Data Science Seminar Final Report

The Brutal Dragons
Hong Hong, Hui Ye
December 3, 2022

Project Introduction

Monitoring and examining complaints against the police regulatory system and various data characteristics, can provide a rare glimpse into the operations of the police department charged with protecting and serving the public at the level of data.

However, as racial issues have evolved, the racial disparities reflected in police enforcement have become increasingly problematic. What was once a police department that the public trusted and relied on heavily has grown increasingly tense with civilians over the years.

According to statistical surveys, police are likely to favor unfair treatment of people of color, which has led to multiple allegations of police misconduct by civilians. It is clear that increasing diversity in the composition of police departments is conducive to better police-civilian relations, and at the same time such police reform is often mentioned in discussions.

Therefore, the theme of our group project was to conduct a study of police composition diversity and discuss the relationship between police diversity and their misconduct rates, where diversity primarily includes the race and gender of police officers. We initially explored the CPDP database to identify some of the classification thresholds needed in the study; followed by further exploration of the data features through data visualization methods, and finally we used two types of machine learning models to make classification predictions on the data. We hope this study will serve as an important reference for those who wish to pursue diversity in police composition. In addition to this, during the experiment, we will focus on communities with similar racial composition to ensure that police misconduct rates are not influenced by the racial distribution of the district.

Database Relational Analytics

Our earliest findings were obtained by querying the CPDP database. In this section, we focused on obtaining some thresholds for classifying police diversity, such as by calculating the rate of on-duty police misconduct in the target area, and the percentage of that non-white police officer as well as non-male police officers.

	avg
1	0.16099523809523809524

Figure 1 - Average of the misconduct rate by police-districts

As shown in the above Figure 1, the average of the misconduct rate by police-districts is around 16%. Therefore, we define mild misconduct as 0%~10%, medium misconduct as 11%~25%, and a misconduct rate more than 25% would be considered as severe misconduct.

In order to classify the level of diversity in the composition of the police, we likewise calculated the average of the share of different ethnic groups in the police department as a whole. It can be seen from Figure 2 that the average proportion of all white police officers is as high as 58.8%, while the average proportion of the remaining non-white police officers is only 41.2%. This ratio is actually quite amazing. We can understand that the average police force in an area is more than half of the white police.

At the same time, among the non-white police, black police have the highest proportion, reaching 23.9%, followed by hispanic accounting for 15.3%, and the proportion of Asian police is the lowest, only 2%.

Therefore, we define low diversity when the proportion of non-white police officers is 0%～30%.When the proportion of non-white police officers is 31%～55%, it is medium diversity, and when the proportion of non-white police officers is higher than 55%, it is high diversity.

avg_nonwhite	avg_white	avg_black	avg_hispanic	avg_asian
0.41217916666666666667	0.58782083333333333333	0.23869166666666666667	0.15278333333333333333	0.0184625

Figure 2 - Average race proportion of police officers

avg_female	avg_male
0.17894166666666666667	0.82105833333333333333

Figure 3 - Average gender proportion of police officers

After determining the relevant classification criteria, we sampled, compared, and grouped the data for the policing areas of primary interest in this study, i.e., 25 districts, using the race composition of the primary resident population as the criterion. Finally, five groups of districts with similar composition were selected, of which two districts in each group had different demographic characteristics between the groups.

Figure 4 lists the subgroups we selected and their specific demographic composition data, while these five groups of districts will be the input to our next study.

	district	non_white	black	white	hispanic	female	district_misconduct_rate
1	8	0.2346	0.0603	0.7654	0.162	0.1441	0.1726
2	11	0.4249	0.2409	0.5751	0.1678	0.1759	0.1509

(a) Black-dominant district 8 and 11

	district	non_white	black	white	hispanic	female	district_misconduct_rate
1	9	0.2952	0.0855	0.7048	0.189	0.1752	0.1596
2	14	0.4439	0.0697	0.5561	0.3367	0.1684	0.1905

