Mijente: Police Disciplinary Action Final Report

April 30, 2021
CS506

Members:
Shelli Gorokhovsky
Rachel Peng
Maya Webb
Tina Wang
Daniel Delijani

Code Base
https://github.com/sgorok/CS506Spring2021Repository/tree/master/PoliceConductProject
** Github didn't allow us to push our csv files. All datasets involved in the project can be found here. For convenience, datasets are also linked as they are mentioned throughout the report. **

Final Presentation
https://docs.google.com/presentation/d/1OCDHgOUaFwWh8jVvckeDjVho4zjOJRiu6Qhh5wQVzIQ/edit#slide=id.gd47b19098d_0_1329

1. Introduction

Our client, Mijente, is a political home for Latinx and Chicanx people who seek racial, economic, gender and climate justice. Our goal with Mijente was to better understand how Boston Police Department officers under disciplinary investigation contribute politically to the Boston city council and Mayor. More specifically, we wanted to investigate the following questions:

1. Do BPD officers under investigation attempt to influence the process via political campaign contributions to city lawmakers and decision makers such as city councilors and the Mayor?
2. Do officers under investigation contribute more in total and more frequently?
3. What are the discrepancies between officer rank in all BPD officers and BPD officers under investigation contributions?
4. How are very severe allegations like rape and shooting affecting officers political contributions?

In this report, we will walk you through how we approached these questions and what we ultimately concluded.

2. Data Collection

Our client and PM provided us with several datasets, of which, we used All Police Contributions, Employee Earnings, and LEAD Blacklisted Police Officers.

In addition to the provided datasets, we also created a web scraper to grab the 10 year database of Boston Police Disciplinary Action from the Boston Globe. The scraper created an automated web browser that went through all the Boston Police officers and pulled the following features for each officer's misconducts: type of misconduct, rank, gender, race/ethnicity, name, year, unique case ID, outcome of the investigation, allegations, and findings.

3. Data Preparation and cleaning

I. *Preprocessing Names*

For preprocessing, we had to prepare the two datasets: AllPoliceContributions.csv and BostonPoliceInternalAffairs.csv. We made sure the names were formatted the same to better merge them with fuzzy matching. We created a python notebook named preprocessing.ipynb, which preprocessed the names. For the BostonPoliceInternalAffairs.csv, it was not as complicated because names were *firstName middleName lastName suffix*. However, some names were capitalized and others not so to preprocess this column we simply put the entire name in lowercase.

For the AllPoliceContributions.csv, it was more complicated. It is formatted as *lastName, suffix, firstName middleName*. So, we transformed it into a list separated by "," and depending on how long the list was we rearranged the names in order to make it formatted the same as the disciplinary action dataset. We additionally made these names lowercase as well to match the disciplinary action dataset.

II. *Merging Using Fuzzy Matching and Filtering*

To merge the two datasets, we used the fuzzy matching template.ipynb provided by our PM, Gowtham. Our final code can be found here. The code first splits each of the data frames by the first character of the last name. We then wrote a function called getLastCh(s) to create a column with the first character of the last name. After, we merged two data frames using their lastName characters and applied a string similarity score. For each row, we filtered the string similarity value to create the final dataframe with name matches. This merged subset is then written to a csv file and we repeat this for all last names that start with each character of the alphabet. Lastly, then merge all these subsets back together for the final merge. After, we filtered the merged dataset for people that had listed “Boston Police” as their employer. This merged and filtered dataset can be found here.

With this cleaned data, we overlaid the LEAD blacklist data, which added 54 new names onto our dataset. Finally, we were able to get information about employee earnings from an earnings report made public by the Boston city government. After filtering out Boston police officers’ earnings, we conducted another round of fuzzy matching to merge the updated dataset we created with earnings data.

After exploration of the original merged and filtered dataset, we noticed that there were many duplicate columns. To filter the dataset further, we dropped all rows that had the same values for: Name, TypeOfMisconduct, Race, Rank, Allegation, Finding, Amount, Recipient, as these rows surely represent repeats of the same transactions. The final filtered dataset we used for analysis is BPIA-APC-LEAD-EARNINGS.csv.

