CSCI 720 - Big Data Analytics - Final Project

New York City Collision Data Analysis

Shashank Rudroju

Department of Computer Science

Rochester Institute of Technology

Rochester, NY 14623

sr1632@rit.edu

May 8, 2018

Abstract

With the rapid increase in the number of deaths and injuries from traffic accidents, the problem of traffic safety is everyone's concern. Everyone suffers from traffic accidents in one way or the other. If either involved in a collision or faced traffic slow down due to an accident, the impact of traffic accidents on the quality of peoples life is huge and addressing it is important. Today, even the government prioritized traffic safety and in order to reduce the damage caused by traffic accidents, data analysis on the accident data can be done. In this paper, we analyze data that contains the breakdown of every collision in the NewYork city. The analysis is used to infer various hidden information, trends in the accident patterns and also contribute to the traffic safety by predicting times, locations in NewYork city that are safe. The predictions are personalized to cyclists, people, pedestrians and motorists.

Overview/ Introduction

The data consists of complete breakdown of all collision details in Newyork city. It has almost 1.25 million records and 29 attributes for each record. Each record represents a collision in Newyork city by borough, street, location, time etc. The data is provided by the public safety of Newyork Police Department to ensure traffic safety by creating awareness among the people about the danger at intersections. This information is collected and updated every month as a part of a legal procedure since 2011. Since this is issued in public interest, it doesn't violate any ethical considerations and doesn't involve any business cases. The data is huge and is difficult to analyze because of its size and number of attributes. The data is clean but it has a lot of null values in it, so preprocessing is required for this data. The data analysis of this data can provide solutions to various issues and this analysis is continued even after the solution is deployed. The conclusions and information that we gathered during this analysis will help in triggering new ideas and the data mining tasks that are undertaken

further will be benefited from the experiments that are done now. Hence, it fits in the CRISP DM methodology perfectly.

Background and Research

Before dwelling into the data analysis, a research on the previous work in the same field has been done. Various studies and data analytics have been done in the past and some of the significant works are described now. In paper [1], data analysis has been done on the same New York city data. The primary focus of this paper was on investigating the spatio-temporal distributions, characteristics of vehicle collision and accident prone area detection. They also proposed an optimal safe route method to travel on a street by defining dangerous index and distance that is calculated for every street. The research was mainly performed by analyzing the data of the main streets of New York City(Figure 1) and have also used motor vehicle collision data(Figure 2) provided by NYPD.