(b) Black-dominant district 9 and 14

	district	non_white	black	white	hispanic	female	district_misconduct_rate
1	2	0.7816	0.7038	0.2184	0.0627	0.223	0.046
2	15	0.4587	0.2698	0.5413	0.1603	0.1841	0.06

(c) District 2 and 15

	district	non_white	black	white	hispanic	female	district_misconduct_rate
1	18	0.3156	0.197	0.6844	0.0947	0.124	0.0756
2	5	0.6243	0.5573	0.3757	0.0587	0.2626	0.0609

(d) White-dominant district 5 and 18

	district	non_white	black	white	hispanic	female	district_misconduct_rate
1	24	0.2491	0.0596	0.7509	0.1318	0.1823	0.1014
2	6	0.6936	0.5554	0.3064	0.1258	0.2229	0.0509

(e) Hispanic-dominant district 6 and 24

Figure 4 - Five groups with similar residents' race composition

From the above figures it is actually clear that districts with a predominantly black residential population tend to have higher rates of police misconduct than other districts. At this stage, we hypothesize that the rate of police misconduct may be related to the number of whites in its composition and also to the race of the population residing in the area it police.

Interactive Visualization

To gain a deeper understanding of the relationship between the diversity of police officers and misconduct rate, we demonstrate two types of interactive visualizations through D3.js. All the code is open source on Observable.[1][2]

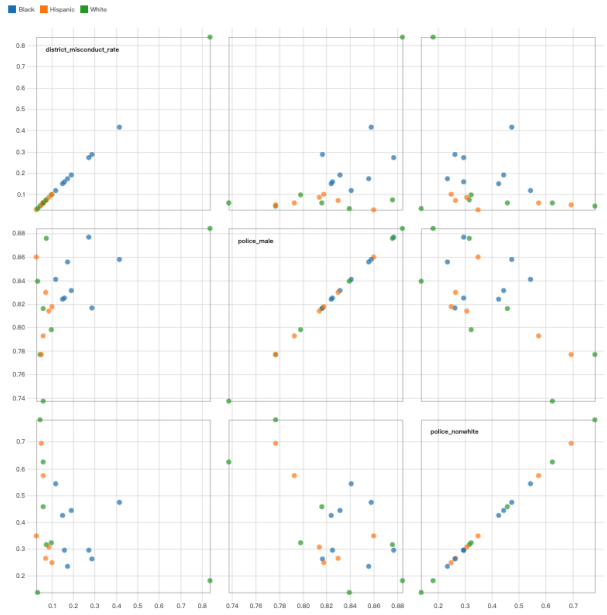


Figure 5 - Overview of brushable scatterplot matrix

[1] https://observablehq.com/@honghong1012/checkpoint-3-interactive-visualization-with-d3-js-1
[2] https://observablehq.com/@honghong1012/checkpoint-3-interactive-visualization-with-d3-js-2

Figure 5 shows the first interactive visualization we have implemented. In fact, as we brush around the graph and make selections, there are many features of the data that corroborate the validity of our hypothesis.

As seen in Figure 6, areas with predominantly black populations tend to have higher rates of police misconduct. Among these areas, the majority of police officers on duty are male, accounting for more than 82%. At the same time, the proportion of non-white police officers in these areas is distributed between 40%-50% and 20%-30%, and according to our previous definition, the diversity of police composition in these areas is at low level and medium level. When we select areas with high level of police composition diversity as in Figure 7, we can see that the misconduct rates in these areas are are at low level.

