Project 1: Risk Assessment Score Analysis

Zora Williams
CS 102: Big Data - Tools and Techniques

Overview:

This analysis will examine data from the commercial pretrial risk assessment tool called COMPAS or the Correctional Offender Management Profiling for Alternative Sanctions developed by the company, Northpointe. COMPAS pretrial risk assessment tool is used to evaluate defendants after arrest to determine their risk of recommitting a crime should they be released. This risk assessment tool is notorious for its black box algorithmic mechanism. The dataset used for this project is called compass-scores.csv which was generated by ProPublica for their investigation into the bias of the COMPAS risk assessment tool. The dataset contains the following columns:

> Name, first, last, compass screening date, sex, date of birth, age, age category, race, decile score (risk of recidivism) ranging from 1-10, risk of violent recidivism scores ranging 1-10, and whether or not each individual recidivated violently or otherwise.

The data also contain a host of other information about the type of crime committed, when the individual was booked into jail etc. All of the information was pulled from individual cases from Broward County, Florida. It contains 11,757 (rows) of individuals who were arrested and assessed in 2013 and 2014.

Objectives

This project will use the data from the scores to look into three main questions:

1) What is the distribution of risk assessment scores across race?
    a. How does the distribution look for violent risk assessment scores across race?
2) Does the risk assessment score correlate with actual recidivism?
    a. Is this true for violent risk assessment scores that predict violent recidivism?
    b. What is the breakdown of recidivism by race?

3) Are there some crimes that correspond to certain risk scores?

    a. Are felony crimes consistently associated with higher or lower risk scores? What about non-felonies?

    b. Do some crimes correspond to a higher recidivism?

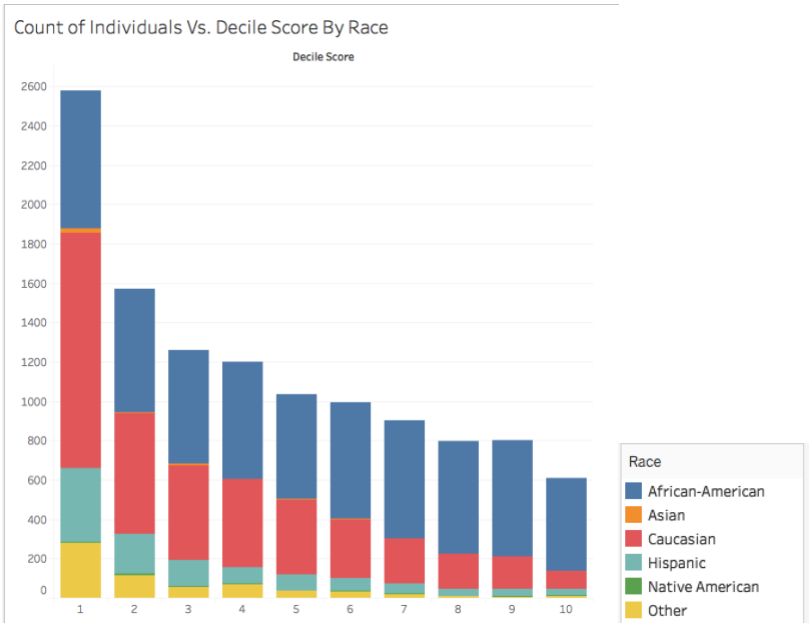
In an effort to answer these questions, this project's analysis focuses mostly on the occurrence of recidivism for each individual and the corresponding scores, crime attributes and race.

Part I : What is the distribution of risk assessment scores across race?

There are two types of scores in this dataset, Risk of Recidivism Score and a Risk of Violent Recidivism Scores. Unless specified as violent, the scores/risk assessment scores in question will be those that predict risk of recidivism. One interesting dimension to investigate is the spread of risk assessment scores across race. In essence, I will try to surface some of the biases revealed in the ProPublica report with less robust statistical analysis. The data used to create the following visualizations were filtered for individuals with missing data.

What does the race distribution look like for each score category?

