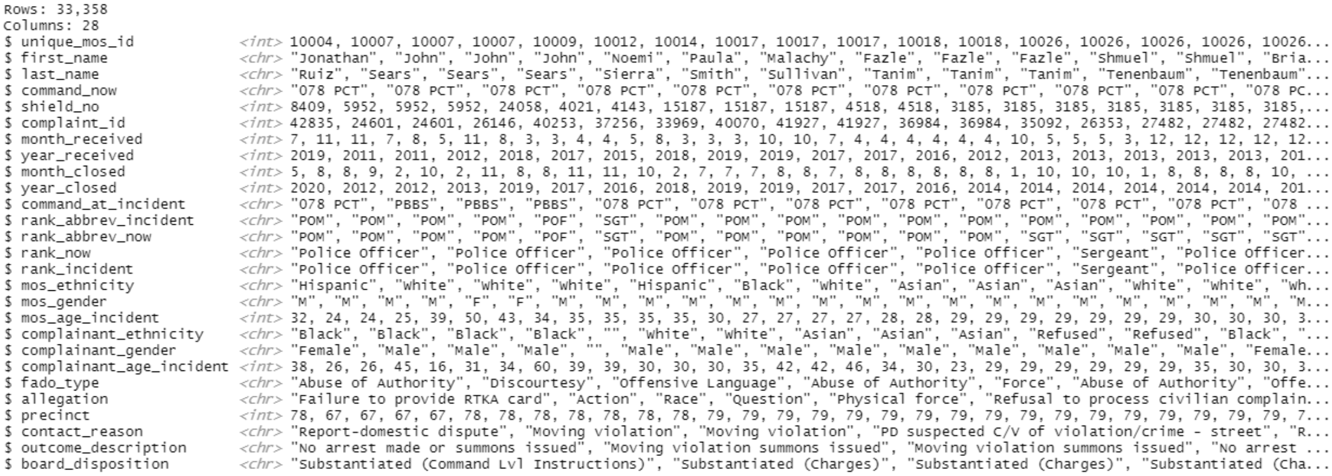
Data Exploration and Visualization:
Civilian Complaints Against New York City Police Officers

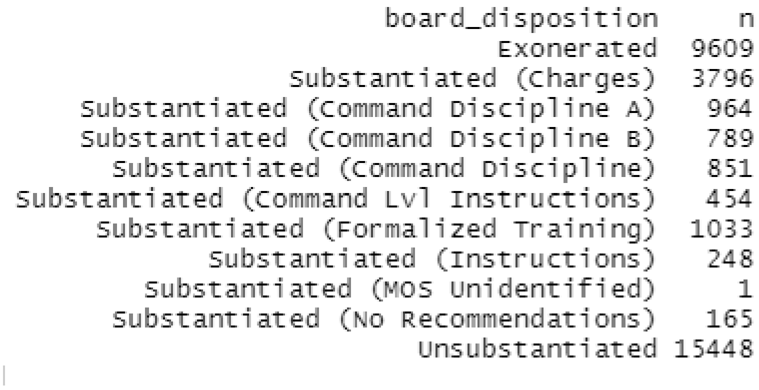
Julia Beilke
COSC 5610 – Data Mining
Marquette University
November 5, 2020

In this report, I will examine ProPublica's dataset, "Civilian Complaints Against New York City Police Officers [1]." The dataset includes over 33,358 civilian complaints against NYPD officers who were currently on the force when the data was released in July of 2020. It provides unique identifiers for cases and officers and details about the officers and complainants. It also includes the type of complaint, type of interaction between the officer and complainant, and the board's disposition after investigating the case. The board disposition can either be substantiated, unsubstantiated, or exonerated. I would ultimately like to create a logistic regression model that calculates the probability of a substantiated outcome based on the complainant and officer's characteristics.



