Jimmy Carpenter
COSC 5610
11/05/2020

Data Exploration and Visualization

I will begin with a quick summary on the ProPublica COMPAS Recidivism Risk Score dataset. The dataset contains 7,215 examples and 52 possible predictive features. The key feature that the dataset contains is if the person returns to jail within two years or not. This feature is called is_recid and is 0 if they did not return to jail and 1 if they did. The first thing I wanted to do was remove bad/unusable data. I filtered out any example which had a charge date that was not within 30 days from when the person was arrested. If the is_recid value was -1, then the compass case could not be found and therefore no label exists. Also, there was two cases where the defendant was charged with ordinary traffic offenses. These offenses do not result in any jail time, so I filtered the two cases out. Any examples that contained a COMPASS score of 'N/A' were filtered out because this is going to be a crucial feature for my model to train on. After the data cleaning I just described, I was left with 6172 examples in the dataset. I then wanted to get an idea of the age distribution of the dataset.

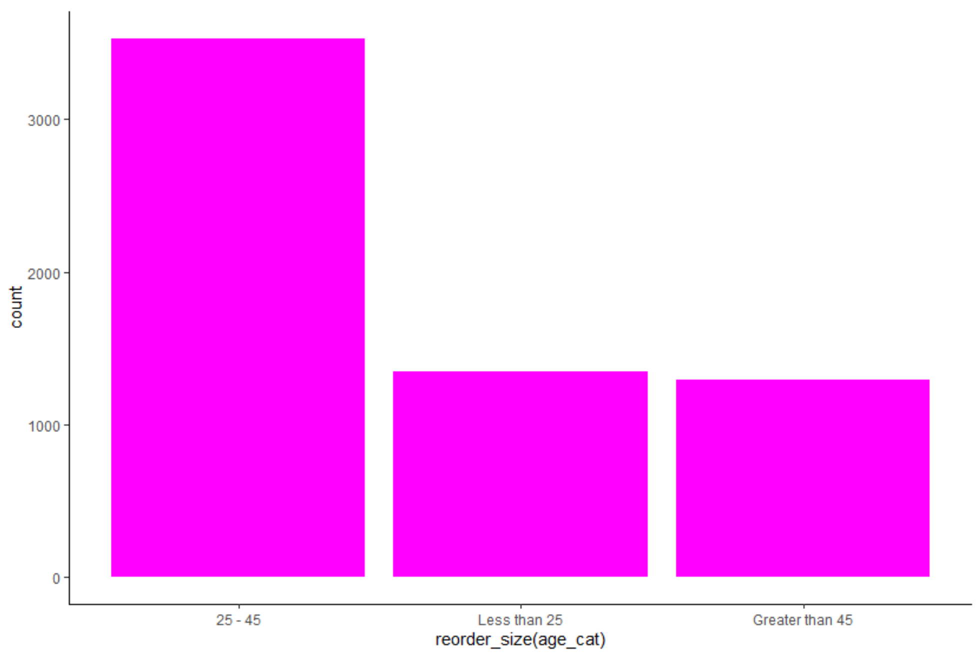

Figure 1. This plot shows the number of people in each age category.

The plot is shown with each category in descending order, according to the number of people in each category. It makes sense that the largest group was between the ages of 25-45 years old, but I was surprised that less than 25 was the second largest. This age category covers the ages 18-25, while the other two categories cover 20 years. I knew that younger people committed a lot of crimes, but I did not know it was to this extent.

