Boston Crime Analysis

1) Authors: Smit Dar, Parv Thakkar, Akshay Krishnan, Nitheesh Koushik Gattu

2) Summary:

We chose a data set that has all the crimes reported by the Boston Police Department in Boston from 2015 to date. This Data comprises columns such as the description of the crime, area of occurrence of the crime, whether a shooting was involved or not and so on.

This project predominantly has two aims, one is to find out how factors such as area and Time affect the crime rate in the city, and the other would be to predict the count of crimes that could happen given all the factors. This not only helps the citizens to travel safely, but also the law enforcement to allocate resources more efficiently and deploy forces accordingly.

3) Proposed Plan:

We have chosen the Crime Report Dataset which is provided by the Boston Police Department. Our aim is to predict the total number of crimes in a particular district on a particular day based on the data (2015 – Present).

In the preprocessing step, a new dataset containing the Date,District, Severity of the Crime (Violent or Non-violent) and a few more time related variables will be created. In addition to this, null values will also be taken care of. Moreover, we wish to do Exploratory Data Analysis to gain further insights on the data. Using visualization techniques, a choropleth of the Boston map will be created which gives information regarding the crimes in different areas.The aim of this idea is to see which area in Boston has the greatest number of crimes reported and also the time or day in which it takes place. We would also want to consider the COVID-19 scenario which may have brought variation to the 2020 and 2021-year data. Finally using Machine Learning methods such as Regression and Time Series we predict the total number of crimes for the year 2022 per district.

Some of the challenges we face would be plotting the choropleth of Boston which we wish to solve using the Folium library of Python. The other challenge would be to predict the number of crimes accurately which we can overcome by improving the modeling approach i.e add more variables to it and also introduce some new variables.

4) Preliminary Results:

There are a total of 8 datasets containing the crime incident reports from **2015** to **2022**. The final dataset is a combination of these datasets and contains **599,072** rows and **17** columns. 10 out of the 17 columns do not have any NULLS in them. However, the **`LOCATION`** column has close to **23,000** rows containing invalid longitude and latitude values. **59%** of the rows in the **`SHOOTING`** column and close to **41%** of the rows in **`OFFENSE_CODE_GROUP`** and **`UCR_PART`** contain NULL values. The remaining 3 columns have less than **4%** of NULL values in them. **2017** had the greatest number of reported crimes whereas **2015** had the least, with most of the crimes being reported on **Fridays.**

5) References: (i) https://www.hindawi.com/journals/complexity/2022/4830411/
(ii)https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system