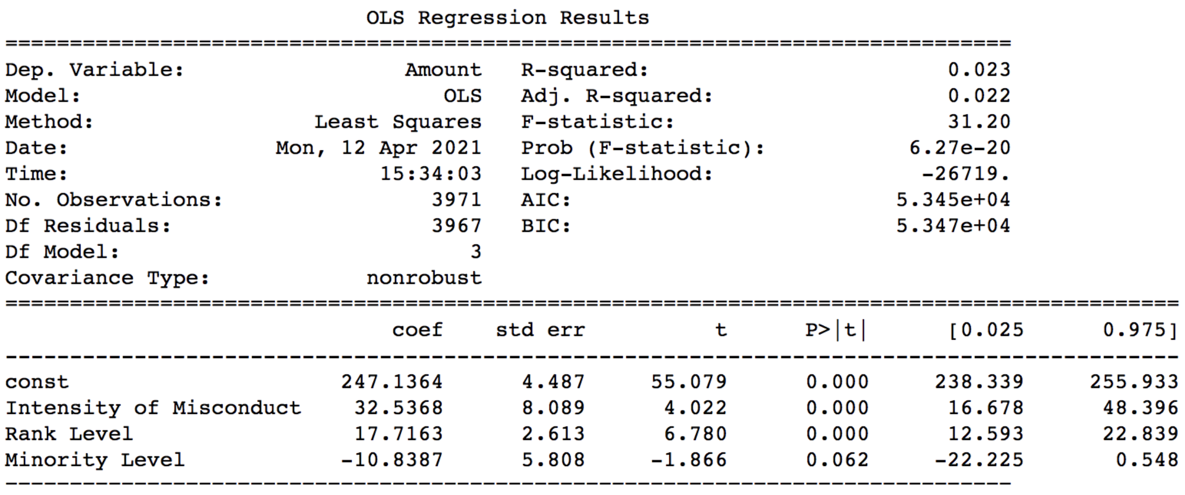
4. Analysis

I. Initial Observation

First, we computed the contributions to political campaigns made by all Boston police officers and all indicted Boston police officers respectively. This gave us some initial insight regarding the patterns of officers who make criminal offenses and their inclination to donate money to campaigns. We found that in our dataset, roughly a third of total contributions were made by indicted officers, which composed over half of the total officers from the set.

