**Arrests vs Crime-Baltimore City**

**Introduction**

The aim of this project is to study the Crime and Arrest public datasets of Baltimore for understanding the pattern through visualization and to derive meaningful insights from these visualizations. Making use of the coordinates of the crimes, the police department can understand the trends of crime with respect to the time and location, these analyses will help them answer key questions like identifying the sensitive areas, the time at which most crimes occur, crime types etc. These studies will help them come up with better strategies to cope with crimes. I have also made a comparison between the crimes and arrests to analyse the efficiency of the police forces using an array of visualizations.

**Data Source**

The data was fetched from the Baltimore Government site by downloading the arrests and crimes data in CSV format. The site also provides a SODA API to get live data, but it has a limitation of fetching only 50,000 records. As both the datasets are big (more than 100000 records), I preferred uploading the dataset rather than API calls, as it provides a faster access and doesn’t impose any limitation on the records.

The datasets have information such as Metadata: Arrest-ID, Age, Sex, Crime Date, Arrest Date, crime type, locations etc.

While all the fields provide significant information about the nature of the crime and arrest, I have filtered certain fields for the visualizations.

The Data sources can be explored through the link <https://data.baltimorecity.gov/>

**Technology Used**

Anaconda Python Notebook was used to do the cleansing, data transformation and visualizations.

**Process**

The project was mainly divided into two parts:

1. Data Cleansing
2. Data Visualization

*Data Cleansing:*

The first step was to load both the datasets into the Jupyter Notebook. After loading the data was analysed in terms of missing rows, data types of columns, relevance of fields etc.

It was observed that the timestamp was in the format of yyyymmdd hhmmss. So, it required processing to extract the data and time details.

The coordinates were also stacked to a single field and a logic was developed to separate them into latitudes and longitudes.

The unnecessary fields were then identified and dropped from the datasets for better clarity while visualizing and to save processing time.

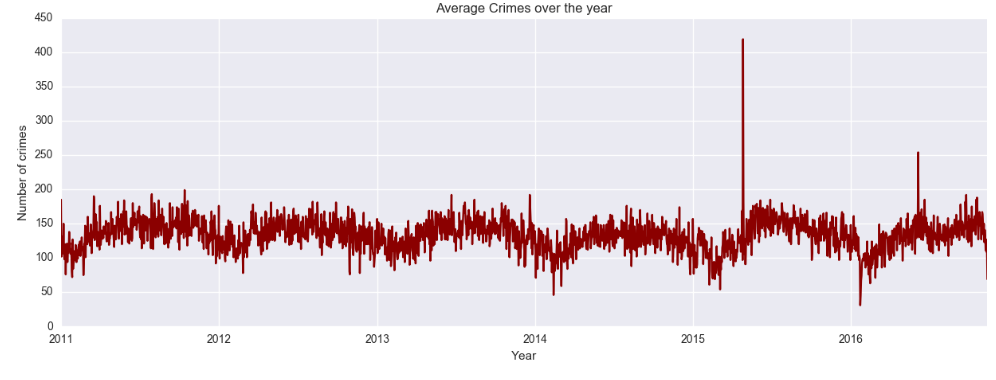
The rows which had null values were then removed to obtain cleansed CSV files, this process was carried out for both the datasets.

*Data Visualization:*

I made the following visualizations to derive insights from the datasets: -

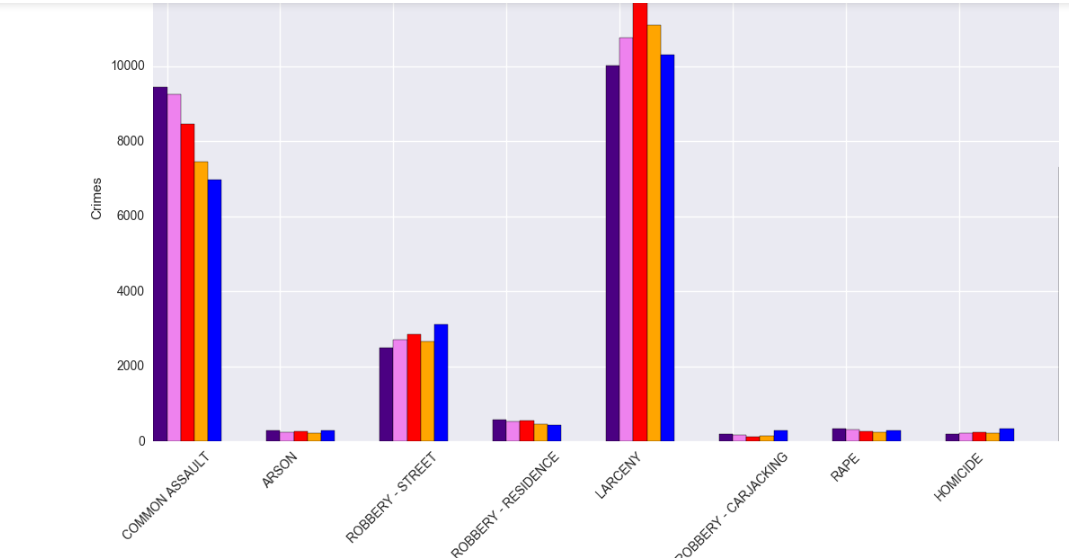
1. Line plot
2. Bar Graph
3. Geocoded Heat Maps
4. Terrain Maps
5. Heat Maps