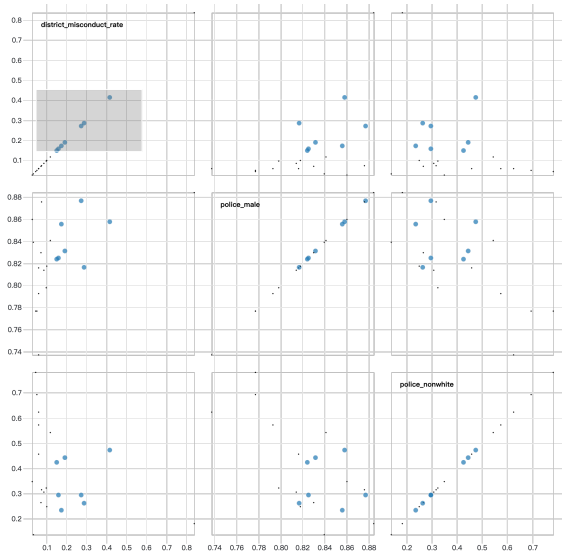


Figure 6 - Brushable scatterplot matrix

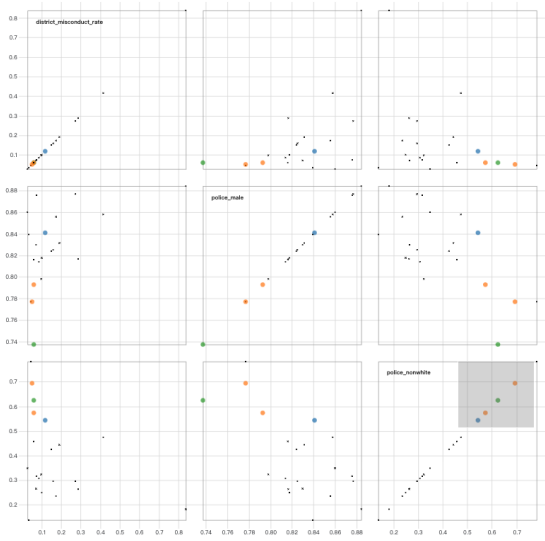


Figure 7 - Brushable scatterplot matrix

Also, in order to further investigate the rate of police misconduct in different ethnic neighborhoods, we conducted the experiment in the figure below. From Figure 8 we can clearly see that predominantly white neighborhoods generally have lower levels of police misconduct. However, there is an interesting finding from Figure 9 that predominantly white areas have the highest rates of police misconduct (as shown in the green circle selected in Figure 9(a) below). This area also shows the highest percentage of male police officers at 88% and a lower percentage of non-white officers at no more than 20%. Also, we can see in Figure 9(b) that one white district has a fairly low rate of police misconduct. In terms of police diversity, this district has a relatively low percentage of male officers and the highest percentage of non-white officers.

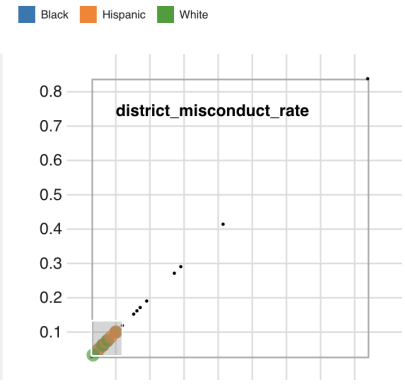
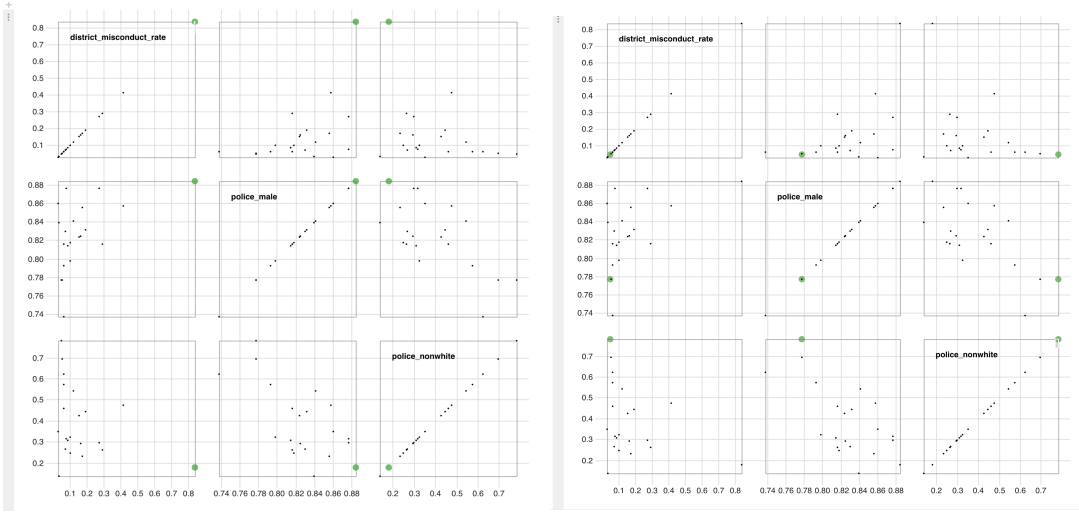


