Tiffany Davidson

Foundational Python for Data Science

Exploratory Data Analysis and Insights of Police Stop Data Set

The data set I will analyze today is the Stop Data set. This set covers every Metropolitan Police Department stop from January 1st of 2023 to June 30th of the same year. The stops include vehicles of all kinds and citizens on foot. I have a history of researching the nutritional status of incarcerated individuals so this set piqued my interest due to the possibility of the people being stopped also becoming arrested. My objective is to see if there is a pattern and how this pattern can help reduce the amount of people and reasonings for police stops in Washington DC. This is also relevant in light of the several unfortunate police stops that have gone wrong like George Floyd. It is possible that our research can deter such horrible acts of social injustice from occurring.

The variables in this data set are many. Some stood out the most to me. They include ethnicity, time of the stop, location of the stop and the duration of the stop. Now separately they may not tell a story but combining them lays out a picture. When these elements combine we see Black and Brown people being stopped for sometimes an hour or more, late at night, and early in the morning. The challenge here is the volume of the data. It would need to be cleaned to tell a more accurate story but also the large number of information in the data set brings difficulty with accuracy tests. Another concern is the large amount of data that is not included. One example is the many times the assignment for the officer is missing. Unfortunately, that leads to questions as to why certain officers are even in certain areas if that is not their normal area.

Cleaning this data would involve removing the data with no information. However, I believe that this information ultimately would need to be provided to show an accurate picture. If we wanted to filter the results to narrow them down by month and remove the NA from division and unit assignment we could use df %>%

filter(month == May) %>%

filter(Division\_Assignment != "NA"

&Unit\_Assignment!= "NA")

The first chart below shows the different police districts and the number of police stops they made. You can see that the 3rd and 5th districts have the most police stops. This is alarming due to the demographics of the people in these neighborhoods. The 5th district serves a large portion of Northeast DC and the 3rd district serves the Adams Morgan, Shaw, and Columbia Heights area of Northwest DC. These areas are full of Black and Brown citizens. The second image shows the stop type along with the location and stop duration minutes.

A bar graph with different colored bars

Description automatically generated

A graph with orange lines

Description automatically generated

This analysis proved the difficulty of dealing with large amounts of information. This proved to be a limitation in the final analysis. Another limitation was the specific data that was missing. It led to the question of why and how long has that type of data been missing. This missing data may have impacted court cases and led to the loss of freedom for members of the community. I feel that missing data should be a priority and it could bring to light other issues with law enforcement that could negatively impact citizens. There needs to be a look into why so many Black and Brown people are stopped and also a look into the times of day they are being stopped. Are more Black people being stopped because they make up a large part of the population or is there something more?