# Checkpoint 3: Interactive visualization and data exploration

Team: The Powerful Turtles

## Introduction

In this project, we would like to study how the demographics comparison between the complainants and the respective police officers reported correlating to the number of complaints. In this checkpoint, we perform interactive data visualization of race matching correlation between officers and complainants relative to the community race distribution. The results reveal uneven race distribution of officers compared to their corresponding community, and the race difference between officer and complainant in their corresponding community may contribute to higher numbers of complaints. The results from this checkpoint provide us some insights for selecting parameters for the machine learning checkpoint.

## Relational Analytics Questions

In this checkpoint, we would like to address the following questions:

- *Stacked-to-Grouped Bars showing the relationship between officer race and the community population race.*

- *A scatter plot matrix showing the relationship between officer race percentage within community vs. complaint filer race percentage within the community.*

## Results

- ***Stacked-to-Grouped Bars showing the relationship between officer race and the community population race.***

  We first used SQL to select the officers who have complaints and who served the communities of the same race. Only officers in white, Black, and Hispanic were left, which means that officers in other races either have no complaints or serve communities of different races. For example, Asian officers are less likely to receive complaints, which in turn indicates that they show less misconduct to the public.

  For the data shown in this and the next question, the officer-community matching and complainant-community matching are quantified by obtaining the percentage of the race in a community, which the race matches with the race of the officer/complainant. We can translate both the officer and the complainant races as continuous data for better evaluation with such filtering.

The Stack-to-Grouped bars is a dynamic interactive visualization that the user can select either to see the distribution as a stacked bar chart (Figure 1) or a grouped bar chart (Figure 2). The observable notebook for this part can be found at: https://observablehq.com/@yunanwu/stacked-to-grouped-bars. As shown in these figures, for each race cohort, we count the number of officers in five groups, i.e., the percentage of people having the same race as that officer in their community.

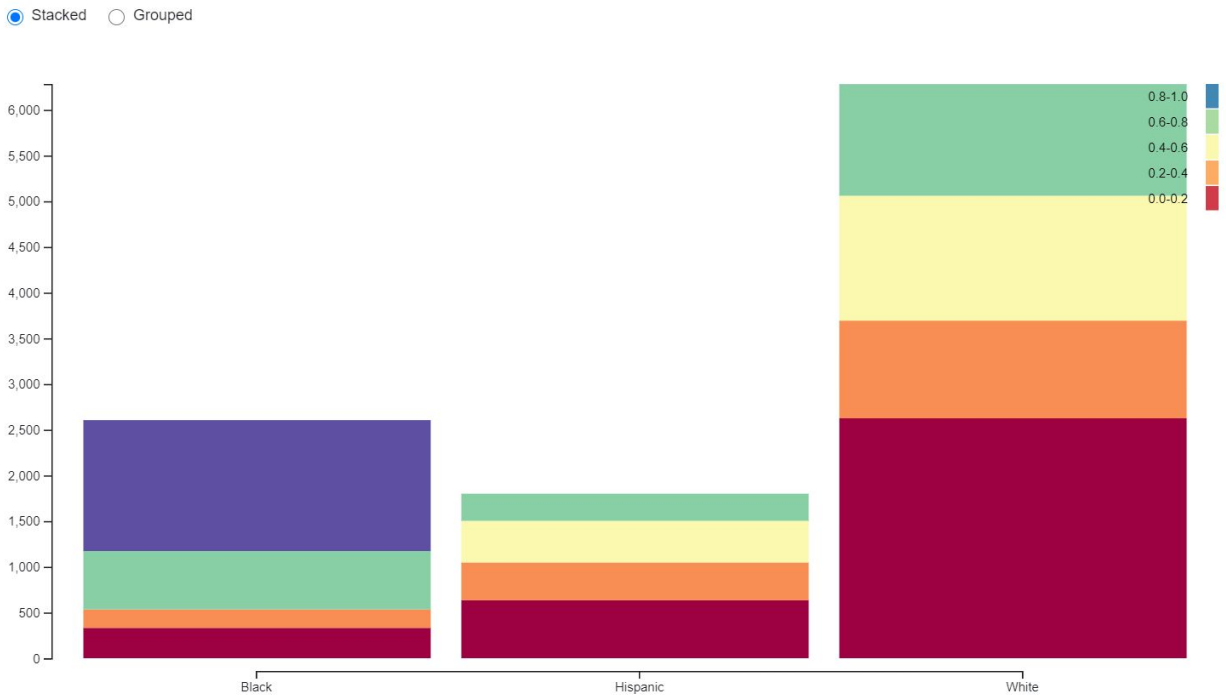


Figure 1. The grouped bar graph

In Figure 1, the total number of officers reveals the imbalanced job distributions among different races. We can observe that the number of white officers is more than the sum of black and Hispanic officers. This result matches with the results from previous checkpoints. White people make up the vast majority of the officers, which tells social inequality.

