represent

November 15, 2019

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
```

All of the data that I am dealing with is publically souced from the Bureau of Justice Statistics, some of it has been compiled by different non-profits that are concered with incarceration, such as the Prison Policy Initiative and the Sentencing Project.

Something that could make this data unreliable are false reports from different county, state, and federal offices. There are many data sets that I have collected that have missing fields, this would make doing any sort of regression difficult this this data. It's unclear to me right now how I want to deal this these missing data points, but these are solvable problems.

Something that I have found is that there is a lot of data about incarceration in the United States. It's very easy to simply download a folder that has over 15 .csv files in it, each with specialized data. In that way, it can be hard to know what questions to ask and how well the data can answer. I think that as I'm cleaning more data and doing visualizations to understand it, I'll get a better idea of exactly what I want to ask.

Here are some samples of the data that I have collected and cleaned.

```
In [9]: df = pd.read_csv('jail_population.csv')
        # df.drop(['Unnamed: 0', 'Unnamed: 0.1'], axis=1, inplace=True)
        df
```

Out[9]:

	Unnamed: 0	Pre-trial (unadjusted)	Convicted (unadjusted)	\
0	0	113984.0	107660.0	
1	1	175669.0	166224.0	
2	2	228900.0	226600.0	
3	3	331800.0	252600.0	
4	4	414800.0	269900.0	
5	5	494200.0	291200.0	
6	6	453200.0	278000.0	

	Held for state prisons	Held for immigration authorities	\
0	9134.0	1304.0	
1	14314.0	1954.0	
2	50966.0	3763.0	
3	24925.0	8544.0	
4	73440.0	13337.0	
5	83497.0	20785.0	

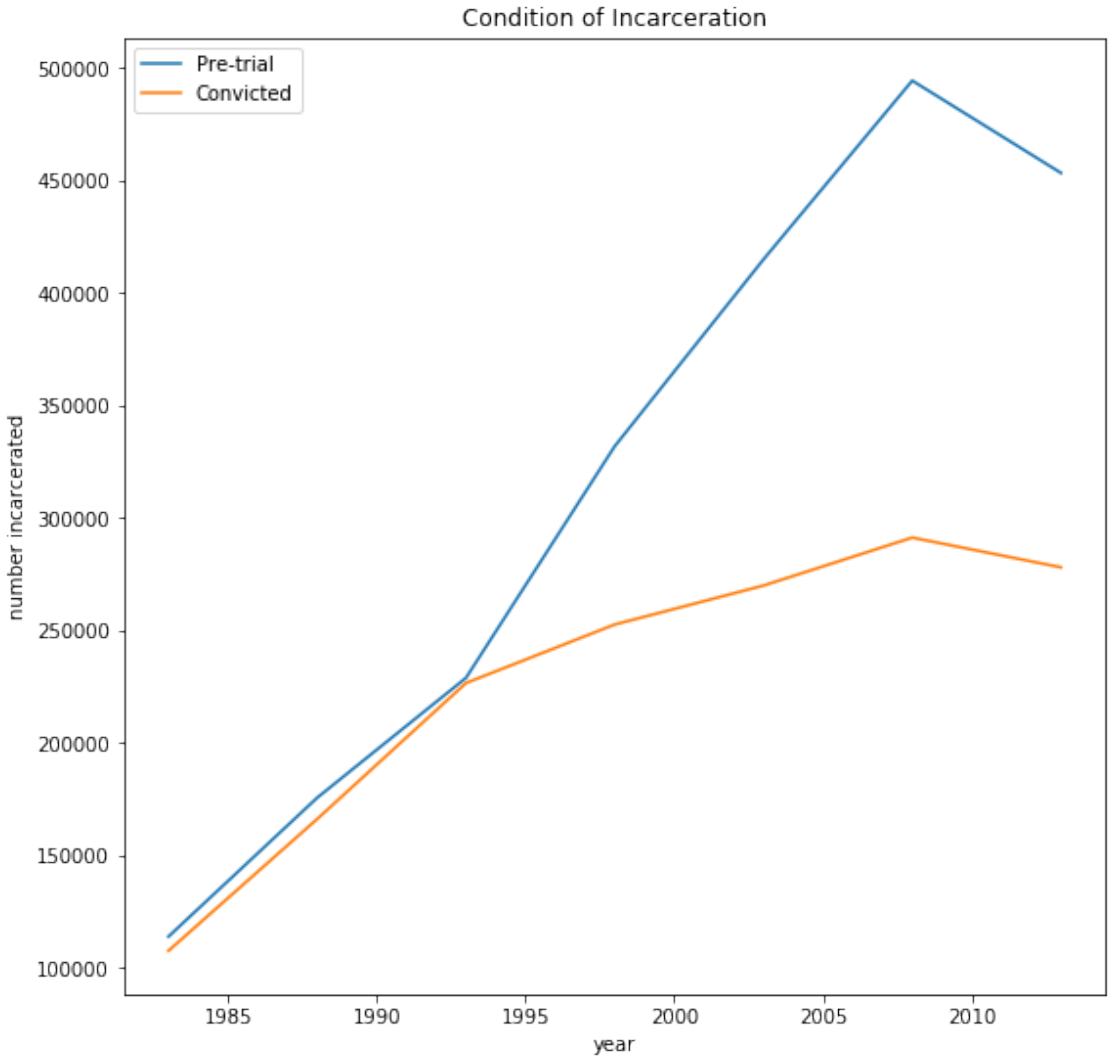
```
6                 85662.0                          17241.0

