Compare neighborhoods of city of Baltimore and predict probability of crime

Pravin Jeyaprakash

February 2021

1 Introduction

1.1 Background

Crime in the city of Baltimore is becoming a concern in the recent times. The crime rate is steadily increasing over years. All the neighborhoods in the city do not have the same level of crime. Crime rate varies across neighborhoods. Baltimore police department releases data on all crimes committed across all neighborhoods of Baltimore. The main reason of releasing this data to public, is to create awareness to people. Staying vigil would help people from staying off trouble.

1.2 Problem

Understanding the crime data of the city, available in the public domain, is very important. The data can be used to understand the neighborhoods which are affected the most, the rate at which crime is increasing, possible reasons for the crime in a particular neighborhood, advisory to local people and travelers to maintain high precaution at certain time & place and to predict the crime. The scope of this project is to understand the crimes at different neighborhoods in the city of Baltimore and to predict the probability of crime based on location and popular places around it.

1.3 Interest

The result of this project is to create general awareness to people and travelers. It is also to help police with the prediction of crime to plan strategies to control crimes. Additional to the crime data issued by the police department the location data of the neighborhoods from foursquare is used to understand the parts of city where crime is high. This project could be further extended by collecting the demographics data of the city to get a thorough insight on the reasons for the crime and to have more accurate predictions on the probability of crimes.

2 Data gathering and processing

2.1 Data gathering

City of Baltimore Crime Data from 2011-2016 is published by police department. For this project, the data is collected from https://data.world/data-society/city-of-baltimore-crime-data. The data is complete with details on all the crimes happened in the city from 2011 till 2016. The data has minute details of the location and date & time of crime which are very much needed for this project to get insight of the crime based on the location. To understand how crowded/busy the place of crime was, the popular venues around that location is gathered using the location data of foursquare. APIs of foursquare are used to explore venues, their locations, and categories.

2.2 Data Cleaning

Crime data from data world consists of the following columns:

CrimeDate	CrimeTime	CrimeCode	Location	Description	Inside/Outside	Weapon	Post	District	Neighborhood	Location 1	Total Incidents
11/12/2016	02:35:00	3B	300 SAINT PAUL PL	ROBBERY - STREET	O	NaN	111.0	CENTRAL	Downtown	(39.2924100000, -76.6140800000)	1

‘CrimeDate’ column contains the date of crime. ‘CrimeTime’ column contains the time of crime in 24 hours format including hour, minute & seconds. ‘CrimeCode’, ‘Description’, ‘Inside/Outside’ and ‘Weapon’ columns provide the code assigned to crime, type of crime, indication on whether a crime happened indoor or outdoor and type of weapon used to commit the crime, respectively. ‘Location’, ‘Post’, ‘District’ and ‘Neighborhood’ columns provide more descriptions on the street where crime occurred. ‘Location 1’ column shows the latitude and longitude data.

I have based this project on the features of a location of the crime, such as distance from the central business district and proximity to popular venues to study the probability of occurrence of crime. So, columns ‘CrimeDate’, ‘CrimeTime’, ‘Description’, ‘Neighborhood’, ‘Location 1’ are retained and rest all columns are dropped.

2.3 Data retained and features added

Latitude and longitude data present in column ‘Location 1’ is separated to different columns as ‘Latitude’ and ‘Longitude’ for ease of handling of the location data. The data must be processed to have consolidated values for each feature at neighborhood level. The data is grouped based on neighborhood and a new feature to hold the total number of crimes in a neighborhood is added. Likewise, latitude & longitude columns are used to calculate the distance of each neighborhood from the center of city and added as a new feature. The data analysis in this project is based on the location of crime and number & type of popular places around it. So, I have added features to include top ten venues in the proximity of crime and the total number of popular venues around it to include the impact of business/crowdedness. To achieve this, I have used foursquare APIs to collect location data. Finally, to produce meaningful results, based on the features considered, I have limited the data to reflect only one type of crime that is ‘STREET ROBBERY’. Rest all crimes such as homicide, rape, burglary etc. are removed from the data. At this stage, the data is good to proceed with clustering and classification of neighborhoods.