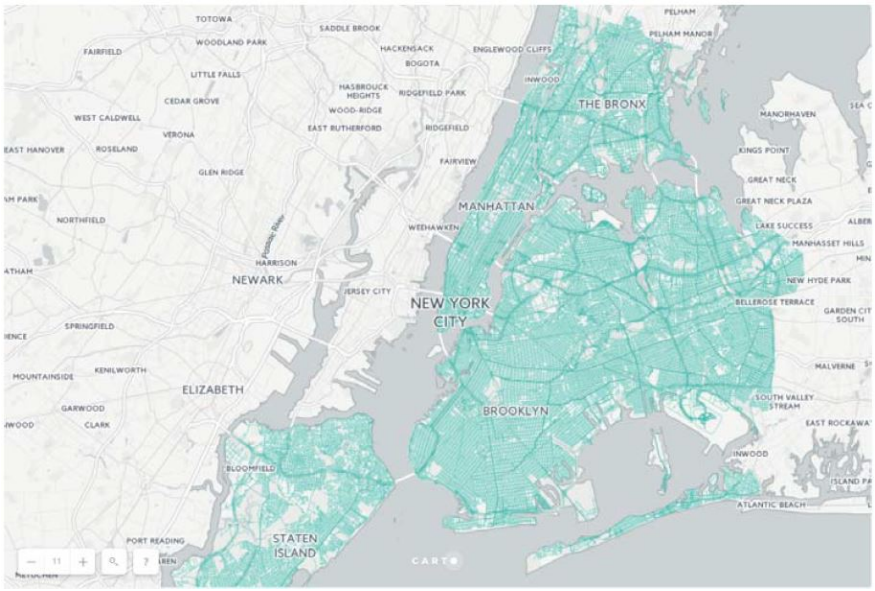

Figure 1: Street Network of NYC

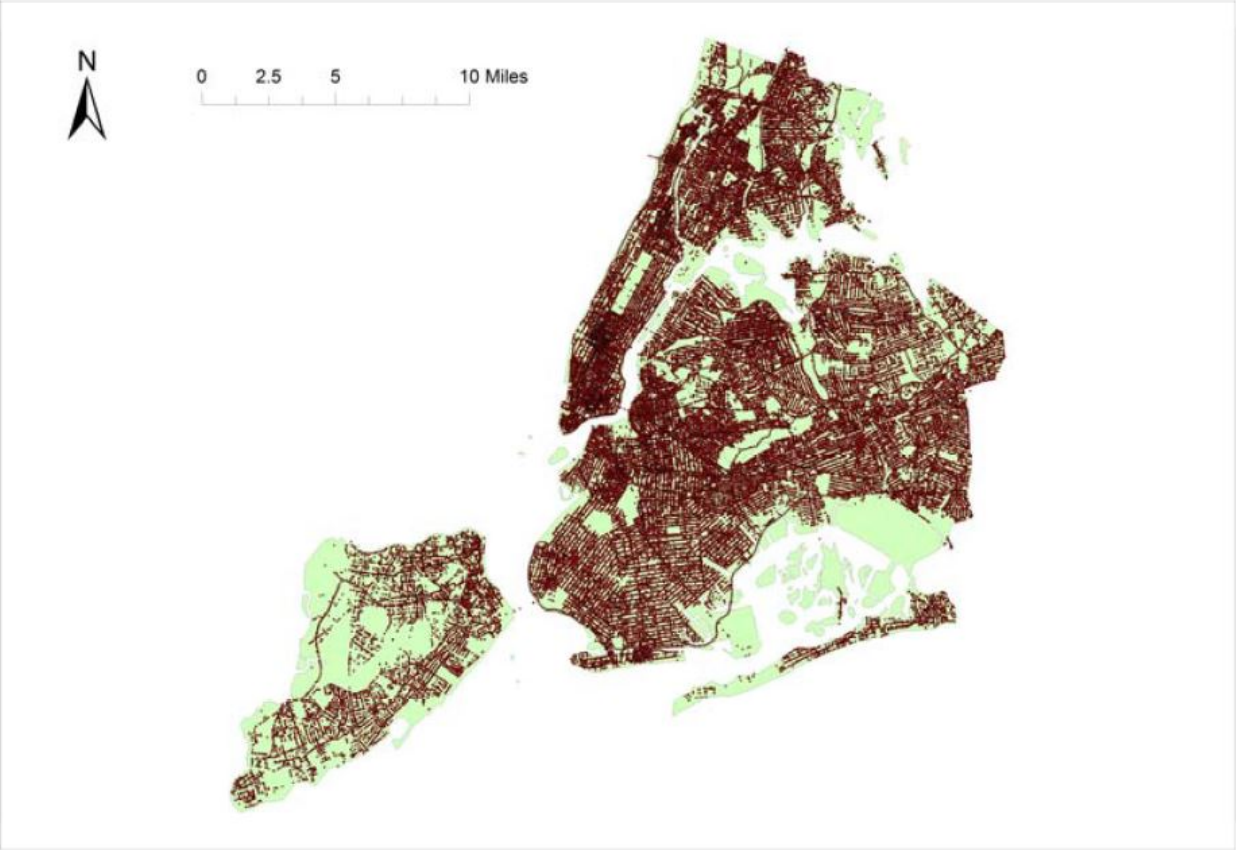

Figure 2: Motor Collisions in NYC

In order to perform the analysis on motor vehicle collision data, the data has to be cleaned. The data is then matched to the streets according to their distance to the streets. Then, a collision curve is generated for every individual street by calculating the number of collisions on hourly basis for 24 hour cycle.The basic regional function of the streets can be seen by observing the patterns on the collision curve. A K-Medoids segmentation clustering algorithm is used to cluster the streets based on collision curves. This algorithm is used with Weighted Dynamic Time Warping for calculating the distance between two curves particularly time series. Apart from calculating collision numbers, there is a need to know about accident prone streets. For this, a dangerous index factor is calculated for every street by considering collision number and street length.

DI = CollisionNumber/StreetLength

By considering the dangerous index, an optimal safety route can be calculated between every two locations using Dijkstra algorithm. There are some limitations associated to this paper as the study considered only traffic accidents without the bike accidents which may not generalize the whole population regarding the collisions accomplished during this study.

In paper [2], the data is taken from New York's traffic collisions dataset to develop the application. The new application includes various functions and web services in order to analyze and visualize the traffic accident scenario.It can store huge loads of data on Hadoop based on Map-Reduce technique. The application sends and receives commands from the server and the server handles all the data processing. Apache Mahout, Apache Hadoop and MYSQL technologies are used for distributed storage, processing large amounts of data and to speed up the processing. These can support variety of data mining techniques and algorithms. Hence, the new application will have an advantage in handling wide amount of data on accidents. The new software works on MySQL database which uses New York's traffic collision dataset. The data mining technique can be applied by extracting the data required from the database and transfer it to Hadoop in order to process the data. The extracted data is then analyzed using Mahout. The proposed application connects to the server using SSH and can provide 6 analysis functions which can then be converted into tabular view, chart view and map view based on the function. For example, the function "Get common cause for accidents" can generate the results by executing the query in MySQL. This function result can be formatted into a table with columns: Borough, Cause and Number of accidents as shown in fig 3.

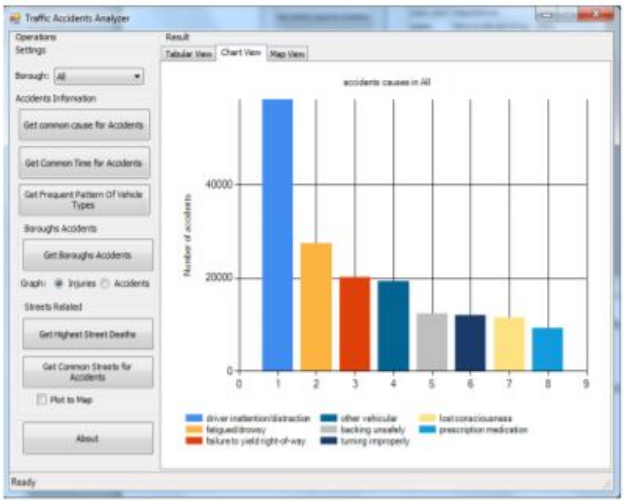

Figure 3: Graphic Distribution for common cause for Accidents

The enhanced edition of the proposed application is created as web services which can be used by various clients, or any device or any application that can send POST requests, or by using SOAP. The application also allows using Hadoop and Mahout online and remote.