Figure 8 - Brushable scatterplot matrix

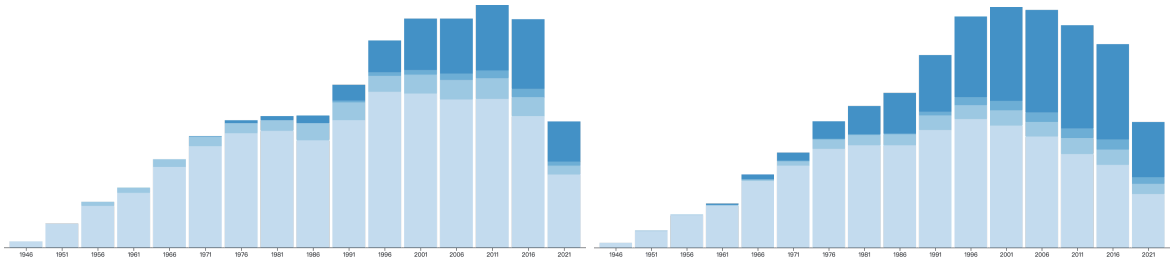


(a) (b)

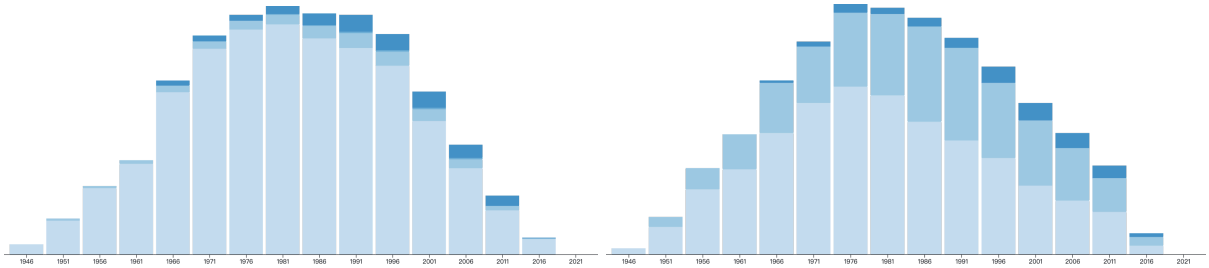
Figure 9 - Brushable scatterplot matrix

Next, we use the stacked bar charts in Figure 10 to look at the proportion of police composition in different districts. In terms of overall proportions, the proportions of police officers actually differ for districts with similar dominant racial composition, from which we cannot discern the exact distribution. However, we can see that in our visual analysis of these 10 districts, white

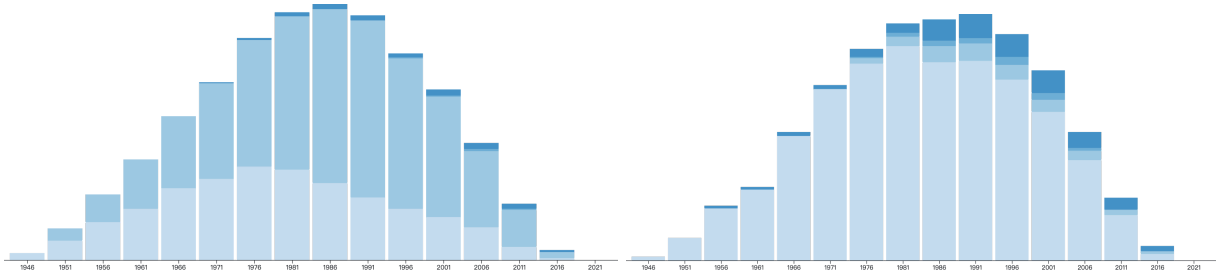
and black police officers are in the majority. And the number of Hispanic officers is very small compared to white officers and black officers, and the number of Asian officers is close to zero. Secondly for the overall trend, the number of police officers rose and fell in roughly similar trends. In particular, for similar areas, we can observe very similar trends in the bar chart. District 9, which has 14 officers, and District 2, which has 15 officers, both show an upward trend in the total number of police officers. For the other districts, the total number of police officers starts to rise in 1946, peaks in 1980-1990, and then declines.