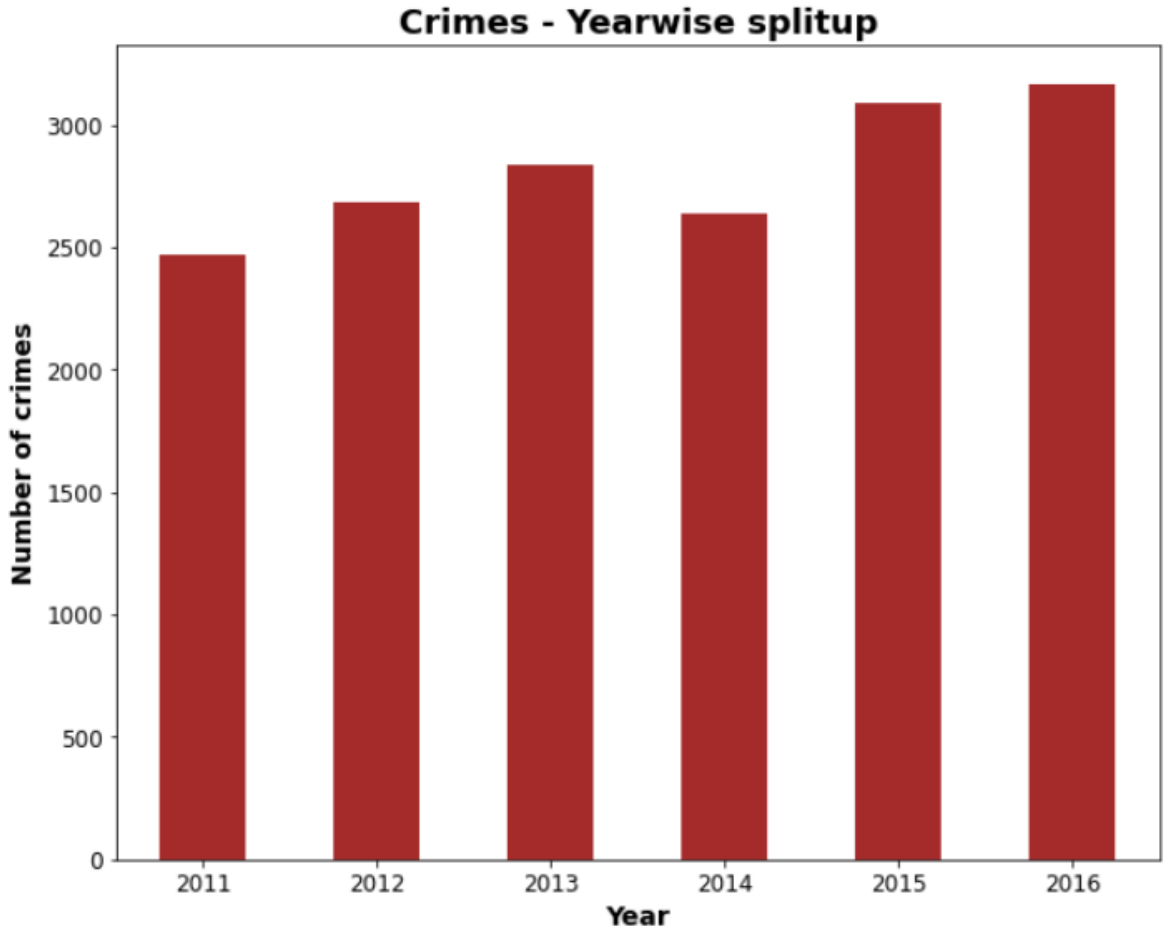
	Bar	Pizza Place	Sandwich Place	Coffee Shop	Convenience Store	American Restaurant	Bus Stop	Theater	Chinese Restaurant	Grocery Store	Distance
0	0.000000	0.033333	0.066667	0.000000	0.066667	0.033333	0.0	0.0	0.033333	0.066667	1.218593

3 Methodology

3.1 Exploratory data analysis

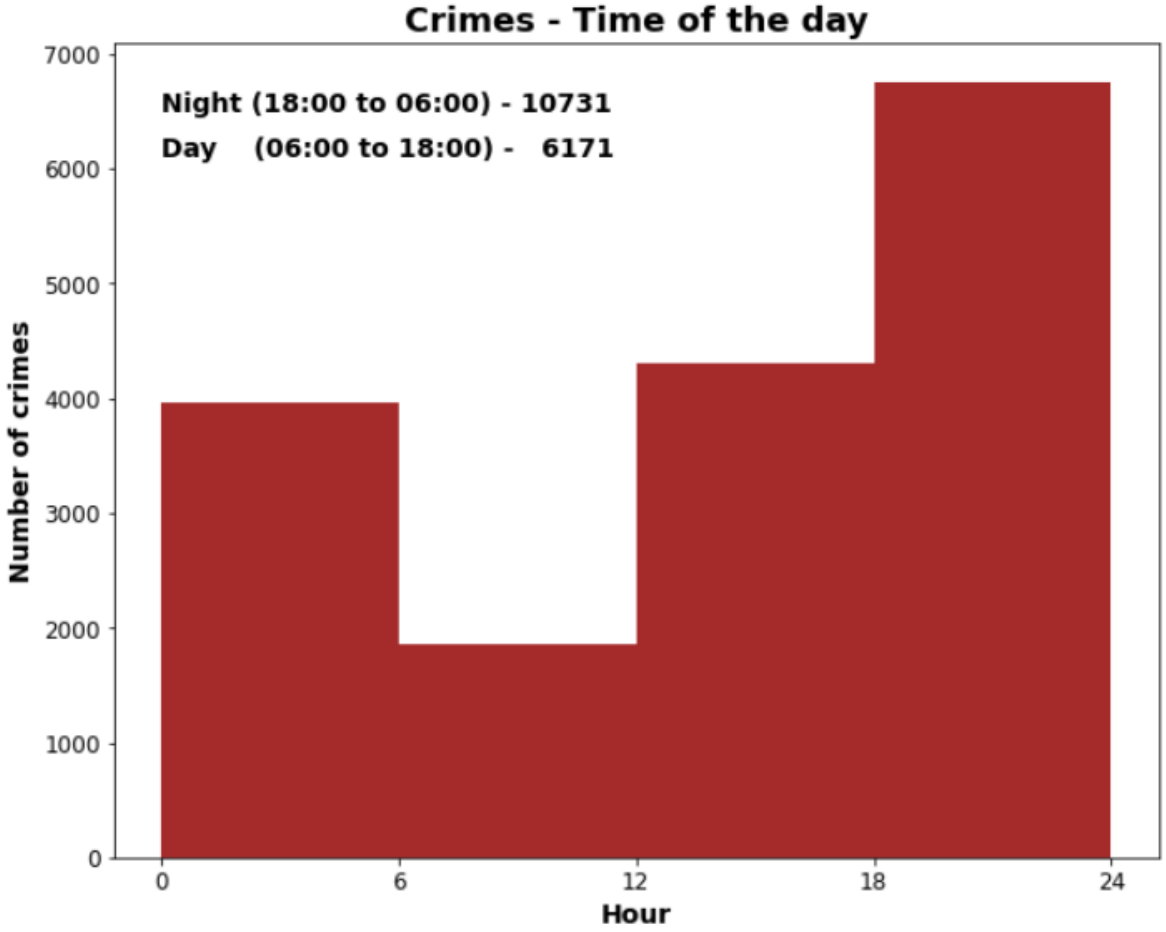
Exploratory data analysis on the crime data of the city is performed to gain insight of the data. This helps with initial understanding of the data and to choose the right set of features. It also helps in choosing the suitable machine learning models. Relationship of the crimes with respect to years, months and time of the day is displayed. Also, the location of neighborhoods is viewed on the map. There is a comparison done between neighborhoods with high and low density of crimes.