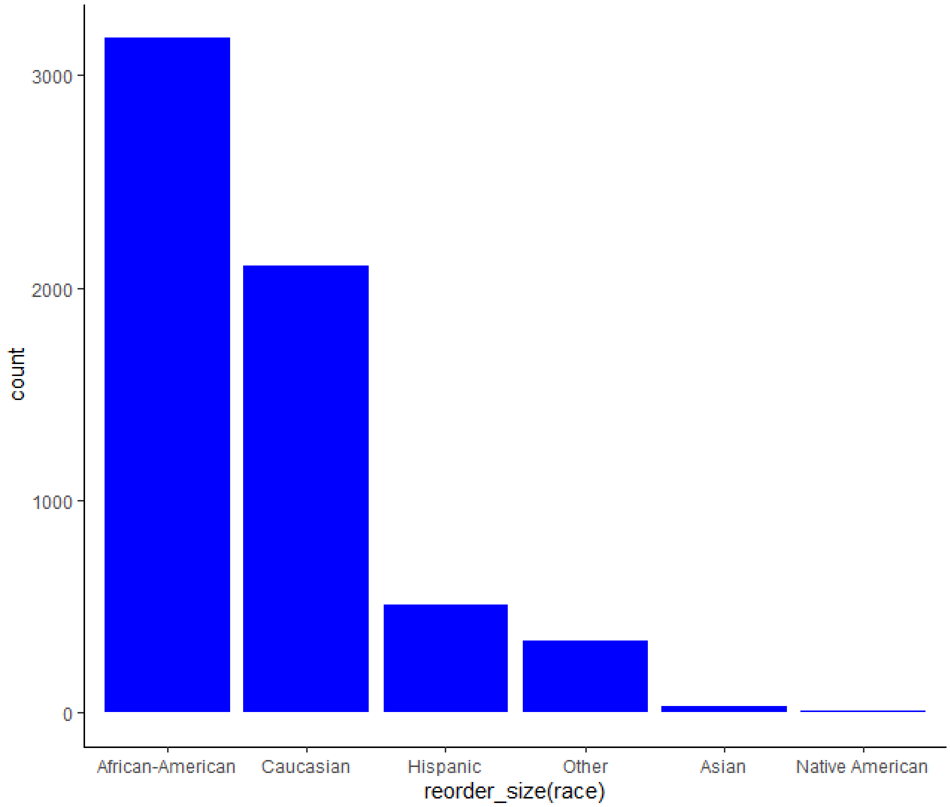

Figure 2. This shows the race distribution within the dataset.

I was not surprised to see the highest population in the dataset being African Americans because the dataset is about racial bias. From my research, there are more African Americans incarcerated then any other race, so this dataset reflects a realistic distribution. Obviously, there is a lot of wrongful reasons that this is the case, but I do not want to make this discussion political in anyway. I do not think the dataset should be balanced racially because it is a realistic representation of the system though. The next thing I looked at was the percentage of people who did recidivate, resulting in another jail time offense. Around 45.51% of the dataset did go back to jail within two years and 48.44% returned to jail. This means that the label we are trying to predict is fairly balanced and does not need any balancing. After seeing that there is an unbalanced distribution of races and a balanced label of recidivism, I wanted to investigate them together.