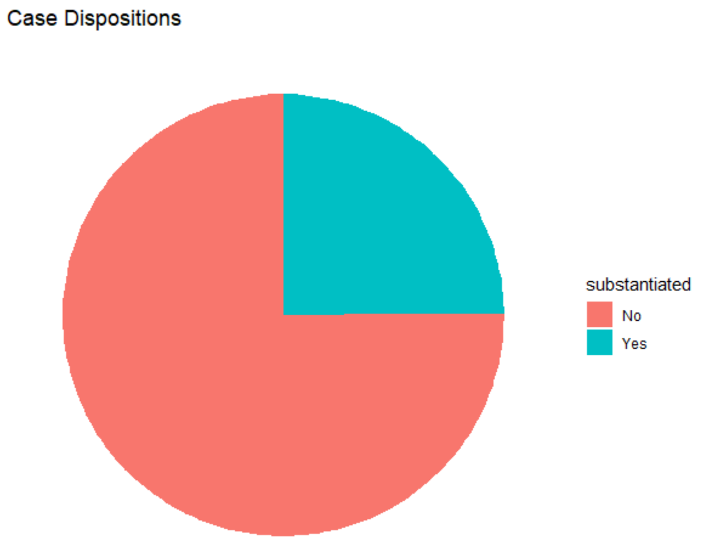
```
Rows: 33,358
Columns: 28
$ unique_mos_id            <int> 10004, 10007, 10007, 10007, 10009, 10012, 10014, 10017, 10017, 10017, 10018, 10018, 10026, 10026, 10026, 10026, 10026...
$ first_name               <chr> "Jonathan", "John", "John", "John", "Noemi", "Paula", "Malachy", "Fazle", "Fazle", "Fazle", "Shmuel", "Shmuel", "Bria...
$ last_name                <chr> "Ruiz", "Sears", "Sears", "Sears", "Sierra", "Smith", "Sullivan", "Tanim", "Tanim", "Tanim", "Tenenbaum", "Tenenbaum"...
$ command_now              <chr> "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PC...
$ shield_no                <int> 8409, 5952, 5952, 5952, 24058, 4021, 4143, 15187, 15187, 15187, 4518, 4518, 3185, 3185, 3185, 3185, 3185, 3185, 3185,...
$ complaint_id             <int> 42835, 24601, 24601, 26146, 40253, 37256, 33969, 40070, 41927, 41927, 36984, 36984, 35092, 26353, 27482, 27482, 27482...
$ month_received           <int> 7, 11, 11, 7, 8, 5, 11, 8, 3, 3, 4, 4, 5, 8, 3, 3, 3, 10, 10, 7, 4, 4, 4, 4, 4, 4, 10, 5, 5, 5, 3, 12, 12, 12, 12, 12...
$ year_received            <int> 2019, 2011, 2011, 2012, 2018, 2017, 2015, 2018, 2019, 2019, 2017, 2017, 2016, 2012, 2013, 2013, 2013, 2013, 2013, 201...
$ month_closed             <int> 5, 8, 8, 9, 2, 10, 2, 11, 8, 8, 11, 11, 10, 2, 7, 7, 7, 8, 8, 7, 8, 8, 8, 8, 8, 8, 1, 10, 10, 10, 1, 8, 8, 8, 8, 10, ...
$ year_closed              <int> 2020, 2012, 2012, 2013, 2019, 2017, 2016, 2018, 2019, 2019, 2017, 2017, 2016, 2014, 2014, 2014, 2014, 2014, 2014, 201...
$ command_at_incident      <chr> "078 PCT", "PBBS", "PBBS", "PBBS", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 PCT", "078 ...
$ rank_abbrev_incident     <chr> "POM", "POM", "POM", "POM", "POF", "SGT", "POM", "POM", "POM", "POM", "POM", "POM", "POM", "POM", "POM", "POM", "POM"...
$ rank_abbrev_now          <chr> "POM", "POM", "POM", "POM", "POF", "SGT", "POM", "POM", "POM", "POM", "POM", "POM", "SGT", "SGT", "SGT", "SGT", "SGT"...
$ rank_now                 <chr> "Police Officer", "Police Officer", "Police Officer", "Police Officer", "Police Officer", "Sergeant", "Police Officer...
$ rank_incident            <chr> "Police Officer", "Police Officer", "Police Officer", "Police Officer", "Police Officer", "Sergeant", "Police Officer...
$ mos_ethnicity            <chr> "Hispanic", "White", "White", "White", "Hispanic", "Black", "White", "Asian", "Asian", "Asian", "White", "White", "Wh...
$ mos_gender               <chr> "M", "M", "M", "M", "F", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M...
$ mos_age_incident         <int> 32, 24, 24, 25, 39, 50, 43, 34, 35, 35, 35, 35, 30, 27, 27, 27, 27, 28, 28, 29, 29, 29, 29, 29, 29, 29, 30, 30, 30, 3...
$ complainant_ethnicity    <chr> "Black", "Black", "Black", "Black", "", "White", "White", "Asian", "Asian", "Asian", "Refused", "Refused", "Black", "...
$ complainant_gender       <chr> "Female", "Male", "Male", "Male", "", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Female...
$ complainant_age_incident <int> 38, 26, 26, 45, 16, 31, 34, 60, 39, 39, 30, 30, 30, 35, 42, 42, 46, 34, 30, 23, 29, 29, 29, 29, 29, 29, 35, 30, 30, 3...
$ fado_type                <chr> "Abuse of Authority", "Discourtesy", "Offensive Language", "Abuse of Authority", "Force", "Abuse of Authority", "Offe...
$ allegation               <chr> "Failure to provide RTKA card", "Action", "Race", "Question", "Physical force", "Refusal to process civilian complain...
$ precinct                 <int> 78, 67, 67, 67, 67, 78, 78, 78, 78, 78, 78, 78, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 79, 7...
$ contact_reason           <chr> "Report-domestic dispute", "Moving violation", "Moving violation", "PD suspected C/V of violation/crime - street", "R...
$ outcome_description      <chr> "No arrest made or summons issued", "Moving violation summons issued", "Moving violation summons issued", "No arrest ...
$ board_disposition        <chr> "Substantiated (Command Lvl Instructions)", "Substantiated (Charges)", "Substantiated (Charges)", "Substantiated (Cha...
```