3.1.1 Trend over years



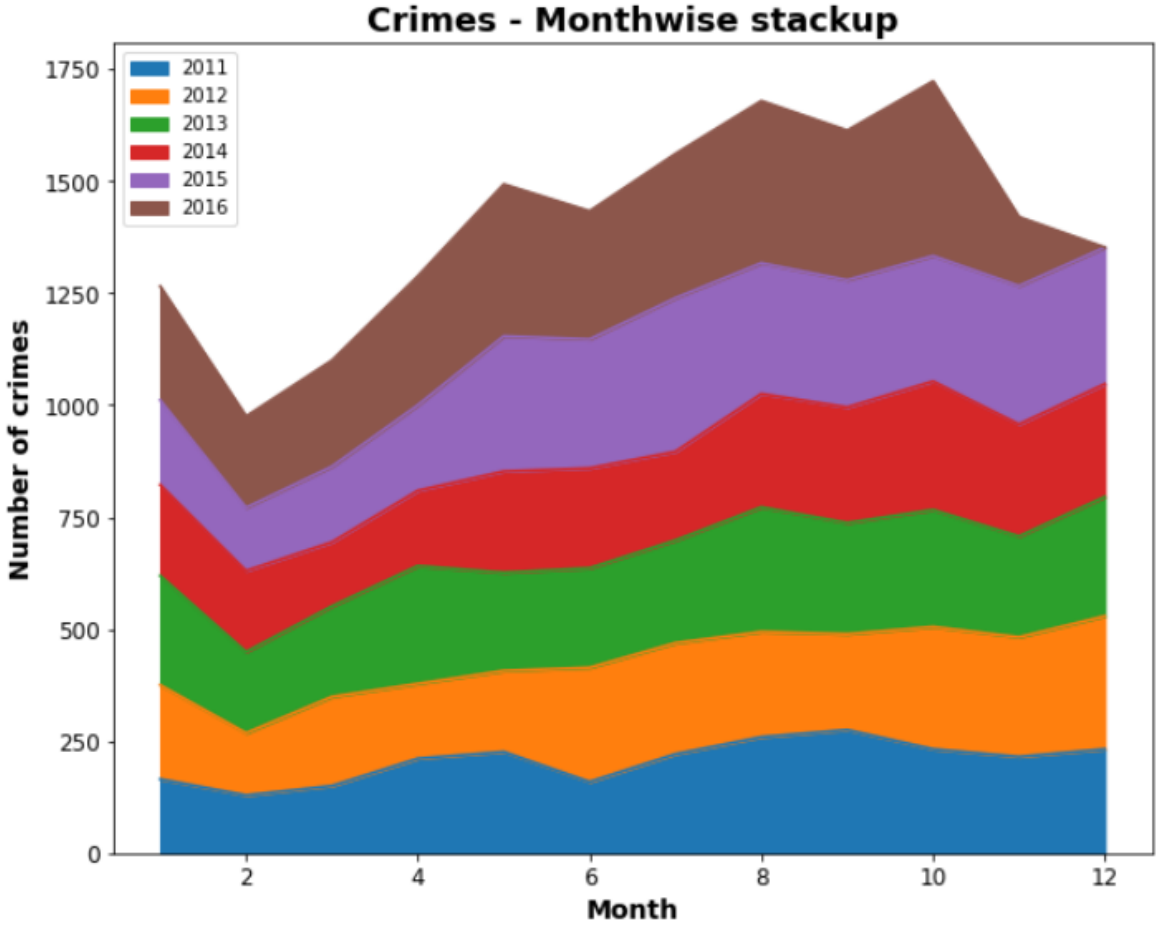
Above chart displays the increasing number of crimes over the years. There is a dip in the crimes in 2014 which was possibly because of the high number of arrests made by police. Anyhow that was not a permanent solution and the crimes started to increase after that. This also proves the need for a system in place to create awareness and predict crimes to establish strategies to reduce the crimes.