Experiments

Before starting the experiments on the data, the data was preprocessed and the records that don't have any location values are removed. Also, the records with no date values are removed. The data had too many attributes, so some of them have been removed and some of them were merged into one. Many attributes like 'CONTRIBUTING FACTOR VEHICLE 2','CONTRIBUTING FACTOR VEHICLE 3','CONTRIBUTING FACTOR VEHICLE 4','CONTRIBUTING FACTOR VEHICLE 5','VEHICLE TYPE CODE 2','VEHICLE TYPE CODE 3','VEHICLE TYPE CODE 4','VEHICLE TYPE CODE 5' had very few values in them. So, I removed them. Instead of having separate columns for 'NUMBER OF PERSONS INJURED', 'NUMBER OF PERSONS KILLED', I merged both the columns into one by adding them and naming the column as 'PERSON CASUALITIES'. Then the original two columns are removed. Same is the case with the columns 'NUM-

BER OF PEDESTRIANS INJURED','NUMBER OF PEDESTRIANS KILLED','NUMBER OF CYCLIST INJURED','NUMBER OF CYCLIST KILLED','NUMBER OF MOTORIST INJURED','NUMBER OF MOTORIST KILLED'. From the data column, three separate columns each for day, month and year are populated and added to the data.

Also there are a lot of records where the longitude and latitude values are invalid. Some of them are (0,0) and some of them are outliers with longitude and latitude values way off from the desired values for New York city. Hence all those records are removed.

After preprocessing the data had 1033940 records and 19 attributes. Of these 19 attributes, first visualization was based on location. The latitude and longitude values are considered as coordinates and a scatter plot of all the location coordinates is drawn.

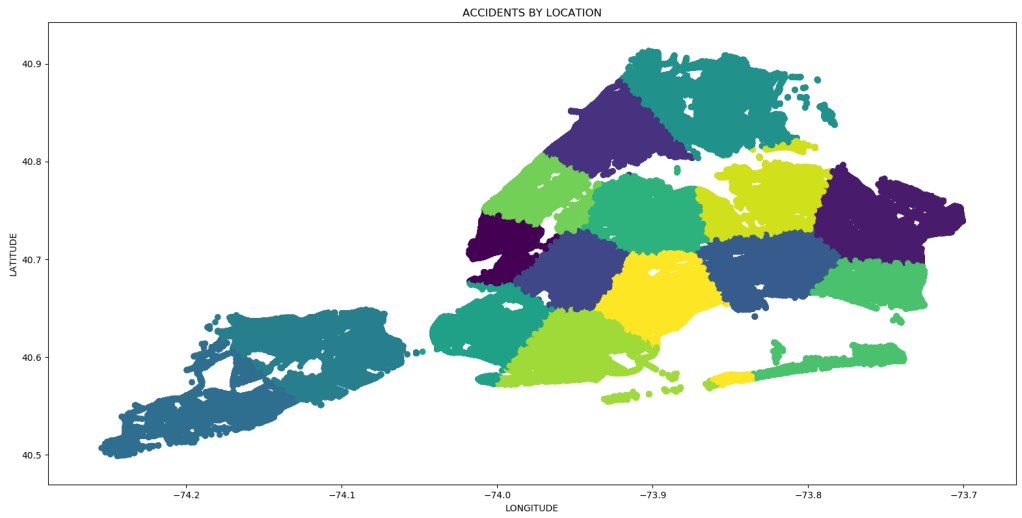

Figure 4: Accidents by Locations

Using the k-Means algorithm, the distribution is clustered into 15 groups. The cluster centers and the number of accidents in that location cluster are as follows:

```
Center : [ -73.995509 40.716114 ] Size of Cluster : 83604
Center : [ -73.754721 40.738092 ] Size of Cluster : 46904
Center : [ -73.92668 40.825001 ] Size of Cluster : 100229
Center : [ -73.951246 40.68722 ] Size of Cluster : 90920
Center : [ -73.82334 40.694582 ] Size of Cluster : 67311
Center : [ -74.184058 40.549391 ] Size of Cluster : 15782
Center : [ -74.107822 40.607946 ] Size of Cluster : 36022
Center : [ -73.86674 40.859819 ] Size of Cluster : 91417
Center : [ -74.002446 40.63346 ] Size of Cluster : 59161
Center : [ -73.904159 40.746205 ] Size of Cluster : 73731
Center : [ -73.761886 40.658223 ] Size of Cluster : 36637
Center : [ -73.974871 40.761981 ] Size of Cluster : 136518
Center : [ -73.952939 40.609582 ] Size of Cluster : 63357
Center : [ -73.833955 40.755196 ] Size of Cluster : 55789
Center : [ -73.900335 40.665684 ] Size of Cluster : 76558
```

Figure 5: Clusters based on location

From the clustering information, we can conclude that the location with longitude and latitude values as 40.761981 and -73.974871 respectively has the highest number of accidents. The location corresponding to the center of the most accidents area is as follows:

Figure 6: The location with most accidents. Img ref : google maps

Then I tried analyzing the accident information based on boroughs and there are a lot of records without borough information. From the information available in the data, by grouping accidents based on the borough, the following bar graph is generated.

