Analysis of Traffic Accidents and Violations in Chicago

Abhay Rajendra Dixit, Pranjal Pandey, Ravikiran Jois Yedur Prabhakar

1. INTRODUCTION

Traffic violations are a common and inevitable problem in every city. Traffic violations include jumping redlights, overspeeding, overtaking from the wrong side, driving on wrong lanes etc. Often, these violations are directly or indirectly linked to accidents. Accidents are one of the main contributors to global mortality rate. In the United States, for people aged 1–54, traffic crashes are the leading cause of non-natural death for healthy U.S. citizens residing or traveling abroad[5]. According to the World Health Organization (WHO), approximately 1.35 million people globally die every year as a result of road traffic crashes[8]. Yet, unfortunately, there is not enough effort made to address this issue. The focus is usually on making stricter traffic rules and not on carrying out a root cause analysis of the problem at hand. Most of the existing analysis is based on survey data obtained from self-reported accidents and violations which are usually prone to errors. Hence, it is important that we study the relationship between traffic violations and car crashes at greater depth to gain meaningful insights which might help in taking preventive measures to avoid accidents.

In this project, we study the relationship between redlight violations in particular and the accidents happening in a particular city. For this purpose, we will be picking three transportation datasets namely, Red Light Camera Violations[3], Traffic crashes[7] and Speed Camera Violations[6] of the city of Chicago obtained from the City of Chicago portal[1]. The chosen datasets consist of 550K (approx.) rows and 10 columns of Red Light Camera violations, 400K (approx.) rows and 49 columns of Traffic Crashes and 200K (approx.) rows and 9 columns of Speed Camera violations. They contain attributes like address, date, time, the traffic control device at the location of crash and the number of red light violations on the respective date, to name a few. The datasets consist of information spanning over the period from 2014 to June 2020. These datasets have different types and formats of attributes like address and street number, missing values, redundant values etc., which would require thorough pre-processing and preparation before it is loaded into our data mining algorithm.

We will be considering various attributes related to violations and accidents such as date, time and location in the aforementioned datasets to achieve our goal. We would be making use of MongoDB as the document-based model as it would be better than a relational model for data analysis because it would be more efficient in handling complex mathematical calculations. Our deliverables include source code to load the data to MongoDB, filter and clean and process the data, mine meaningful information from the data and a detailed report describing the procedures and techniques involved in implementation of data cleaning and mining algorithms used in this project.

The paper is organised as follows. We are going to talk about the motivation for taking up this topic. This is followed by the design section. We then move on to the implementation section that provides details about the techniques used in cleaning and analysing data. This section is followed by the inferences from the analysis, current status and future work.

2. MOTIVATION

Jumping red lights is one of the main forms of traffic violation that is linked to road accidents[4]. Although stricter rules can reduce the number of red light violations, it is still not sufficient. People lack awareness about road safety and they continue to violate rules unless they understand the importance of following them. Another prominent form of violation is the speed camera violation which has a significant role in the increase of traffic crashes. Most speed violations occur due to drunk or reckless drivers. It can also be due to negligence[2]. In both these cases, the consequences of violations are serious which include severe injuries and fatalities resulting from crashes. In this project, we study, analyse and detect correlation between the redlight and speed violations with the number of crashes.

For this project we have focused our study on the city of Chicago owing to its size and the abundant data that is available that is suitable for our study. The datasets consist of important information such as number of accidents, their location, dates, violations etc.

3. METHODOLOGY

3.1 Data Collection and Prospecting

Data prospecting is the foremost step in understanding data. It will help us explore our large datasets which come in different forms. It is important to keep track of the sources and employ different data collection methodologies to ensure better understanding of the data.

For this project, the data has been taken from the city of Chicago open data portal[1]. Three datasets have been collected from the portal namely Traffic crashes, Red Light violations and Speed Limit violations. All the three datasets provide the latest records which are collected in the form of CSV format. The traffic crashes dataset provides information regarding crashes like street condition, weather condition and posted speed limits. The collected records are based on the documented entries reported by the reporting officer at the crash location. The other two datasets, the Red Light violations and the Speed Limit violations lay out the information regarding the violations' like street details, location details and number of violations. The collected records are based on the entries generated by the camera and radar systems in the city of Chicago.