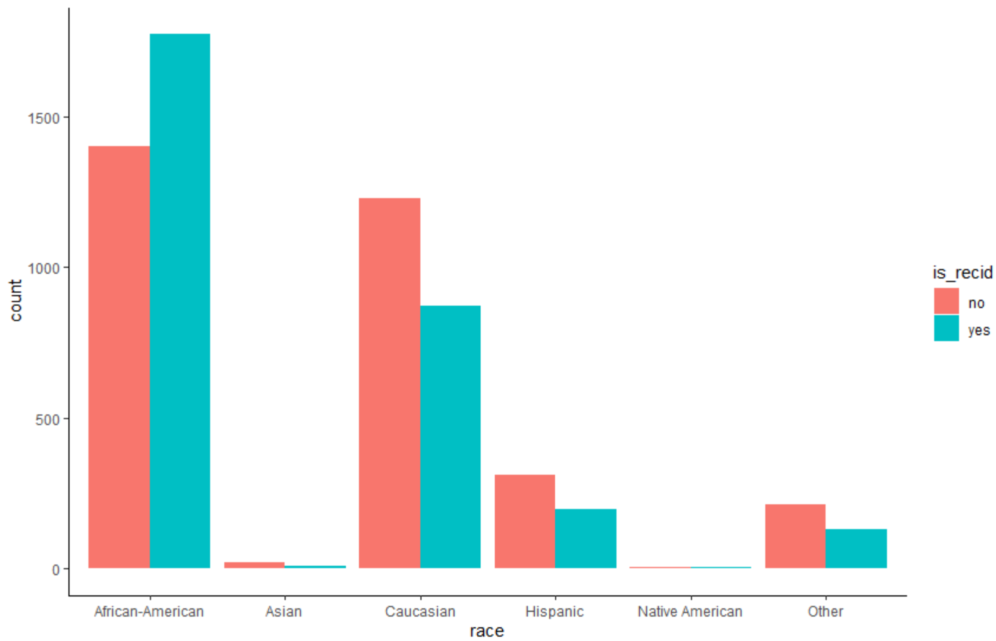

Figure 3. This shows the distribution of races and whether they returned to jail within two years or not.

I found this plot quite interesting and presents some very controversial debate. The COMPAS algorithm gives higher scores to defendants that are likely to result in recidivism. The dataset shows that 56% of the African American defendants did return to jail after their release. The data also shows that around 40% of Caucasians, Hispanics, and the other race category reoffended. Obviously, no direct correlation can be made between race and recidivism, but African Americans did recidivate more then any other race. This made me want to look at the distribution of COMPAS scores for African Americans and Caucasians.

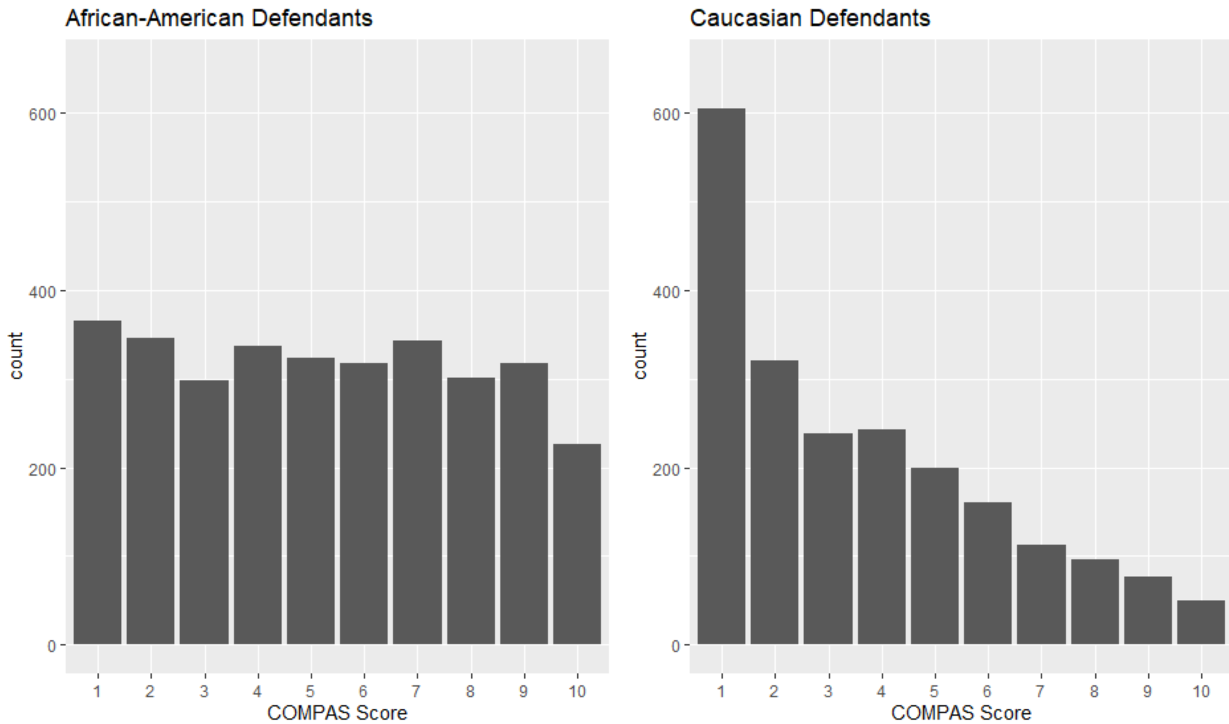

Figure 4. These plots show the distribution of COMPAS scores for African-American and Caucasian defendants.