**Line plot** :- The crime data was grouped in-terms of Day, Month and Year and then plotted to analyse the trends of crime over a period of 5 years.



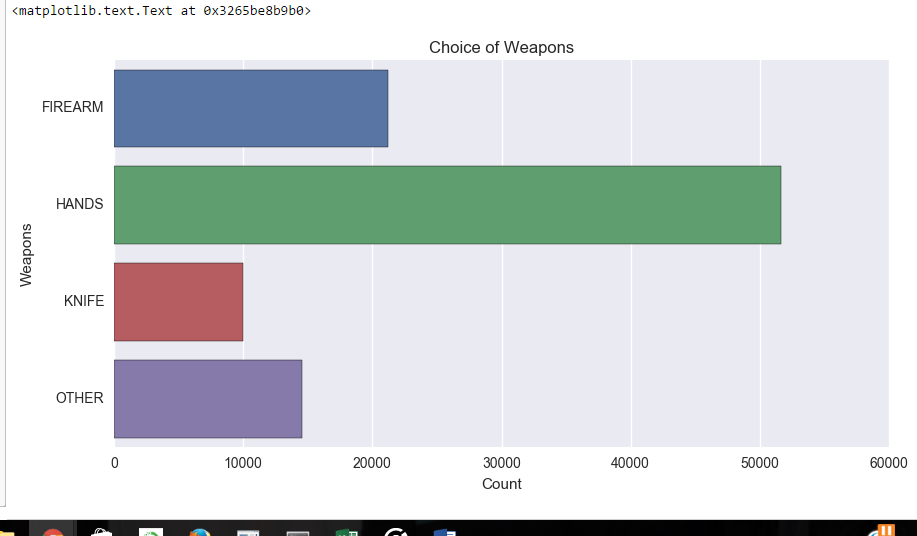
It is clear from the plot that most crimes occur during the summer months as people tend to stay indoors more often in Winters.

**Bar Graph**: A bar graph was used to categorize and summarize the occurrence of different kinds of crimes over the years, here each bar in a cluster denotes a particular year.

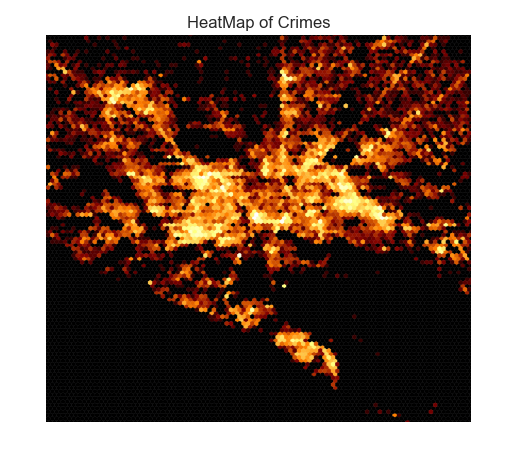


With this graph, it was understood that Larceny and assault accounts for the bulk of the crimes occurring in Baltimore province, also a declining crime rate is observed for assault which indicates stricter action by the law enforcement.

Using similar logic, a vertical bar graph was also plotted which details the weapon on choice used in various crimes.

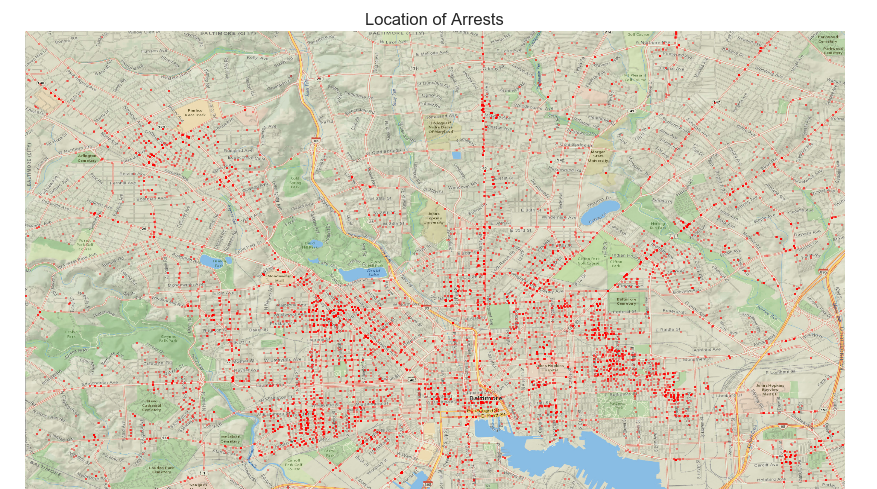


**Maps:** These are used to provide a visual representation of areas with high and low density occurrence of an event. In our case, a dense area will represent an area with high crime rates and vice versa.



