Group 3: Sohaib, Haider, Nisha

Deliverable 1: Preprocessing and Exploratory Data Analysis

As part of initial exploration, we explored the summary statistics and types of all attributes.

Missing Values:

There were significant number of missing values in the dataset (Total missing values = 626535).

Shooting: is a binary variable, so we just filled all the missing values with 0.

Lat & Long: Doing backward fill, mean/mode fill will create bias in these cases. So, we just dropped the entries with missing values of these attributes.

We preferred backward fill method to fill missing values in the other attributes.

Using maps to visualize a particular offence group:

We have used folium library to visualize the distribution of three offense groups (Harassment, Robbery, Drunkenness) separately on the map. It is a good way to see the intensity of that offense group in a particular area.

Changing types of attributes:

We have changed the string attributes to numeric for further use.

Correlation:

We have found correlation matrix for the dataset using two methods. The heatmap shows a visual of which attributes are correlated. We found that the correlation between following attributes is significant:

Shooting, Incident number (corr = 0.209005)

Shooting, Occurred on date (corr = 0.401727)

Location, Shooting (corr = 0.152571)

Most Frequently Occurring Incident:

Most frequent occurring offense code is 3115, which corresponds to the offense “Investigate Person”.

District Wise analysis:

We have plotted a histogram to see the number of offenses in each district. From that histogram we can see that district B2 has highest number of offenses reported.

Outliers:

We made box plot for different attributes to detect outliers. There were no significant outliers found.