation of sentence lengths:
 RACE
AMER IND     3208.584729
ASIAN        5122.011265
BLACK       37306.146883
HISPANIC    35763.282330
WHITE       46882.705552
Name: SENTENCE DAYS, dtype: float64


Max sentence lengths:
 RACE
AMER IND     14610
ASIAN        27394
BLACK       368897
HISPANIC    368897
WHITE       368897
Name: SENTENCE DAYS, dtype: int64
```

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.

```
White Kurtosis: 54.01344161453346
Black Kurtosis: 86.11479232153843
Hispanic Kurtosis: 96.03075354970996
American Indian Kurtosis: 4.0085627062942955
Asian Kurtosis: 7.45297417968173
```

In [20]:
```
race_regr_df.dropna(inplace=True)
sentence = race_regr_df['SENTENCE DAYS']
race_regr_df.drop(
    ['SENTENCE DAYS','Unnamed: 0'], axis=1, inplace=True
)
X_train, X_test, y_train, y_test = model_selection.train_test_split(
    race_regr_df.astype(float), sentence, test_size=.3
)
res_gen = sm.OLS(y_train, X_train).fit()
print(res_gen.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SENTENCE DAYS   R-squared:                       0.036
Model:                            OLS   Adj. R-squared:                  0.036
Method:                 Least Squares   F-statistic:                     6017.
Date:                Tue, 10 Dec 2019   Prob (F-statistic):               0.00
Time:                        22:03:58   Log-Likelihood:            -9.7052e+06
No. Observations:              809642   AIC:                         1.941e+07
Df Residuals:                  809636   BIC:                         1.941e+07
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
AGE          642.6980      3.731    172.241      0.000     635.385     650.011
ASIAN       1947.5280   1061.851      1.834      0.067    -133.665    4028.721
BLACK       7312.0117    852.800      8.574      0.000    5640.551    8983.472
HISPANIC    5578.0856    854.307      6.529      0.000    3903.671    7252.500
WHITE       4787.5315    853.636      5.608      0.000    3114.434    6460.629
0          -2.191e+04    862.310    -25.403      0.000   -2.36e+04   -2.02e+04
==============================================================================
Omnibus:                  1125746.686   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        193819619.671
Skew:                           8.505   Prob(JB):                         0.00
Kurtosis:                      76.865   Cond. No.                     1.75e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.75e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [ ]: