PJ Samuels, Samantha Shih, Vedika Srivastava, Kevin Martin

Prof. Lance Galletti

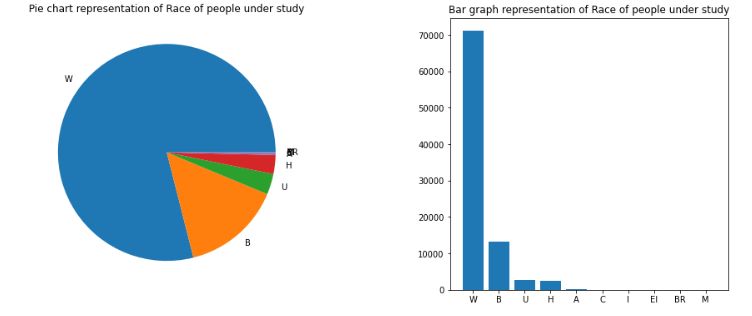
CS 506 Project: CPCS Bias Prosecution

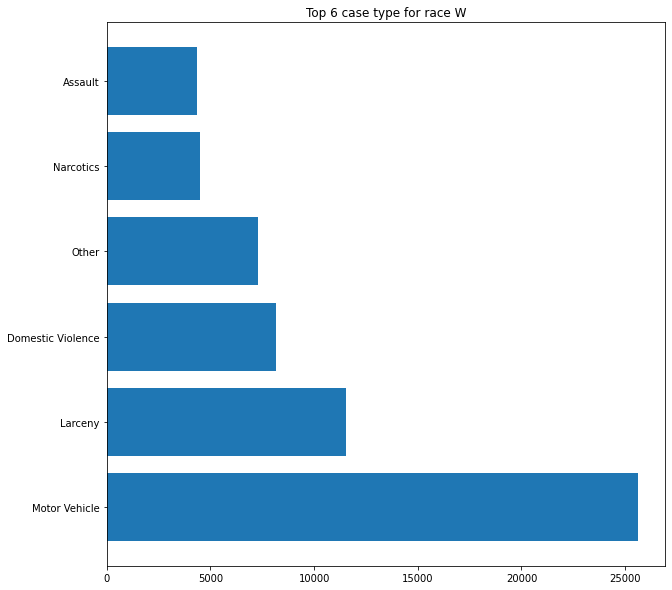
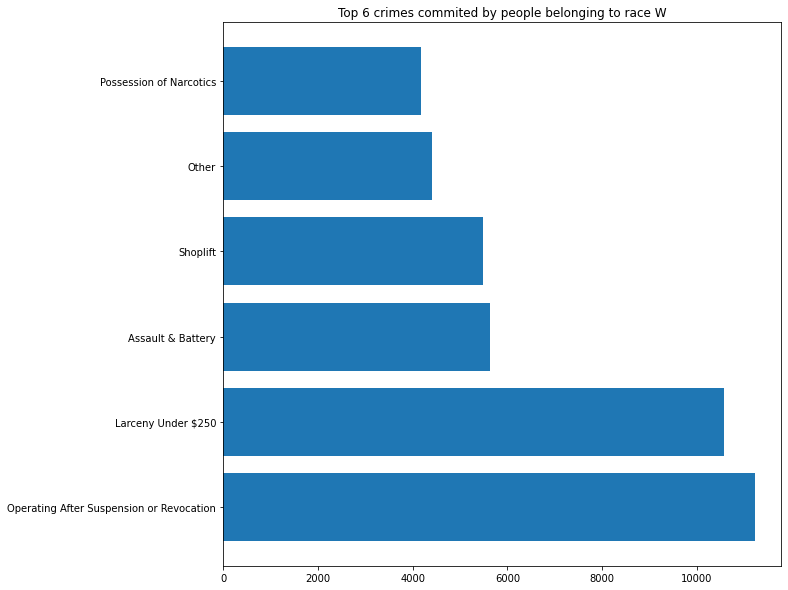
October 17, 2023

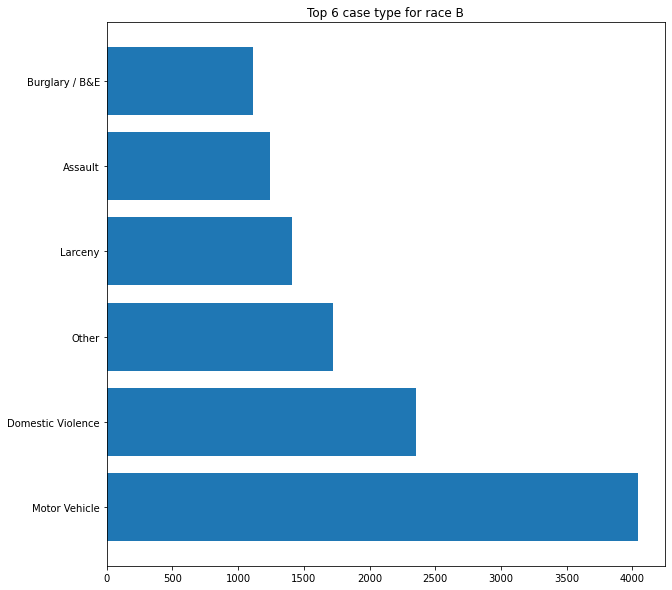
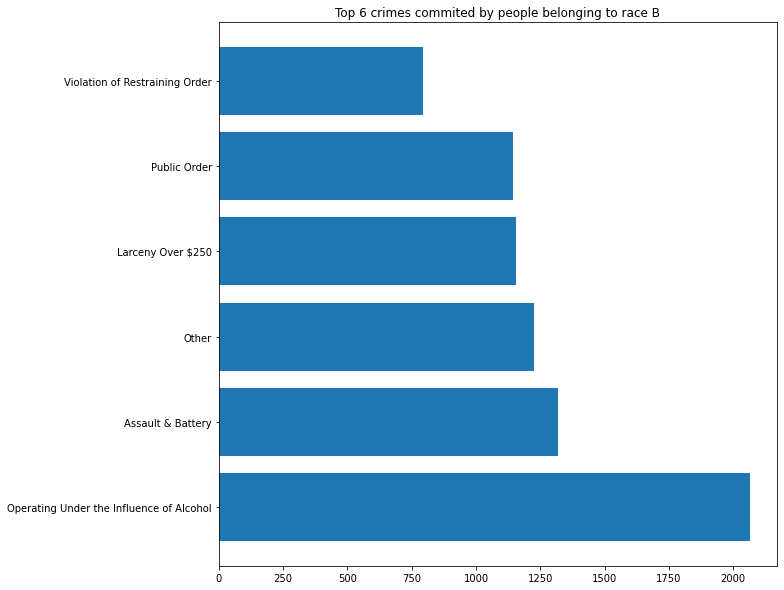
**Deliverable 1: Preliminary Analysis Report**

