Checkpoint 2: Data Exploration

The Freedom Deer: Tianchang Li, Hualiang Qin, Qingwei Lan

October 21, 2021

1 Data Visualizations

1.1 Boxplot for Subject Age Distribution Grouped by Race and Gender

We plotted two box and whisker plots, one for subject age over subject race, and one for subject age over subject race and gender.

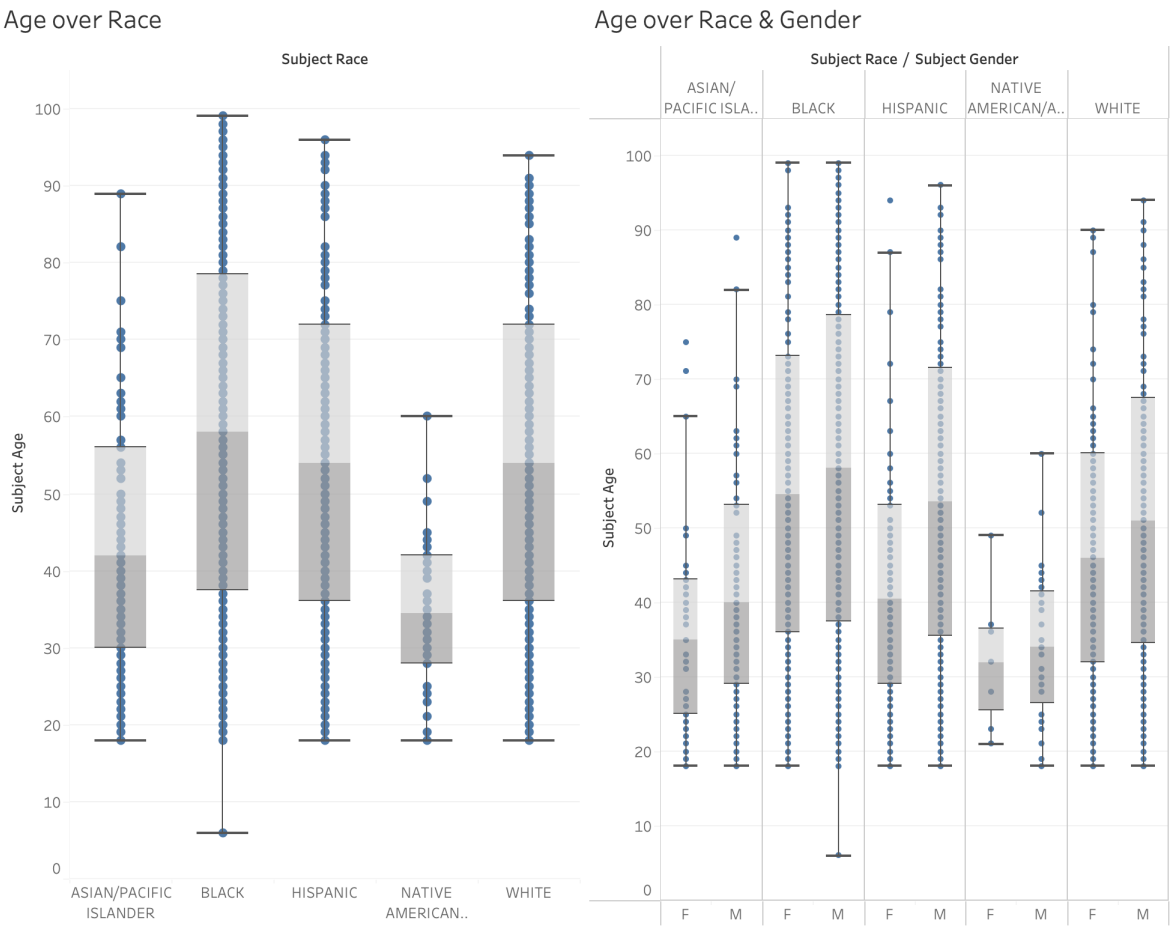

Figure 1: Box and whisker plot of subject age over race and gender.

From Figure 1, we can see that among all the races, subjects of use of force cases that is Black, White, and Hispanic have higher age median than that of Asians and Native Americans. When we look into the gender, we can see that Hispanics have the largest difference between male and female. In the future, we can explore if the higher age median of race correlate with their larger proportion of population. Also for the Hispanic population, we would like to dig deeper to see if culture factor played a role.

1.2 Influence of Different Environmental Conditions on Use of Force Cases for each Subject Race

In this section, we plot a horizontal barchart for each environmental condition. In each barchart, we plot the percentage of cases happening under different conditions for each race.

