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[9]. 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. DESIGN

To achieve sensible results from analysis, the raw data must be converted into a form which could be fed to a data mining model. This process is a multi phase process where the data goes through different phases to get a better understanding of the nature of the dataset. Data preparation and preliminary analysis of data are essential steps for discovering patterns and deducing various relationships which must be done before feeding data into the data mining model. These steps provide information and understanding of the dataset which also helps in deciding the data mining model. This section of the project discusses the numerous phases as stated in Figure 1 and the technology used in each phase. First, the three collected datasets namely red light violation, speed limit violation and traffic crashes goes through the preprocessing phase. In this phase, different preprocessing techniques like replacement of missing values, elimination of less important attributes and formatting data values have been implemented. The second phase involves statistical and preliminary analysis of all the three datasets. In the final phase, different analysis and mining techniques are carried out which are discussed in detail in the following sections of this paper.

5. IMPLEMENTATION

5.1 Analysis based on Date

The analysis based on date has been done to bolster the statement that road indiscipline increases the chance of traffic crashes. Ideally, the number of violations occurring in different areas should not affect the occurrence of crashes in a given area. For instance, if there is an increase in traffic vi-

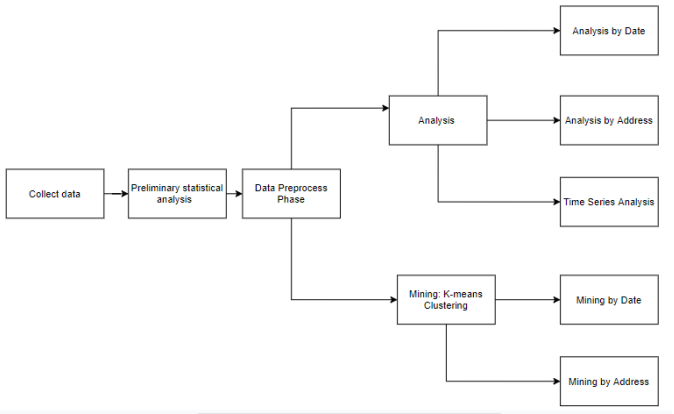

Figure 1: The Design

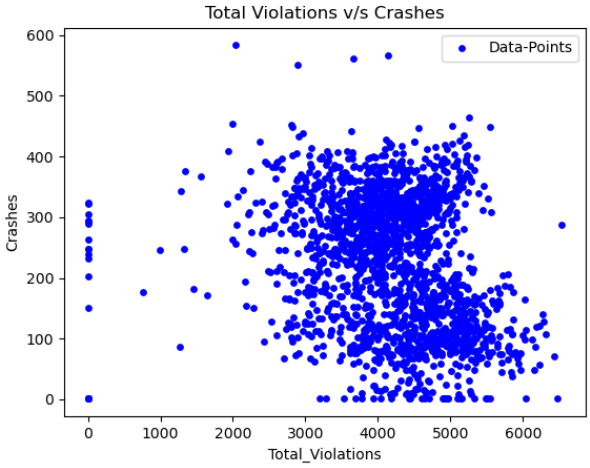

Figure 2: Total Violations v/s Crashes

olations in south Chicago, it should not cause more crashes in the north of Chicago. To establish this claim, we have performed the following experiments.

In this analysis, the three datasets are merged based on the dates. Various quantitative analyses are performed using date as a binding factor to find the total number of violations and crashes that occurred on the particular day. This is done by grouping the dataframe by date. The three datasets are merged doing a left join on date. This merged dataset gives information about the total number of red light violations, speed light violations and crashes for any particular day. If there is no crash or violation on a particular date, the corresponding values would be set to zero. In order to establish a relationship between both violations and crashes, Pearson's correlation coefficient is used. The correlation coefficient for the number of crashes and red light violations is -0.306571, for the number of crashes and speed limit violations is -0.018954 and that for the number of crashes and total number of violations for a given date is -0.239416. The result of the analysis shows that there is very weak negative correlation exists between the number of violations and crashes occurring on a particular day if the street address is not taken into consideration. The results can be observed in Figure 2. The graph is plotted taking x axis as violations and y axis as the total crashes.

5.2 Analysis based on Location

In this case, we make a detailed analysis of red light violations, speed camera violations and traffic crashes with respect to a given location. We have considered only street names as our location and omitted other details for simplicity. We have used *groupby* function provided by pandas to group the rows of the data by the street name and get the count of red light violations, speed limit violations and crashes. We have then used this data to find out if any of the violations i.e., red light violations, speed camera violations and total violations is related to the number of crashes. As mentioned before, we used Pearson's correlation coefficient for this purpose. On calculating these values, we can observe that red light violations has the highest positive correlation with the number of crashes with a value of 0.780 (Shown in Figure 4), followed by total violations with the value of 0.635. Speed camera violations is moderately correlated with a coefficient of 0.595 (Shown in Figure 3).