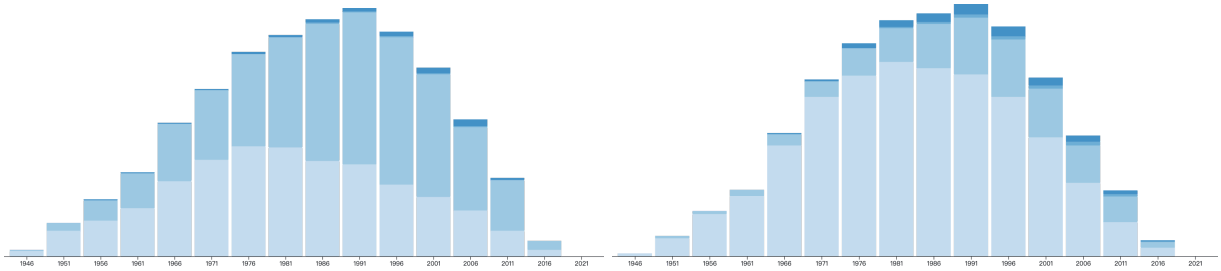
(a) Black-dominant district 9 and 14



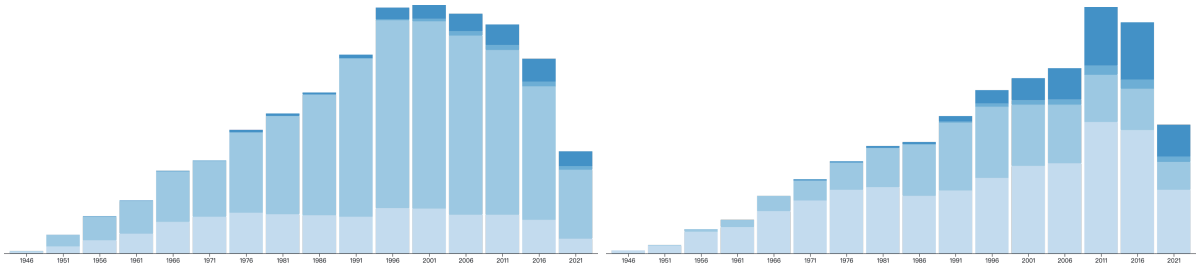
(b) Black-dominant district 8 and 11



(c) Hispanic-dominant district 6 and 24



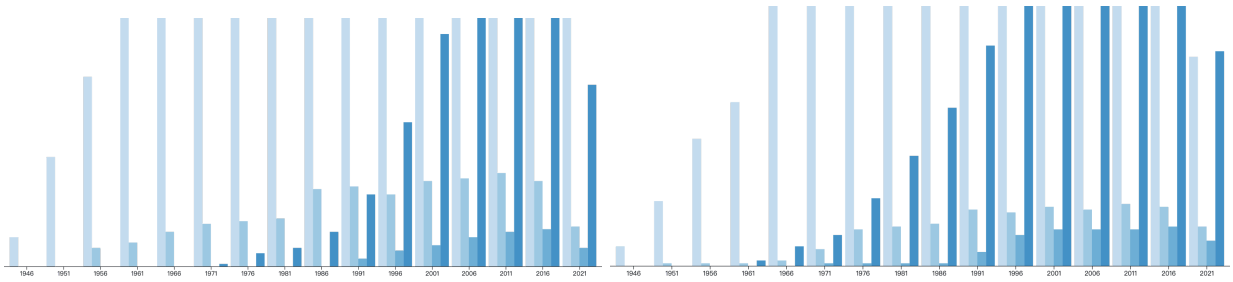
(d) White-dominant district 5 and 18



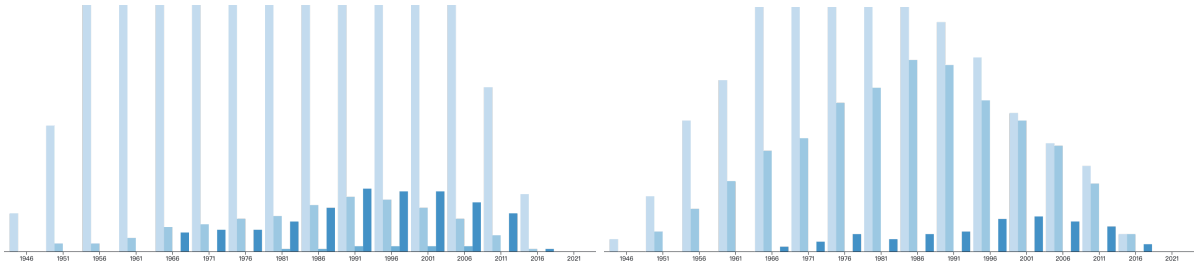
(e) District 2 and 15 (Asian population relatively higher)

Figure 10 - Stacked bar charts

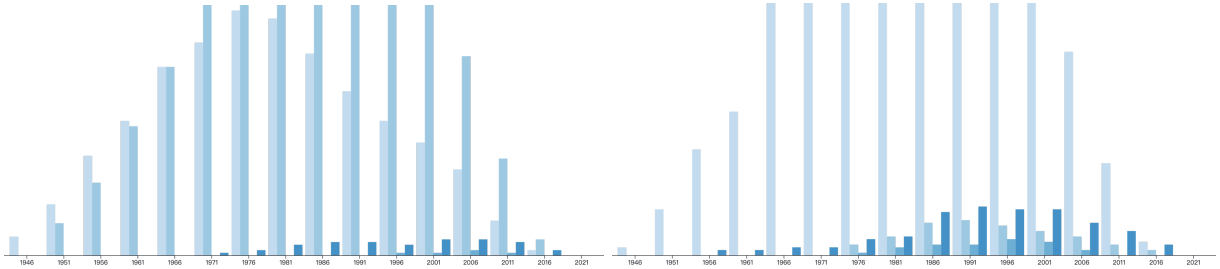
Finally, we can look at the grouped bar charts in Figure 11, (the four groups from left to right are white, black, Asian, and Hispanic). It is reassuring to see that the diversity of police proportions is mostly increasing over time. In predominantly black areas, the proportion of black police officers has also increased over the years. Correspondingly, the number of white officers has remained stable in most years, which may be a positive indication that police departments consciously increase the diversity of their officers when recruiting new officers or assigning officers to duty in the corresponding areas.