I initially expected that there would be far more unsubstantiated outcomes than substantiated. This expectation was based on a ProPublica report showing that the NYPD often withheld critical evidence from the CCRB, which prevented the CCRB investigators from substantiating complaints [2]. To check for class imbalance in board disposition, I began by aggregating the board disposition data.



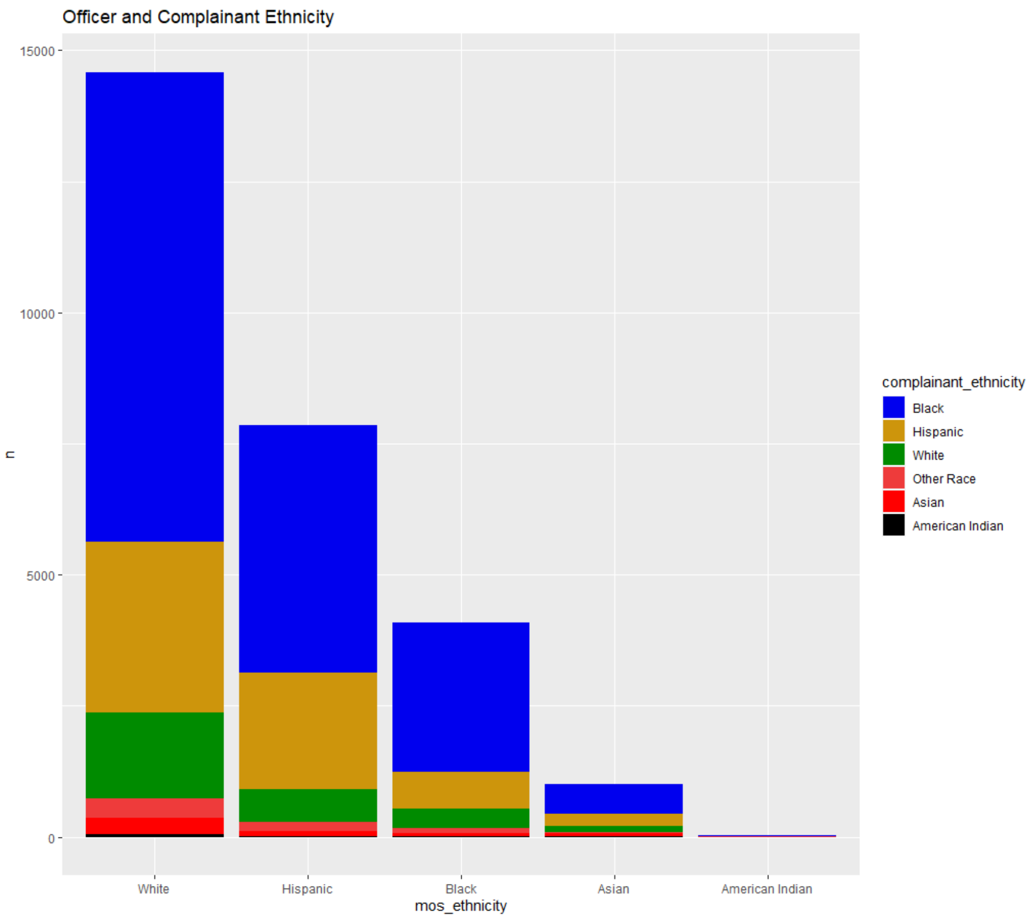
```
                           board_disposition     n
                                 Exonerated  9609
                      Substantiated (Charges)  3796
          Substantiated (Command Discipline A)   964
          Substantiated (Command Discipline B)   789
            Substantiated (Command Discipline)   851
      Substantiated (Command Lvl Instructions)   454
           Substantiated (Formalized Training)  1033
                  Substantiated (Instructions)   248
              Substantiated (MOS Unidentified)     1
            Substantiated (No Recommendations)   165
                             Unsubstantiated 15448
```