3.1.2 Crimes at night



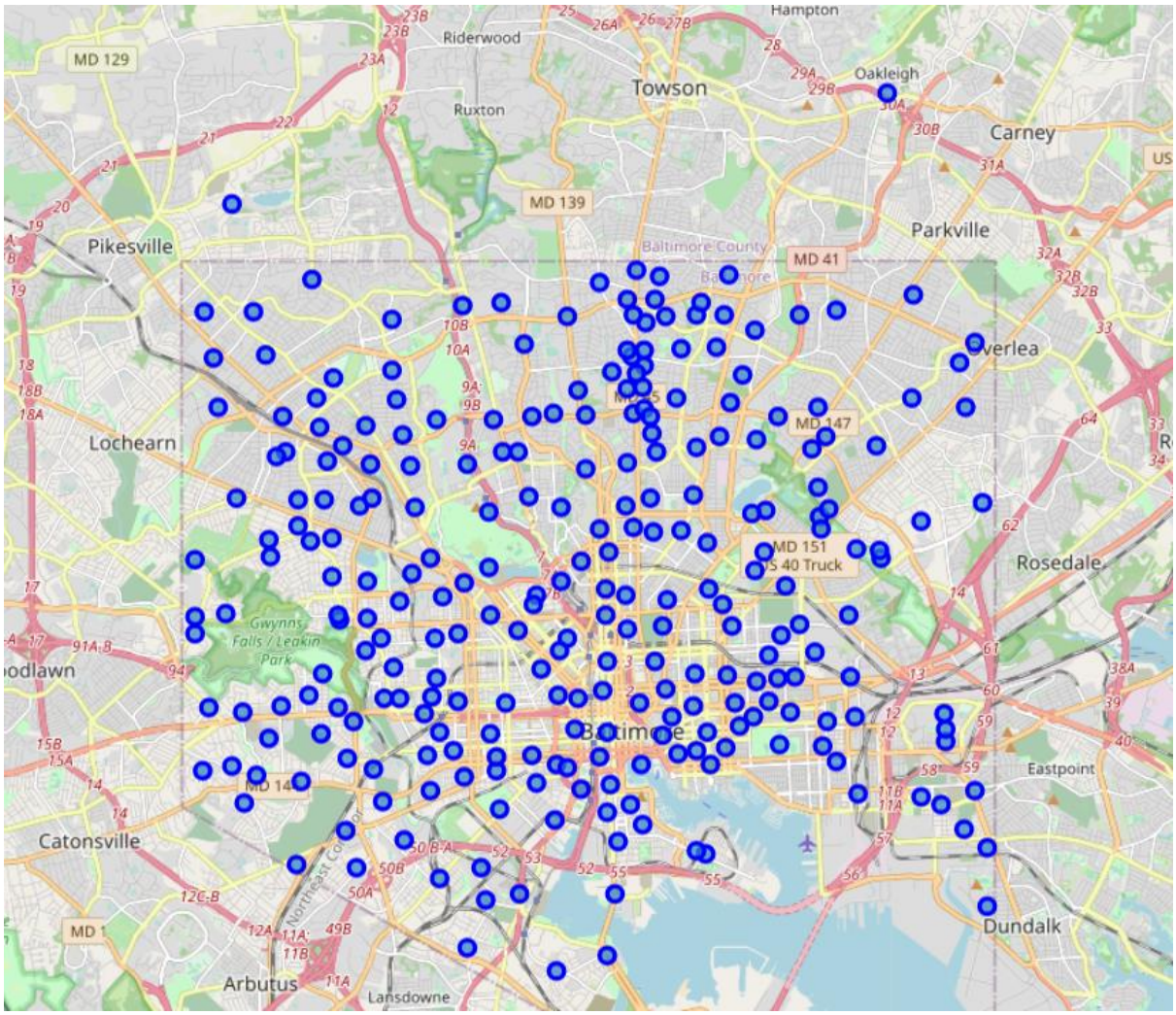
Most of the crimes happened in the darkness. Above chart shows the time of the day when crimes happened. About 70% of the crimes happened after sunset. This shows people and travelers must be extra cautious while visiting some parts of the city after sunset.