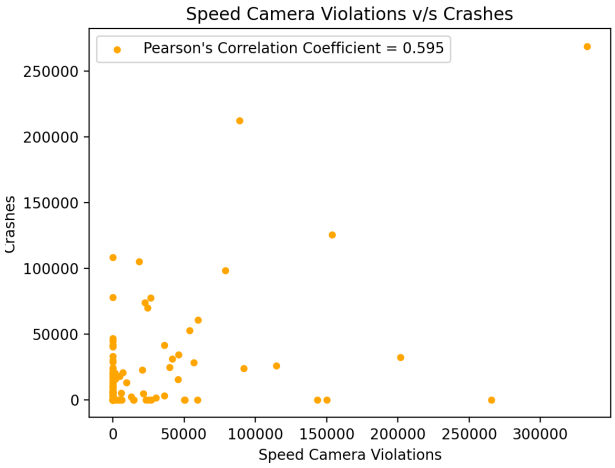

Figure 3: Speed Violation v/s Crashes

5.3 Time Series Analysis

The violations and crashes at different time intervals have been analysed and the data has been plotted. This is done to recognize possible relationship of crashes with time. The time interval in this case is each month of the last six years.

The number of violations per month has been calculated by making use of mongoDB. The violations attribute of the speed camera violations and the red light violations collections gives us the number of crashes with respect to each day. In traffic crashes collection, the number of crashes per month has been calculated by grouping the data based on

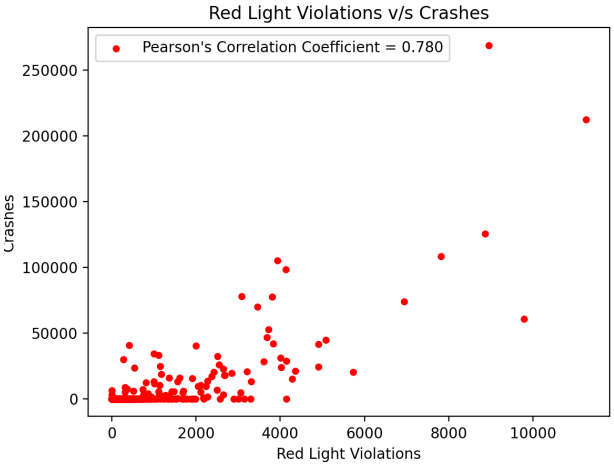

Figure 4: Red Light Violation v/s Crashes

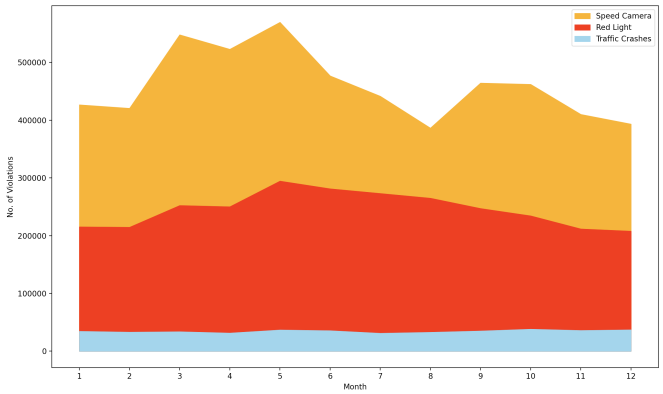

Figure 5: Monthly Crashes and Violations

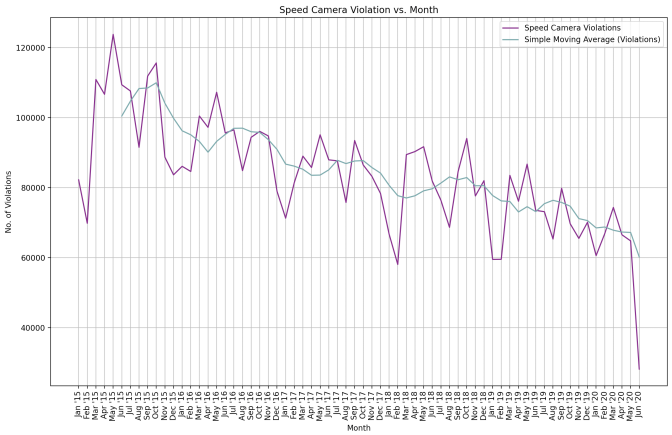

Figure 6: Speed Camera Violations

each month. This data has been plotted by using the *pandas* plot function by converting the results from mongoDB to a dataframe.