Clearly, there are many cases where the board disposition is exonerated or unsubstantiated. However, because there were so many different versions of substantiated outcomes, it was difficult to visualize the distribution of substantiated versus unsubstantiated or exonerated cases. To simplify the dataset for visualization, I created a new binary variable, substantiated, which indicates whether the case disposition was substantiated. This will also be necessary for eventually creating a logistic regression model that predicts board disposition. I set the new variable to 1 when the disposition is some version of substantiated and 0 when the disposition is exonerated or unsubstantiated. For the remainder of this report, I will include exonerated cases as part of the unsubstantiated class.

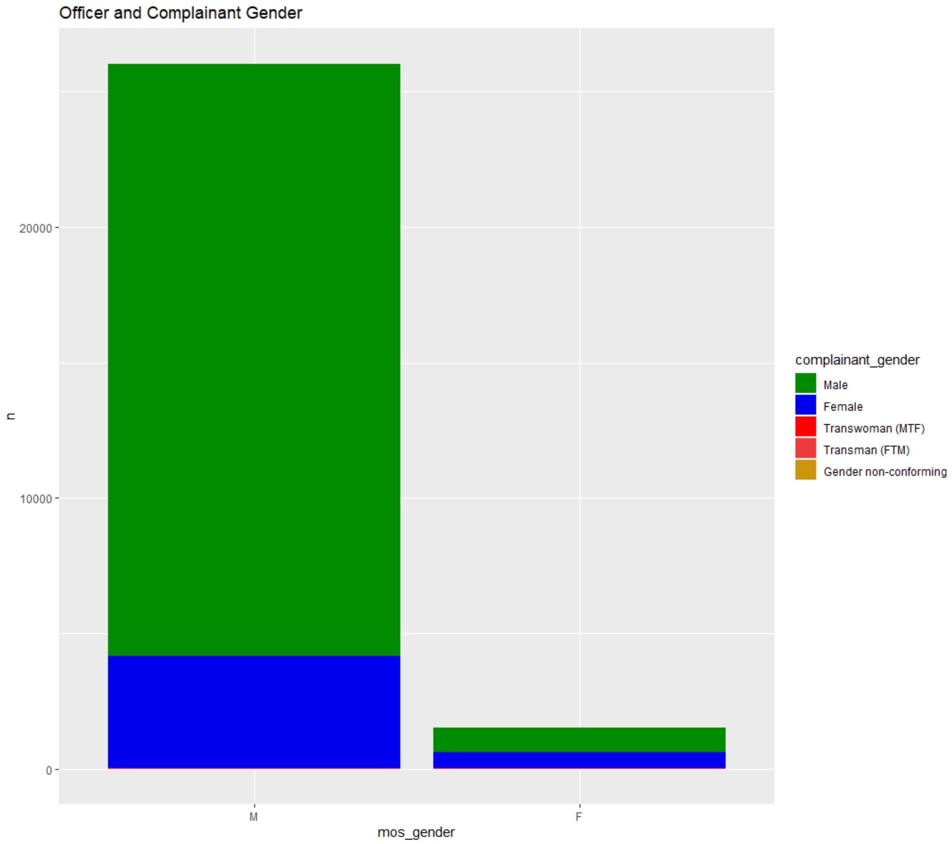


I found that there is about a 3:1 ratio of substantiated to unsubstantiated cases (25057 unsubstantiated and 8301 substantiated). This is a bit more balanced than I was expecting, but I will still be considering rebalancing with Undersampling, Oversampling, or SMOTE to see if these techniques will improve my model.

I was expecting that the majority of the officers in the dataset would be White males because this demographic makes up the majority of the NYPD [3]. I also expected that the majority of complainants would be Black because an overwhelming majority of arrests made in New York City are of Black people [4]. There were no instances where the officer's ethnicity or gender was unknown, but there were several instances where the complainant's ethnicity or gender was not included. I dropped these instances from the dataset.

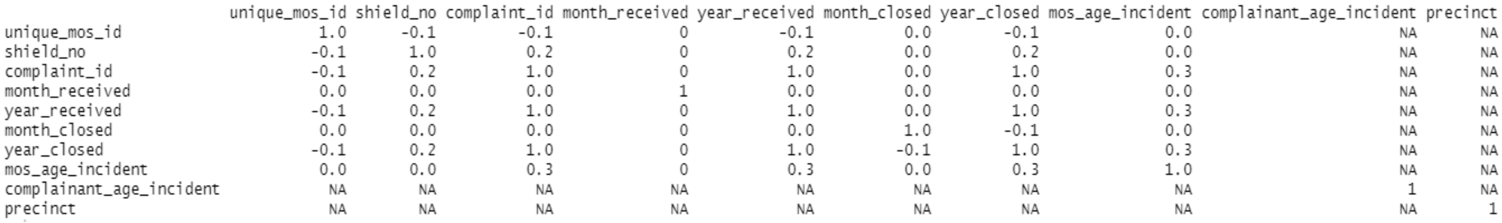


I can see that the complaints recorded in this dataset are largely against White, male officers by Black, male complainants. However, it doesn't look like there are disproportionately more or fewer complaints per ethnic against White officers versus complaints against Hispanic, Black, or Asian or American Indian officers.



As was expected, the officers in the dataset are overwhelmingly male. The complainants are also overwhelmingly men. Similarly to the ethnicity breakdown, there is a fairly proportionate amount of complaints from male, female, and other genders against male officers versus female officers.