Figure 1.1



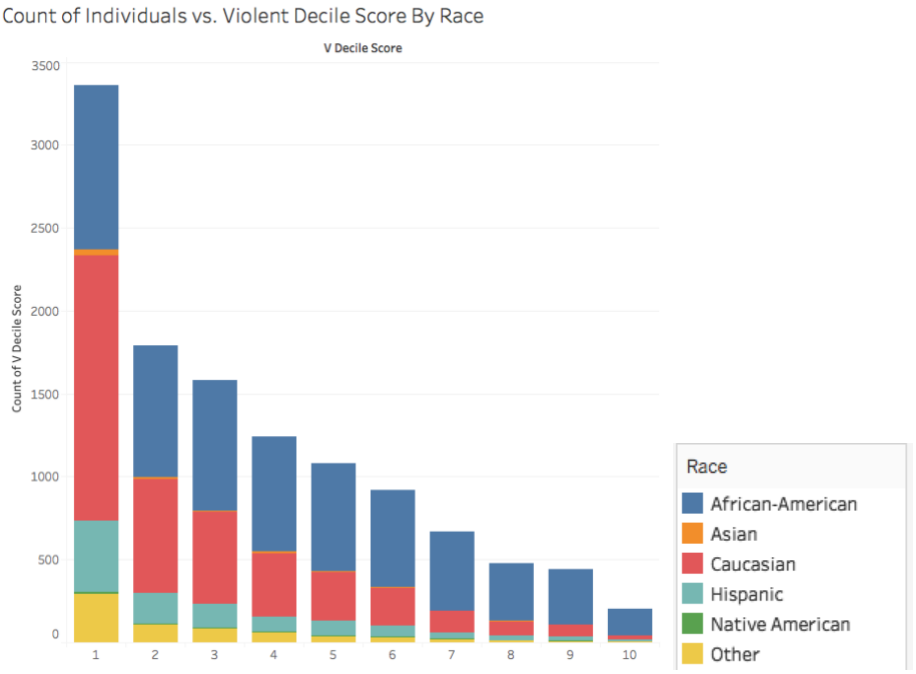
In Figure 1.1 we see the raw counts of all individuals for every single decile score (score for Risk of Recidivism). Generally, we can see that in the lowest score bracket, Caucasians have the highest total and in the highest score bracket they have the smallest total. Whereas African Americans appear to account for the majority of the highest score bracket and a smaller portion (as compared to Caucasians) of the lowest score bracket. In order to normalize the counts to do an adequate comparison of the proportion of race for the lowest and highest score bracket, I calculated the ratios of black and white individuals amongst each score group in SQL. The values are:

Rate_Of_AA_Low	Rate_of_AA_High	Rate_Of_C_Low	Rate_Of_C_High
0.269305393869	0.770491803279	0.462553356616	0.149180327869

From these numbers we see that African Americans are represented half as much in the lowest score bracket (Score of 1) as compared to Caucasians (.26 vs .46). In the high score bracket (Score of 10) African Americans make up 5 times the proportion of Caucasians. There is a clear discrepancy between race and the score categories. I would expect that the delineations between score boundaries to be based on factors of the crime at hand or previous violent behavior. The skew in race indicates that there are some score predictors that could be correlated with race. Different factors could also be associating certain races to certain behaviors, as seen in this case where Blacks are associated with high risk behavior and Caucasians more associated with low risk behavior.

How does the distribution look for violent risk assessment scores across race?

Figure 1.2



The distribution of violent risk assessment scores over race resembles the distribution of the risk of recidivism scores across race. Figure 1.2 shows the raw count of all individuals by race for each score decile. The same trend holds from Figure 1.1 where Caucasians accounts for most of the scores in the lowest category of "1" while African Americans account for most of the scores of "10." Similar to the recidivism scores, we see almost exactly the same difference in the ratios of African Americans and Caucasians who have violent scores of "1" and "10".

