**CIS 602-01**

**Data Visualization**

**Project Report**

Submission By

***Ankita Anil***

*01592435*

**Project Report**

**Data Analysis of Crime Reports in City of Boston, MA**

***Link to Dataset***

<https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx>

We can download the data as either .csv file or json file from the export button on the top right corner of the page.

***Description***

The above dataset describe the crimes reported in and around the capital city of Boston Massachusetts.

The original dataset contains the following attributes:

1. CompNos Categorical
2. Nature Code Categorical
3. Incident Type Description Categorical
4. Main CrimeCode Categorical
5. RepDistrict Categorical
6. ReportingArea Categorical
7. FromDate Quantitative
8. WeaponType Categorical
9. Shooting Categorical
10. Domestic Categorical
11. Shift Ordinal
12. Year Categorical
13. Month Categorical
14. DayWeek Categorical
15. UCRPart Ordinal
16. An attribute X Quantitative
17. An attribute Y Quantitative
18. StreetName Categorical
19. XStreetName Categorical
20. Location (a pair of longitudes and latitudes) Quantitative

As we can see there are not much quantitative attributes in the data set, this would restrict visualizations limited to only Bar Charts, Line Charts, and Scatterplots. But we do have geographical data, which will help in creating a map and marking certain locations on it.

Since the data chosen was very large (greater than 25 MB) and around 40,000 of rows, it was not even possible to upload and host data on sites like Github, etc… So, the cleaning of data was required. Unnecessary attributes and some rows from the end were omitted at Pandas using Python.

The final data used after cleaning had around 25000 rows and only selected attributes which were needed for the visualization. The selected attributes included :

1. ObjectID
2. CompNos
3. IncidentType
4. ReportingArea
5. DateTime
6. Year
7. Month
8. WeekDay

9. X

10. Y

11. StreetName

12. Location

Initially the dataset had data for for 4 years (2012- 2015), but due to cleaning of the dataset and removing the rows from the end, the final data was left with only two years 2012 and 2013 details.

Following is the link for the final data:

<https://raw.githubusercontent.com/anki24/DVproject/master/BostonCityCrimeData.csv>

***Questions that can be addressed to the dataset***

1. *Time period when the most crime takes place.*
2. *Streets with the highest number of crimes.*

The top 5 streets with the highest number of crimes in descending order are Washington St, Blue Hill Ave, Boylston St, Commonwealth Av, Massachusetts Av.

1. *Streets with the lowest number of crimes (Safest neighborhood).*

Since every street mentioned in the dataset have had at least one crime incident, so all the streets with exactly one crime incidents can be considered as safest streets which includes Hallam St, Hillock St, Maxfield St, Perry St, Oxford St, etc…

1. *Most common type of crime in Boston City.*

According to visualizations and its analysis, the most common type of crime found in the city of Boston is “Other Larceny” which counted upto 2504 incidents in two years.

1. *The day of the week with more number of crimes.*

As we can see from the first bar chart, there is not much of a pattern in number of crimes as per the day of the week, but still Friday has slightly more number of crimes than other days.

1. *Time difference between the time crime took place and the time it was reported.*
2. *The month with more number of crimes.*