Initially, the graph was plotted for all the collections combined to notice any pattern. However, the patterns were not very evident in the graph as the range of values of crashes and violations were different for each collection. Thus, separate time series graphs are created for each collection.

Moving averages were calculated for each plot and the results were plotted to see the rise or fall in the number of violations with respect to time. This was calculated for all three collections by converting the data obtained from the MongoDB queries to pandas dataframes. After looking at the red light traffic violations collection, some kind of seasonality was noticed in the graph. So, the required averages and metrics were calculated to deseasonalize the data.

We can notice from the graph in Figure 5 that the number of crashes is very less when compared to the number of violations that have occurred. Every traffic violation need not result in a crash. It might be possible that the crashes were not reported.

From the speed camera violations graph in Figure 6, we can notice that there is a downward curve in the number of violations. This downward tendency is confirmed when we look at the moving average line plotted on the graph. Thus, we can conclude that the number of speed camera violations has reduced over the course of the last 6 years.

The moving averages were also calculated for the red light violations data. However, it gave little information about the trends in the number of red light violations in general. If we notice carefully, most violations in each year occur in the months of June or July. We can also see that these violations are less every year in the winters (December to February).

From the red light violations graph shown in Figure 7, at first glance, we can see what appears to be a seasonal trend. To reduce this effect of seasonality, the deseasoning operations have been performed and the graph is shown in Figure 8. However, the graph representing the deseasoning operations does not convey a lot. The graph is smoothened slightly and no significant upward or downward trend is seen.

The number of traffic crashes with respect to each month has been plotted as a time series plot (See Figure 9). In this graph, we notice that the number has increased through the course of the last 6 years. However, we can see a clear decrease in the number of crashes since January, 2020. One reason might be the lockdown and thus, fewer vehicles on the streets due to the Coronavirus pandemic. The moving averages calculated for the traffic crashes data is also plotted and this confirms the increase in the number of crashes till January, 2020.

5.4 Heat Map

A heat map representing the number of traffic crashes has been plotted [8] using the longitude and latitude values in Tableau. The locations marked in orange/red on the heat map represents the crashes that have occurred in the respective areas. We can see that the most number of crashes have occurred in the area that is marked with a bright red and the lesser number of traffic crashes can be noticed in the