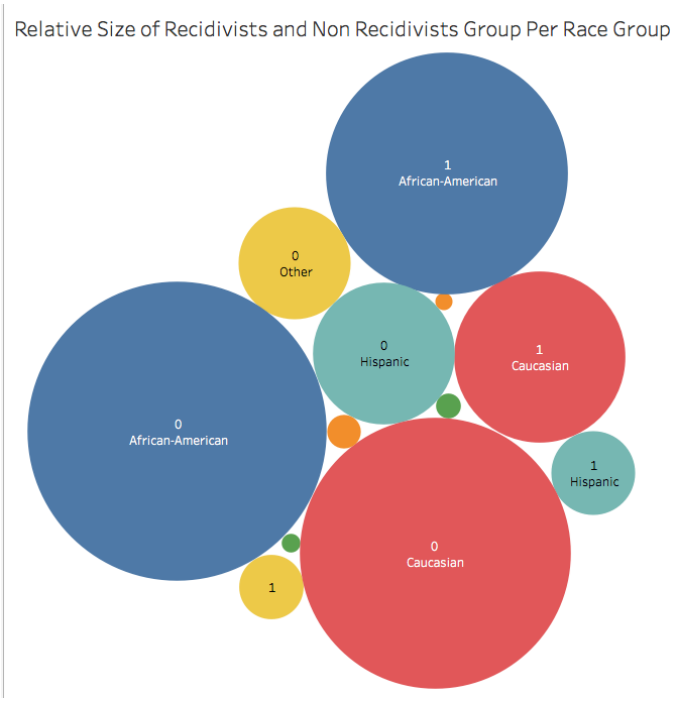
Rate_Of_AA_Low_V	Rate_Of_AA_High_V	Rate_Of_C_Low_V	Rate_Of_C_V_High
0.295028282227	0.78	0.476332241739	0.14

Part II: Does the risk assessment score correlate with actual recidivism?

After looking into the distribution of race across each score category, the efficacy of the tool is also in question. It is important to know if the tool predicts recidivism with some degree of accuracy. In Figure 2.1 we see the breakdown of recidivists and non-recidivists by race. The orange and green bubbles are Asian and Native American respectively. Given the small number of individuals in each group, the label failed to show.

The number of recidivists in the bubble chart mirrors the ratio of each racial group in the dataset. Figure 3.2 (page 11) shows the number of all individuals in each racial group and the corresponding ratio of each racial group in the dataset. In this table, we see that African Americans account for nearly half of the dataset, so it makes sense why they have the largest bubbles and why Caucasian has the second highest bubbles and so on. Yet, it looks as though within groups the distribution of those who do and do not recidivate is evenly spread.

Figure 2.1



The next two visualizations reveal that the algorithm is accurate to a certain degree. Figure 2.2 shows the number of individuals in each score bracket who did (noted by "1" Category) and did not (noted by "0" Category) recidivate. It illustrates that out of the total number of individuals who earned low scores few recidivated given the significant difference in the height of the bars from the non-recidivist and recidivist graphs. People labeled with scores 1-5 generally (low to medium scores) were generally correctly predicted to not recommit a crime. It is worth noting however that African Americans do make up a greater proportion of the higher end scores for both those who do and do not go on to recommit crimes. This observation is reminiscent of ProPublica's statistical conclusion that the false positive rate for African Americans is twice that for Caucasians.