Figure 2 shows a more detailed officer race matching with the communities for their service. More than half of the black officers served the neighborhoods that contain 80%~100% black people. Additionally, it can be observed that the distribution of black officer community matching is bimodal. On the other hand, it is rare for the Hispanic and white officers to serve in communities with more than 80% population in their race. Most of them served the areas with 0%~80% of the people of the same race as them. The Hispanic officer matching has more or less uniform distribution in the 0~80% matching range, but slightly less than half of the white officers serve in communities with less than 20% white population. This is a really interesting finding, which shows that officers in

black are more likely to work and live with their race. In contrast, most white officers chose to serve the communities with fewer people the same as their races. Hispanic officers have no specific preference. From the results, we can indicate that black people prefer living in groups, where people feel more comfortable than living in other environments, which also reflects some inequality to black people.

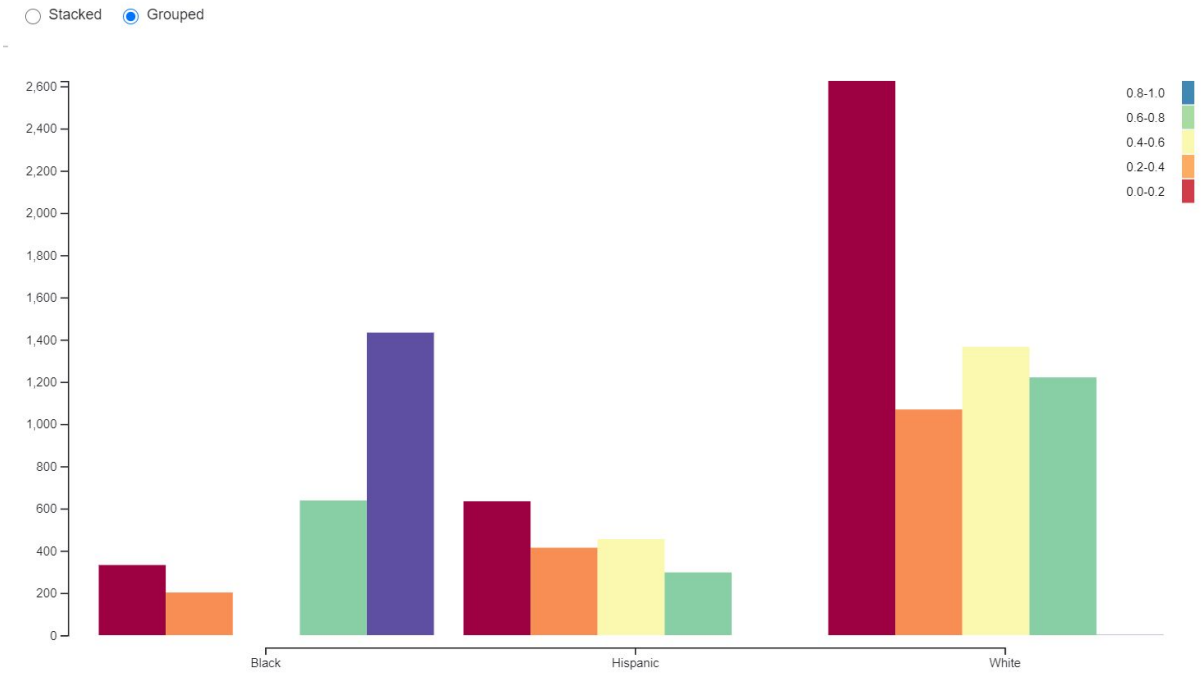


Figure 2. The stacked bar graph

Advantages of using stacked-to-grouped bars:
From the stacked figure 1., we can intuitively compare the differences among each cohort and see which races have the majority number. From group figure 2., it is easier to see the difference within each cohort. The interactive visualization is reflected in the interfaces that people have the option to see stacked or grouped figures and showing the details in each color bar. Observable is easy to use, and the notebooks are straightforward to show any changes in the results without compiling locally on your own computer.

- ***A scatterplot matrix showing the relationship between officer race percentage within community vs complaint filer race percentage within the community.***

