Exploratory Analysis

Mark Powell

The dataset I have is “Civilian Complaints Against New York City Police Officers.” This CSV included officer information, the complainant allegation, their gender/race, the outcome of their altercation, and the result of the board against the officer. My goal with this information is to see how the gender and race of an officer might affect their interaction with the complainant. Then to see if a complainant is more or less likely to be arrested based off their gender or race. Finally, look to see if an officer is more or less likely to be prosecuted by the board based off their race or gender. My initial assumptions are that white, male officers are going to be more likely than other officers to be the subject of some sort of complaint especially against minorities. Then after that they are more likely to be let off by the board. Another thing I am guessing I will see if black male officers be more likely to be given some corrective action by the board. My guess is that gender will have a much lesser impact than race when looking at overall trends, but I am curious if that will hold up.

I chose bar and stacked bar graphs to represent my data. This dataset does not deal with a ton of different values for the most part. I am pretty sure the columns I will be focusing on do not go over ten unique values. Bar graphs are perfect because it allows you to visualize each chunk of data easily. I thought stacked bar graphs also worked well with my dataset. It allowed me to compare for example officer ethnicity with the board depositions. You get a general idea right away the various results of the board compared to the officer’s ethnicity.

Honestly, I was somewhat surprised by the breakdown in figure one. I would have guessed that whites would have had a higher share of the complaints mostly based of having higher numbers in the police force but I do not really know the actual racial breakdowns of police officers so maybe not. Figure two is not too surprising. Males make up a higher percentage of force. Figure three deals with the ethnicity of the making complaints against officers in the NYPD. Minorities having the highest share here does not surprise me whatsoever. Minorities interact with police the most often. Also, I am coming from the view that there is racial biases that might cause those interactions and this figure might be able to support that. The non-labeled bar is no race because it was not reported. That much data missing was surprising to me. I would have guessed under 5 percent missing data, but its closer to 1/6. I wonder if the complainant purposely did not report it or some sort of "admin error." There is enough missing which makes me think it was purposeful in some way. This won't affect my model at all because there's still plenty of other data, but it did pique my interest. Figure four isn't surprising. Males are going to be involved with police officers way more often than women. Figure five is where the data starts to get interesting. The board found in over half its cases that the officer was exonerated or unsubstantiated. Exonerated means that the alleged conduct occurred but did not violate the NYPD's rules, which often give officers significant discretion over use of force. Unsubstantiated is the CCRB fully investigating but could not affirmatively conclude both that the conduct occurred and that it broke the rules. I find it hard to believe over half of the cases that the officer was found at no fault, especially since someone had to take time out of their day to make a complaint. A problem for me here is how little the board came to some of the

conclusions. That leaves almost no training or testing data for some of those options. This is something I will have to consider when designing my model and I will need to think about the over/under sampling techniques discussed in our readings.

Figure six is the result of the altercation between the officer and the complainant. Pretty much the action the officer took after what the complaint alleges happened. It is close to a 50-50 split between the person getting arrested or not, so not too surprising to me. Figure seven shows ethnicity breakdown per board disposition. I don't think there's any major conclusions to draw from it at first glance. It looks even overall. Figure eight is the same thing but with gender instead of race. Not too much data for females so something to make sure to consider when looking at my model. Figure nine is police officer ethnicity vs. complainant ethnicity. It breaks down the ethnicity of complainants and the ethnicity of the officers they make complaints against. This is an important area I want to explore for my model.

References

Anon. ggplot2 barplots : Quick start guide - R software and data visualization. Retrieved November 5, 2020 from http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

Datavizpyr. 2020. How To Rotate x-axis Text Labels in ggplot2? (September 2020). Retrieved November 5, 2020 from https://datavizpyr.com/rotate-x-axis-text-labels-in-ggplot2/

ProPublica. 2020. Civilian Complaints Against New York City Police Officers. (July 2020). Retrieved November 5, 2020 from https://www.propublica.org/datastore/dataset/civilian-complaints-against-new-york-city-police-officers