Figure 2.2

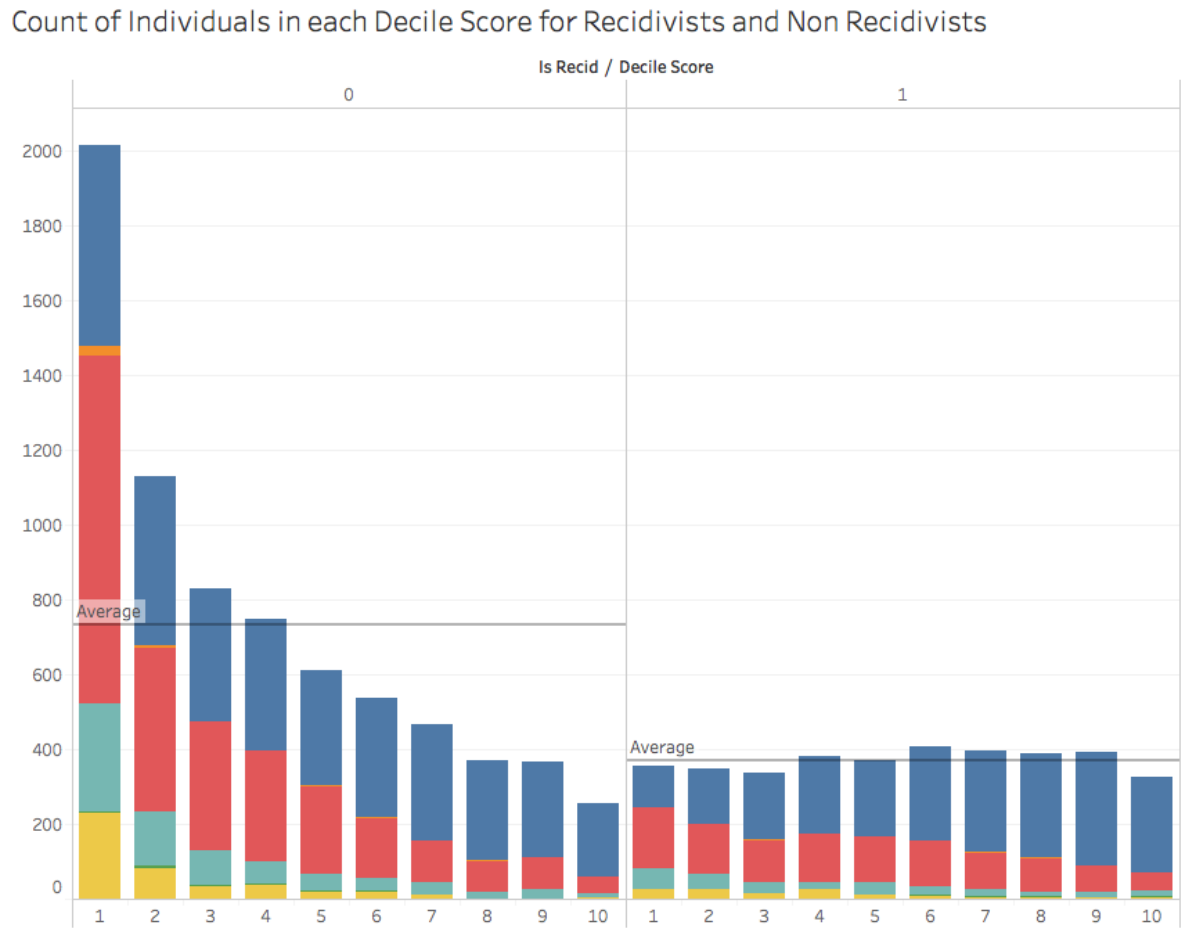
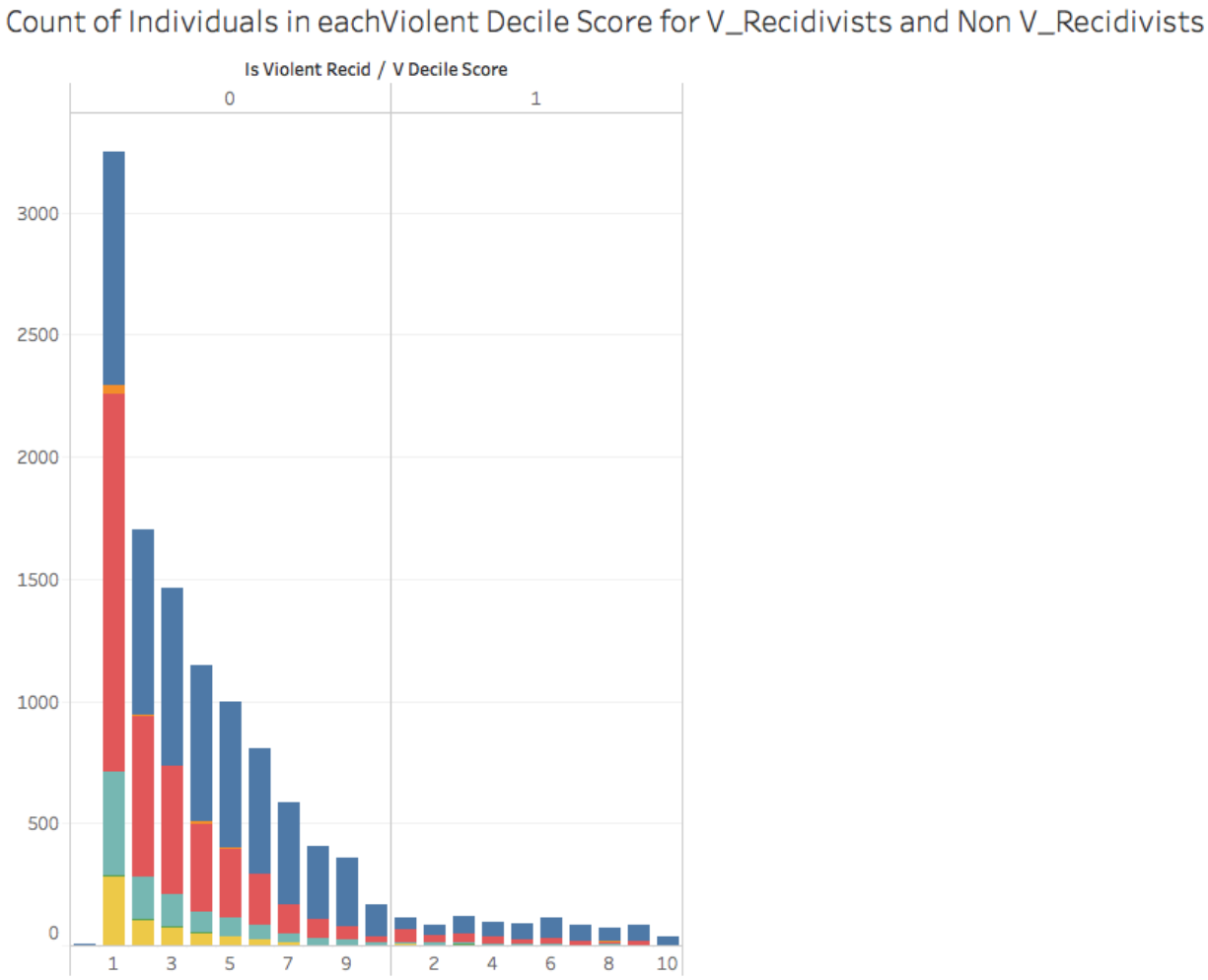