   Held for Bureau of Prisons or U.S. Marshals Service  \
0                                             2044.0
1                                             6302.0
2                                            11833.0
3                                            16690.0
4                                            25522.0
5                                            32761.0
6                                            31745.0

   Total held for other authorities  Pre-trial (adjusted)  \
0                           12482.0              112632.0
1                           22570.0              172863.0
2                           66562.0              223568.0
3                           50159.0              322147.0
4                          112299.0              399860.0
5                          137043.0              472607.0
6                          134648.0              434188.0

   Convicted (adjusted)  year
0               96530.0  1983
1              146460.0  1988
2              165370.0  1993
3              212094.0  1998
4              172541.0  2003
5              175750.0  2008
6              162364.0  2013
```

In [19]:
```
plt.figure(figsize=(9, 9))
plt.plot(df.year,df['Pre-trial (unadjusted)'], label='Pre-trial')
plt.plot(df.year,df['Convicted (unadjusted)'], label='Convicted')
plt.legend()
plt.ylabel('number incarcerated')
plt.xlabel('year')
plt.title('Condition of Incarceration')
plt.show()
```




```
In [21]: df = pd.read_csv('incarceration_by_race.csv')
         df.columns

Out[21]: Index(['Unnamed: 0', 'GEOID', 'GEOID2', 'Geography',
                'Total : In Correctional Facilities for Adults',
                'White alone : in Correctional Facilities for Adults',
                'Black or African American alone : in Correctional Facilities for Adults',
                'American Indian and Alaska Native alone : in Correctional Facilities for Adult
                'Asian alone : in Correctional Facilities for Adults',
                'Native Hawaiian and other Pacific Islander alone : in Correctional Facilities
                'Some other race alone : in Correctional Facilities for Adults',
                'Two or more races : in Correctional Facilities for Adults',
                'Hispanic or Latino : in Correctional Facilities for Adults',
                'White alone, not Hispanic or Latino : in Correctional Facilities for Adults',
                'Total Population', 'Total Population: White alone',
```

```
       'Total Population: Black or African American alone',
       'Total Population: American Indian and Alaska Native alone',
       'Total Population: Asian alone',
       'Total Population: Native Hawaiian and other Pacific Islander alone',
       'Total Population: Some other race alone',
       'Total Population: Two or more races',
       'Total Population: Hispanic or Latino',
       'Total Population: White alone, not Hispanic or Latino',
       'Incarceration rate', 'Incarceration rate: White alone',
       'Incarceration rate: Black or African American alone',
       'Incarceration rate: American Indian and Alaska Native alone',
       'Incarceration rate: Asian alone',
       'Incarceration rate: Native Hawaiian and other Pacific Islander alone',
       'Incarceration rate: Some other race alone',
       'Incarceration rate: Two or more races',
       'Incarceration rate: Hispanic or Latino',
       'Incarceration rate: White alone, not Hispanic or Latino'],
      dtype='object')
```

```
In [26]: df.boxplot(
             column=[
                 'Incarceration rate: White alone, not Hispanic or Latino',
                 'Incarceration rate: Black or African American alone'
             ], grid=False, vert=False,
         )
         plt.xlabel('incarcerated per 100,000')
         plt.show()
```