Crime Analysis of Cities in the USA

Abhay Rajendra Dixit, Adya Shrivastava, Pranjal Pandey

1. INTRODUCTION

Infrastructure, gross national income, per-capita income, public health are not the only things that count toward the development of a country. What's also necessary is to see how safe the citizens are and how safe they feel to be residing in that country. A high crime rate is not only indicative of bad governance but also a very bad mark on the country's global reputation. The US has seen some interesting trends in crime rates since World War II[4]. The crimes reported were very high during the period 1970s to 1990s and gradually the numbers dropped[8] after that due to the introduction of stricter laws, expansion of the police force, rise in per-capita income, etc. Crime depends on a lot of factors like the location, population of the area, unemployment, education, etc. A good analysis of these crimes helps us understand why the crime occurred and what measures the government can take to prevent it. Effective crime analysis can help the police department determine what areas are the most impacted and what laws can be enforced to decrease crime rates.

In this project, we would like to study in detail the crimes that have been committed in the last 10 years with an objective to uncover interesting patterns that would help take measures to prevent these crimes. For this purpose, we will be picking 5 prominent cities in the US namely Chicago[3], Austin [1], Baltimore[2],Los Angeles[7] and Rochester[9] . The datasets obtained from different government websites of the aforementioned cities add up to over 2.2 million records containing interesting attributes such as location, type of crime, description, date, action taken which will be our primary focus. However, the data in these datasets have varied formats, contain irrelevant attributes, missing values, redundant values, and abbreviations which would require a thorough pre-processing. After this crucial step, we plan to analyze the trend in committed crimes over the years, figure out the most common type of crime, look for areas worst affected by crimes, and apply mining techniques like clustering to extract hidden patterns and regression to see the closeness of predictions with reality. And finally, we will be leveraging visualization software, tools, and libraries like Tableau, Minitab, Matplotlib, etc., to support our findings.

To store the datasets, we will be using MongoDB owing to its flexibility in data insertion, retrieval, and updation and also because of the ease of its use in data analysis. After a careful analysis of the attributes, we will be creating a common collection in MongoDB having only the attributes of our interest. Our deliverables will include source code to pre-process the data, load the data on to MongoDB, and carry out all the necessary analysis of the data. Along with source code, we will provide a detailed report of the techniques and algorithms implemented to pre-process and analyze data.

The rest of the study is organized as follows. The next section provides the motivation for taking up this topic. This is followed by the design section. In the next section, we will talk about the implementation and mining techniques that are used to discover patterns and trends. Then we will move on to discussing inferences from the analysis, current status, and future work. The last section concludes the study.

2. MOTIVATION

Crime is a pressing issue in most of the societies. According to the [5] survey over 63% of Americans believe that there is more crime in the city than the previous year. Although it has declined since the year 1990, some types of crime has seen a sharp increase like murder in 2020[6]. The crime trends tell a lot about how the society has evolved and statistics play a huge role in analyzing the trends. It also provides valuable insights about various causes of crime. Certain types of crimes may exhibit patterns in the times and the places that they occur. In this project we try to discover such patterns, analyse and compare the various types of crime and their trends in five major cities of the USA.

For this project we have focused our study on the cities of USA namely Chicago, Baltimore, Austin, Los Angeles and Rochester. The datasets consist of important information such as date, address, offense type, weapon used victim age, time of the day etc. This will be used to discover interesting trends and patterns.

3. METHODOLOGY

3.1 Data Collection and Prospecting

As mentioned earlier, for this project, we have selected 5 prominent cities of the USA which are LA, Chicago, Baltimore, Austin and Rochester. We obtained the crime datasets from open data portals of the respective cities. The datasets on each of these portals have a plethora of information related to the crime. However, we have focused only on a few common important attributes such as date and time of crime, type of crime, weapons used, location of the crime incident, latitude, longitude, etc., on which we plan to carry out a detailed analysis. Other attributes in the datasets like block, arrest status, crime codes, ids, ward number, etc. are discarded.

3.2 Data Preparation

Data preparation and cleaning is a crucial part of data mining and analysis and the stage that requires maximum efforts in terms of time. It is a process that involves formatting, correcting, combining and refining of data to make the data ready to use. Like any other real-world datasets, the datasets we have chosen also contain several issues like inconsistencies, missing values, empty rows, abbreviations, etc. which need to be addressed. For instance, date attribute has different formats across the 5 datasets. To achieve uniformity, we have split the date into day, month and year. An added advantage of splitting the date is that data retrieval and analysis of each of these attributes becomes very easy. Another issue was that each of these datasets were of different sizes and covered crimes from dates of varying range. The Chicago dataset was massive, having over 7 million records with crimes dated back to 2001 while the Rochester crime dataset had crimes that started from 2011. Again, to have a common timeframe for all the datasets, we retained only the records that fell in the time window of 9 years from 2011 to 2019 and eliminated the rest. We also deleted records whose dates were missing as dates are a crucial part of our analysis and crimes without dates add little value to our goal. To address the problem of abbreviations, we discovered common short forms used in the datasets and designed a common function to replace all of these short forms with complete and meaningful words. For instance, 'ST's were replaced with 'STREET', 'RD's with 'ROAD', 'AV's with 'AVENUE' and so on. After all the mentioned modifications and few other minor refinements, we finally obtained datasets that were ready to be pushed into the database.

3.3 Data Data Storage

For our project, we decided to use a NoSQL database, specifically MongoDB. The main reason for choosing MongoDB is that insertion, updation and data retrieval is absolutely hassle-free and way faster than SQL databases. Also because Mongo offers fewer constraints for data storage as compared to SQL databases. We merged datasets of different cities into one common collection and added another attribute (city) to distinguish data of different cities. From a programming perspective, we extensively leveraged the functionalities of the Pandas library for data pre-processing and used Pymongo library to carry out all the operations pertaining to MongoDB.

4. ANTICIPATED CHANGES IN THE FUTURE

This phase of the project was primarily focused on data pre-processing. In the next phase, we plan to uncover interesting patterns in the data by applying various data mining techniques such as clustering, classification, regression, etc. Depending on the kind of patterns we try to discover, we might add a couple of more attributes such as description of crime, premise, arrest status. Also, from what we observed, the "offense type" attribute has too many distinct values. We plan to group these values to form a common category. For instance, all the offenses resulting in death of person can be categorized as murder. This is only an example and not the exact name that we would be using.

5. REFERENCES

[1] *Austin Dataset https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu.*

[2] *Baltimore Crime Dataset https://data.baltimorecity.gov/Public-Safety/BPD-Part-1-Victim-Based-Crime-Data/wsfq-mvij.*

[3] *Chicago Crime Dataset https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.*

[4] *Crime in United States https://en.wikipedia.org/wiki/Crime$_i$n$_t$he$_U$nited$_S$tates.*

[5] *Gallop Crime Survey https://news.gallup.com/poll/1603/crime.aspx.*

[6] *It's Been 'Such a Weird Year.' That's Also Reflected in Crime Statistics. https://www.nytimes.com/2020/07/06/upshot/murders-rising-crime-coronavirus.html.*

[7] *Los Angeles Crime Dataset https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-2019/63jg-8b9z.*

[8] *Reported violent crime rate in the United States from 1990 to 2018 https://www.statista.com/statistics/191219/reported-violent-crime-rate-in-the-usa-since-1990/.*

[9] *Rochester Crime Dataset https://data-rpdny.opendata.arcgis.com/datasets/rpd-part-i-crime-2011-to-present/.*