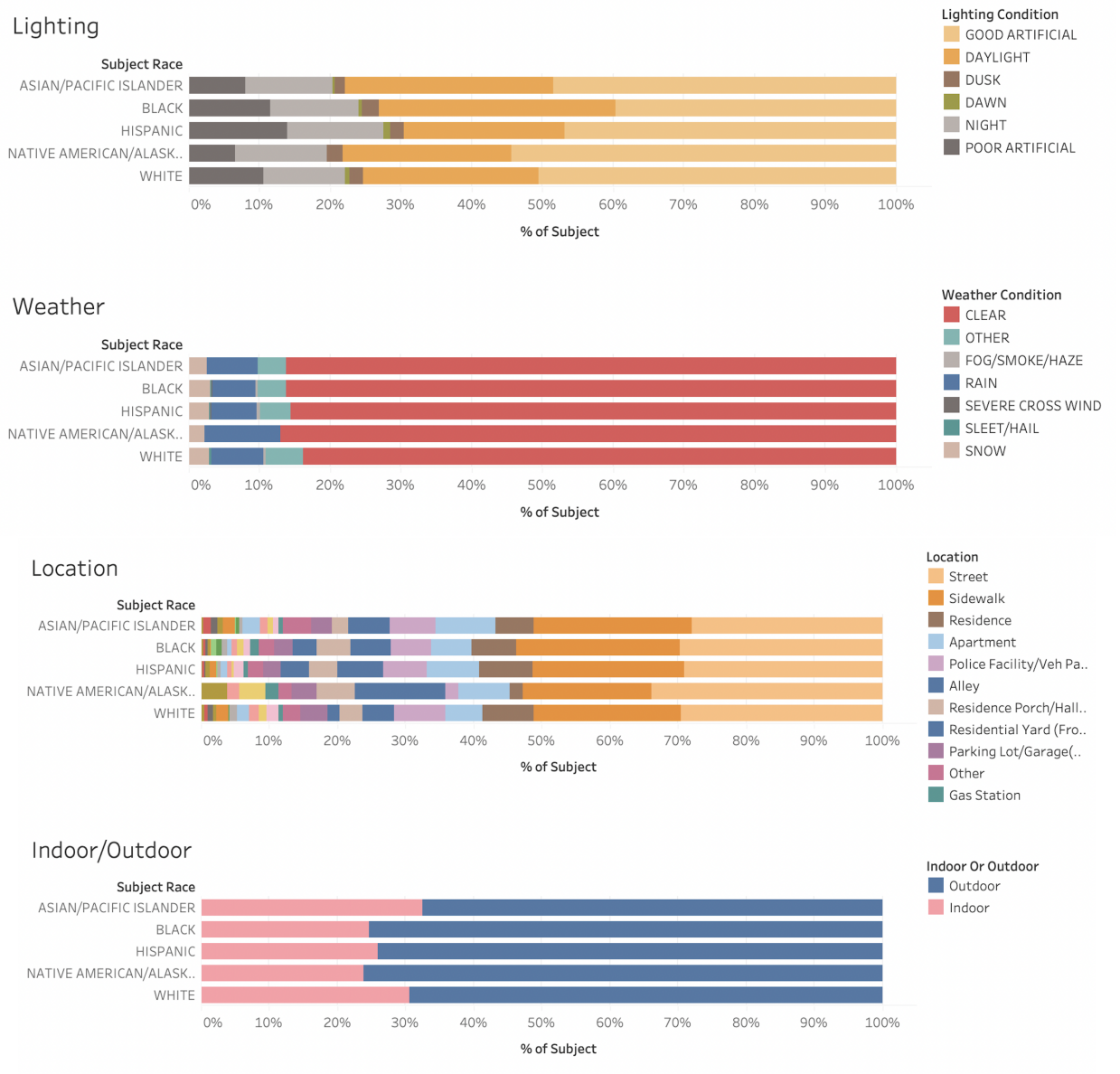

Figure 2: Barchart of the frequency of police use of force cases in different environmental conditions (lighting, weather, location, indoors/outdoors) over each race

The four charts in Figure 2 reflect the occurrence of use of force cases under different four conditions (Lighting, Indoor/Outdoor, Weather, Location). In terms of location. we can see that the majority of use of force cases happened in the street and on sidewalks. Most cases happen outdoors rather than indoors. Daylight and good artificial light, both of which can provide a better vision (brighter conditions) outrace the lighting conditions that provided worse vision. (darker conditions) The weather condition is almost dominated by clear weather. We don't know if these conditions are independent. Intuitively, clear weather and (daylight + good artificial) both imply better ability to see, while (sidewalk + street) and outdoor both imply a broader space. Further PCA (principal component analysis) might have us draw a better conclusion on this. What's more, race seems independent to the effect of these environmental factors.

