Baltimore Crime Prediction Analysis

By Team 22

Programming and Database Fundamentals for Data Scientists **(EAS 503)**

A Project By:

1. Hitesh Rambhaskar Rachuri (50464898)
2. Yashwanth Seshavarapu (50464540)
3. Jayanth Sajja (50463766)
4. Sri Venkata Sai Karthik Rakurty (50476124)

ABSTRACT

Baltimore is one of the cities where residents face high risk. The crime data analysis provides insights that can help residents and police understand the crime in terms of time, day, month, location, and so on, and take appropriate action. A random forest algorithm is used to build a model that predicts the risk factor based on location and time. This model and data analysis could aid in the prevention of future occurrences.

INTRODUCTION

Baltimore has a violent crime rate of 1,858.7 per 100,000 people. This is 406.88% higher than the national rate of 366.7 people per 100,000. Which makes Baltimore one of the cities with a high risk for residents. It would be fascinating to examine crime statistics in Baltimore and draw conclusions from them. The objective of this project is to present a crime analysis, visualize the crimes, and develop a model that predicts risk based on location and time.

DATA:

In this project, the crime dataset of Baltimore city was downloaded from the official Baltimore Police Department. The data file gives an overall view of a crime, defining the location of the crime, characteristics of person like age, gender, race etc. Also, the data describes the type of crime (Larceny, Theft, etc..), weapon used (Rifle, Knife, etc..) and section of crime (Crime code). The data file contained around 543617 rows and 23 columns.

DB SCHEMA:

The data was loaded into our system from the csv file and was then preprocessed. The columns which were repeated were eliminated. The data has been normalized into three tables, such as crime, location, and person. Crime table holds the precise information of the crime such as Crime Date, Crime Time, Crime Code (section under Crime registered), Crime Description (Ex: Burglary, Rape.,) and Weapons (Detail of weapon used). The table has the primary key of CrimeID. The Location Table contains information about the location of the crime and has attributes such as Latitude, Longitude, Location, District, Neighborhood and Location ID as a Primary key. Person table contains PersonID as a Primary Key, Gender, Age, Race, Ethnicity. Person Table holds the connection between the remaining two tables with the foreign keys CrimeID and Location ID.

Analysis:

With the help of a normalized database, for instance, analysis is carried out by removing the data necessary for a few scenarios such as: total offenses broken down by weekday, total criminal activity by calendar month, highest crimes in a single day, period of time when crime rate is high, crime rates over last decade, locations where crime is most prevalent, genders most affected and often used weapons, most attempted crimes in the district.

For easier interpretation, the data is transformed into bar, tree, line, and pie charts. To identify the areas of Baltimore with a higher crime rate, a circle marker is placed on the city map. A random forest model is developed to forecast the risk level of a person is resting in a neighborhood at any moment by using data such as the neighborhood, time, and number of crimes and adding additional column of risk component.

Results:

For a clearer understanding of the analysis, the raw results of the graph query results are shown using the Python packages matplotlib, seaborn, and folium. Plot depicts the number of crimes that have occurred in the city in chronological order by weekday, month, and hour of the day. According to the afore mentioned plots, it can be deduced that every weekday has a comparable crime rate, with Friday having a little higher rate. Crimes have also been more common between 3pm and 8pm, with January and February appearing to have lower crime rates than June to September. The murder of Freddie Gray, a 25-year-old black man, while in the custody of the Baltimore Police Department sparked outrage in Baltimore, Maryland, and later across the country, making April 27, 2015, the day with the highest crime rate from the previous ten years. Though there is a slight difference, the proportion of female victims of crime is higher than that of male victims. The top 25 crime types are shown on a tree map, with assault and theft ranking as the top two offenses. To determine which type of crime is more common in a certain district, a district vs. category of crime plot is created. The degree of risk (low, medium, and high) that a person will confront at a specific time and place is predicted using a random forest model that is fitted to data containing neighborhood, crime time, and risk level. The model's accuracy was close to 80%.

CONCLUSION

This project's main goal was to use effective Python strategies to deal with large data sets and to formulate results based on the prepared database schema. According to the data presented, every weekday has a comparable crime rate, with Friday having a slightly higher rate. Crimes have also been more common between 3 p.m. and 8 p.m., with January and February appearing to have lower crime rates than June to September, and female victims of crime outnumber male victims. As a result, these insights can be used to raise public awareness, and police can use them to track crimes and take preventative measures in the future.

Future Scope:

Furthermore, the random forest model developed in this project can be turned into an application that people can use to predict crime risk based on their location. The application's concept can be expanded in the future scope of this project.

References:

https://data.baltimorecity.gov/datasets/baltimore::part-1-crime-data-/explora (Data Repository)