Glossary

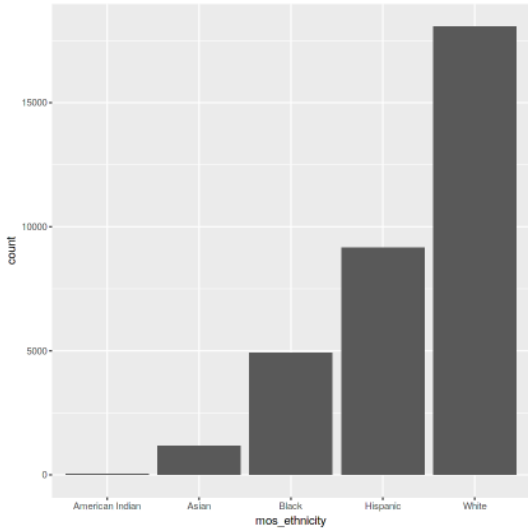

Figure 1: Officer ethnicity

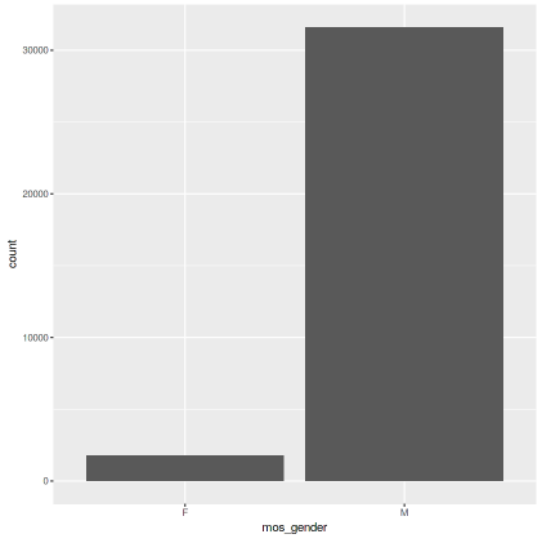

Figure 2: Officer gender

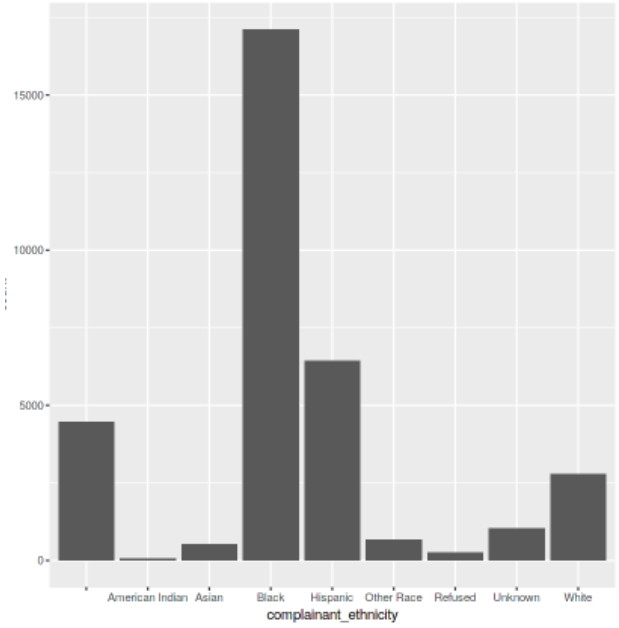

Figure 3: Complainant ethnicity

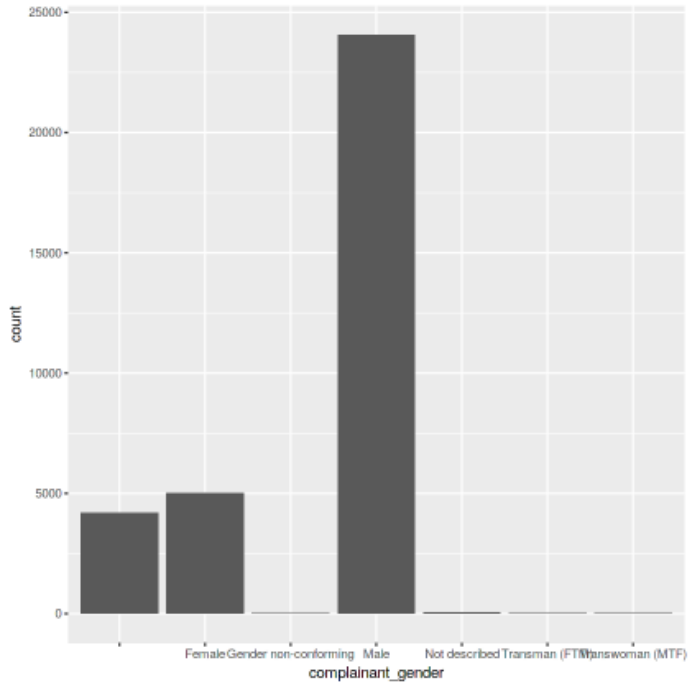