3.1.3 Monthly stack up of crimes over years



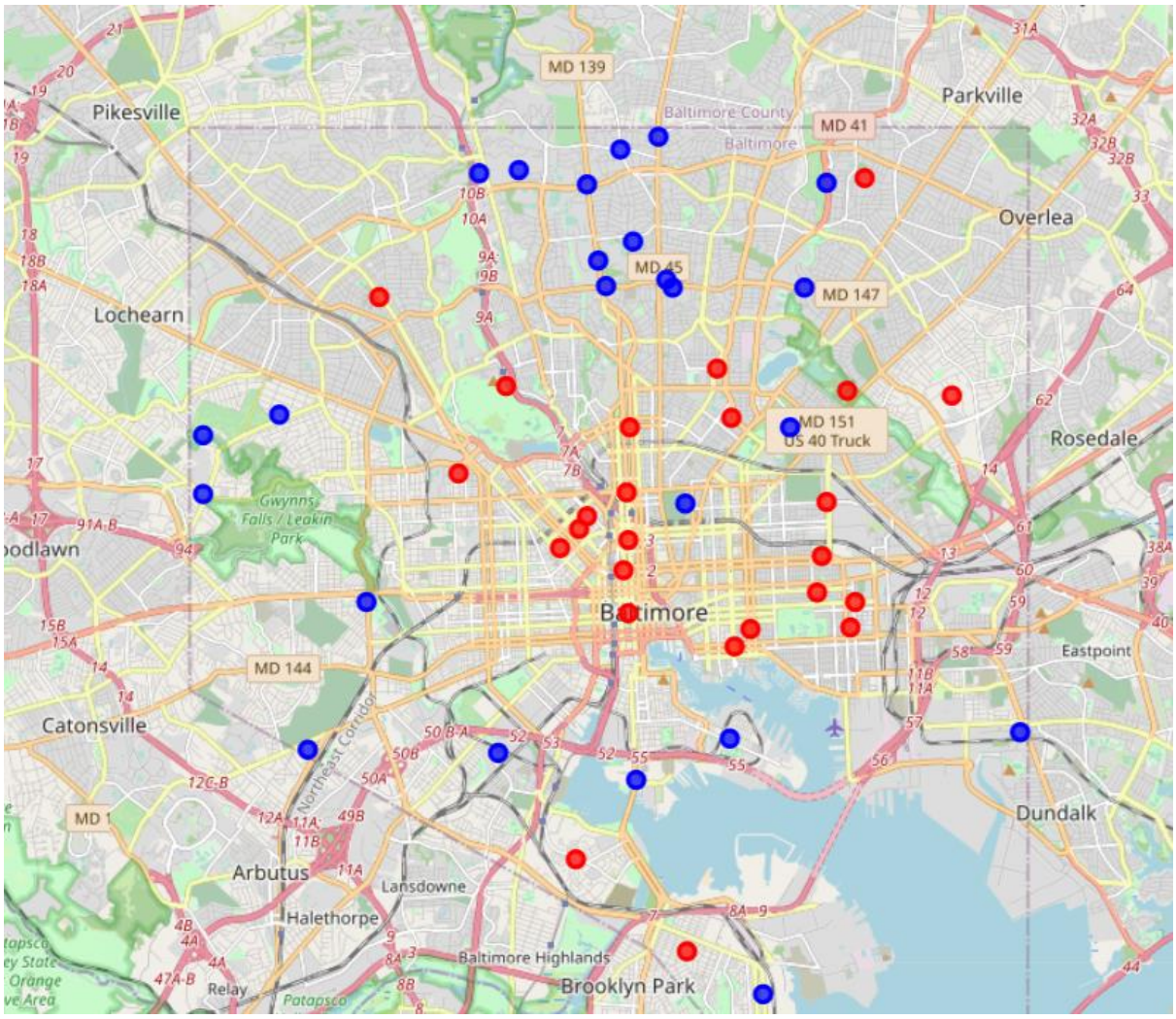
Above area chart shows the monthly split up of crimes stacked up over each year. It is clear the crime rates are higher and increases from June till October every year and decreases from November till February. February accounts for the lowest accumulation of crimes. This could be because of the cold weather during those months. This representation would help the police department to step up more surveillance measures during the summer months due to increased movement of people and travelers.

3.1.4 Location of neighborhoods



This plot shows all the neighborhoods of Baltimore city where crime is reported. This shows how the crimes are spread across all neighborhoods of the city.

3.1.5 Comparison of neighborhoods – High density vs Low density



Above plot shows high- and low-density regions of one crime that is 'STREET ROBBERY'. The main intention of the project is to create awareness to people and travelers so that the crime on the streets is considered. Plot shows the top twenty-five neighborhoods with the greatest number of crimes shown in red and bottom twenty-five neighborhoods with least number of crimes. From the plot, it is clear, many crimes happened in areas that are close to central business district (CBD) of the city. Findings from this help in choosing the features for further analysis of data.