II. Linear Regression

We ran a linear regression model on our final merged dataset, filtered for unique contributions. We decoded categorical variables to obtain meaningful features for our dataset, ultimately using them to develop data points for Intensity of Misconduct, Rank Level, and Minority Level. We ranked some by intensity and others by 0 and 1 (dichotomous). Our code and results could be found more in depth here.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Amount   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.022
Method:                 Least Squares   F-statistic:                     31.20
Date:                Mon, 12 Apr 2021   Prob (F-statistic):           6.27e-20
Time:                        15:34:03   Log-Likelihood:                -26719.
No. Observations:                3971   AIC:                         5.345e+04
Df Residuals:                    3967   BIC:                         5.347e+04
Df Model:                           3
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                     247.1364      4.487     55.079      0.000     238.339     255.933
Intensity of Misconduct    32.5368      8.089      4.022      0.000      16.678      48.396
Rank Level                 17.7163      2.613      6.780      0.000      12.593      22.839
Minority Level            -10.8387      5.808     -1.866      0.062     -22.225       0.548
------------------------------------------------------------------------------
```

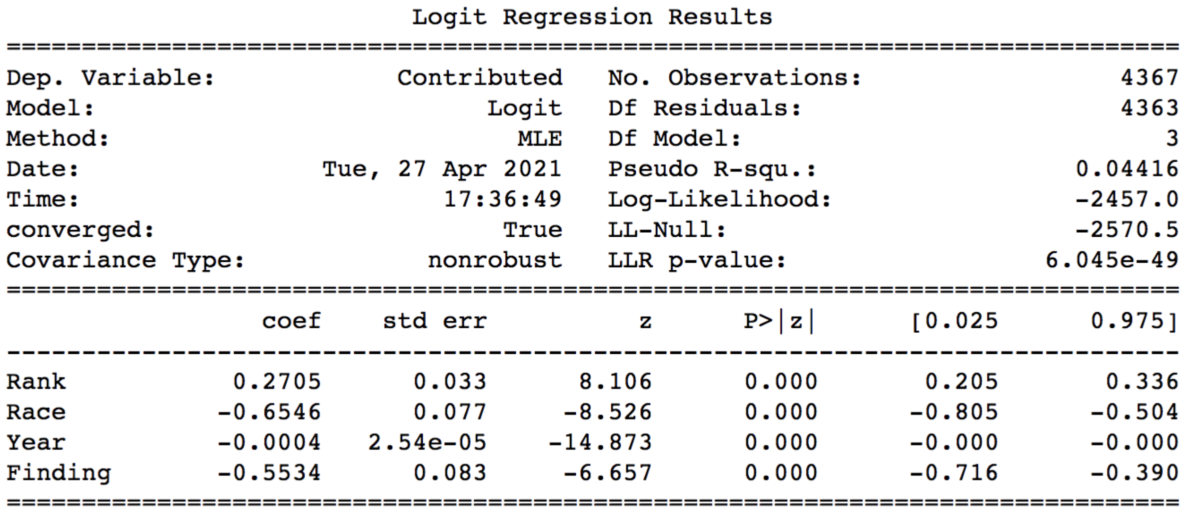
From this linear regression, we found these variables (which were renamed) to be significant for significance level of 0.05: Intensity of Misconduct, Rank Level, Minority Level (if we used a significance level of 0.10). The Rsq is very low -- 0.023, indicating that the model might not be the best predictor for the response variable, amount contributed.

III. Logistic Regression

After discussing with our client, they wanted to see a logistic regression instead of the linear regression we had. We built a logistic regression to predict the likelihood of a person to have contributed based on a number of attributes (code and detail results found here). First, we loaded two datasets, BostonPoliceInternalAffairs.csv (data containing Boston Police Disciplinary Action) and BPIA-APC-LEAD-EARNINGS.csv (data we merged and filtered containing contributions). We then found the names in common in both datasets. Our results show that there were 398 unique names in common with 27% of the BostonPoliceInternalAffairs.csv dataset who contributed.

With these names, we created a new attribute within BostonPoliceInternalAffairs.csv dataframe called "Contributed" with 1 if the person did contribute and 0 otherwise. We then re-encoded the categorical variables to prepare the dataset for logistic regression modeling.

Initially, we created a logistic regression model with 6 attributes: Rank, Race, Gender, Year, TypeofMisconduct, and Finding. The results from this model showed that 4 of these variables had significant predictors. Next, we created a logistic regression model with only these 4 variables: Rank, Race, Year, and Finding.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:            Contributed   No. Observations:                 4367
Model:                          Logit   Df Residuals:                     4363
Method:                           MLE   Df Model:                            3
Date:                Tue, 27 Apr 2021   Pseudo R-squ.:                 0.04416
Time:                        17:36:49   Log-Likelihood:                -2457.0
converged:                       True   LL-Null:                       -2570.5
Covariance Type:            nonrobust   LLR p-value:                 6.045e-49
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Rank           0.2705      0.033      8.106      0.000       0.205       0.336
Race          -0.6546      0.077     -8.526      0.000      -0.805      -0.504
Year          -0.0004   2.54e-05    -14.873      0.000      -0.000      -0.000
Finding       -0.5534      0.083     -6.657      0.000      -0.716      -0.390
==============================================================================
```

From these results, we see that the logistic regression equation that could predict whether the person contributed or not is: probability contributed = 0.3075Rank -0.6816Race - 0.0004Year - 0.6297Finding. From these coefficients, we can see that higher the rank, the more likely the person is to contribute. Race is encoded 0 for white and 1 for minority. From the coefficient predictor, we can see that it is negative indicating that if the person is white, they are more likely to contribute. Additionally, Year's coefficient is very close to 0 indicating that it most likely does not influence whether or not a person contributes.