I began looking for relationships between the numerical variables in the dataset by building a correlation matrix for these variables.



```
                         unique_mos_id shield_no complaint_id month_received year_received month_closed year_closed mos_age_incident complainant_age_incident precinct
unique_mos_id                      1.0      -0.1         -0.1              0          -0.1          0.0        -0.1              0.0                       NA       NA
shield_no                         -0.1       1.0          0.2              0           0.2          0.0         0.2              0.0                       NA       NA
complaint_id                      -0.1       0.2          1.0              0           1.0          0.0         1.0              0.3                       NA       NA
month_received                     0.0       0.0          0.0              1           0.0          0.0         0.0              0.0                       NA       NA
year_received                     -0.1       0.2          1.0              0           1.0          0.0         1.0              0.3                       NA       NA
month_closed                       0.0       0.0          0.0              0           0.0          1.0        -0.1              0.0                       NA       NA
year_closed                       -0.1       0.2          1.0              0           1.0         -0.1         1.0              0.3                       NA       NA
mos_age_incident                   0.0       0.0          0.3              0           0.3          0.0         0.3              1.0                       NA       NA
complainant_age_incident            NA        NA           NA             NA            NA           NA          NA               NA                        1       NA
precinct                            NA        NA           NA             NA            NA           NA          NA               NA                       NA        1
```

In most cases, there was little or no correlation between these variables. For example, month_closed, the month that the complaint was closed, had no relationship with any of the variables. There was a slight negative correlation with year_closed, but this is likely coincidental.

