CSCI-721 Data Cleaning and Preparation

Analysis of Traffic Accidents and Violations in Chicago

Team: Hakuna MaData

Abhay Rajendra Dixit | Pranjal Pandey | Ravikiran Jois Yedur Prabhakar

Introduction

- Traffic violations are a common and inevitable problem in every city. Often, these violations are directly or indirectly linked to traffic crashes.
- Approximately 1.35 million people globally die every year as a result of road traffic crashes
- The focus is usually on making stricter traffic rules and not on carrying out a root cause analysis of the problem at hand
- In this project, we study the relationship between traffic violations and the accidents that have happened in Chicago for the last 6 years

Why traffic analysis?

- People lack awareness about road safety and they continue to violate rules unless they understand the importance of following them.
- Jumping red lights and speed camera violation are the prominent forms of traffic violation.
- The consequences of these violations are serious which include severe injuries and fatalities resulting from crashes.
- 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

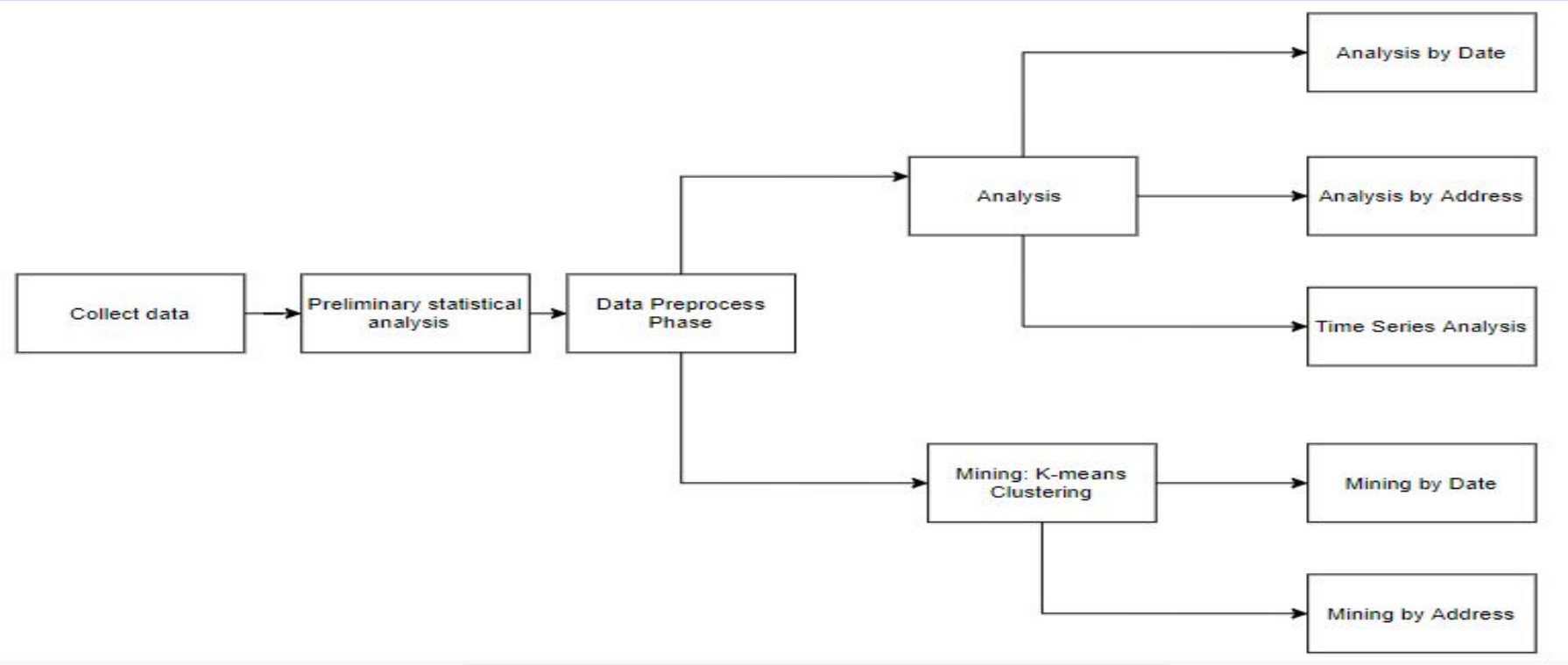
Datasets

- Three different datasets have been used for the analysis
  - Traffic crashes
  - Red light violations
  - Speed limit violations
- The datasets are based on the region of chicago, and are taken from https://data.cityofchicago.org/
- The datasets contain information about the traffic crashes and violations which have occured in various streets of Chicago.
- For this project, data from 2015 onwards has been taken into consideration.

Data Preprocessing

- The chosen datasets have many different inconsistencies
- Some of them are:
  - Null values in some attributes
  - Dates are in string datatype as well as in the wrong format
  - Address has many spelling mistakes and the format is not uniform in all three collections with spelling mistakes
- For example, the word “Street” has different abbreviations in and across collections. They have been converted to “Street” everywhere
- These inconsistencies have been corrected using the pandas package and the dataframes are reconstructed before saving them to MongoDB
- Once these fields are corrected, they are inserted to MongoDB

Design


Collect data
Preliminary statistical analysis
Data Preprocess Phase
Analysis
Analysis by Date
Analysis by Address
Time Series Analysis
Mining: K-means Clustering
Mining by Date
Mining by Address

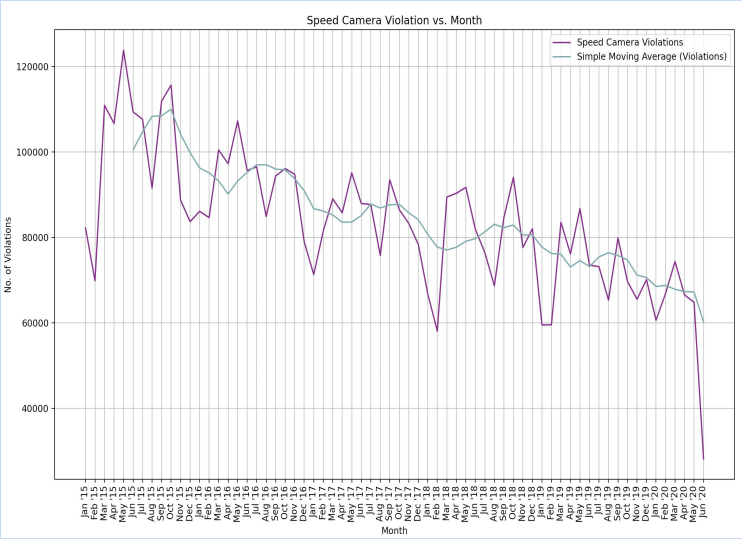
Database

- MongoDB has been chosen for data storage.
- We have made use of this document based model to achieve better flexibility for storing and accessing content in the right formats
- We have created one database by name *traffic_analysis*
- This database has three collections namely,
  - speed_violations_cn
  - redlight_violations_cn
  - traffic_crashes_cn
- During the analysis, the collections are joined, queried and integrated and the results are plotted and examined

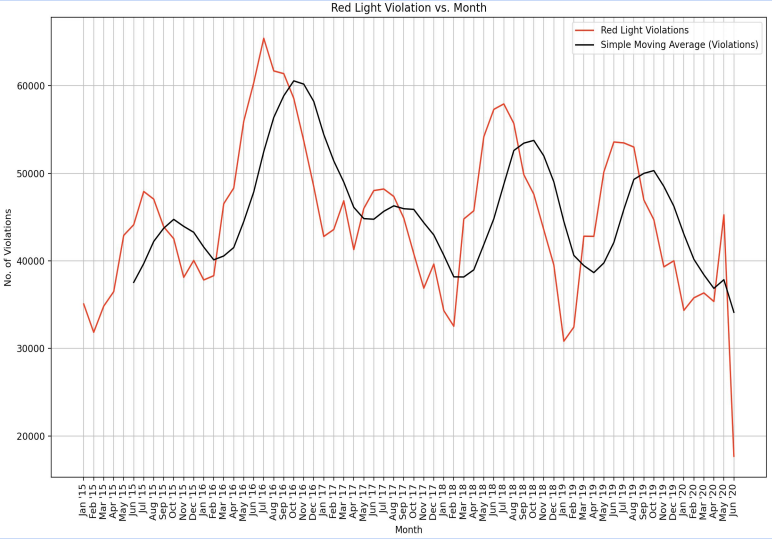
Analysis

- The analysis has been divided into three parts:
  - Analysis based on Date
  - Analysis based on Address
  - Time Series Analysis
- For analysis based on date, all the datasets have been merged and grouped by “Date” to get information of violations and crashes on each date since 2015
- For analysis based on location, all datasets have been merged and grouped by “Street Name” to get information of violations and crashes on each street since 2015
- For time series analysis, the number of violations for each month for 6 years of data have been plotted and analysed

Graphs for Analysis



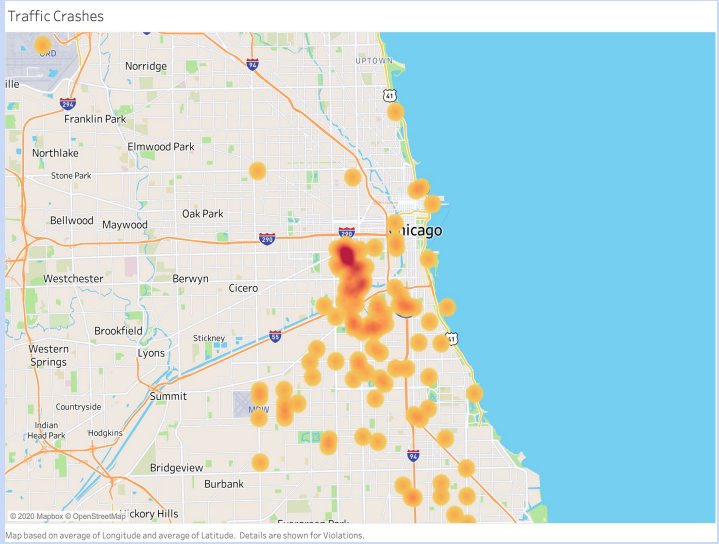
- Speed Violations v/s Month



- Red light Violations v/s Month

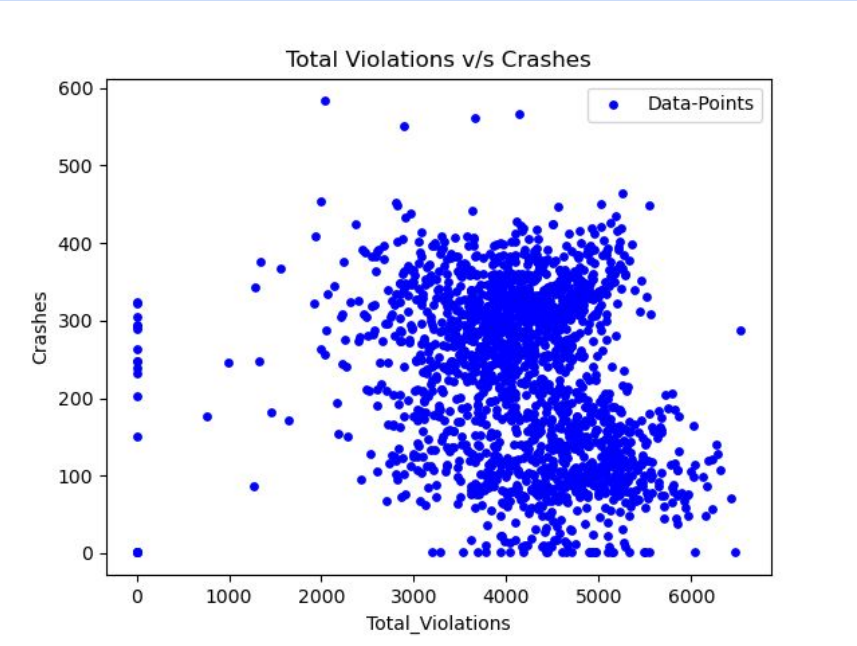
Graphs for Analysis

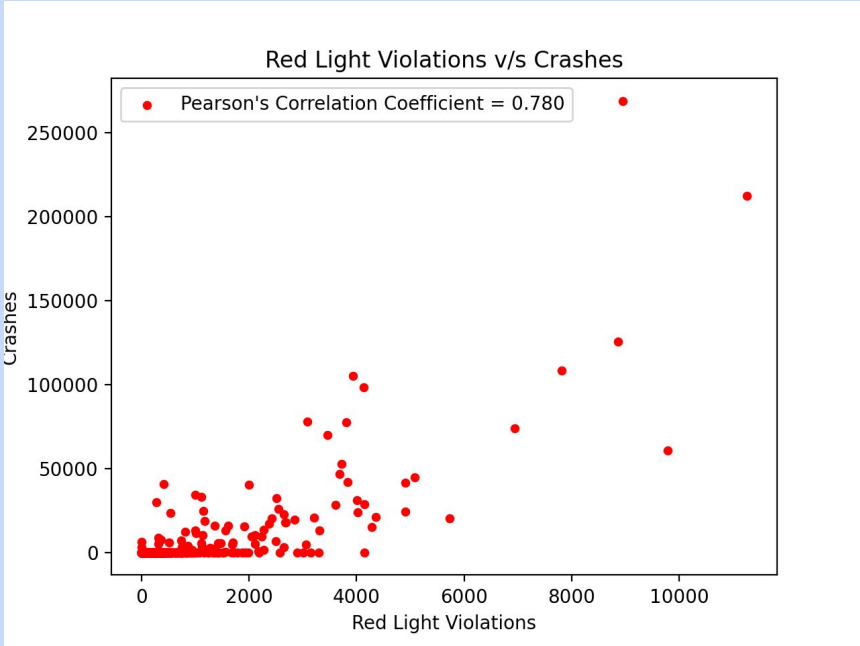
- Traffic Crashes v/s Month



- Heat map for Traffic Crashes

Graphs for Analysis





- Total Violations v/s Number of Crashes grouping by Date
- Red light Violations v/s Number of Crashes grouping by Location