Figure 4: Complainant gender

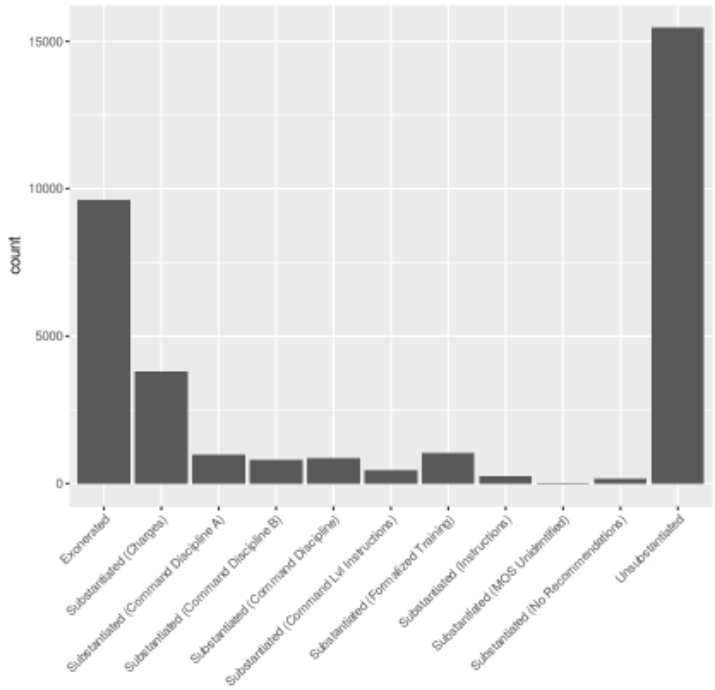

Figure 5: Total Board Dispositions

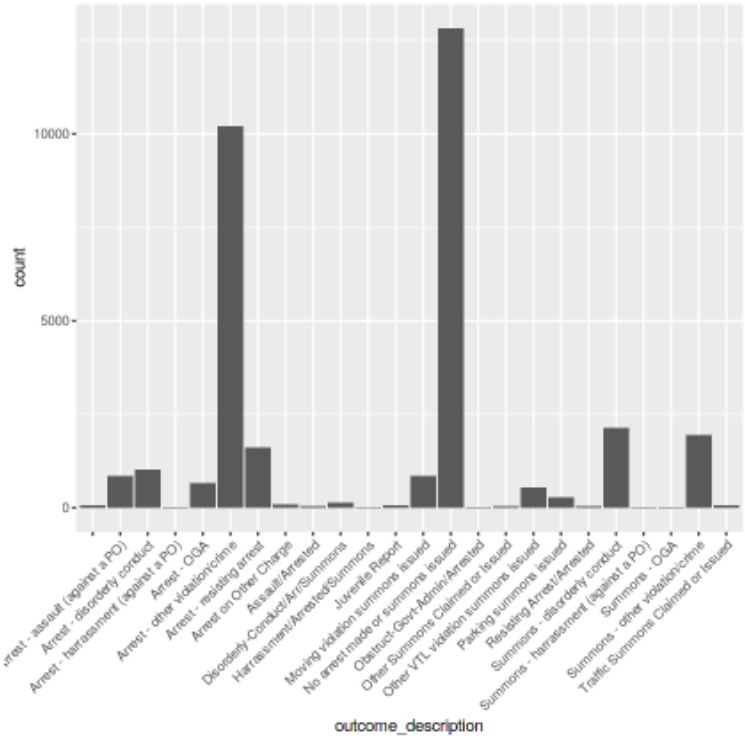

Figure 6: Outcome of complainant interacting with officer

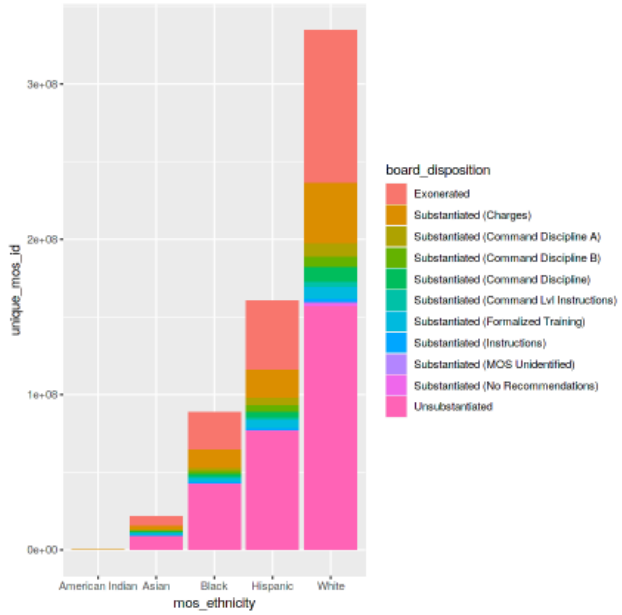

