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.

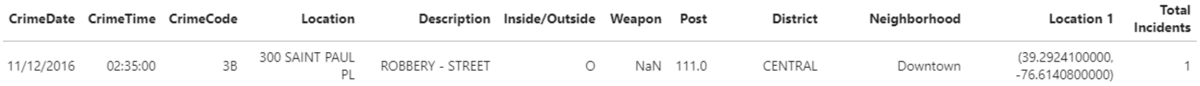
2 Data gathering and processing

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:



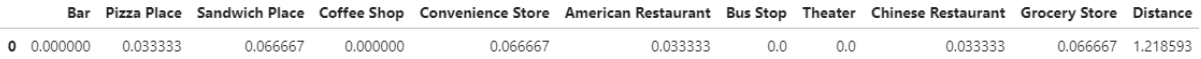
CrimeDate	CrimeTime	CrimeCode	Location	Description	Inside/Outside	Weapon	Post	District	Neighborhood	Location 1	Total Incidents
11/12/2016	02:35:00	3B	300 SAINT PAUL PL	ROBBERY - STREET	O	NaN	111.0	CENTRAL	Downtown	(39.2924100000, -76.6140800000)	1

‘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