Data prospecting is an important step to achieve an effective data model. For efficient data prospecting, the process should always be aligned to the defined goals of data modelling. When we collect the data, it does not always fit the needs of a mining algorithm. It usually contains certain attributes which would be of high importance and some attributes which would be completely irrelevant to the defined goal. In the datasets chosen for this project, there are a few attributes which give the impression of being useful in order to find the correlation. For instance, in the Traffic crash dataset, the attributes like weather information, device information and injury are not aligned with our goals meanwhile, the attributes like street address and crash date are of utmost importance. In the other two datasets, the information about the camera device and geographical location serve us no purpose. Hence, in both the datasets, there are unnecessary attributes which should be discarded.

In this project, before selecting or discarding the attributes, we have created logs which rank the attributes based on how important they are to our analysis as high, medium and low. The attributes like street address and date might serve as a binding factor later in our data modelling. Hence, they are ranked high. The attributes with low rank are discarded from the dataset in the data preparation phase.

3.2 Data Preparation

Data preparation and Cleaning is the phase where we transform the raw data into clean data that is suitable for mining and analysis. This is a tedious process that requires formatting, correcting, combining and refining the datasets that result in enriched datasets.

The datasets that we have chosen for this project have a number of inconsistencies. For example, the dates and addresses have different formats in each dataset which makes it necessary to clean them. Another important aspect is the presence of missing values. They are present either as a missing value or as a string value that depicts a missing value. These datasets also have many attributes that are irrelevant for our analysis. This is taken care of by reviewing and ranking them and concentrating on the goal of the project at hand. Some of these unnecessary attributes are, the Report number, the Device condition and the Weather condition.

Some of the attributes like Address also have many 'na' values which have been currently converted to an arbitrary number (-9999) for our reference. We will be replacing them with appropriate values through the course of the project.

The address attributes in the three datasets are in different formats. Two of them have the street number, the street name and the street direction in a single attribute whereas, the other dataset has these three features as separate attributes. Many inconsistencies in the names of these addresses existed namely, "ST" for "Street" in one of the datasets and "STREE" in another. The same kind of inconsistency was found for other values like Avenue, Drive, etc. We have corrected these inconsistencies in all the three datasets using the functions of Pandas Library.

The date attribute, which is what we are likely to use as a binding attribute was also inconsistent across the datasets. It was in the date-time format which was not desirable for our analysis. We have converted this attribute into the date format. We did an analysis of all the datasets by using pandas and found that there are only 3 values in all the attributes that have null values in them.

3.3 Data Storage

For storing data in this project, we have chosen MongoDB as the document based DBMS as it is more suitable for our analysis. We have created three collections, each containing the Red Light violations, the Traffic Crashes and the Speed Camera violations datasets respectively. We have chosen the attributes that provide information about the area, location and cause of the violations and crashes.

From the programming standpoint, we are using the PyMongo library to load, retrieve and update the data and also perform other database related operations.

4. ANTICIPATED CHANGES IN THE FUTURE

The primary focus of this phase has been data extraction and preparation. For the next phase of our project, we will be explaining mining algorithms that we use to extract meaningful information from the datasets. Since we are not sure about the mining algorithm we will be using, we anticipate a few changes in the dataset that we are using. They are: Addition of more attributes to our datasets like latitude and longitude to analyse the accidents or traffic violations happening in a particular area, creation of new collection to include specific attributes of the dataset based on the needs of the mining algorithm, introduction of redundant attributes to make the mining algorithm discover the pattern faster. An example for our case could be to create a new attribute called Number of Crashes by grouping the traffic crashes based on day, month or a year. Lastly, we will be using One-hot encoding for the analysis and processing of categorical data.

5. REFERENCES

[1] The Chicago Data Portal. `https://data.cityofchicago.org`.

[2] Defending Motorist Rights in the Free State. `http://www.mddriversalliance.org/p/arguments-against-speed-cameras.html`. Maryland Drivers Alliance.

[3] Red Light Camera Violations. https://data.cityofchicago.org/Transportation/Red-Light-Camera-Violations/spqx-js37. Chicago Data Portal.

[4] Red Light Running. https://www.iihs.org/topics/red-light-running. Insurance Institute for Highway Safety.

[5] Road Traffic Injuries and Deaths—A Global Problem. https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=Road%20traffic%20crashes%20are%20a,citizens%20residing%20or%20traveling%20abroad.

[6] Speed Camera Violations. https://data.cityofchicago.org/Transportation/Speed-Camera-Violations/hhkd-xvj4/data#Manage. Chicago Data Portal.

[7] Traffic Crashes - Crashes. https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if. Chicago Data Portal.

[8] W. Pietrasik. Road Traffic Injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.