IV. One-Hot Encoding

Next, we used One-Hot Encoding to relabel 3 categorical variables -- Race, Finding, Outcome -- using dummy variables. For the model with Race encoded, we can see that the predictors for Black, Hispanic, and White categories for Race are all significant. All 3 predictors are negative, but black has the largest quantity of negativity indicating that if the person if black, they are less likely to have contributed. For the model with Finding encoded, we can see that there are 4 significant categories for Finding: Exonerated, Not Sustained, Sustained, Unfounded. All predictors are negative ranging from -2 to 0. Lastly, from the model with Outcome encoded, we see that only 4 of the 12 categories for Outcome have significant predictors: Oral Reprimand, Retired/Resigned, Suspension, and termination. (Detailed results can be found here.)

5. Visualizations

We began our visualizations with charts showing the data we have, and if there are any relationships already showcased. To begin, we went through our filtered dataset to get the total amount contributed by each officer, and removing any duplicate dates an officer contributed. We added this new value as a column to the data set called "Total Amount." Then, we wanted to see if there existed a pattern of any particular race, type of misconduct, or rank having an effect on the total contribution. Our results can be found in Figures 1-5 on the following pages.

Figure 1 demonstrates the relationship between officer demographics and prevalence in our datasets. The chart on the left is the distribution by race of all Boston area officers under investigation, and the chart on the right is the distribution by race of Boston area officers who are both under investigation and have made a financial contribution. As one can see, the proportion of white officers is greater in the second chart than in the first and in contrast the proportion of black and hispanic officers is reduced; additionally, it appears that none of the Asian officers under investigation made any financial contributions.

Figure 2 demonstrates the relationship between the rank of the officers under investigation and their likelihood to make financial contributions. It appears that those whose roles are just defined as “police officers” make up most of the officers under investigation. However, detectives and higher ranking officers who are under investigation are more likely to make contributions. You can also view the ranks of all the officers who made financial contributions in the final chart in this figure. This was much less convincing of a diagram, mainly because the donors self-reported their role under “occupation,” and there were 570 different responses to this. As such, a large chunk of the occupations were not classified as any of the five categories we created, but instead fell into “other.”

Figure 3 shows the frequency of each of the six possible findings for a case: Sustained, Not Sustained, Exonerated, Filed, Withdrawn, and Unfounded. Again, on the left is the distribution of these case findings for all Boston officers accused of misconduct, and on the right is the distribution for Boston officers who were accused of misconduct and who also made contributions. It appears that cases of officers who made financial contributions are more likely to be unfounded than the cases of all officers. Still, in both the general set and the subset, the most likely outcome is that the case was not sustained.

Figures 4 and 5 show the median number of times that a contribution was made by accused officers grouped by race and rank, respectively. This allowed us to see that white and black officers both contributed a median of 2 times, while hispanic officers contributed a median of one time. It is important to note that there are almost 9 times more white officers than hispanic officers in this dataset, and about 5 times as many white officers than black, so that has an effect on this analysis. We can also see that accused higher ranking officers, especially Detective Lieutenants, contribute more frequently than lower ranking officers. However, again, officers with these higher ranking titles appear less in our dataset; therefore, this disparity may be caused by very few individuals.