1.3 Cross Race Use of Force

In this section, we created a bubble plot depicting the different combinations of officer race and subject race for police use of force cases.

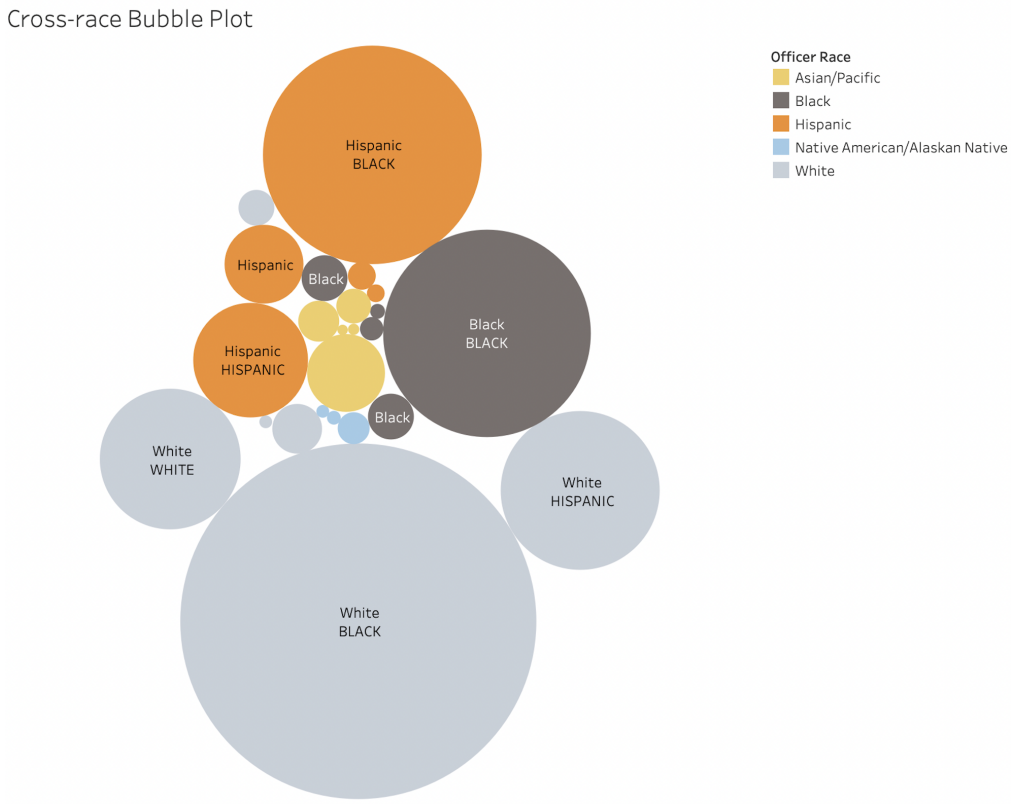

Figure 3: Bubble plot depicting cross-race use of force cases. Each bubble contains the officer race (at the top) and the subject race (capitalized, at the bottom). The size of the bubble shows a visualization of count of use force cases. The plot filtered out cases where officer race is Null. Furthermore, the plot only kept the top 29 most common combinations.

By looking at Figure 3, we can see that top three cross race combinations are white officer to black subject, black officer to black subject, and hispanic officer to black subject. Surprisingly and unexpectedly, we found that black officer to black subject is one of the top three combinations. This leads us to consider the case of discrimination happening in the same race. Alternatively, this phenomenon may be explained by the fact that black subjects have the higher crime rates, so the occurrence of this situation is solely reflecting the portion of each race of officers in the population of officers.

1.4 Barchart Comparing Use of Force Cases on each Race and Percentage of Population for each Race

In this section, we plot a barchart of the percentage of the population of each race compared the total population. We also plot a barchart of the percentage of subjects of each race in police use of force cases compared to the total use of force subjects. We combined these two into a single barchart, making it easy to visualize the percentage comparisons.