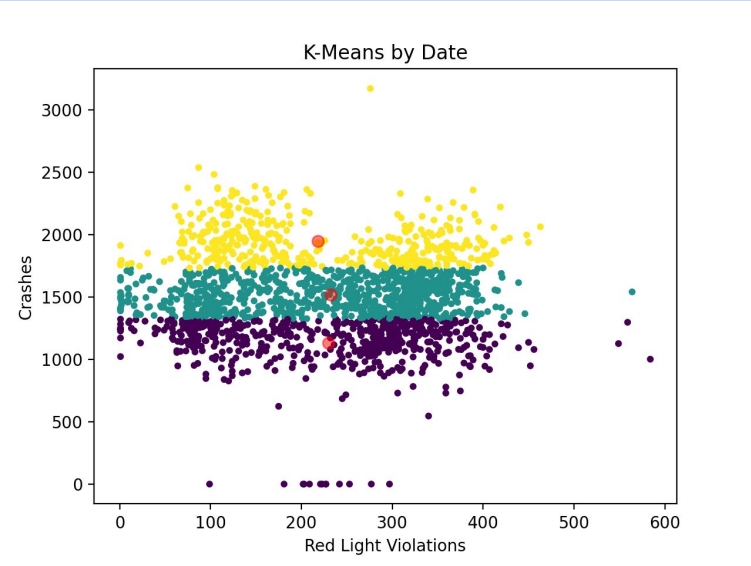
Experiments

- To understand more about the pattern in the time series, the moving averages were calculated and plotted for each collection
- For the red light violations dataset, there appeared to be seasonality Deseasonalization was performed and the graph was plotted
- A heat map demonstrating the number of crashes in Chicago was also plotted using Tableau
- Grouping the data by date and finding the relationship of the number of violations with crashes gave no useful information.
- Grouping the data by address

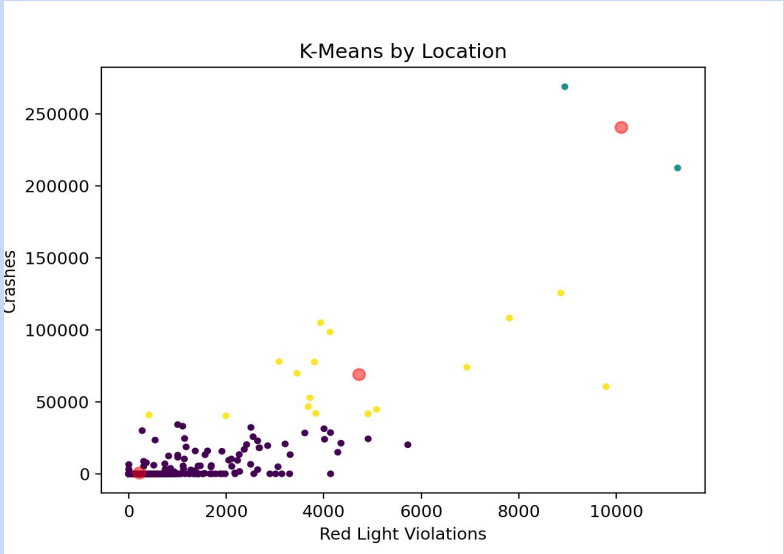
Data Mining

- We have used K-means clustering technique that partitions data into k different non-overlapping clusters.
- For our project, we use this algorithm on two different cases
  - Dataset obtained by grouping the rows by date
  - Dataset obtained by grouping the rows by location
- This dataset is fed into a k-means algorithm with 3 clusters
- We notice that there are no significant patterns in the graph when grouping by date, but for data that was grouped by location, we could see 3 clusters that a vague representation of severity of violations and crashes on each street.

Data Mining



- No clear distinction between clusters.



- Clusters vaguely representing severity of violations and crashes.

Results

- For data based on location, a correlation coefficient of 0.68 was obtained for total violations that occurred and the number of crashes, depicting moderate correlation
- For data based on location, a correlation coefficient of 0.77 was obtained for red light violations that occurred and the number of crashes, depicting high correlation
- The time series analysis showed that the number of crashes has been reducing over time
- The time series analysis also showed a decrease in the number of speed camera violations that occur in the city of Chicago
- Grouping the data by date and finding the correlation coefficient, we can infer that the number of violations occurred in one area does not affect the occurrence of a crash in another area

References

- https://data.cityofchicago.org/Transportation/Speed-Camera-Violations/hhkd-xvj4
- https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if
- https://data.cityofchicago.org/Transportation/Red-Light-Camera-Violations/spqx-js37
- W. Pietrasik. Road Traffic Injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.
- Red Light Running.https://www.iihs.org/topics/red-light-running.Insurance Institute for Highway Safety.

References

- Defending Motorist Rights in the Free State. http://www.mddriversalliance.org/p/arguments-against-speed-cameras.html. MarylandDrivers Alliance.
- Red Light Running. https://www.iihs.org/topics/red-light-running. Insurance Institute for Highway Safety.

Thank You!