Figure 1. Demographic of Officers Under Investigation

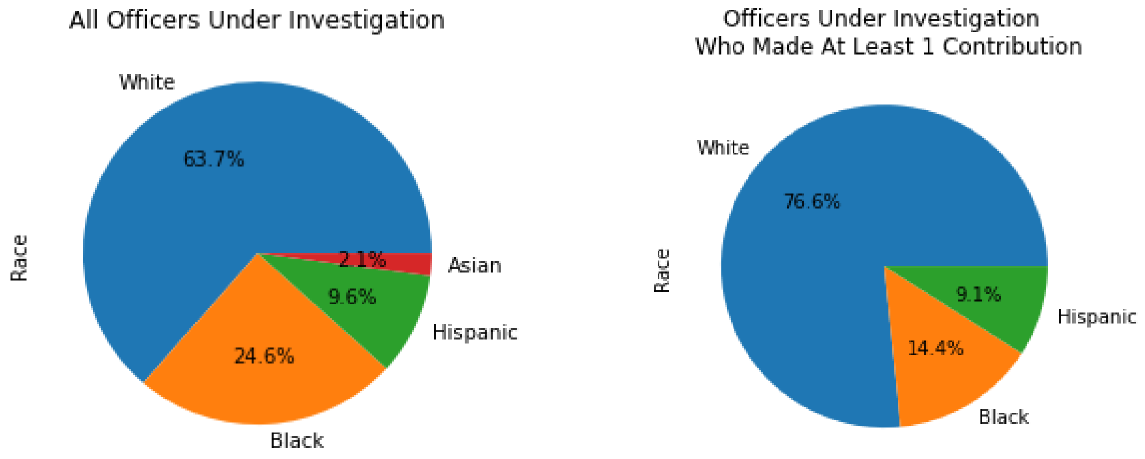


Figure 2. Ranking of Officers Under Investigation

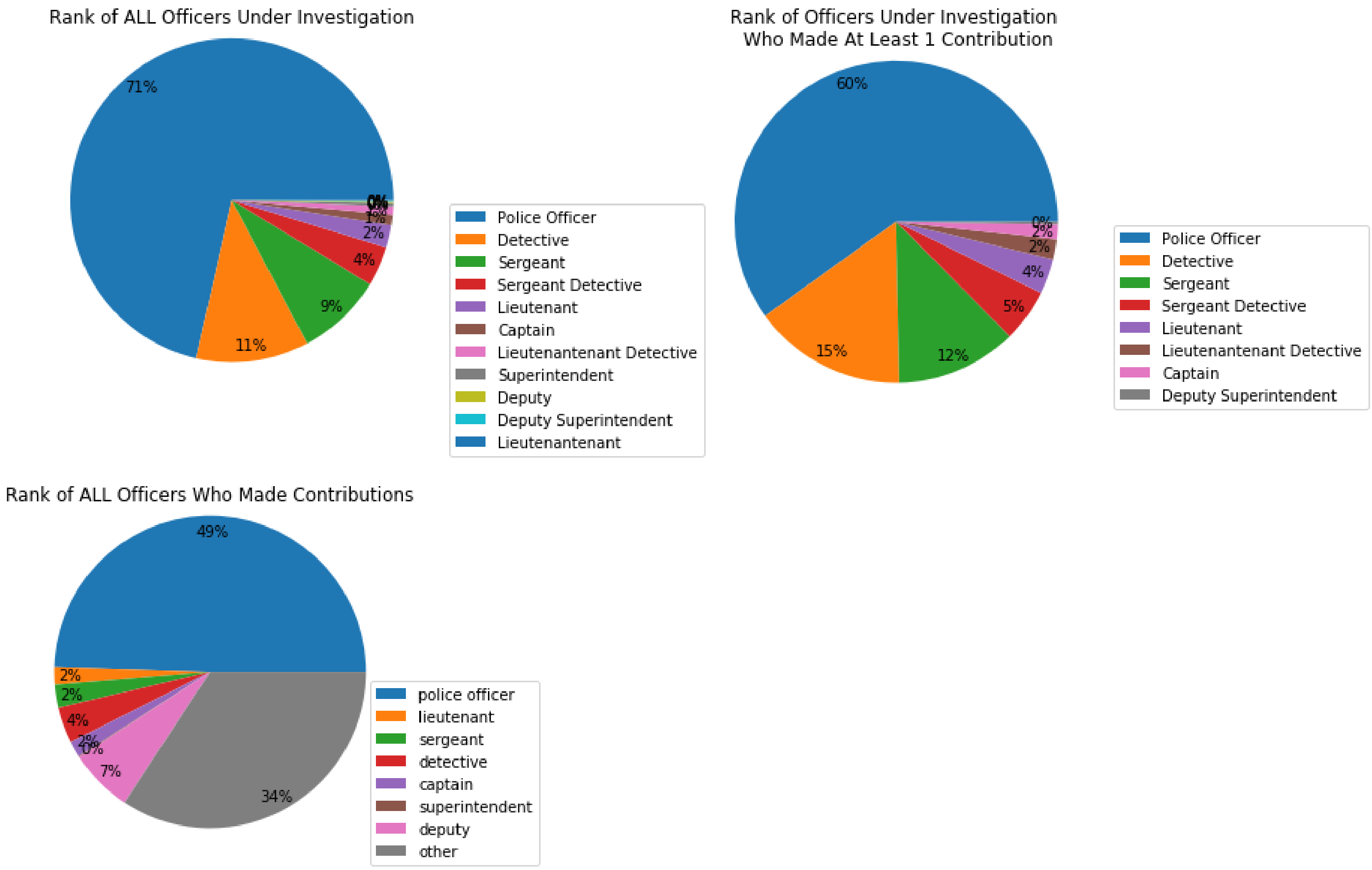


Figure 3. Frequency of Case Findings of Officers under Investigation

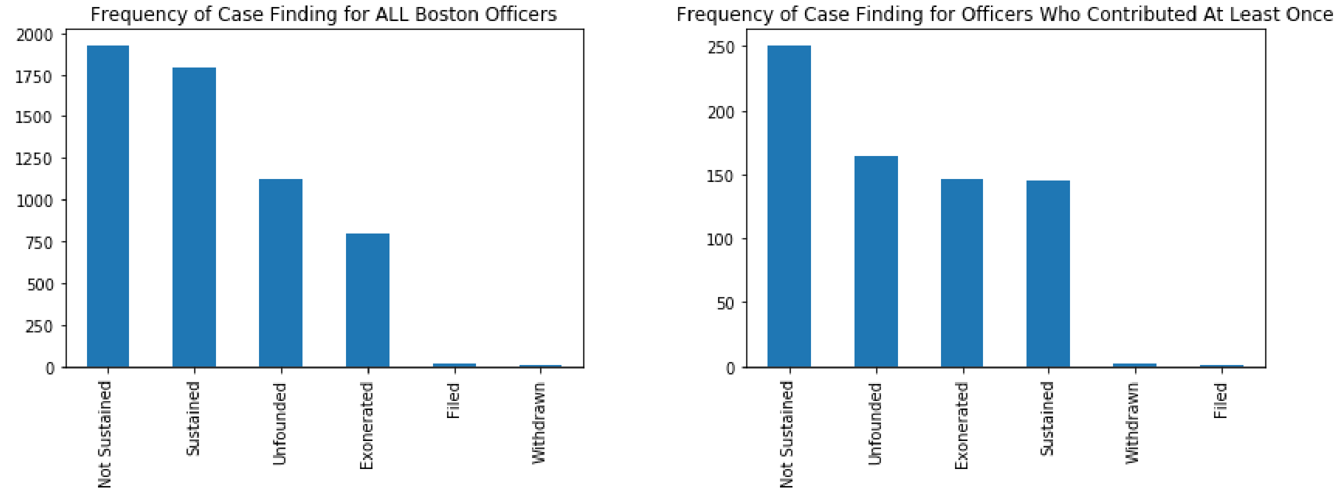


Figure 4. Median Contributions of Each Race

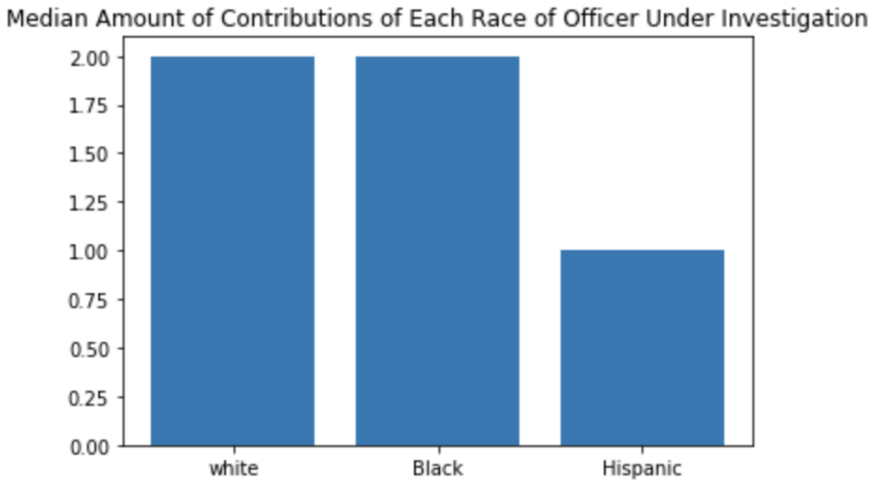
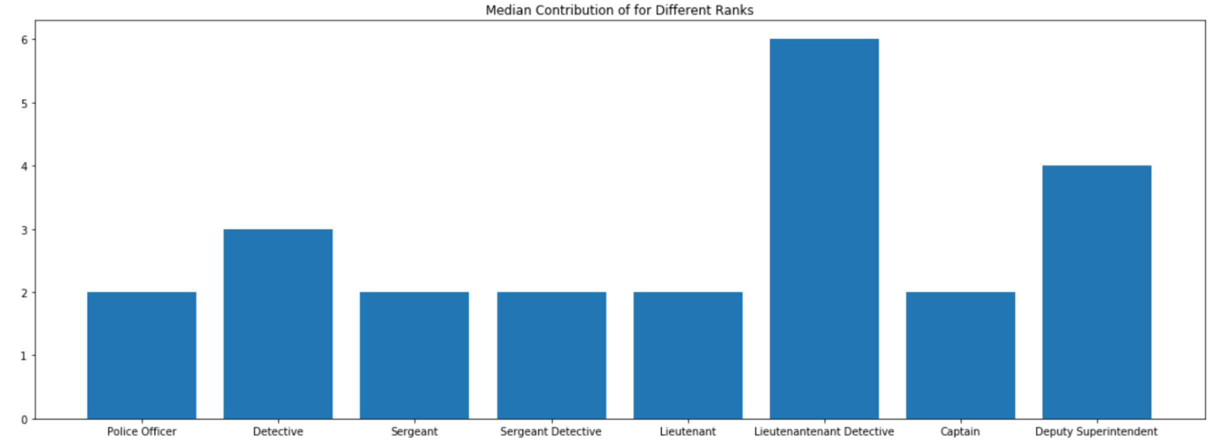


Figure 5. Median Contribution by Rank



6. Conclusion

I. Summary of Results

After our analysis and visualizations, we can say with high confidence that rank is a significant predictor of the likelihood that an officer contributed. Our logistic regression indicated a coefficient of 0.27 for rank. Our analysis also showed correlation between race and the likelihood that the person contributed. We saw a negative predictor of -0.65 for Race in the logistic model in which categories were relabeled 0 for white and 1 for minorities (Black, Hispanic, and Asian). When using one-hot encoding, we saw that the predictor for Black is slightly lower than White and Hispanic, with the predictor for Asian being not significant. Additionally, we found that 27% of all disciplined Boston area officers made at least one contribution and that 