Report

December 2, 2019

```
In [1]: import pandas as pd
        import numpy as np
        import scipy.linalg as la
        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
```

```
In [21]: import warnings
         warnings.filterwarnings('ignore')
```

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 a natural in our criminal justice systems are quite abnormal among justice systems world wide. Since the 1990s we America has seen a drastic increase in the incarcerated population. Many Americans believe, as some politicians would have it, that this drastic increase in incarceration is a reaction 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.

In this project I am interested in understanding more about the criminal justice system and the ways in which the law is being applied to different groups in America. With this particular data set I will be examining the effects of different factors on 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 of 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 the 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.
```

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. This was the data that I spend the most time working to engineer as it is the data set that I intend to use for different regression-related analyses. This file is

```
Sentenced_Inmates_in_Correctional_Facilities.csv.
```

```
In [2]: inmates = pd.read_csv('Sentenced_Inmates_in_Correctional_Facilities.csv')
```

Here are the columns of the data obtained from the Connecticut Department of Corrections

```
In [3]: pp = pprint.PrettyPrinter(indent=4, width=80)
        pp.pprint(inmates.columns.values)
```

```
array(['DOWNLOAD DATE', 'IDENTIFIER', 'LATEST ADMISSION DATE', 'RACE',
       'GENDER', 'AGE', 'END SENTENCE DATE', 'OFFENSE', 'FACILITY',
       'DETAINER', 'SENTENCE DAYS',
       'SPECIAL PAROLE END DATE                                                    '],
      dtype=object)
```

I also created an engineered version that has each of the offences one hot encoded

```
In [4]: # reg_df = pd.read_csv('regression_df.csv')
        # pp.pprint(reg_df.columns.values)
```

These are data sets that I collected from the BJS. I'll show some of the columns of these data sets

```
In [5]: 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')
```

```
In [6]: pp.pprint(incar.columns.values)
        print('\n')
        pp.pprint(pop.columns.values)
        print('\n')
        pp.pprint(trend.columns.values)
        print('\n')
        pp.pprint(state.columns.values)
```

```
array(['Unnamed: 0', 'Year', 'State prisons', 'Federal prisons',
       'Local jails'], dtype=object)


array(['Unnamed: 0', 'Pre-trial (unadjusted)', 'Convicted (unadjusted)',
       'Held for state prisons', 'Held for immigration authorities',
       'Held for Bureau of Prisons or U.S. Marshals Service',
       'Total held for other authorities', 'Pre-trial (adjusted)',
       'Convicted (adjusted)', 'year'], dtype=object)


array(['State  Jail incarceration rate (2013)', 'Jail growth (1983-2013)',
       'Percent pre-trial (2013)',
       'Percent held for all state and federal authorities (2013)',
       'Percent held for state prisons (2013)',
       'Percent held for immigration authorities (2013)',
       'Percent held for U.S. Marshals Service (2013)',
       'Percent held for other agencies (2013)  ', 'Unnamed: 8'],
      dtype=object)


array(['Unnamed: 0', 'CONFPOP', 'MALE', 'MALE_PERC', 'JUVMALE',
       'JUVMALE_PERC', 'FEM', 'FEM_PERC', 'JUVFEM', 'JUVFEM_PERC',
       'WHITE', 'WHITE_PERC', 'BLACK', 'BLACK_PERC', 'HISP', 'HISP_PERC',
       'ASIAN', 'ASIAN_PERC', 'ICE', 'ICE_PERC', 'BIA'], dtype=object)
```

```
In [66]: incar['State prisons'].values
         # # incar.rename(columns={'Unnamed: 0.1':'Year'}, inplace=True)
         # # incar.to_csv('incarceration_trends.csv')
         # # incar.drop('Unnamed: 0', axis=1, inplace=True)
         # incar.to_csv('incarceration_trends.csv')
         incar
```

Out[66]:

	Unnamed: 0	Year	State prisons	Federal prisons	Local jails
0	0	1925	85239.0	6,430	NaN

```
1            1  1926        91188.0          6803.0          NaN
2            2  1927       101624.0          7722.0          NaN
3            3  1928       108157.0          8233.0          NaN
4            4  1929       107532.0         12964.0          NaN
..         ...   ...            ...             ...          ...
87          87  2012      1315856.0          196574     744524.0
88          88  2013      1325305.0          195098     731208.0
89          89  2014      1316407.0          191374     744592.0
90          90  2015      1298159.0          178688     727400.0
91          91  2016      1286691.0          171482     740700.0