The COMPAS score is a likely hood on a scale of 1-10 that the defendant will recidivate. A high score (8-10) means that the algorithm predicates the defendant will reoffend, thus a harsher sentence being needed. The sublot on the right of Figure 4 shows a downward trend for the number of Caucasian defendants and the COMPAS score they received. As you can see, over 600 Caucasian defendants received a COMPAS score of 1 and only 50 Caucasians received a COMAPS score of 10. The subplot on the right shows a fairly equal distribution of COMPAS scores for African American defendants. The next thing I wanted to know was why these distributions are so different. On one hand this dataset showed that African Americans did recidivate more than Caucasians and on the other hand race should not define the actions someone will take. The issue is very complex and not the subject of this paper, but I thought the plots were interesting and crucial for understanding the data.

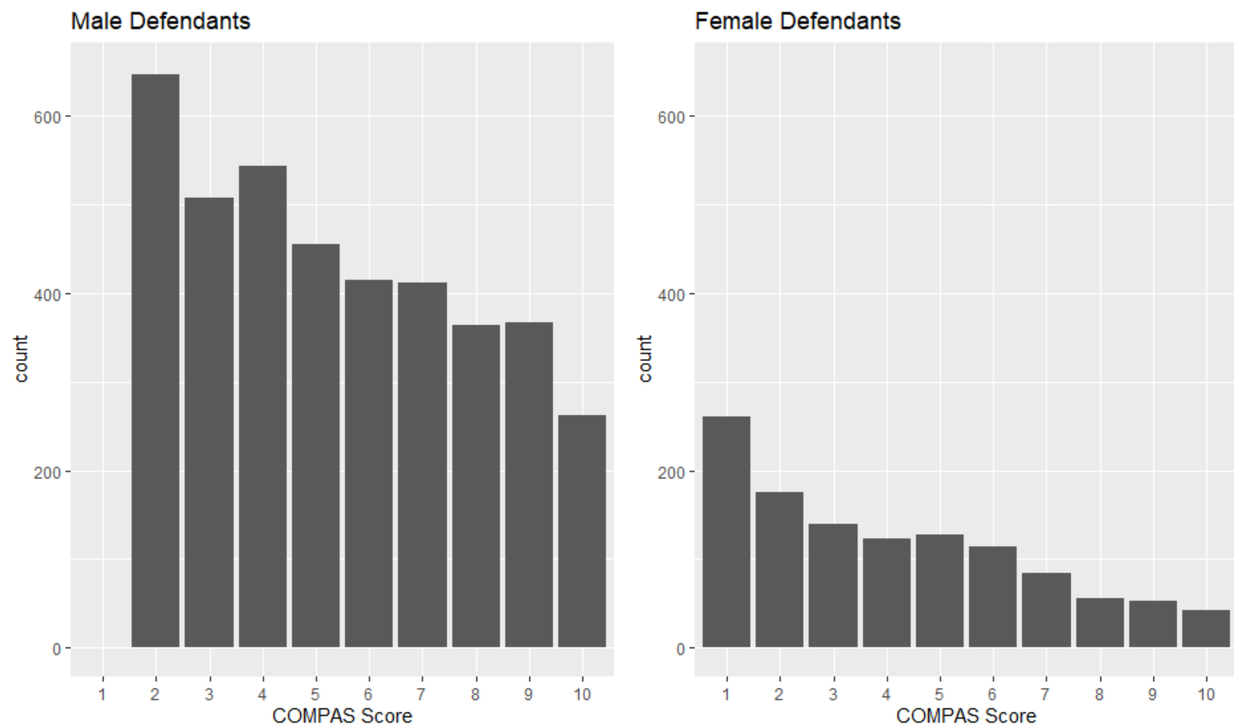

Figure 5. These two plots show the distribution of COMPAS scores for male and female defendants.