***Introduction to Data Visualization***

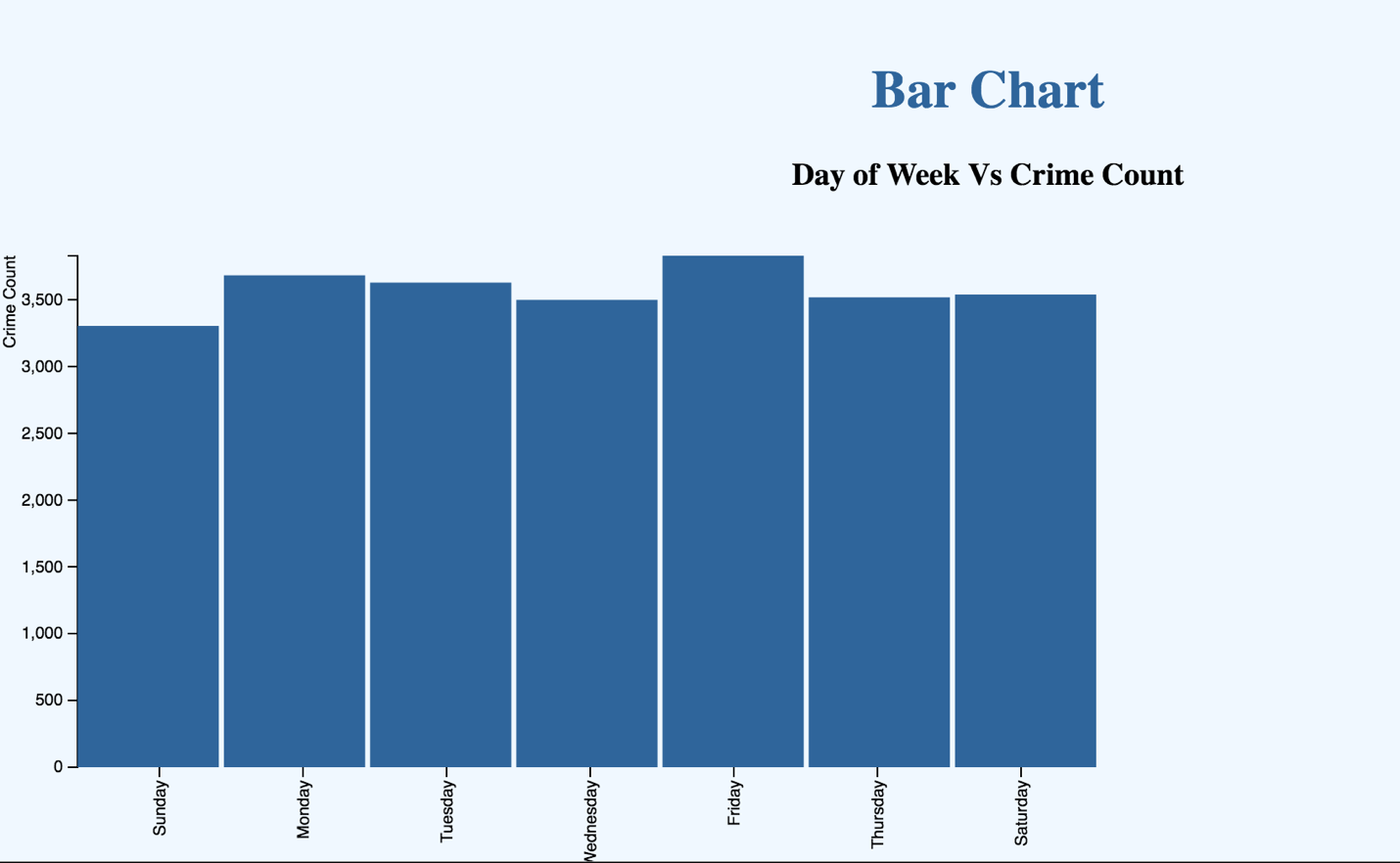
Data visualization can be de defined as the presentation of data in some pictorial or graphical format which helps in analyzing the data more easily than the data in text format. It also enables us to identify the outliers, anomalies, patterns and trends in datasets. Interactive data visualization further helps in decision making.