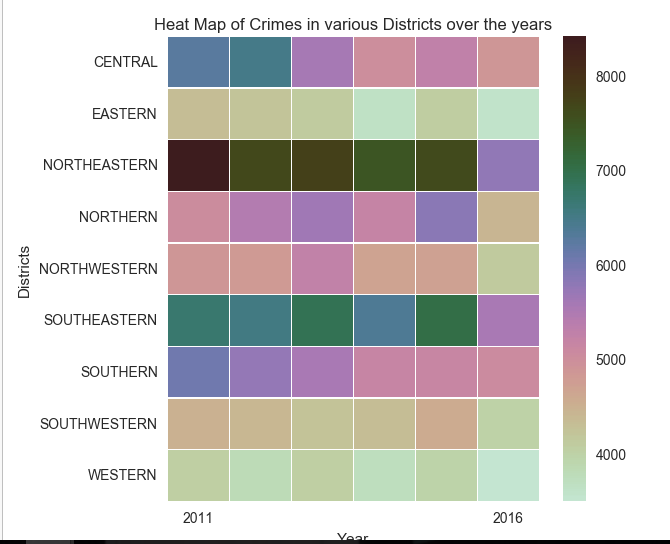
The heat map of Baltimore which was plotted using the coordinates of crime locations indicates that the bulk of the crime happens in the city center and it decreases when we move to the outskirts.

After analysing the crime rates using Geo Heat Maps, the arrest data was also plotted to show the coordinates of arrest locations across the city, this was visualized through ‘Miller Projection’ available with the BaseMap package.

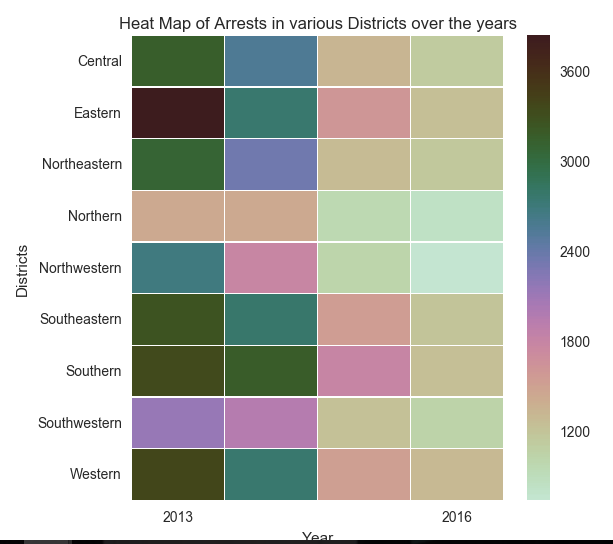


This visualization also depicts a similar story, most arrests are made in the center of the city. This indicates the prompt action of the police department in the wake of a criminal incident.

**HeatMaps**: A heatmap was used to visualize the most vulnerable districts over the years in Baltimore province.



It is seen that the Northeastern region is the most ill-famed district in terms of crime, whereas Western District seems to be the safest, also it is seen that the crimes of all the districts have declined over the years.



Interestingly the arrests have declined for all districts over the years. Comparing the heatmaps of Crimes and Arrests, it is seen that the arrests have declined more rapidly compared to Crimes, this may suggest that the fugitives have become smarter in evading the police.

**Challenges Faced**

1. Merging the datasets proved difficult as it didn’t have unique entries and a lot of duplicate entries were getting created which would have affected the visualizations.
2. In Big data sets a lot of cleansing is required, so it is better to load the data into technologies such as Google Refine, Fusion Tables etc which provides an array of tools to do fast cleansing as compared to cleaning it programmatically.
3. Due to API restrictions, large data sets could not be loaded through JSON API calls.

**Attachments**

The attachment below is the copy of the code used to develop the visualizations in python notebook.  


**Conclusion**

Using different kinds of visualizations, I could realize that there is a clear connection between the crime and arrest statistics. With few improvements, this approach can be implemented to showcase the crime patterns through visualizations as it will provide immense insights to determine the crime pattern and improve the planning of law enforcement.