Figure 2.3



In the second graph, Figure 2., we see the same trend of accuracy holds for the violent recidivism scores. The graph shows the number of individuals in each violent risk score category who did ("1") and did not ("0) recommit a violent crime. The non-recidivist graph has a right skew meaning that most people earned lower violent risk assessment scores. Those predictions proved to be accurate to some extend given that the violent recidivism graph is very even and small in comparison. This means that most people who were predicted not to recidivate did not. We even see in this graph that those who were predicted to recidivate at high rates (score of 6

and above) did not recidivate. This leads me to believe that perhaps the false positive rate for the violent risk assessment is a bit higher than that for the risk of recidivism scores.

Part III: How does race map onto the type of crime?

This section will look into the distribution of crime across race and risk assessment scores to determine if there is any correspondence or association between these attributes. The data used in this portion were filtered to exclude those individuals who did not have recidivism scores nor information about whether or not they recidivated not had a label about the description of their crime.

If we first examine the distribution of charges across race we see that African Americans have the most distinct charges and are thus the most represented in this criminal dataset. Out of the 531 distinct charges that individuals were evaluated for, African Americans and Caucasians have been charged for around 60% of all crimes as noted in Figure 3.1. Moreover, in Broward County, African Americans make up 29.9%[1] of the population whilst in this data set they account for nearly half (49.4%) of the crimes followed by Caucasians who make up 34.7% of the data set and yet 63.5% of the Broward County population. All of the other races are represented at a significantly lower rate in both the county and the data set as illustrated in Figure 3.2.

Figure 3.1

Count_of_Charges	race
377	African-American
30	Asian
310	Caucasian
177	Hispanic
20	Native American
124	Other

[1] https://www.census.gov/quickfacts/fact/table/browardcountyflorida/RHI225217#RHI225217

Info: This table shows the count of all the distinct crimes that an individual from each race has been charged with

Figure 3.2

race	Ratio	Count
African-American	0.494428850897	5813
Asian	0.00493323126648	58
Caucasian	0.347452581441	4085
Hispanic	0.0935612826401	1100
Native American	0.00340222845964	40
Other	0.0562218252956	661

Do some crimes correspond to a higher recidivism?

Keeping in mind the distribution of races across the dataset and charges, we can reasonably expect that African Americans will be over represented in the association of crimes to recidivism score. This reveals itself with a closer look at the data. Figure 3.3 shows the number of individuals who actually recidivated (Count_Recid) and initially committed felony crimes and who were predicted to recidivate at high rates (or receiving a score of 8-10). African Americans commit the most felony crimes before recidivating having the most rows in the table which is logical given their overrepresentation in the dataset. However, the number of Caucasians and African Americans who receive a high score after committing felony petit theft seems about equal (9 and 11 individuals respectively). Even though there are more African Americans who commit felonies before recidivating it, the numbers are spread pretty evenly across felony crime type where typically 5 or fewer individuals recidivate after committing any given crime. The distribution looks similar amongst the other races too.