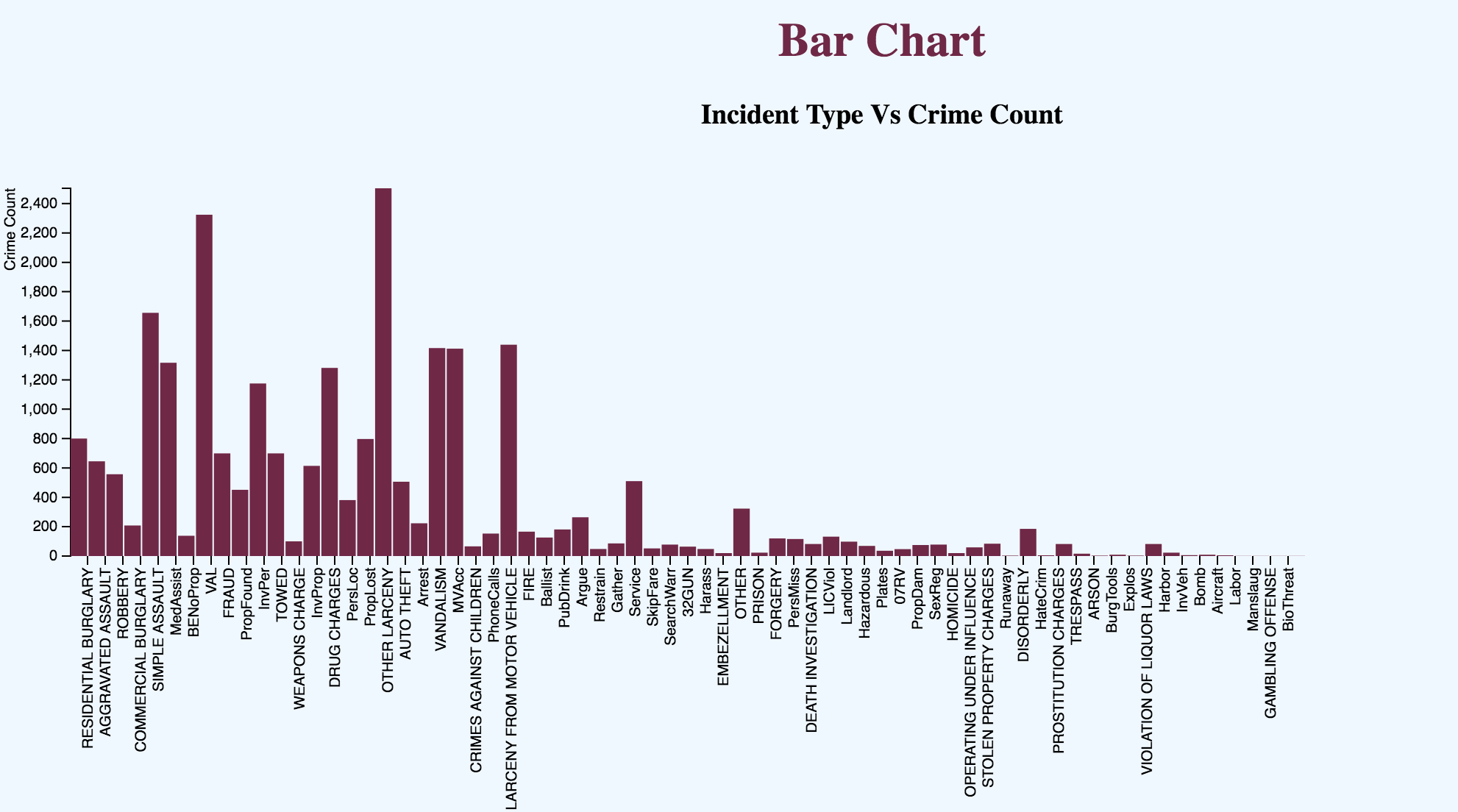
***Visualizations Presented***

*Bar Chart*

Bar Charts are representation of data with one quantitative value, generally on Y axis; and one categorical value generally on X axis. We can have vertical as well as horizontal bar charts. Bar charts used in the projects are all vertical. They helped in identifying the most common type of crime, week days with most number of crimes, and months with most number of crimes.

Following are the screen shots of the bar charts presented in the projected:





*Line Charts*

Line charts are similar to bar charts with quantitative value attribute on one axis (Y axis) but has one ordered value attribute on the other axis (X axis). And instead of bars, it has points representing the (x, y) value and a graph line joining all these dots.

I have tried to use the line chart in the project to identify the trends of crime through the months in year 2012 and 2013.