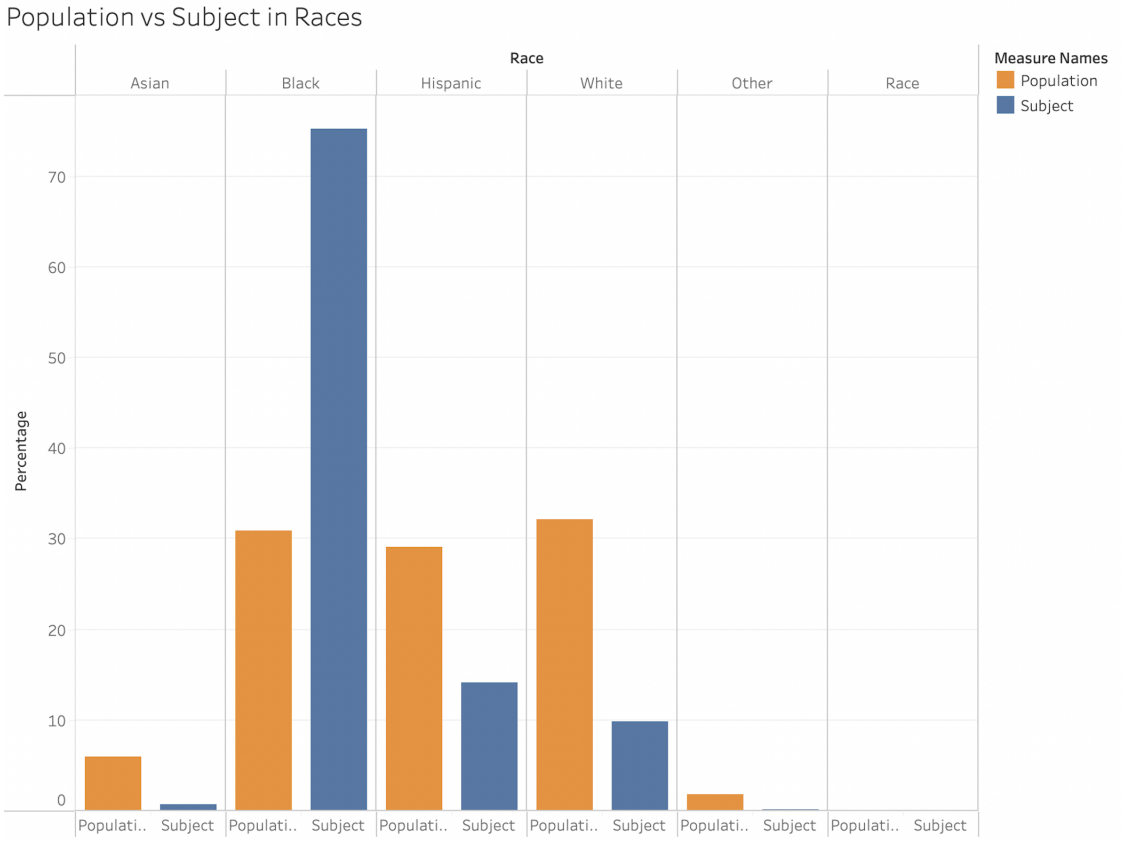

Figure 4: Barcharts of (1) the percentage of the population of each race compared to the total population and (2) the percentage of use of force cases on subjects of each race compared to total use of force cases. The orange bar represents the portion of race in the total population and the blue bar represents the portion of race in all subjects.

From Figure 4, we can observe that Black, Hispanic, and White make up the majority of the population, and they have almost the same portion in the population but at the same time, Black subjects account for more than 70 percentage of all subjects, while hispanic and white each accounts for only about 10-15 percentage of subjects. This result brings us to the existence of racial discrimination among police officers. We still need to do some investigation on the crime rate of each race in reality before we can make a solid conclusion on the existence of racial discrimination.

2 Tableau Review

In this section we will present an overview of the experience of our usage of Tableau, including what we believe to be advantages and disadvantages of the software.

2.1 Advantages

1. Tableau provides great visualiations and an intuitive user interface for handling these data tasks. The built-in graphs are easy to use and the software makes it easy to switch between different graphs. The software also makes sure that we are using the correct data values for different types of graphs.
2. Once connected, exploring and processing the data becomes an easy task. The software provides an overview of all the tables and the fields of each table. Table joining and field filtering is also straightforward.
3. Tableau makes it hard to make mistakes because it restricts the way we can visualize data. In other words, it won't allow us to use a certain type of graph unless our data satifies all constraints.
4. Tableau provides the capability to export workbooks. These workbooks are represented as structured XML data internally and this allows it to be checked into source control, making it easy to track version history.

2.2 Disadvantages

1. Tableau's license is rather difficult to setup, even being a student. It requires us to upload our student IDs and wait for verfication. This process can fail unless the photo is taken clearly.
2. Connecting Tableau to a remote server is easy, but the connection relies heavily on the internet bandwidth. In most cases, the connection is not stable and the visualization tasks are extremely slow.
3. Connecting Tableau with local Postgres is hard. It requires us to download a specific driver to connect the software with the database and requires us to enter specific information for the connection. The error messages that are associated with the failures to connect are counter-intuitive and provide no information as to why the connection is failing, making it impossible to debug. We spent hours on the connection problems alone.