Figure 3.3

score	race	Count_Recid	c_charge_desc
8	African-American	1	Felony Battery
8	African-American	3	Felony Battery w/Prior Convict
8	African-American	5	Felony Driving While Lic Suspd
8	African-American	2	Felony Petit Theft
9	African-American	1	Felony Battery
9	African-American	2	Felony Battery (Dom Strang)
9	African-American	5	Felony Battery w/Prior Convict
9	African-American	1	Felony Driving While Lic Suspd
9	African-American	8	Felony Petit Theft
10	African-American	2	Felony Batt(Great Bodily Harm)
10	African-American	1	Felony Battery
10	African-American	1	Felony Battery (Dom Strang)
10	African-American	2	Felony Battery w/Prior Convict
10	African-American	1	Felony Committing Prostitution
10	African-American	1	Felony Driving While Lic Suspd
10	African-American	1	Felony Petit Theft
8	Asian	1	Felony Petit Theft
8	Caucasian	1	Felony Battery (Dom Strang)
8	Caucasian	1	Felony Battery w/Prior Convict
8	Caucasian	2	Felony Petit Theft
9	Caucasian	1	Felony Battery
9	Caucasian	1	Felony Battery w/Prior Convict
9	Caucasian	1	Felony Committing Prostitution
9	Caucasian	7	Felony Petit Theft
10	Caucasian	1	Felony Battery (Dom Strang)
8	Hispanic	1	Felony Battery (Dom Strang)
8	Hispanic	1	Felony Petit Theft
10	Hispanic	1	Felony Petit Theft

Are felony crimes consistently associated with higher or lower risk scores? Non felonies?

To answer the core question, *do risk scores correspond to certain crimes*, after the analysis it does not seem that there is a strict correspondence. There are individuals who commit felonies initially but score under a low for their risk to recidivate while there are others who commit misdemeanors and can score a 10. Figure 3.4 below illustrates this point. This graph depicts the total number of cases for each score decile (1-10) for felony and misdemeanor cases. Though there are some misdemeanors with high scores, bar chart (on the right) has a right skew showing that most misdemeanor cases earn low risk scores. The felony chart is a bit more evenly distributed with a surprising number of individuals who score a "1" on their risk of recidivism score. Generally, felony cases have higher risk assessment scores than the misdemeanor cases That may because those individuals do not have a prior criminal history or come from stable family units making them appear as a lower risk. Figure 3.5 paints a similar picture showing the number of individuals in each race that has been charged with either a felony or misdemeanor. We can infer that since African Americans account for most of the felony charges in the dataset they also have an association to high risk assessment scores.

This becomes evident also when investigating non-felony or misdemeanor crimes. Figure 3.6. This table shows the number of non-recidivists who were arrested with no charge and still labeled as high risk of recidivism (score of 8-10). Out of the 263 individuals in this category African Americans account for 78.3% of those mislabeled non-recidivists. I believe this could be due to the overrepresentation of African Americans in the dataset. These non-recidivists individuals are probably being compared to the characteristics of the entire group of African Americans who account for most of the felonies and thus high risk assessment scores. It is an unfortunate consequence of the nature of the data set.