32% of all Boston area police donations were made by officers who had at least one misconduct. This gives us direct insight into our strategic questions of whether officers under investigation contribute more. Specifically, we saw that the higher the rank of the officer, the more likely they had contributed and that if the officer is White, they are also more likely to have contributed.

II. Additional Notes Regarding Challenges, Confidence, & Calculations

Over the course of our project, our team worked very hard to accurately define our data, recognize and tackle any flaws, and validate our work. Even so, we are aware of potential misguided assumptions that were made in our analyses. Due to the difficulties posed by complex data extraction and cleaning, we are not 100% confident in all of the values given by some of our analyses. For example, towards the end of our project, we discovered that CaseIDs are not unique in the dataset; repeated values likely signify multiple individuals involved in the misconduct and multiple allegations. This could have contributed to some misrepresentations of data distribution, since in some cases the Case ID was used to filter other features. Likewise, some of our analysis was deterred by something as simple as spelling errors that were not identified early on. You can see in the pie charts in Figure 2 that the spelling of Lieutenant, for example, was misspelled in some rows and likely led to classification errors since it created a separate category for the misspelled roles.

III. Going Forward

Despite having compiled concrete datasets and results, much more can be done to investigate our question of whether officers under disciplinary action or misconduct contribute politically more. There are many categorical variables of interest that we did not have the time to investigate. Specifically, Allegation had many categories and we had difficulty ranking them. If more research and investigation can be done regarding how to rank Allegations, more interesting relationships can possibly be found. Additionally, we could broaden our datasets and analysis to the Massachusetts state as a whole and eventually the nation as a whole instead of only Boston itself. Overall, we were able to successfully clean and preprocess our multiple datasets for analysis, merge datasets of interest together, and conduct analysis and modeling of those finalized datasets. We are satisfied and proud of the project we have produced this semester and hope our project helps Mijente going forward.