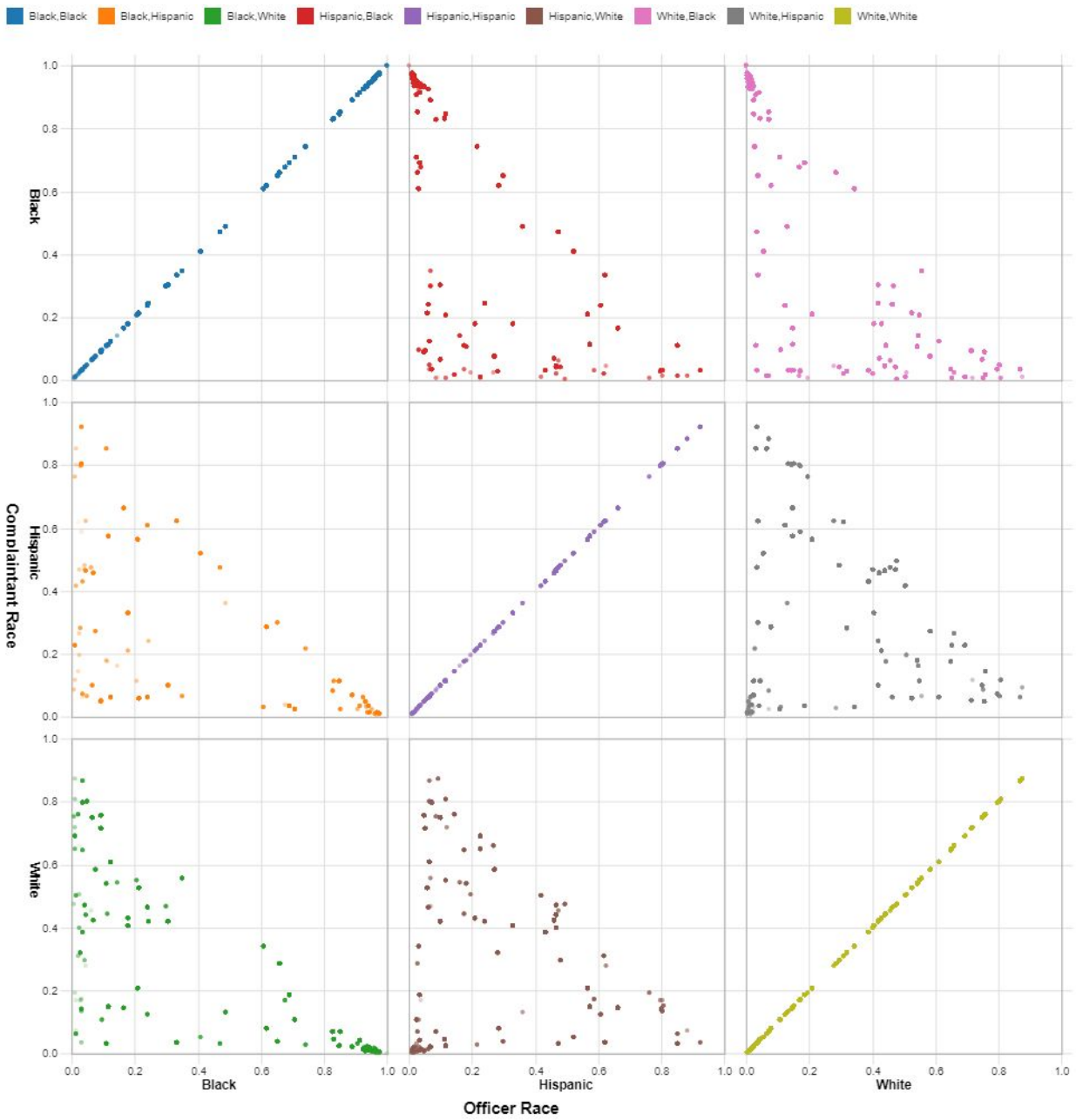


Figure 3. Scatter plot matrix

A Scatter plot matrix shows an array of scatter plots, and these plots can be used for easy visualization of correlations between different sets of data. Figure 3 shows the scatter plot matrix for this question. In this plot, each point represents one complaint. The D3.js version of the plot also has a brushable feature, where users can highlight a range of the area in one scatter plot, and all data points in the same range in the matrix are highlighted. This can provide a more straightforward comparison between different scatter plots in the matrix. The observable notebook for this part can be found at https://observablehq.com/d/a210438ec4686c6e.

From the scatterplot graph, we can see the relationship between officer race, complainant race, and race percentages in the location of the complaint. We can see that when the officer and complainant have the same race, the graph shows a linear relationship because they represent the same race component in the community; however, when the officer and complainant have different races, all points are within the lower-left corner. Since the sum for the distribution cannot exceed 100%, it is expected that data points should not appear on the top-right half of the plot for the off-diagonal plots.

In the off-diagonal graphs, although there is not a linear relationship, the data points seem to not lie in the mid-low/mid-low range in the scatter plots, and holes can be observed in this range. This might be an interesting feature of the data. We might be able to regard such distribution as a combination of three linear models: the first one is along the x-axis, the second one is along the y-axis, and the last one is a negative correlation between the x- and y-values. We might be able to attribute this feature to three different behaviors of the officers and complainants: 1) if the complainant is in the minority group in the community, they are likely to complain about the officer; 2) if the officer is in the minority comparing to the community, the community are likely to complain about the officer; 3) there is a negative correlation between the officer racial representation and complainant racial representation when there is a complaint. We might need additional information (e.g., officer rank) to help us better separate these three cases.

Additionally, in the off-diagonal scatter plots, in the cases involving black officers or black complaints, we see clusters when either the black officer representation or the black complainant racial representation is high. This could mean that black complainants usually receive misconduct (and file complaints) when they are in their own communities, and that black officers seem more likely to commit misconduct when in black communities. In contrast, in the Hispanic vs. white we see clusters where the racial representation of both the officers and the complainants is low.

The features observed from the data can guide our machine learning model.