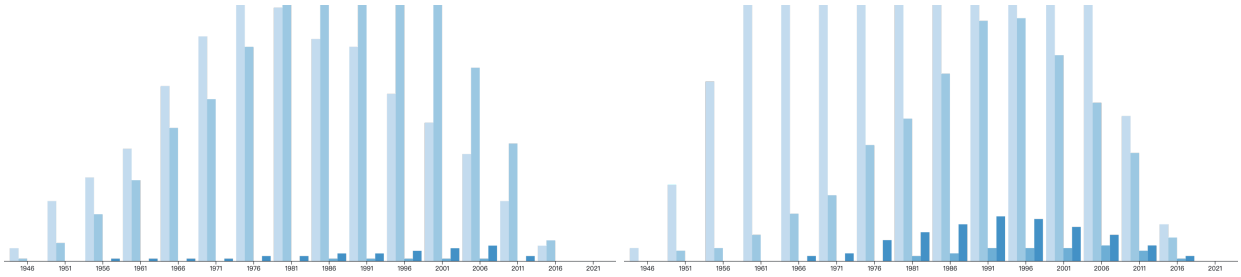
(a) Black-dominant district 9 and 14



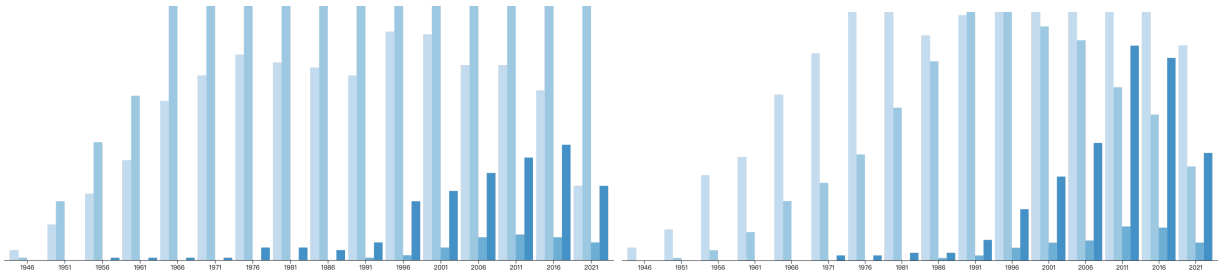
(b) Black-dominant district 8 and 11



(c) Hispanic-dominant district 6 and 24



(d) White-dominant district 5 and 18



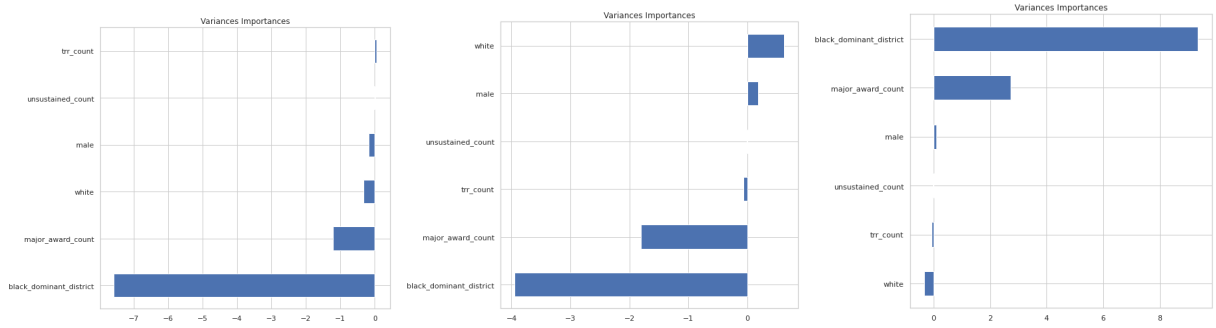
(e) District 2 and 15 (Asian population relatively higher)

Figure 11 - Grouped bar charts

At this stage, we could tell from the interaction with the graph that the predominantly black districts had the highest level of misconduct rate on average. And looking at the racial and gender diversity in them, the diversity level of these districts is at low level or at medium level, reflecting to some extent the relationship between the misconduct rate and the diversities of the police officers on duty.