Figure 7: Officer ethnicity vs. board disposition

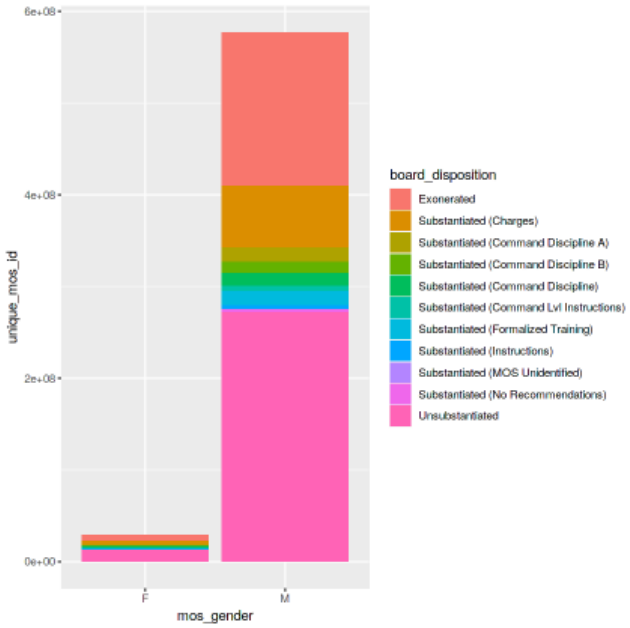

Figure 8: Officer gender vs. board disposition

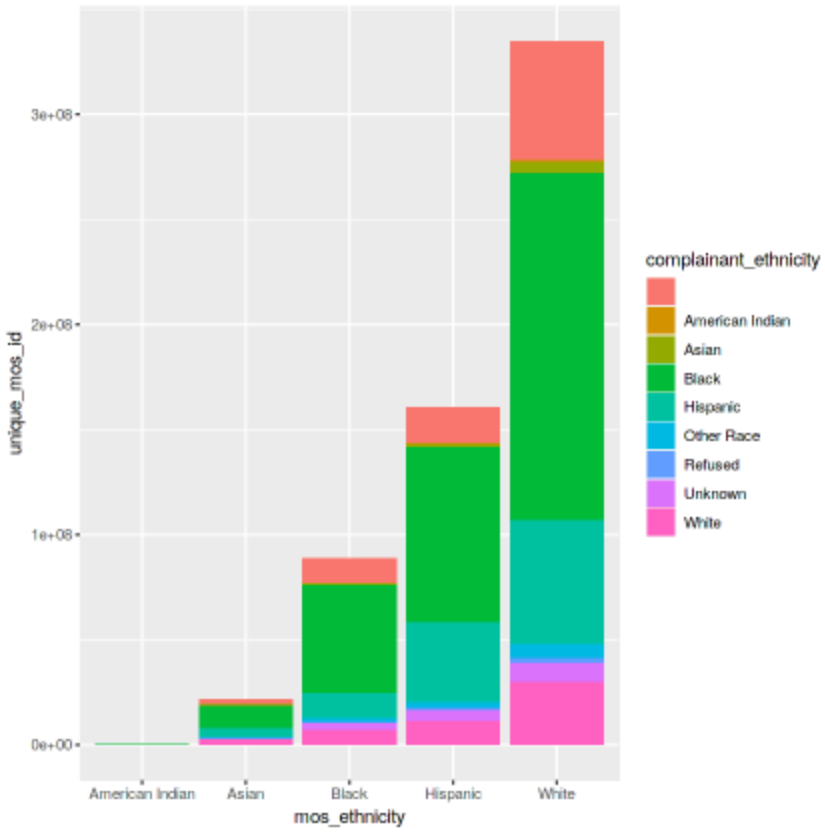

Figure 9: Officer ethnicity vs complainant ethnicity