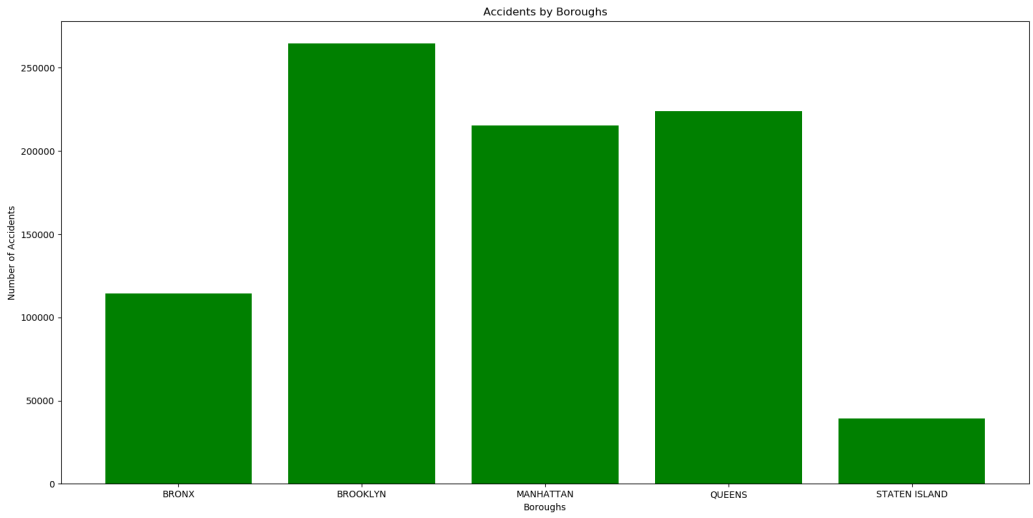

Figure 7: Accidents by boroughs

This bar graph is just the count of accidents in different boroughs. Brooklyn has the highest number of accidents whereas Staten Island has the least number of accidents. Brooklyn is much densely populated and experiences high traffic compared to Staten Island and hence the results are justifiable.

The number of collisions and the total number of deaths/injuries in boroughs have different distributions. The total number of a deaths/injuries for every borough is analyzed and visualized as follows:

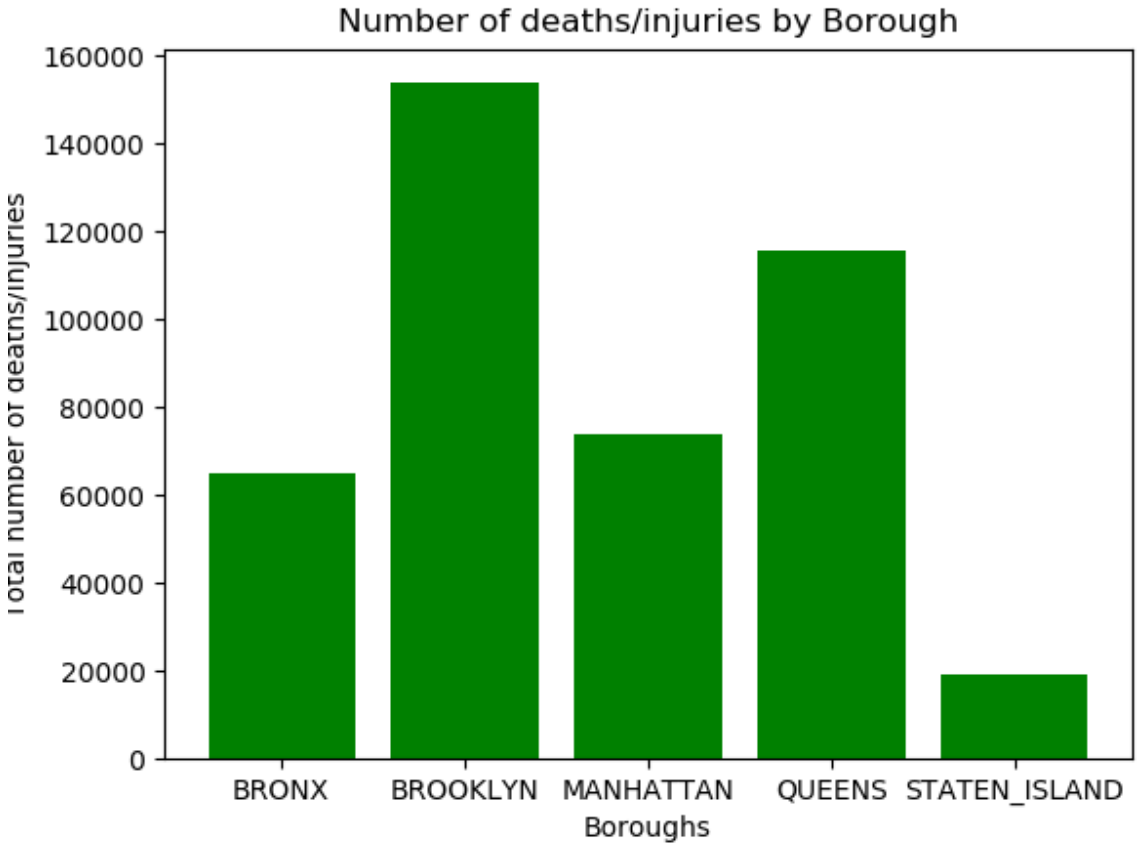

Figure 8: Total Deaths/Injuries by boroughs

The deaths in Brooklyn are significantly much higher than other boroughs even though the number of collisions are closer.

In order to see the accident trends through out the year, the data is analyzed based on months. The count of all the accidents in each month is calculated and the results are as follows:

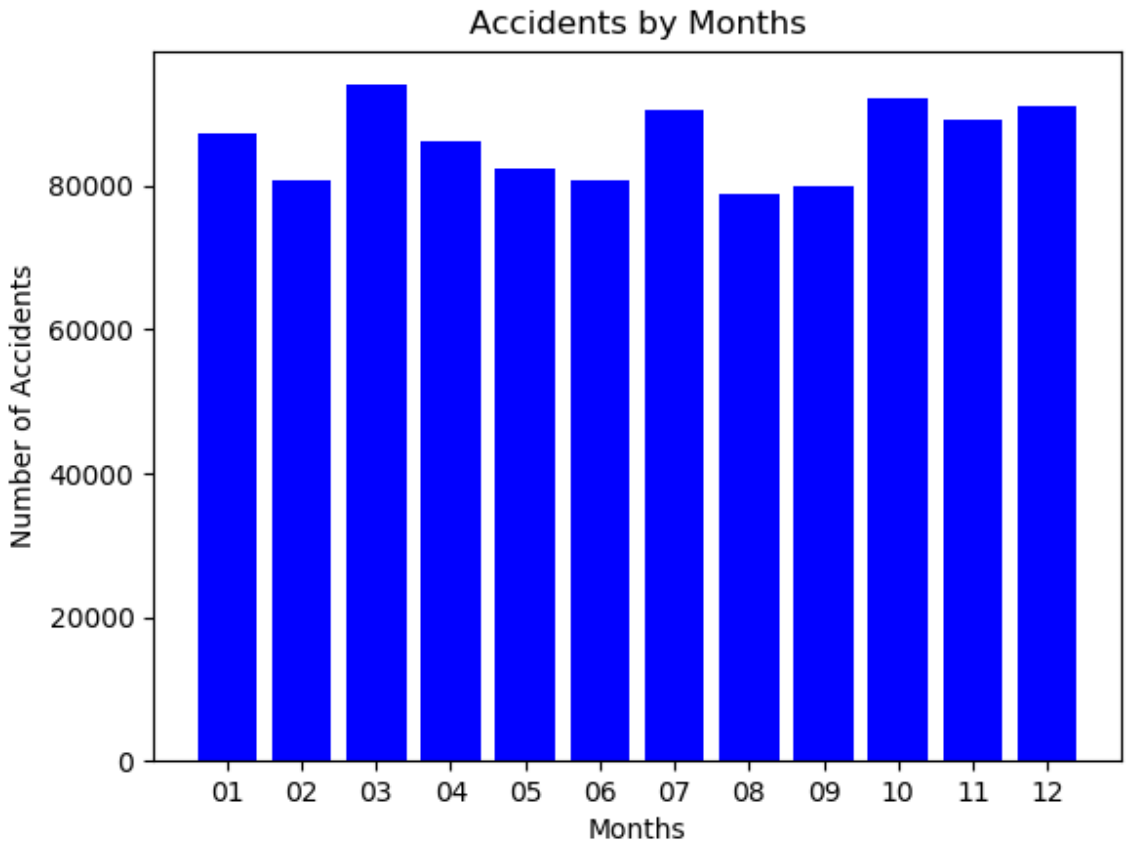

Figure 9: Accidents by months

It is observed that most number of accidents occurred in the month of March, October and December. Before the data was cleaned, the accidents in the month of October were significantly higher than all other months. The reason behind this could be the start of winter season and snowy weather conditions are prone to more accidents. The spring daylight savings is in March and hence there is a rise in the number of accidents in that period. Also the fall daylight saving is in November and there is a rise in the number of accidents in that period too.

The rate of accidents through out a day based on the time is also analyzed. The number of collisions by hour from 0 to 23 is calculated and a graph is plotted.

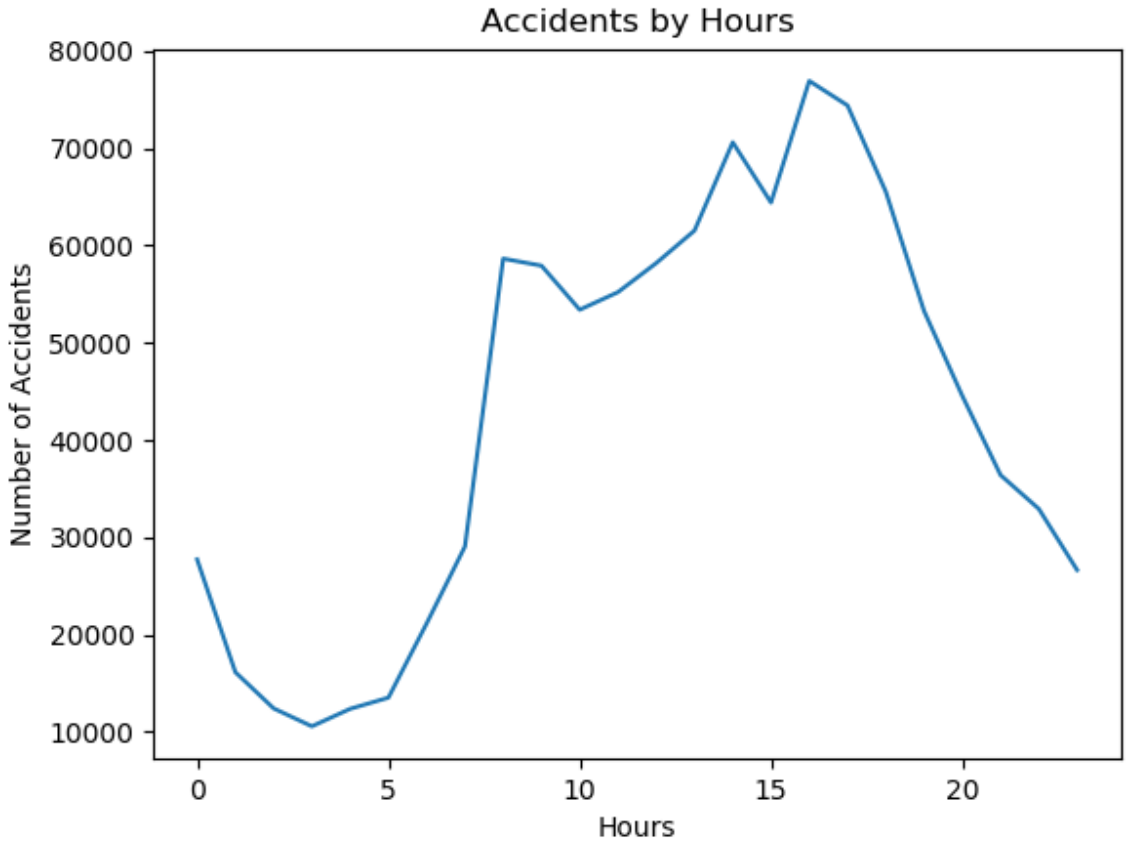

Figure 10: Accidents by months

The number of collisions drastically increase during early mornings and reach a peak at 17 hours i.e 5 in the evening. This could be due to the high traffic of vehicles of people going from work to home. They drastically increase in the morning in between 8-10 AM as that is the time people will be commuting to their work place.

To extract more specific information from the time and make more conclusions, the time is divided into slots of 15 minutes and the number of accidents in that slot is drawn as a bar graph. It can be observed that the 15 minutes slot before the exact times is a very accident prone time slot. People tend to rush during that time to reach on time and hence the accidents are more. Every hour, the last 15 minutes slot is the one with high accidents.

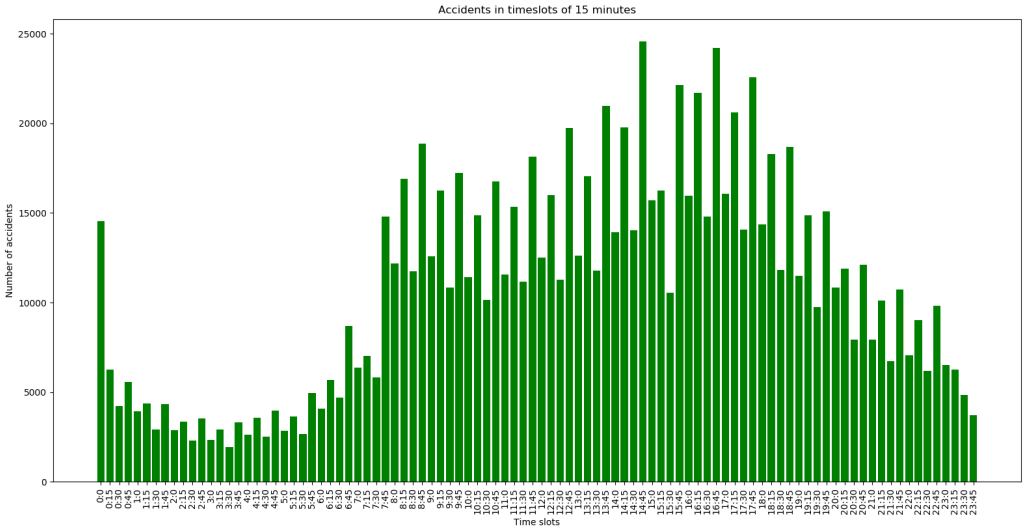

Figure 11: Accidents in every 15 minutes

The number of accidents in each year is counted and visualized as follows:

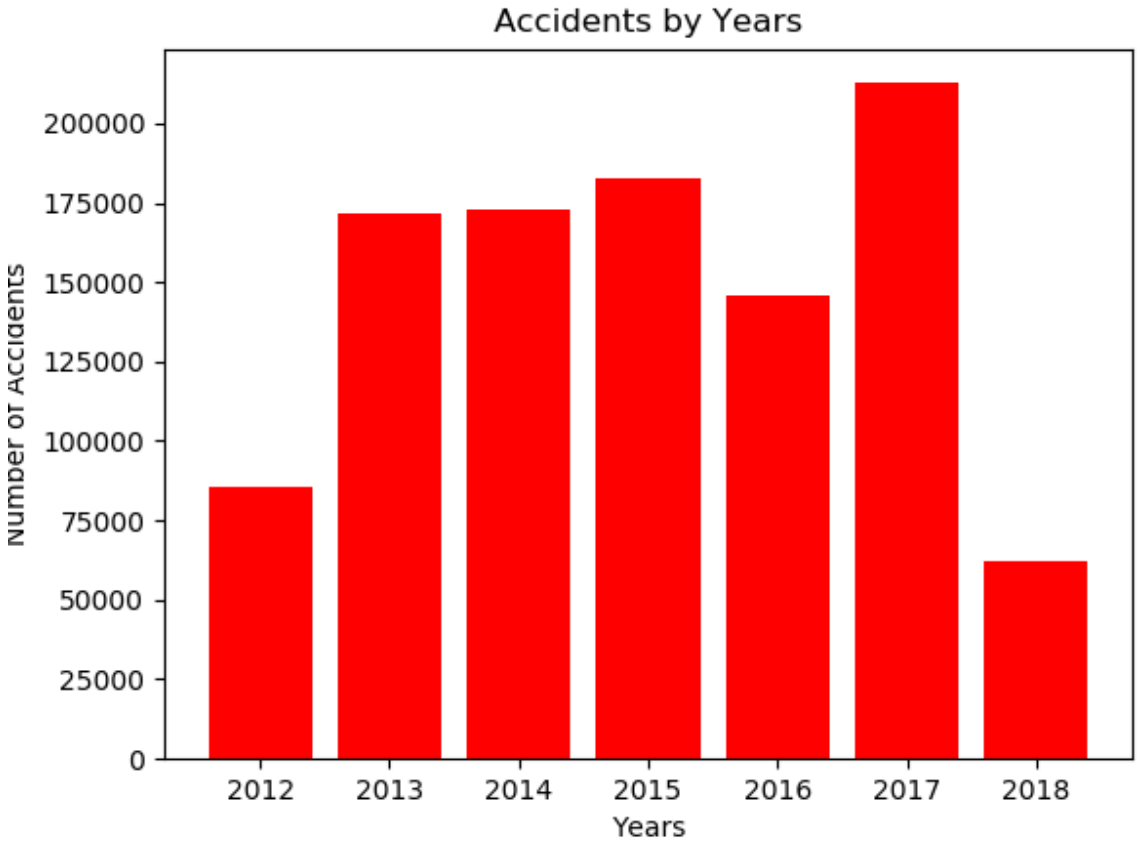

Figure 12: Accidents every year

If anyone wants to know which month is better for cycling, the data analysis of all the bike accidents in June, July and August has been done.