Figure 3.4

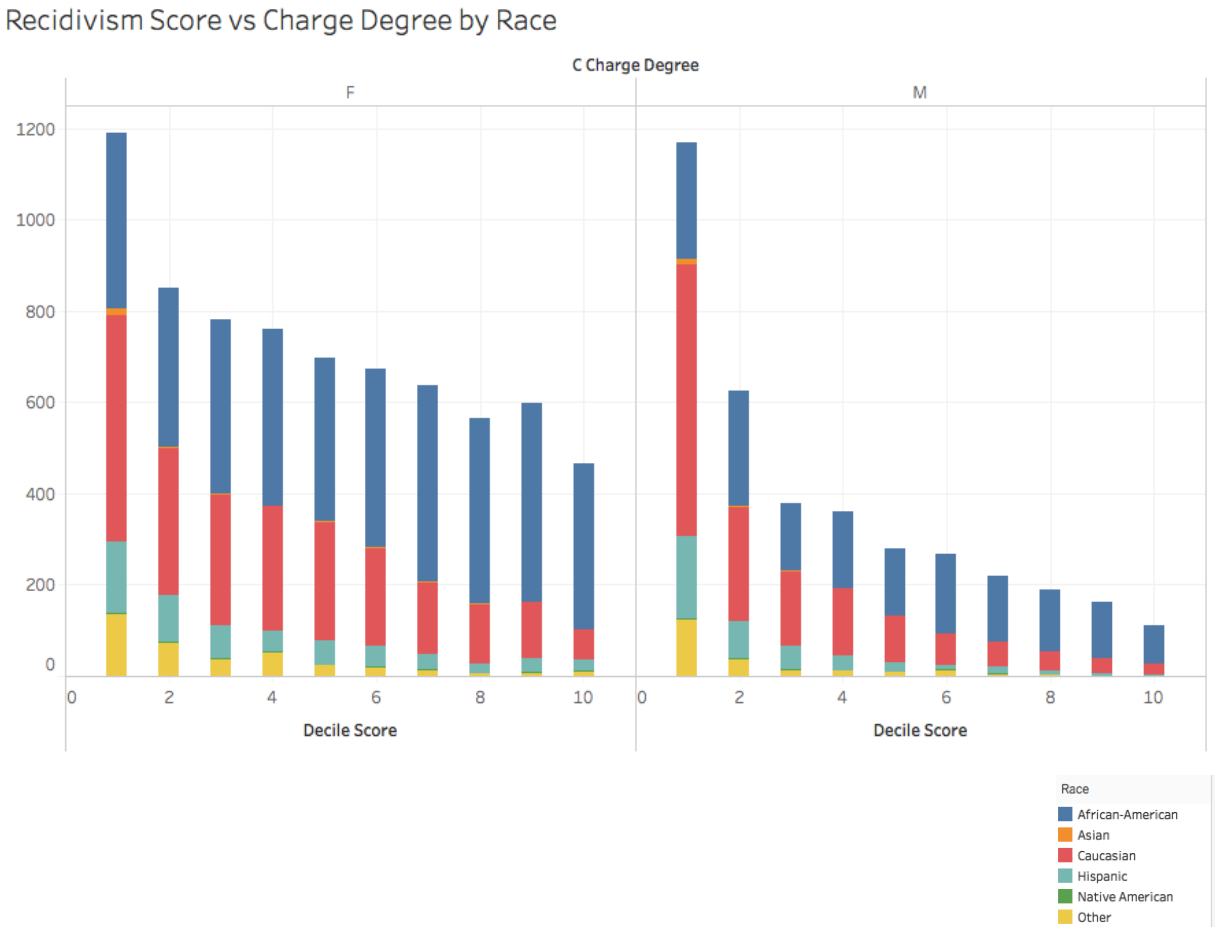

Recidivism Score vs Charge Degree by Race
C Charge Degree
F
M
1200
1000
800
600
400
200
0
0
2
4
6
8
10
Decile Score
Race
African-American
Asian
Caucasian
Hispanic
Native American
Other