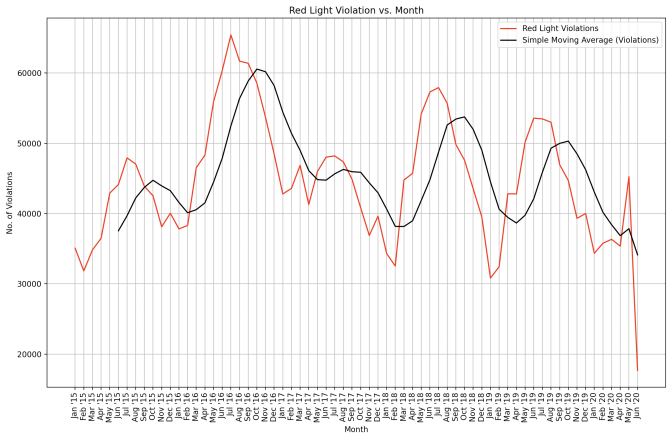

Figure 7: Red Light Camera Violations

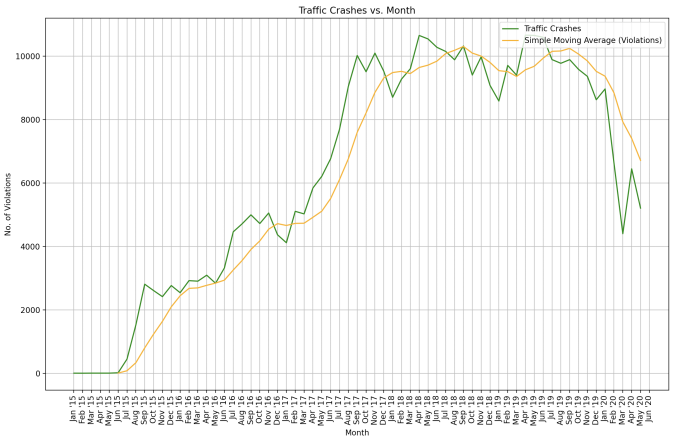

Figure 9: The Analysis Design

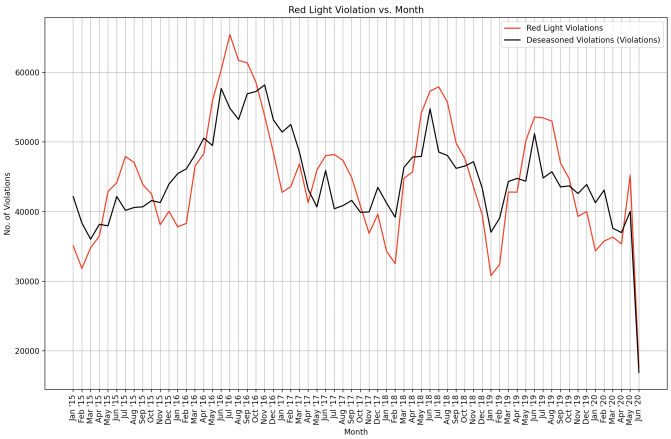

Figure 8: Deseasonalized Red Light Camera Violations

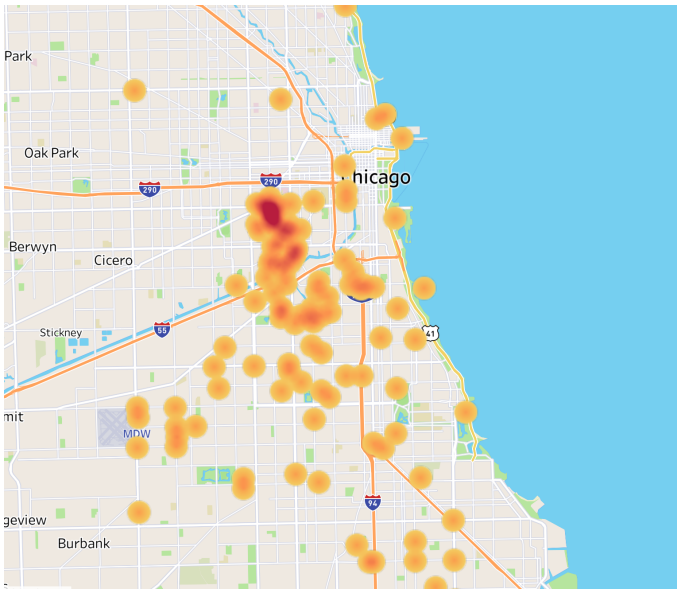

Figure 10: Heat Map

area marked with orange. This can be viewed in the Figure 10.

As the exact block information is not too visible in the above heat map, the heat map in Figure 11 tells us a more detailed story as it is zoomed in.

From the above graph, we can see that the highest number of crashes have occurred near *South Damen Avenue*.

6. DATA MINING AND ANALYSIS

6.1 K-Means Clustering

K-means clustering[10] is one of the prominent data mining techniques that partitions data into k different non-overlapping clusters. All data points in a given cluster have something in common that makes them form a cluster. This algorithm is very helpful in finding any pattern that is not visually discoverable.

For our project, we use this algorithm on two different cases similar to those discussed before.

For the first case, we use the dataset obtained by grouping the rows by date. This dataset is fed into a K-Means algorithm with 3 clusters to obtain results. We notice that there are no significant patterns in the graph. The plot does not give us any new information. The plot for Red Light Violations when data is grouped by date is shown in Figure 12.

For the second case, we use the dataset obtained by grouping the rows by location, which in our case is the street name. On feeding this to the k-means algorithm with 3 clusters (Shown in Figure 13), the data points are clustered into three regions. These three regions depict different levels of severity in terms of violations and crashes. The cluster with most density in the graph has the data points that correspond to locations which have minimum violations and crashes. These locations have a lesser number of violations and crashes reported than those in the other two clusters. The locations in the second cluster (green) have moderate

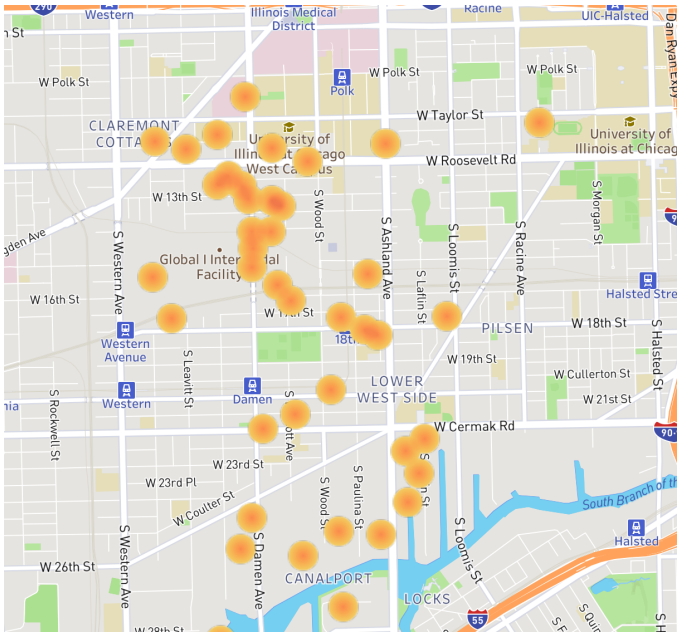

Figure 11: Heat Map: Zoomed

to high violations and crashes. Finally, the third cluster has two locations that have an alarmingly high number of red light violations and crashes. These two locations are *Cicero Avenue* and *Western Avenue* which need immediate attention.