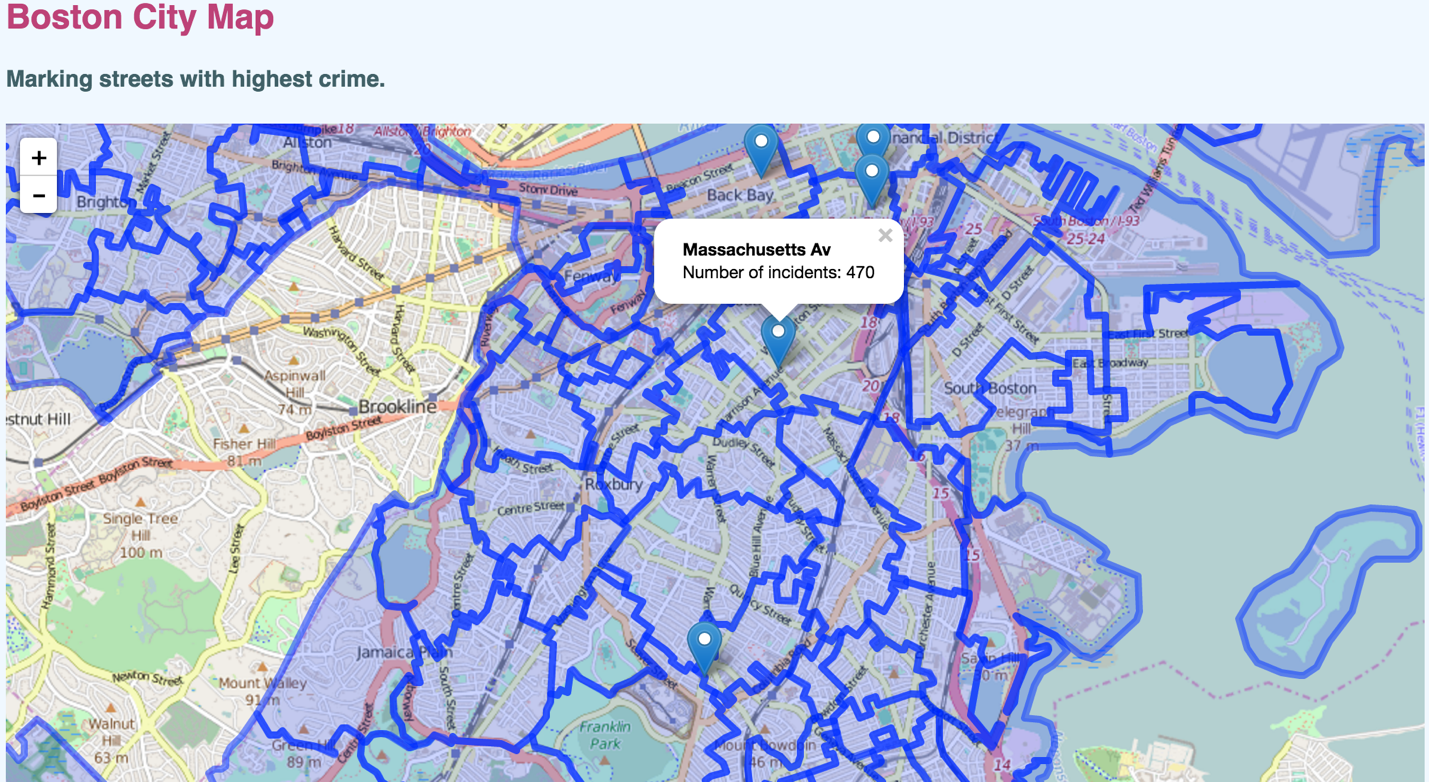
*Leaflet Map*

Leaflet is the open-source JavaScript library which help developers in creating interactive and mobile friendly maps.

I have created a leaflet map for Boston City by using data from <http://mass.gov/mgis> , and downloaded the neighborhood boundaries layer from <http://zetashapes.com/> and further uploaded the geojson file on Github at <https://github.com/anki24/DVproject/blob/master/25025.geojson> .

I have used markers to mark the top 5 streets with most number of crimes and also a pop up which appears on clicking the markers.

I wanted to show a Chloropleth Map, but I did not have any Precinct attribute value and I was unable creating it.



***Marks and Channels Used***

*Bar Charts*

As we know bar charts are used to encode two attributes (one quantitative and one categorical).

We use ***line mark*** with the horizontal spatial position channel for the categorical attributes and vertical spatial position channel for the quantitative attributes.

*Line Charts*

We use ***line mark*** with the horizontal spatial position channel for ordered attributes and vertical spatial position channel for the quantitative attributes. We can also use ***dot marks*** to present specific points on graph. We can also use Color channel to represent different categorical attributes across the ordered attributes. Eg. in the project I have tried using two different colors of graph lines to represent two different years across the orders value attribute of months along the X axis.

*Boston City Map*

We use ***containment mark*** to represent an area on a geographical map along with ***color hue channel***.

***Interactions Used***

Tooltips, Brushing, Interactive leaflet map with Markers and Popups.

***References***

<http://maptimeboston.github.io/leaflet-intro/>

<http://zetashapes.com/>

http://leafletjs.com/

<http://bl.ocks.org/d3noob/b3ff6ae1c120eea654b5>

<http://bl.ocks.org/Caged/6476579>

http://bl.ocks.org/d3noob/e5daff57a04c2639125e