For some variables, there was a correlation, but this won't be relevant in my model. For example, there is a correlation of .3 between mos_age_incident (the age of the officer at the time of the incident), and complaint_id, year_received, and year_closed. This might be because many of the officers on this list have worked for the police department for several years and have had multiple incidents. Since complaint_ids and years are incremental, officers who've had multiple complaints over the years would be younger for incidents with lower complaint_ids and earlier years and older for incidents with higher complaint_ids and later years. Similarly, shield_no (the officer's shield number) has a correlation coefficient of .2 for complaint_id and year opened/closed. There is a correlation coefficient of 1 between complaint_id, year_received, and year_closed. This makes sense because cases are typically closed within a year of opening, so year_closed is usually equal to year_opened plus 1 or 2. Similarly, case_id values increase as the years increase.

Most of the variables that I expect will be more relevant to my model are categorical variables in the dataset. Many of these variables have more than two potential values. For example, fado_type (category of complaint) includes Abuse of Authority, Discourtesy, Force, and Offensive Language. To prepare for modeling, I used One-Hot Encoding to expand each categorical variable into multiple binary, numerical variables. From there, I created another correlation matrix on the categorical variables, including the "substantiated" variable. Using this method, I found no correlation between most variables and a substantiated disposition. However, there was a very slight positive correlation between female officers and a substantiated disposition and a negative correlation for male officers. To follow up, I ran a Chi-Square test on the one-hot encoded value for female officers and substantiated outcomes. The test resulted in a p-value very close to zero, 2.2e-16. This indicates to me that officer gender might be statistically significant in my model, and I'd like to examine the relationship further going forward.

Based on the correlation matrix, I found no correlation between the race of the officer or complainant and the case disposition. However, the Chi-Square test on officer ethnicity versus disposition and complainant ethnicity versus disposition both resulted in very low p-values, which indicate that a relationship between these variables exists. I also found that White officers and any ethnicity of complainants. However, there was a slight positive correlation of .1 between Asian officers and Asian complainants, Black officers and Black complainants, and Hispanic officers and Hispanic complainants.

The correlation matrix also indicated a slight correlation between the rank of Police Officer and an outcome of substantiated. There was also a negative correlation between the rank of Police Officer and a complaint type of Abuse of Authority, as well as a positive correlation between the rank of Sergeant and a complaint type of Abuse of Authority.   It would not be surprising to find that lower-ranking officers are less likely to receive Abuse of Authority complaints than high ranking officers.

Moving forward, I'd like to continue examining potential relationships between officer gender, ethnicity, and rank, and complainant gender and ethnicity with case disposition. I would also like to spend more time tidying my dataset and resolving class imbalance before beginning modeling.

Sources

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

[2] Eric Umansky and Mollie Simon. The NYPD Is Withholding Evidence From Investigations Into Police Abuse. (August 2020). Retrieved from https://www.propublica.org/article/the-nypd-is-withholding-evidence-from-investigations-into-police-abuse.

[3] Wikipedia. The New York City Police Department. Retrieved from https://en.wikipedia.org/wiki/New_York_City_Police_Department.

[4] Bill Hutchinson. ABC News. Blacks account for nearly half of all NYC arrests 6 years after end of stop-and-frisk: NYPD data. (June 2020). Retrieved from https://abcnews.go.com/US/blacks-account-half-nyc-arrests-years-end-stop/story?id=71412485#:~:text=Data%20from%202020%2C%20shows%20that,are%20targeting%20Blacks%20and%20Hispanics.