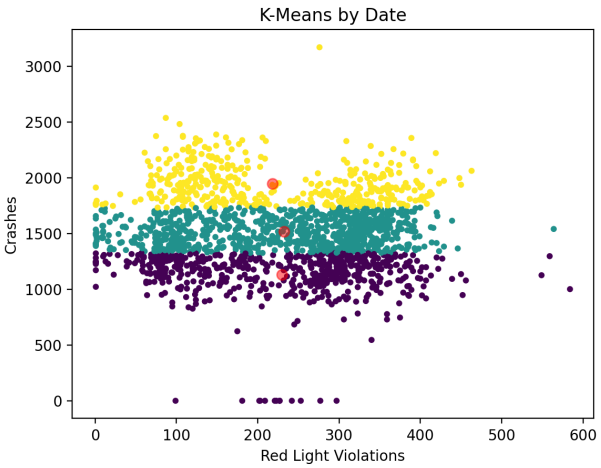

Figure 12: K-Means Clustering for Red Light Violations (Group by Date)

7. LESSONS LEARNT

Through the course of this project, we have understood the significance of data cleaning in data analysis. We have discovered interesting patterns and different relationships between violations and crashes.

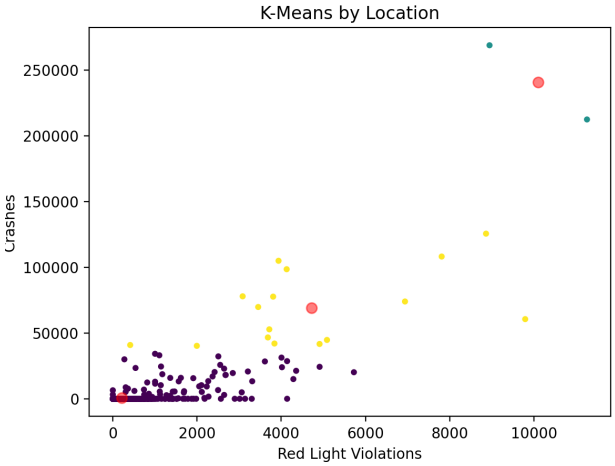

Figure 13: K-Means Clustering for Red Light Violations (Group by Location)

Grouping the data by date and finding the relationship of the number of violations with crashes gave no useful information. In the time series plot for red light violations and crashes, there seemed to be seasonal relationship between them. However, when deseasonalization was performed on the data, there was very little change to seasonality and almost no depiction of upward or downward trends.

8. CONCLUSION

In this project, we have chosen three different datasets from different sources. We have programmatically cleaned these datasets by applying different cleaning techniques learnt in the course. These include elimination of unnecessary and less important attributes, replacing missing and null values with appropriate values, universalizing attribute formats in and across datasets, correcting spelling mistakes, replacing abbreviations with apt words, etc. We have experimented with different data visualization techniques using Matplotlib and Tableau. Also, we have performed in-depth analysis of datasets that include time series analysis, establishing correlations, K-Means clustering and discovered interesting patterns that have been mentioned. Finally, as mentioned before, we established that traffic violations have indeed a good correlation with the number of crashes for any given location.

9. CURRENT STATUS AND FUTURE WORK

Our project is based on the data limited to the city of Chicago. For our future work, we can analyze other cities of the United states or any other country. We can carry out analysis by taking other interesting attributes like weather conditions, traffic type, road type, traffic congestion, etc. We can also try out more sophisticated techniques to figure out if violations are actually the cause of traffic crashes. If so, we could build a model to predict the occurrence of crashes given the data related to violations.

10. 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] R. Nétek, T. Pour, and R. Slezakova. Implementation of heat maps in geographical information system – exploratory study on traffic accident data. *Open Geosciences*, 10:367–384, 08 2018.

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

[10] W.-L. Zhao, C.-H. Deng, and C.-W. Ngo. K-means: A revisit. *Neurocomputing (Amsterdam)*, 291:195–206, 2018.