Figure 3.5

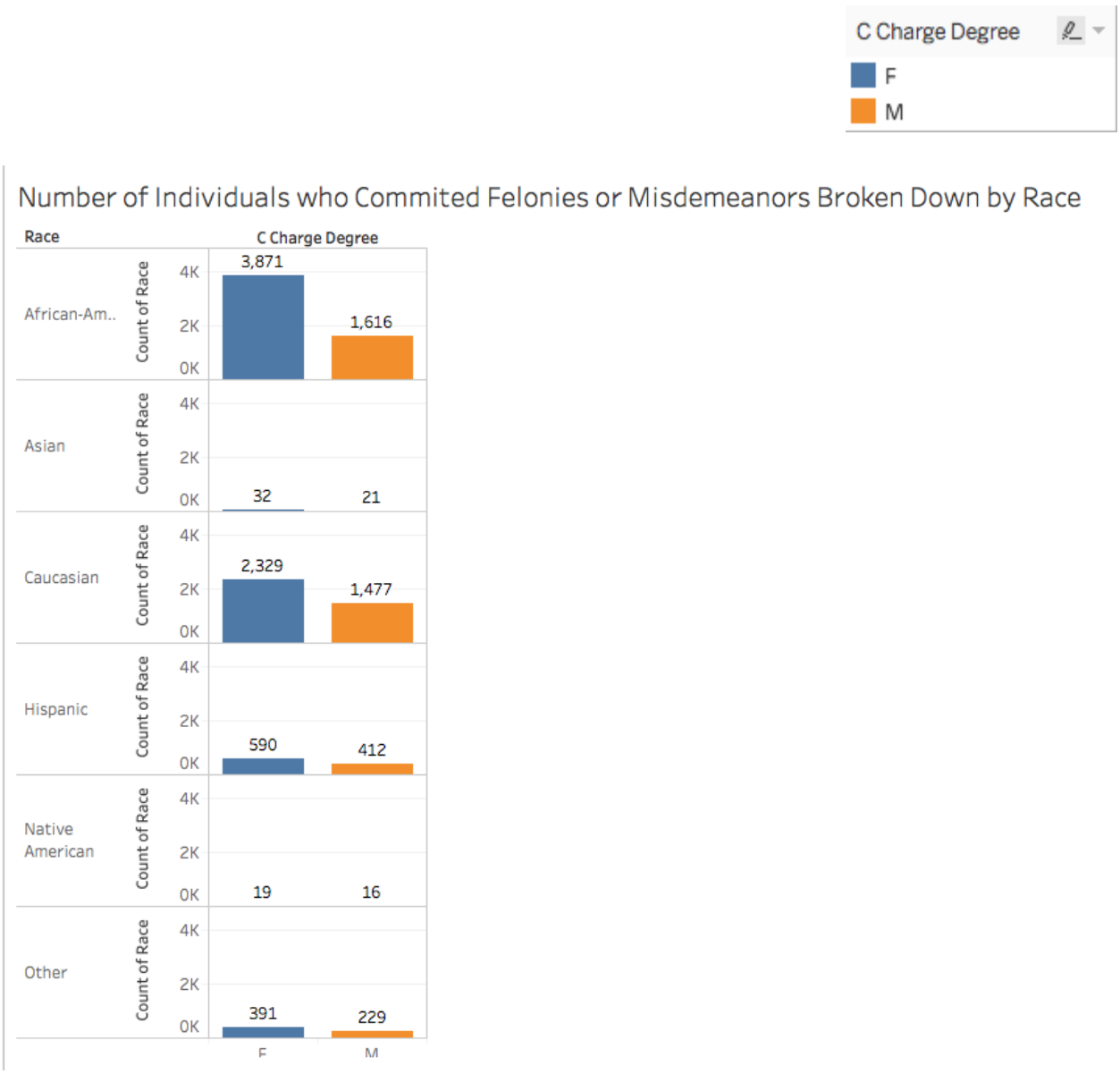

C Charge Degree
F
M
Number of Individuals who Commited Felonies or Misdemeanors Broken Down by Race
Race
C Charge Degree
African-Am..
Asian
Caucasian
Hispanic
Native American
Other
Count of Race
4K
2K
0K
3,871
1,616
32
21
2,329
1,477
590
412
19
16
391
229
F
M

Figure 3.6

race	score	c_charge_desc	Count_Non_Recid
African-American	8	arrest case no charge	65
African-American	9	arrest case no charge	74
African-American	10	arrest case no charge	66
Caucasian	8	arrest case no charge	13
Caucasian	9	arrest case no charge	22
Caucasian	10	arrest case no charge	6
Hispanic	8	arrest case no charge	6
Hispanic	9	arrest case no charge	4
Hispanic	10	arrest case no charge	3
Native American	10	arrest case no charge	1
Other	8	arrest case no charge	1
Other	10	arrest case no charge	2

Conclusions

Overall, this data has shown us that the distribution of race in amongst the scores is really a reflection of the overrepresentation of certain groups in the dataset, namely African Americans. This lead to a discrepancy in who was consistently assigned violent risk scores or high risk of recidivism scores. There was also not a strict association between crime and risk score. Although those who committed misdemeanors tended to have lower scores seen in Figure 3.4. Furthermore, in part II we see that generally the score accurately dubs individuals who will likely recidivate. These analyses were all drawn using SQLite in Jupyter Notebooks and the visualizations were generated in Tableau. You can find the notebook and Tableau sheets attached with the full analysis.