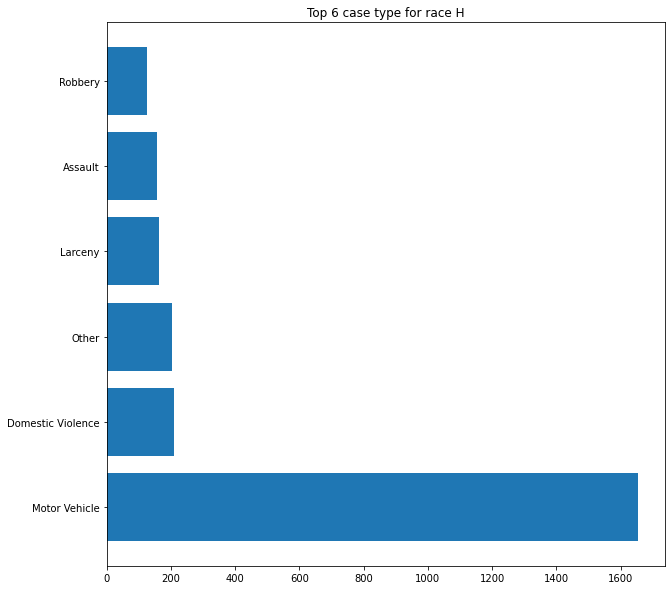
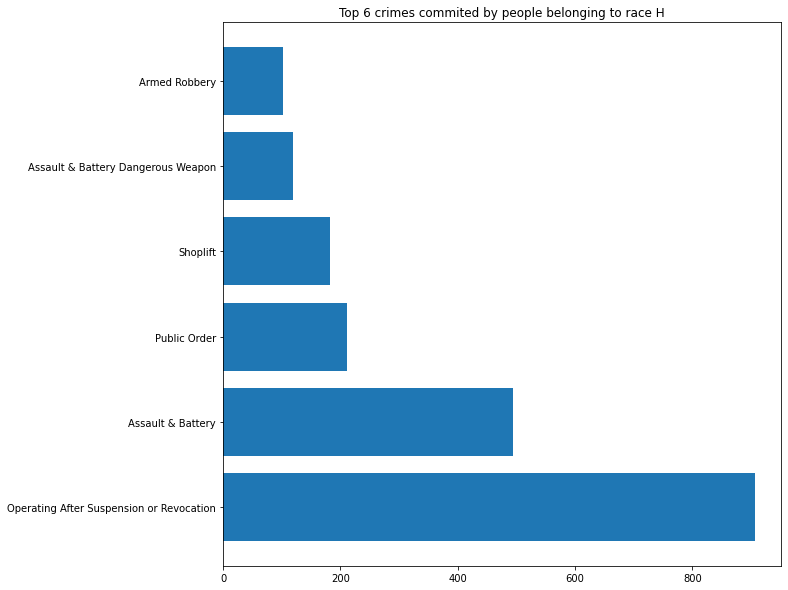
**Analysis of Data**

One of the key questions we are trying to answer is if prosecutors are more likely to use their discretion to offer diversion or generous pleas to certain people. Though we don’t have as much visual representation in our conclusions, the data available to us still shows very clear trends in more generous pleas towards White defendants. Currently our analysis is showing data of total and types of crimes being covered by prosecutors by race, not necessarily their relation. Although we are able to compare statistics on surface level, it isn't as accurate as we would like it to be. There are a fair amount of other details and variables to take into account like the total population and population of each race in comparison to the total number of defendants since there is a clear skew in how many White defendants and cases there are. All things considered we were still able to see and concluded that proportionally speaking many minorities were being charged, found guilty, and sentenced to higher percentages and to harsher punishments on average. Some of our data visualizations are shown below:









**Revised Project Scope**

As of now, we only have data for one location, Berkshire, MA so we are forced to narrow our project scope to only that area of Massachusetts. Our next steps will be to look into gathering data and analyze data from more locations. For the datasets we are given however, we will need to go more in depth into which specific crimes and races (ethnicities) have more guilty verdicts and how harsh the punishments may be. Once we are able to get more than one primary data set, currently being Berkshire, we will be able to compare the differences in location and prosecution disparities. By narrowing down which specific data set columns and correlations within said data sets will be able to get a better look at any potential harshers charges based on race. Currently we are analyzing sentencing data in relation to race but as we progress we will be able to expand to include gender relation as well as comparing said data to other races and other locations.

**Limitations with data & risks of achieving project goal**

Most of the cases in our dataset have no value for the sentence columns. After analyzing the data, we noticed that many of the rows with no sentence were dismissed trials. So, we had to make some assumptions that the lack of sentence data meant that the defendant was given no sentence. Also, a vast majority of our features are categorical so we would need to transform the columns into numerical in order to run data analysis models on the dataset, but we have multiple categories for each column and that would result in a very large dataset.

There is also a lot of vague data and ambiguity, making analysis of the whole dataset very difficult. For example, for the “Crime Type” column, there are 12,878 rows which are under the category of “Other,” making it the second-largest unique value in the “Crime Type” column. We also found limitations in understanding the meaning behind many of the columns, such as the difference between “Disp Type” and “Disp Desc” (disposition type vs. disposition decision). For the “Race” column, we could not find the meaning behind some of the values. For example, we were unsure what the “BR” and “EI” meant in terms of race. Without meaning behind the features, it will be challenging to translate correlations in the dataset to prosecution bias.