[92 rows x 5 columns]
```

3 Visualization

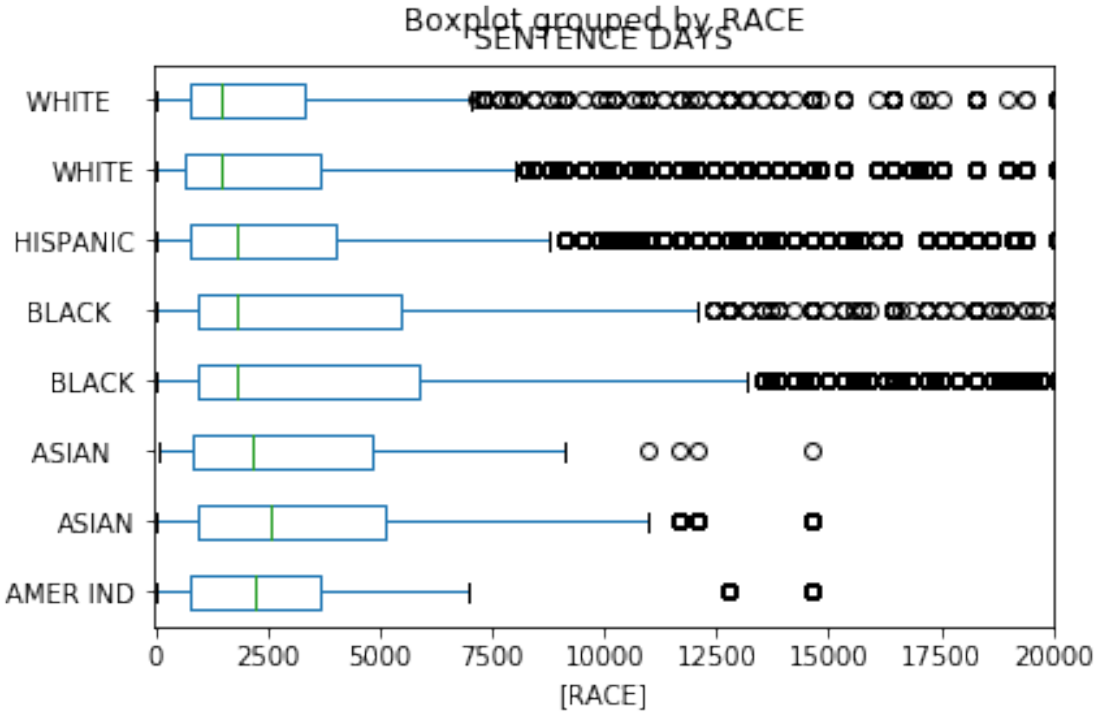
```
In [27]: fig = go.Figure(
             layout={'title':{'text':'incarceration trends over time'},
                     'showlegend':True}
         )
         fig.add_trace(
             go.Scatter(
                 x=incar['Year'].values, y=incar['State prisons'].values,
                 name='State Prisons'
             )
         )
         fig.add_trace(
             go.Scatter(
                 x=incar['Year'].values, y=incar['Federal prisons'].values,
                 name='Federal Prisons'
             )
         )
         fig.add_trace(
             go.Scatter(
                 x=incar['Year'].values, y=incar['Local jails'].values,
                 name='Local Jails'
             )
         )
         fig.show()
```

3.1 Increased rates

Here we can see the drastic increase of the rate and 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.

```
In [87]: fig, ax = plt.subplots(1)
         ax.set_xlim(-50,20000)
```

```
bp = inmates.boxplot(
    ['SENTENCE DAYS'], by=['RACE'],
    vert=False, grid=False,
    ax=ax
)
plt.show()
```

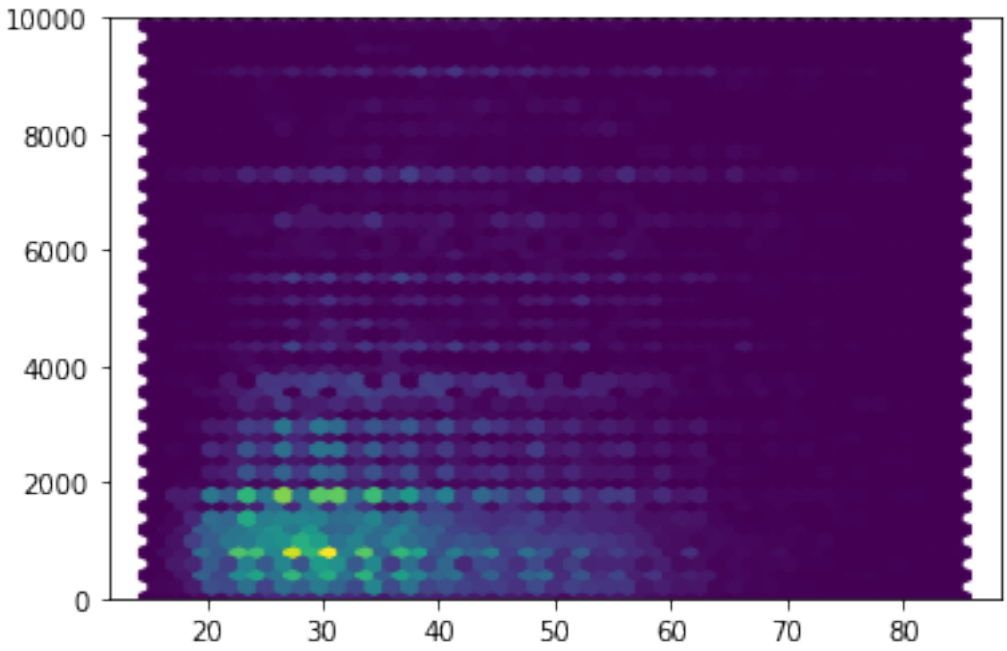


3.2 Racial Disparity

Here we can see some preliminary findings related to racial disparities in the criminal justice system. Though we do see comparable median sentence lengths, the variance in sentence length among black inmates is much higher and the spread of the top two quintiles is much wider.

```
In [26]: # fig = go.Figure(
         #     layout={'title':{'text':'Comparing Age and Sentence Lenght'},
         #             'showlegend':False}
         # )
         # fig.add_trace(
         #     go.Scatter(
         #         x=inmates['AGE'].values, y=inmates['SENTENCE DAYS'].values
         #     )
         # )
         # fig.show()
         fig, ax = plt.subplots(1)
```

```
ax.set_ylim(0,10000)
mask = inmates['SENTENCE DAYS'] < 10000
ax.hexbin(
    x=inmates['AGE'][mask].values,
    y=inmates['SENTENCE DAYS'][mask].values,
    gridsize=45
)
plt.show()
```



3.3 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.

3.4 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.

In [ ]: