# Final Report Draft - Mijente: Police Disciplinary Action

By: Shelli Gorokhovsky, Rachel Peng, Maya Webb, Tina Wang, Daniel Delijani

Code Base:
https://github.com/sgorok/CS506Spring2021Repository/tree/master/PoliceConductProject

## Data Collection

To collect the data, we created a web scraper for the 10 year database of Boston police disciplinary action from the Boston Globe (code and csv found in Deliverable 1). The features of this dataset includes: type of misconduct, rank, gender, race/ethnicity, name, year, unique case ID, outcome of the investigation, allegations, and findings.

For the web scraper, which is found in our repository under "webscraper.ipynb", we used the Selenium webdriver package in Python, as well as the webdriver_manager.chrome package. These worked together to create an automated web browser that is able to get the URL of the Boston Globe database from which we want to collect the data, and extract the information necessary based on the features we are interested in analyzing. The data collected by the webscaper can be found as a comma-separated value file in our repository titled BostonPoliceInternalAffairs.csv.

Additionally, our client and PM provided us with several datasets, which can be found here. The datasets include campaign finance data. In particular, we used the All_Police_Contributions.csv. We also downloaded the Employee Earnings Report found here.

## Data Preparation and cleaning

For preprocessing, we had to prepare the two datasets: disciplinary action database and the BPD financial contributions data. We made sure the names were formatted the same in order to allow us to better merge them with fuzzy matching. We created a python notebook named "preprocessing.ipynb", which preprocessed the names. For the disciplinary action dataset, it was not as complicated as names were *first name middle name last name suffix*. However, some names were capitalized and others not so to preprocess this column we simply put the entire name in lowercase. For the contributions dataset, it was more complicated. It is formatted as *last name, suffix, first name middle name*. 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.

For fuzzy matching, we used the "fuzzy matching template.ipynb" provided by our PM, Gowtham. 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. Then, 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 and then we repeat this for all last names that start with each character of the alphabet. We then merge all these subsets back together for the final merge. Finally, 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 this, we noticed that there were many duplicate columns. So, 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 can be found here.

**Analysis**

*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.

*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.

*Next Steps*

To build a better model, we plan to use one-hot encoding for the “Finding” variable instead of dichotomous encoding we have now. We are also working with our client to get the best ranking for the “Allegations” variable. Additionally, we will build a logistic regression to predict the probability of a police officer to donate given a set of attributes. For this, we will use the original dataset, which included data for police officers who did not donate.

**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 were any particular race,  the type of misconduct, or rank had an effect on the total contribution. Our results can be found in the figures on the next page.

Going forward we plan on continuing the visual analysis by interpreting the relationships between variables such as the frequency of political contributions and severity of misconduct, the median amount of money contributed to campaigns by indicted officers, and the amount contributed by all police officers compared to the amount contributed by indicted officers. As we work to understand whether officers accused of misconduct use politics to influence their punishments or lack thereof, there are many factors to consider. Understanding relationships

between these variables will allow us to recognize the underlying patterns and, possibly, confirm this hypothesis.

**Figure 1. Average Contribution Made By Each Race (White, Black, Hispanic)**

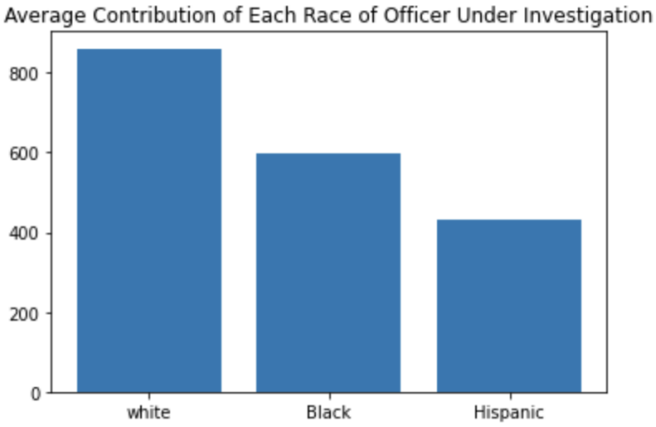


**Figure 2. Average Contribution by Type of Misconduct**

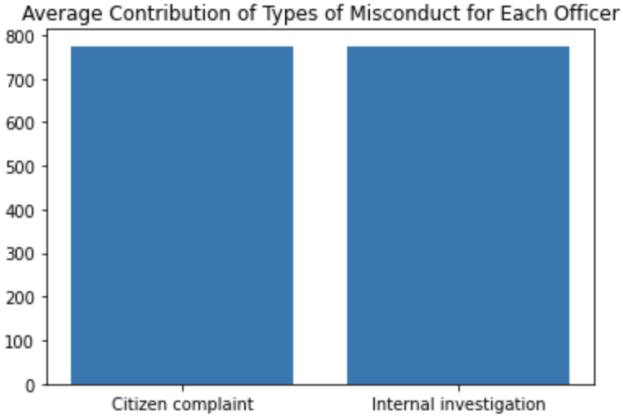


**Figure 3. Average Contribution by Rank**