3.2 Clustering model

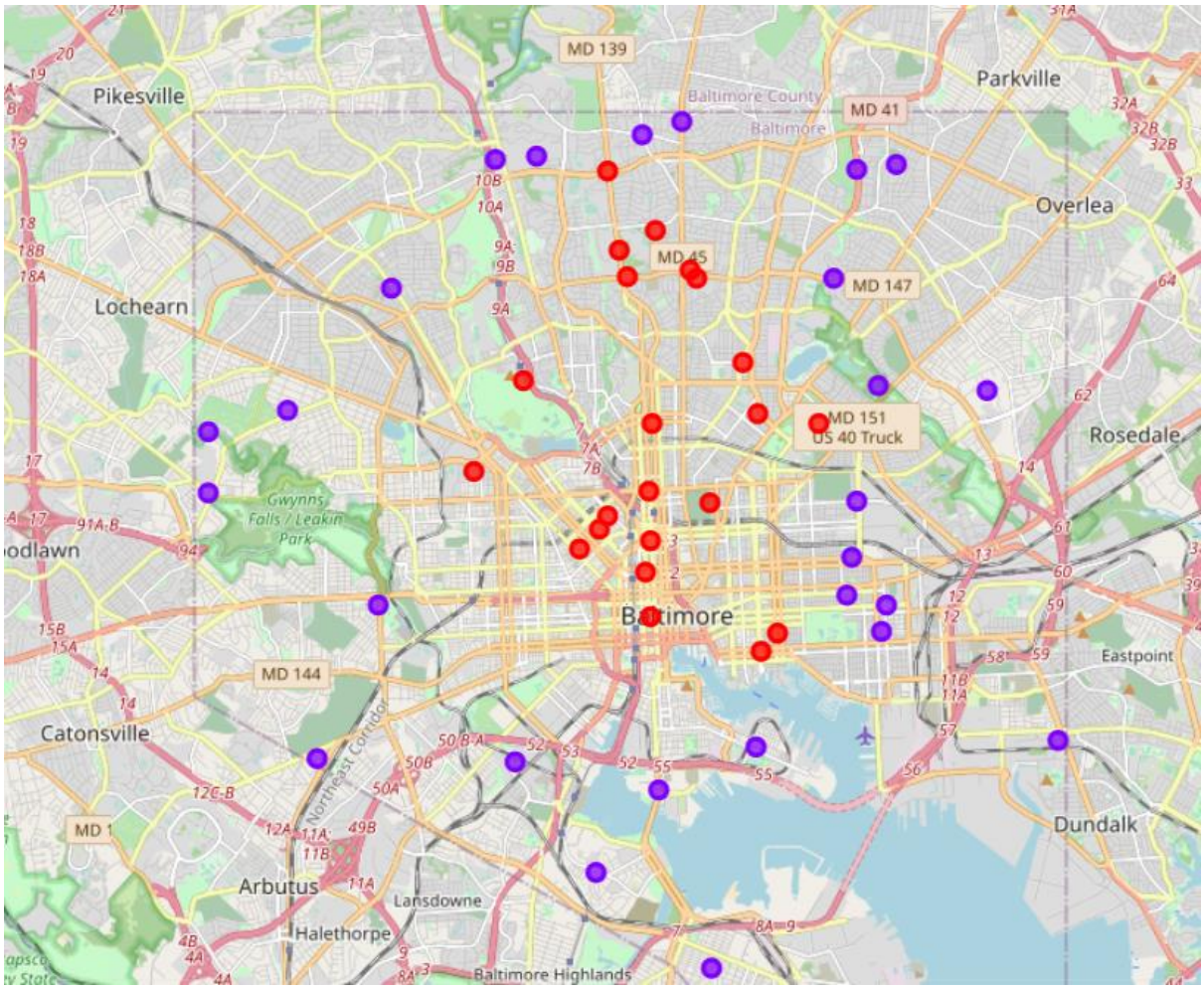
Exploratory analysis of data suggests that crime on the street happens mostly in the central locations of the city. To further understand this k means clustering machine learning model is used to group the neighborhoods. Top features considered for each neighborhood are the distance from central location and the availability of top ten venues consolidated across all the neighborhoods. This data on venues is collected using the foursquare APIs. The top ten venues include bars, liquor stores, restaurants & cafes, and parks. Clustering is done to group neighborhoods into two groups to show the most crowded or busy neighborhoods and less crowded ones.

3.3 Classification model

A logistic regression classification model is created to predict the probability of a crime to occur in neighborhoods. Features included to train the model include the availability of top ten popular venues, total number of venues available in the neighborhood and the distance of the neighborhood from the center of the city. The data included all the neighborhoods where street robbery is reported. A new feature 'High-Low' is introduced to label all the neighborhoods as 1 for high and 0 for low. Number of crimes at the 50% mark when sorting the neighborhoods by number of crimes is used to label the neighborhoods. Further the data is split into train and test data using model selection function of scikit learn. The data is split to have 80% samples for training the model and 20% for testing the created model. Logistic regression of scikit learn is created by setting the solver to 'liblinear' and a regularization strength co-eff of 0.01.