Machine Learning Prediction

In this section, we used a multicategorical logistic regression model as well as a decision tree model for classification prediction. In addition to this, we also extended the input data to 25 districts in an attempt to find greater commonality in the data. The accuracy of our logistic regression model and decision tree model is 87.83% and 70.89% respectively, which proves the reliability of our model in classification.



(a) Misconduct_rate_level = ‘low’ (b) Misconduct_rate = ‘medium’ (c) Misconduct_rate = high

Figure 12 - Variance importance in multicategorical logistic regression model

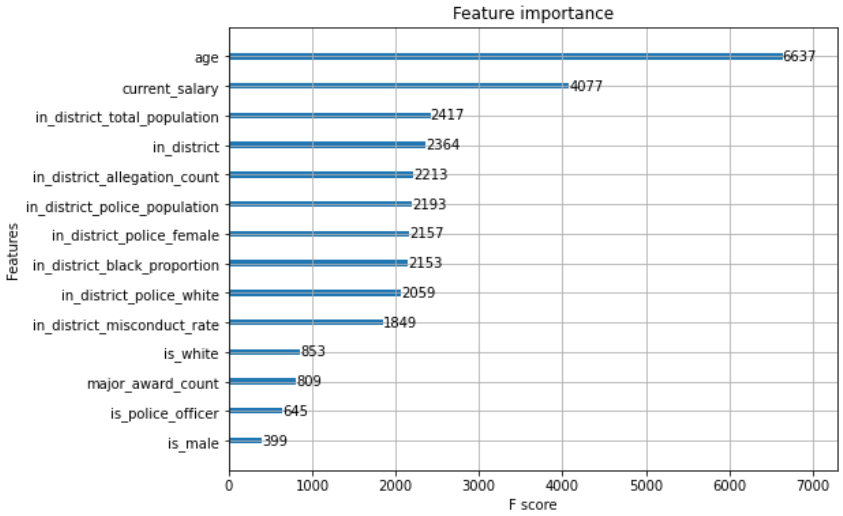


Figure 13 - Feature importance in decision tree model

We focused on the impact of the model input data on the weighting of the model results, in which we were able to argue that our hypothesis in checkpoint1 is reasonable: the race and gender of the police composition has an impact on the area in which they are on duty. For example, we can see from Figure 12 that the higher the percentage of male police officers. The higher the percentage of male police officers, the higher their level of misconduct rates.

The most important finding is that misconduct rates are higher in predominantly black areas compared to other areas, which in turn is the most important factor in the model predictions, confirming what we have seen reported in class and consistent with what we concluded in the interactive visualizations section.

Combining the top factors with higher weights in Figure 13, we were able to surprisingly find that whether a police officer is white or male is less important than the race and gender ratio of the police unit they work in. Surprisingly, the occurrence of misconduct is more likely to be influenced by the group rather than the individual.

Discussion and Future Work

Through the project, we found that the diversity of police composition have a significant impact on the misconduct rates in the area they police. In particular, when their management area is a predominantly black residential area, when the police in that area tends to be less diverse (a lower percentage of non-white officers, cf. database relational analysis section), the misconduct rate in that area is high, at a moderate or severe level.

Although simple data visualization and machine learning cannot give very strong evidence, to some extent we found a positive effect of police diversity on reducing misconduct rates in the governed area. In addition to this, we were even more surprised to find that whether a police officer is white or male is less important than the race and gender ratio of the police unit they work in.

In this project, we also ended up with some phenomena that we have not yet explored in depth, such as our data reflecting that the emergence of police misconduct is likely to be influenced by groups rather than individuals, which could be taken as directions for future work. Second, we also found an unusually high misconduct rate (Figure 14) for some specific districts when computing queries on the database. With more time, we can conduct a deeper analysis by combining the geographic location of these districts, police departments, and specific misconduct events.

21	59458	6770	0.1139	274	0.4157
23	62781	13617	0.2169	259	0.8375

Figure 14 - District 21 and district 23