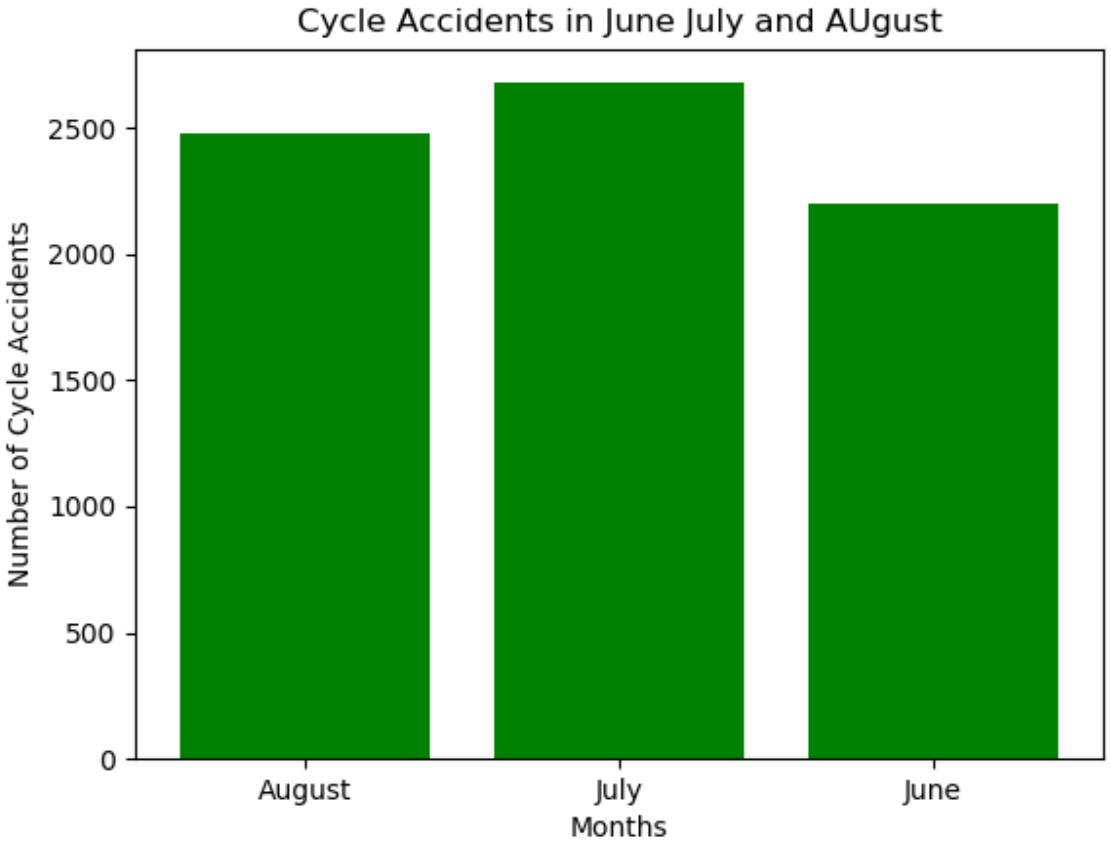

Figure 13: Cycle Accidents in June, July and August

To visualize what contributed more for accidents, the contributing factors for all the accidents is aggregated and shown in a bar graph.

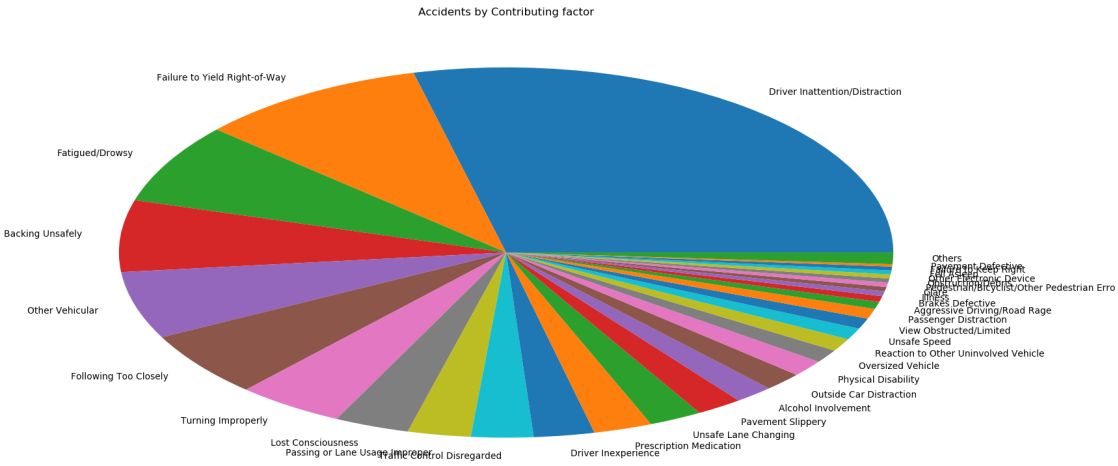

Figure 14: Contributions to accidents

Discussions

In figure 4, when all the location coordinates are scatter plotted, all the points are densely populated in 5 areas. From the scatter plot it is evident that the data can be clustered into 5 clusters. Using KMeans clustering, I extracted 5 clusters of uneven sizes and the sizes are compared. In figure 10, the bar graph of the number of collisions in each year from 20012 to 2018 is drawn. We can see that the accidents have increased in the recent years. In 2017 the number of collisions hit an all time high of over 200000. We cannot consider 2018 year because the data is incomplete as the data is collected only till April. In figure 11, the number of collisions involving injuries to cyclists is analyzed during the months of June, July and August. For cyclists June seems to be a safe month to ride in Newyork city.

The cause for the accidents is an important attribute. Hence data mining is done on the contribution to the accidents and a pie chart is drawn in figure 12. The major cause for accidents is Driver Inattention/Distraction.

Conclusion

In this project, I learnt how to work on real data. How to infer various hidden details by trying various data mining techniques and applying the right technique. I also learnt how details from one technique will help us resolve other issues. The data was huge and had so many irrelevant attributes and overcoming such challenges was a good experience. Identifying the important attributes and getting rid of the records in the data that might cause problems in the future is an important aspect in data mining. By data mining on the Newyork city, the following information is discovered. Most of the accidents occur near the location with coordinates (40.68 ,73.98). Brooklyn is the most unsafe borough among the 5 boroughs. March month and December month has the most accidents. This might be due to the change in the day light savings time. In a day, accidents drastically increase during 10 in the morning and reach a maximum at 5 in the evening. This might be due to the heavy traffic of vehicles returning from work. Accidents due to Driver Inattention/distraction are the most common type of accidents. If anyone likes to ride their bike through NY City, of the months: June, July, or August, June is the safest. In recent years the number of accidents reached an all time high in 2017.

References

[1] Enbo Zhou, Shanjun Mao, Mei Li. Investigating Street Accident Characteristics and optimal safe route recommendations: a case study of New York City. 25th International Conference on Geoinformatics, November, 2017.

[2] Eyah Abdullah, Ahmed Emam. Trac Accidents Analyzer Using Big Data. International Conference on Computational Science and Computational Intelligence, December, 2015.