4 Results

4.1 Clustering



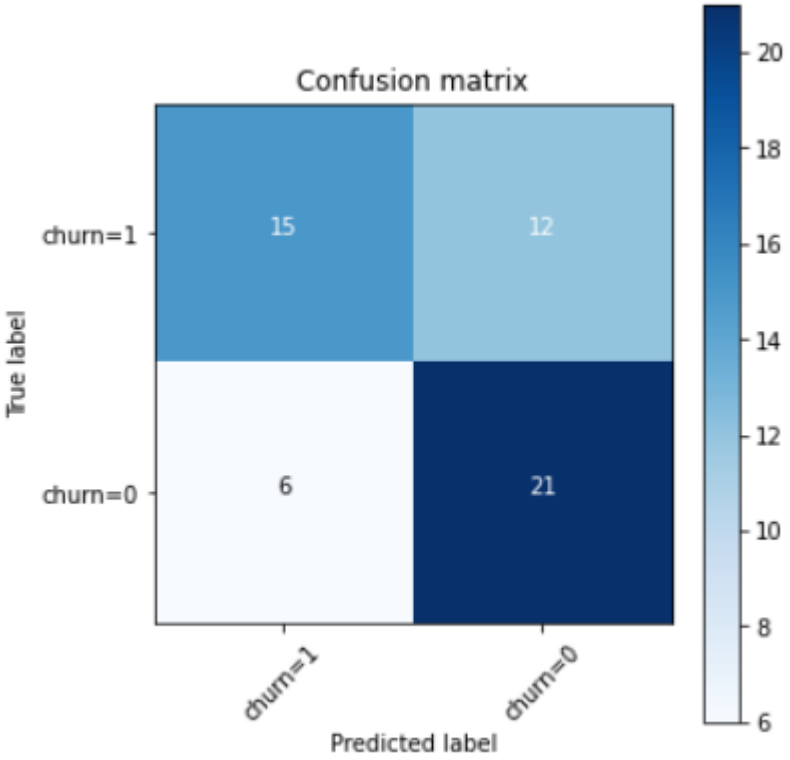
Clustering the samples into two clusters produces a result set that is plotted above. These red dots represent neighborhoods with high availability of top ten popular venues and their proximity to center of the city. The red dots are concentrated in the center of the city. The peripheral violet dots represent neighborhoods with the lesser number of popular venues and the long distance from the center of the city. This plot resembles the one shown in the exploratory data analysis section where high number of street robbery are reported in the central part of the city.

4.2 Classification

The classifier developed is used to predict the probability of crimes in neighborhoods. The trained logistic regression model is tested on the test data. Various evaluation techniques are used to validate the accuracy of the model. Jaccard index, F1 score and log loss values are calculated to evaluate. Values from these techniques are listed below. Results show the models are quite reliable to predict.

Evaluation techniques	**Values**
Jaccard index	0.67
F1 score	0.63
Log loss	0.69

Below confusion matrix is created to show the accuracy of prediction.



This shows good correlation in predicting the neighborhoods. F1 score indicates the accuracy of classifier is 0.69.

5 Discussion

5.1 Observation

Clustering model reveals that the neighborhoods in the central region of the city with popular venues have more chances of street robbery than the neighborhoods in the peripheral of the city. Also, exploratory analysis reveals there is a high chance for crime at nighttime and during summer months. This data would help the local people and tourists to stay vigil accordingly. Classification model created can be used to predict the probability of crime in a neighborhood based on the location and popular venues in it. This model could be used by police to plan strategies and implement measures to control crime in the city.

5.2 Way forward

The accuracy of prediction currently depends only on the location and popular venues. The demographics data for each neighborhood such as population, education, unemployment, poverty etc. can be added to the dataset. Obtaining reliable data on these features would drastically improve the accuracy of the prediction and help in police implementing efficient measures to control crimes.

6 Conclusion

I have taken crime data of city of Baltimore to get good understanding of crimes happened from 2011 till 2016 and to predict crimes in future. The data is cleaned up and studied to understand different neighborhoods their location and proximity to popular places. This reveals a trend in the data with high street robberies occurring in central part of the city. Thus, I created clustering model to group the neighborhoods based on distance and availability of popular venues. This confirmed the observation. I also created a classifier to predict the crime. I used logistic regression to predict the probability of crime in a neighborhood. These models can be used to create awareness and guidance to people and travelers. Classifier can be used by police department to predict the crime, plan strategies and control crime. Further, the accuracy of the models can be drastically improved by including demographics data of the neighborhoods.