I thought there might be a difference in the trend for male and female defendants but there is not. Both subplots in Figure 5 look like the distribution of COMPAS scores for Caucasian people. This furthers my confusion on why the plot for African American defendants and their COMPAS scores is so different. This led me to create a simple logistic regression model that was given gender, age, race, number of prior, misdemeanor or felony, recidivism within two years, priors count, and a binary output of a low or high COMPAS score.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -3.30957    0.15716 -21.059  < 2e-16 ***
gender_factorFemale          0.22127    0.07951   2.783 0.005388 ** 
age_factor25 - 45            1.35563    0.09908  13.682  < 2e-16 ***
age_factorLess than 25       2.66402    0.11548  23.069  < 2e-16 ***
race_factorAfrican-American  0.90560    0.12331   7.344 2.07e-13 ***
race_factorAsian             0.17398    0.48897   0.356 0.721986    
race_factorCaucasian         0.42839    0.12813   3.344 0.000827 ***
race_factorNative American   1.82260    0.77283   2.358 0.018356 *  
race_factorOther            -0.39795    0.19121  -2.081 0.037416 *  
priors_count                 0.26895    0.01110  24.221  < 2e-16 ***
crime_factorM               -0.31124    0.06655  -4.677 2.91e-06 ***
two_year_recid               0.68586    0.06402  10.713  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. This is a screenshot of the regression coefficients from the logit model.

As shown in Figure 6, the coefficient for African Americans is positive, meaning it contributes to a high compass score. This is true for Caucasians for as well, but it is about half the size of the coefficient for African Americans. Interestingly enough, being less than 25 years old is the most significant predictor for a high COMPAS score. This makes sense because they are younger and might not be mature enough to not reoffend. The output also shows that receiving a misdemeanor charge will help contribute to a low COMPAS score. I think fitting this basic model can give many interesting insights on how the COMPAS algorithm works and the predictors they use. Essentially this logistic regression model I created is a basic version of the COMPAS algorithm. From this I can conclude that race is used as an input feature for calculating the defendants COMPAS score. I think this is why the plot for African Americans and the distribution of COMPAS scores they received did not follow the same trend as any other race. The COMPAS algorithm most likely was trained on biased data, and I do not think that race should be used as a predictor for recidivism. Now, I feel quite confident with my dataset and have some ideas on what inputs I would like to use. Also, this gave me ideas on various tests and parameters I would like to tinker with.

References:

Anon. 2020. Criminal Justice Fact Sheet. (July 2020). Retrieved November 3, 2020 from https://www.naacp.org/criminal-justice-fact-sheet/

Anon. Bar Chart & Histogram in R (with Example). Retrieved November 2, 2020 from https://www.guru99.com/r-bar-chart-histogram.html

Anon. Categorical Data Descriptive Statistics. Retrieved November 2, 2020 from https://uc-r.github.io/descriptives_categorical

Anon.Retrieved November 3, 2020 from https://towardsdatascience.com/a-guide-to-data-visualisation-in-r-for-beginners-ef6d41a34174

Ed Yong. 2018. A Popular Algorithm Is No Better at Predicting Crimes Than Random People. (January 2018). Retrieved November 3, 2020 from https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

Julia Angwin Jeff Larson. 2016. How We Analyzed the COMPAS Recidivism Algorithm. (May 2016). Retrieved November 2, 2020 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Karen Grace-Martin et al. 2020. Interpreting Regression Coefficients. (January 2020). Retrieved November 3, 2020 from https://www.theanalysisfactor.com/interpreting-regression-coefficients/

ProPublica. 2019. COMPAS Recidivism Risk Score Data and Analysis. (March 2019). Retrieved November 3, 2020 from https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis