CSCI-720 Project: New York City Motor Vehicle Collisions - Data Analytics

Sanjay Haresh Khatwani
Department of Computer Science
Rochester Institute of Technology
Rochester, NY, USA
sxk6714@g.rit.edu

April 30, 2018

Abstract

The purpose of this work is to perform various data mining activities on the New York City Motor Vehicle Collisions data. These activities are aimed at exploring the hidden trends and knowledge in this huge data-set. This knowledge can be used by the state government to take measures for reducing the number of accidents and make the city safer.

1 Introduction

The goal of data mining is to process data in order to uncover the hidden knowledge in raw data. This knowledge is present in the form of co-relations between attributes, patterns etc. which are not directly observable by just looking at the raw data. This knowledge can then be used to inform effective decisions.

Road safety is a very critical issue which impacts everyday lives of general population. Not only is there an emotional impact associated with an accident, but also there is an economic loss to deal with. AAA estimates that crashes cost society $299.5 billion annually [2]. There have been various studies that even show that societal loss due to accidents outnumbers the societal costs due to any type of crimes. This suggests that an effective plan needs to be in place for reducing and eventually getting rid of any motor vehicle crashes to save lives and losses. Data mining techniques can be employed to study the past incidents and come up with effective measures to achieve this goal.

In this project the data-set used is the open-source data provided by The city of New York containing details of motor vehicle collisions [1]. This data-set is very rich and regularly updated by the New York Police Department. Effective data mining on the data-set can reveal a good deal of knowledge such as 'Intersections prone to accidents', 'Times of day with highest accidents', 'Days with highest accidents', 'Popular causes of accidents', 'Patterns of collisions', etc. And this knowledge can then be used by the local government to take measures like, installing stop signs, increasing time on signal at some intersections, installing pedestrian signals etc to reduce the number of collisions. Data mining can also be later used to assess how these measures were working.

2 Ethical Considerations

Even though the data is freely provided by the state of New York, ethical considerations should be thought of before using the data for analysis. The knowledge that will be derived from data would identify trends and correlations between various attributes associated with a motor vehicle crash. This knowledge will not contain any personal information about any individual, or group. The knowledge can be used by government organizations to inform safety measures or some other decisions for the benefit of population. Hence, there are no foreseeable ethical problems with analysis of this data set [15].

3 Data

The data is freely provided by the New York State and regularly updated by New York Police Department [1]. This data has details about all the motor vehicle crashes that took place in New York City from July 2012 till date. The different attributes provided in the data-set are as follows:

- Date and Time.
- Borough of collision in NYC.
- Zip, Latitude, Longitude and Location.
- On street name, Off street name and Cross street name.
- Numbers of people, pedestrians, cyclist and motorists injured and killed.
- Contributing factors of upto 5 vehicles.
- Vehicle type codes of upto 5 vehicles.

This is a live data which is updated regularly. For the purpose of this project the data was accessed on April, 13 2018. This version was last updated on April, 12 2018. This version has $10,48,575$ records. This project will only focus on the data from last 3 years. That reduces the data-set to $6,65,213$ records.

4 Previous Work

A considerable amount of work has been done in this area by data scientists. Almost all the research starts by analyzing the trend over the years of the number of crashes, number of deaths, injuries, number of crashes due to drink and drive, red light running etc. All of them use interesting plots to show the correlations and patterns in the data. Some of this work has been reviewed in this section along with the interesting discussions in them.

4.1 Analysis of NHTSA data set

There is a similar data set like the one used in this project provided by National Highway Traffic Safety Administration. This organization's goal is to reduce crashes, fatalities and injuries caused due to motor accidents. NHTSA provides crash reports similar to what is present in NYPD data set and also provides research notes on this data. In [3] NHTSA analyze the 2015 crash reports in their research notes. In this they discuss the statistics of these reports and also identify some interesting trends in the data. They discuss the percentage increase in crashes and how it is the largest since 1995. They also provide charts to discuss these numbers and rates. They discuss something called as Fatality composition, which is basically a pie chart of class of people involved in crashes i.e. *Car Occupants, Light Truck Occupants, Bicyclists, Pedestrians etc.* They discuss percentage change in number of crashes by person category e.g < *16 year olds, Male, Female, 65+ year olds.* In this they identify maximum increase of 12.4% of cases involving drivers that are lesser than 16 years old. They also provide a similar analysis for human choices like distraction affected cases, alcohol impairment etc.

In [4] the author has also used the openly

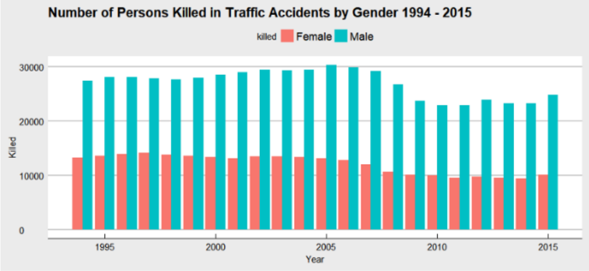


Figure 1: Male vs Female collisions

available NHTSA data set and provided some interesting observations. The author analyzes data across various states in the US and and over 20 years. The author discusses how the fatalities have been almost double for cases involving Male drivers versus Female drivers as depicted in Figure 1. A novel thing in this article is the analysis based on the time of the day. This uncovered the trend that on weekdays most crashes occur between 3:00 pm to 9:00 pm and on weekends most crashes occur between Midnight and 2:59 am. The author then breaks down the crashes on weekend in this time range by the reasons of crash. Most of these fatalities are due to Alcohol impaired driving and most of crashes due to Alcohol impaired driving occur between midnight and 2:59 pm as in Figure 2 .

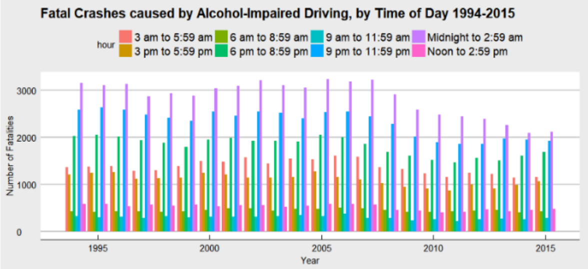


Figure 2: Fatalities due to motor vehicle collision by time of day

4.2 Analysis of NYPD data set

There is a significant amount of work done using the NYPD data set too. And most of them start by making some straightforward analysis of crashes over the year, by borough, etc.

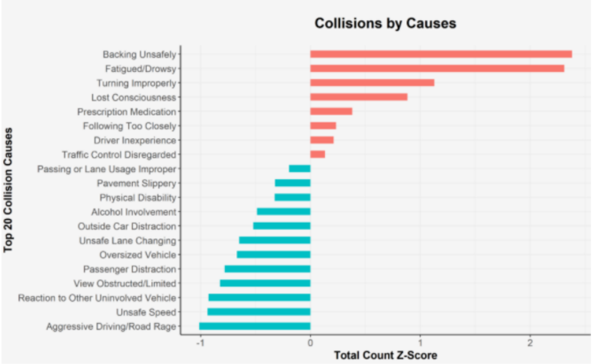


Figure 3: Collisions by causes

In [5] author discusses various trends and compares them for the 5 boroughs. The author points out the Friday has a visibly more number of crashes than any other day for all of these boroughs. It is also pointed out that the number of accidents reduced in the year 2016 and states that it is because of the successful vision zero campaign [6]. This is how measures taken by the government can be gauged for success or failure as discussed earlier. The author also points out that Manhattan has a very high number of pedestrian accidents as compared to any other boroughs and directly links this to the highly crowded skyscrapers in Manhattan. An interesting plot of the z-score as in Figure 3, of causes of crashes is provided in the article. The author also discusses major takeaways from this exploratory data analysis. The most major cause of accidents is drowsy and alcohol induced drivers. The author suggests some measures for the government and individuals to reduces motor vehicle crashes.

[7] is a work of continuation from [5]. In this article author discusses the results of using NYC hourly weather data along with the NYPD motor vehicle collision data to understand correlations between weather and collisions. for this the author scraped the NYC hourly weather data from

Weather Underground [8] using Python. From the weather data, author fetched 8 new attributes for every data point in NYPD data set. These attributes include 2 categorical variables, Weather conditions and wind directions and 6 numeric values, visibility, temperature, humidity, pressure, dew point, and wind speed. The author then performed some cleaning and preprocessing activities on NYPD data set like, consolidation of causes into 1 column, converting them to numeric values for correlation analysis, etc.

Then the author performed frequency analysis on total number of accidents and by hour across different weather conditions. It was found that snow and humidity have larger impact than temperature and visibility. Although there were no strong correlations observed. However, the author does point out some correlations between different collision causes and significant positive correlation between temperature and bicycle or motorcycle.

In [13] too, author combines weather data with the NYPD motor collision data and observes some good correlations. The author uses the two public data sets on Google BigQuery [14] for this. It is shown that isWeekend and alcohol involvement have very high positive correlation, people also ignore traffic lights at this time. snow depth is highly correlated with slippery pavement. Over sized vehicles and isWeekend are negatively correlated.

4.3 Kaggle Kernels

A version of the NYPD motor collision data set is also available on Kaggle [9] and a significant amount of exploratory data analysis is done. In [10] the author shows that mean number of accidents reduce on wednesday. The author also provides locations on map that are unsafe for cyclists or pedestrians, etc. In [11] the author, shows a plot of accidents by hour and a significant peak is observed at 5:00 pm, shown in Figure 4, verfing the conclusion in [4]. Another peak is observed at 8:00 am. In [12] the author shows that Brodway and Atlantic Ave. top the 10 most dangerous streets in NYC list.

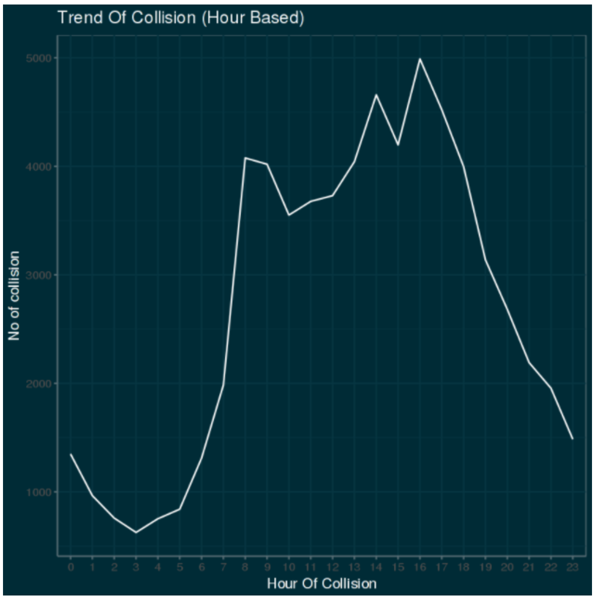


Figure 4: Collisions by hour

5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach in which various analytical techniques are used, or various graphs are developed from a data set to maximize the insight into the data. EDA can also be used to uncover underlying structure of the data, test assumptions/hypotheses, detect outliers etc. For this project, EDA will be performed to get more insights, get some hints on correlations among different attributes and test some hypotheses.

For the purpose of EDA, original data-set has been filtered to obtain a version of data, that can be loaded on to a desktop system and be worked with. This project only uses data from the last three years to April 12th, 2018.

R has been used to perform EDA. *dplyr* [17] library has been used to perform database like select and filtering on the orginal data-set. This library was used to create different filtered subsets of data to generate graphs using *ggplot* [18] in R. Apart from these, *lubridate* [16] library was used for working with date and time columns. *lubridate* was mainly used to compute day of week, month from the date column and hour from the time column. These attributes are used in some graphs. Apart from these *ggmap* [19] was also used to obtain a map of the New York City on which heat-maps were generated.

5.1 Number of accidents by Borough

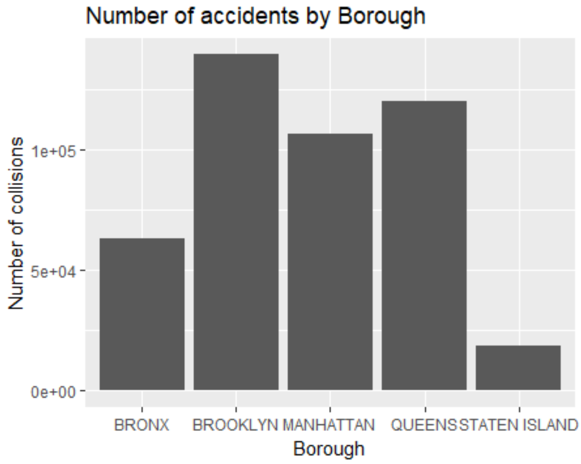


Figure 5: Accidents by Borough

One of the first things that a person would think when looking at the data would be to compare the number of accidents across various boroughs in NY. This shows that Brooklyn needs more attention than Staten Island in terms of accidents. Bronx is doing better than Queens and Manhattan.

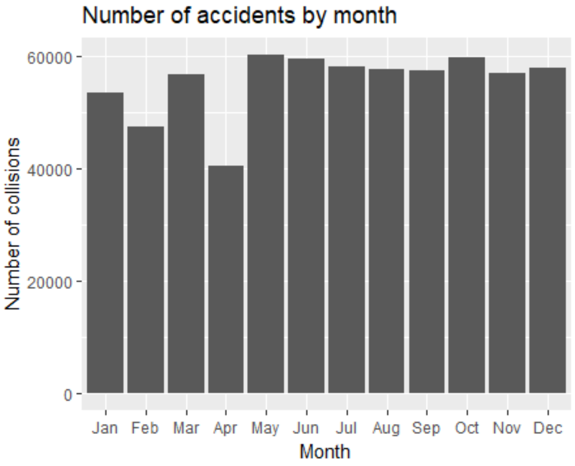


Figure 6: Accidents by Month

5.2 Number of accidents by Month

Figure 6 shows the total number of accidents that happen in any month over all the years. The number is pretty much uniform except for April which has fewer accidents. There is no directly observable reason for this. Maybe its just a gap in reporting the accidents that happen or something similar.

5.3 Number of accidents by Day

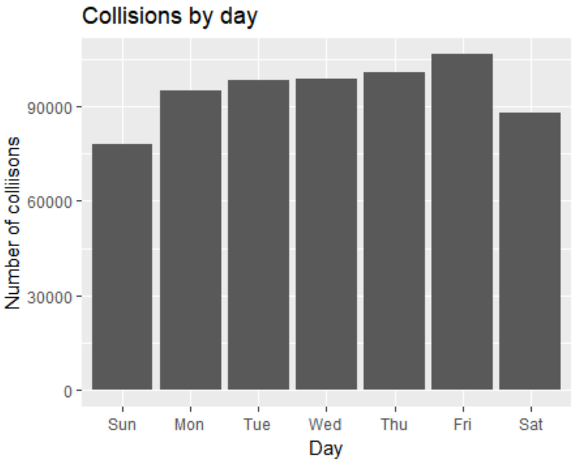


Figure 7: Accidents by Day

Hypothesis: *Friday has more than average accidents in week while Saturday and Sunday have lesser.*

Figure 7 shows the variation in the number of accidents on a day of the week over the years. More accidents occur on Friday, because it's the end of the week and people are out at night, most offices have outings on Friday and so on. While Saturdays and Sundays are off and lesser people are heading to work or any appointments and thus less accidents.

5.4 Number of accidents by Hour of the day

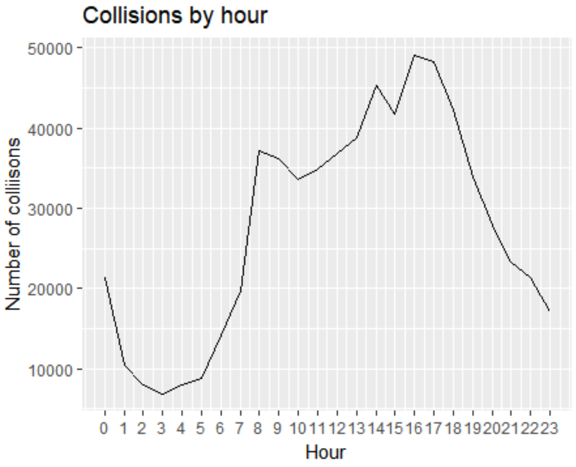


Figure 8: Accidents by Hour

Hypothesis: *There are very less accidents from 12:00 am to 6:00am and more around 8:00 am or 5:00 pm*
Figure 8 shows the trend in accidents by the hour of the day. There are very less accidents at night from 12:00 am to 6:00 am since people are sleeping. There are two peaks observed one when the people usually leave for work or appointments in the morning and other in the evening when they reurn from work. After this it gradually reduces. This trend is pretty much the same on all the days of the week including Friday, Saturday and Sunday. This is shown in Figure 9

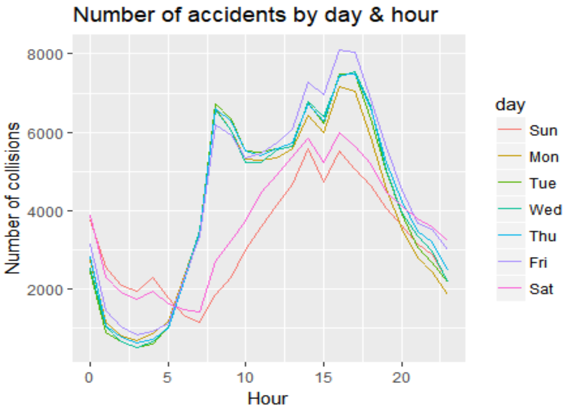


Figure 9: Accidents by Hours on different days

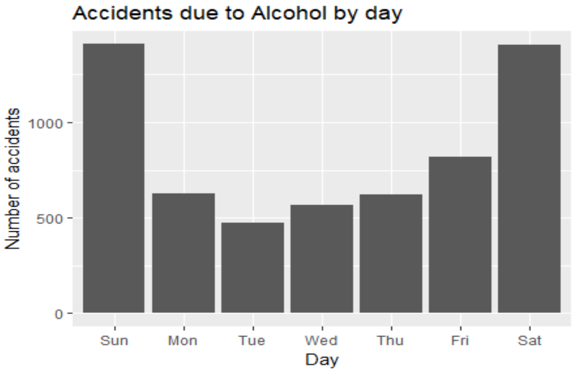


Figure 10: Accidents due to Alcohol involvement by day

5.5 Number of accidents by contributing factors

Figure 11 shows the causes of the accidents and how frequently occuring they are. There need to be some measures in place for keeping drivers in check in NY, such that they are not distracted by their cell-phones, eating or drinking, talking to people in the vehicle, fiddling with music or GPS systems etc. The rate of accidents due to driver distraction is alarmingly high. Other popular causes are due to inefficiency of the driving task and ignorance towards rules. Drowsy driver is at number 5. Alcohol involvement does not make it to the top 10.
Hypothesis: *There are more accidents on weekends due to alcohol* Figure 10 shows this.

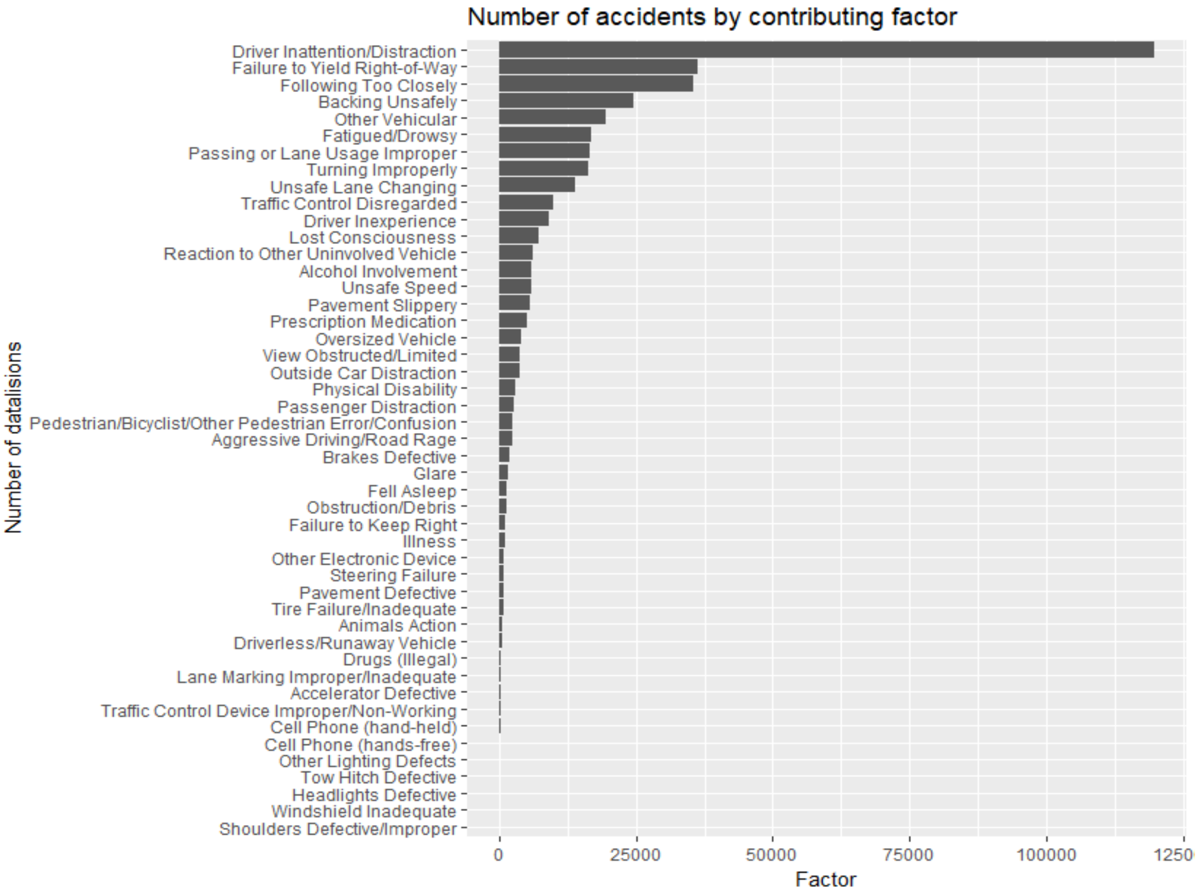


Figure 11: Contibuting factor

5.6 Number of cyclist accidents by month

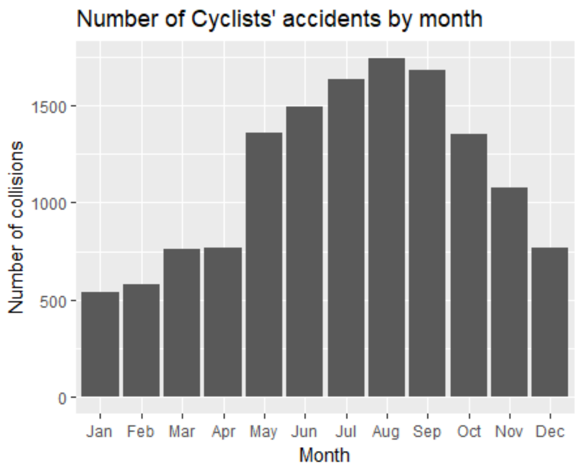
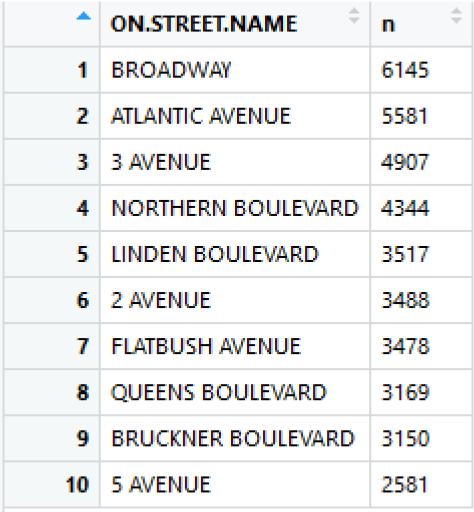


Figure 12: Accidents involving cyclits by month

Hypothesis: *Accidents where cyclists are killed or suffer injury happen more in Summer or Spring.*
Figure 12 shows the number of accidents involving cyclists in each month over the years. These accidents happen more in the Summer or Spring months when there are more cyclists on the street than in winter.

5.7 Top 10 most dangerous streets in NY

Figure 13 shows a list of 10 most dangerous streets in NY going by the number of collisions that happen on these streets. *'n'* is the number of accidents that took place on these streets over all the years. This list can be used with other data, like most popular factor of accidents on them to inform some safety measures to minimize the number of accidents that happen on them. The governing bodies can make NYC safer one street at a time using this kind of analysis.



	ON.STREET.NAME	n
1	BROADWAY	6145
2	ATLANTIC AVENUE	5581
3	3 AVENUE	4907
4	NORTHERN BOULEVARD	4344
5	LINDEN BOULEVARD	3517
6	2 AVENUE	3488
7	FLATBUSH AVENUE	3478
8	QUEENS BOULEVARD	3169
9	BRUCKNER BOULEVARD	3150
10	5 AVENUE	2581

Figure 13: Top 10 most dangerous streets in NY

5.8 Heat-maps for accidents involving pedestrains, cyclists and motorists

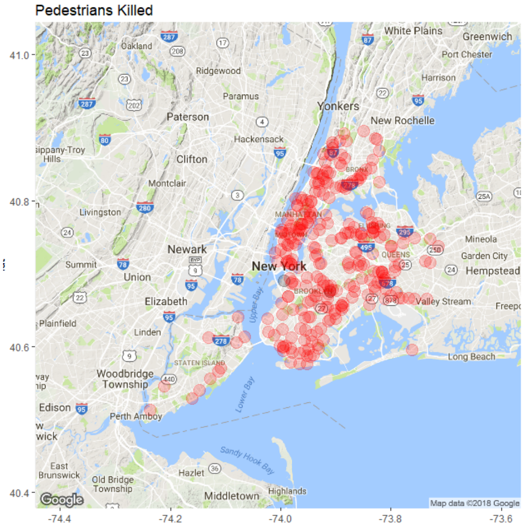

Pedestrians Killed

Figure 14: Heat-map for number of pedestrians killed

A heat-map is a kind of graph where the number or extent is represented using colors. So in our case, the location where a lot of accidents occured will be represented by a darker colour than others on the map of NYC.
Figure 14 shows a heat-map for accidents in which Pedestrains were killed. Figure 15 shows a similar map for cyclists and Figure 16 shows a heat map for motorists.

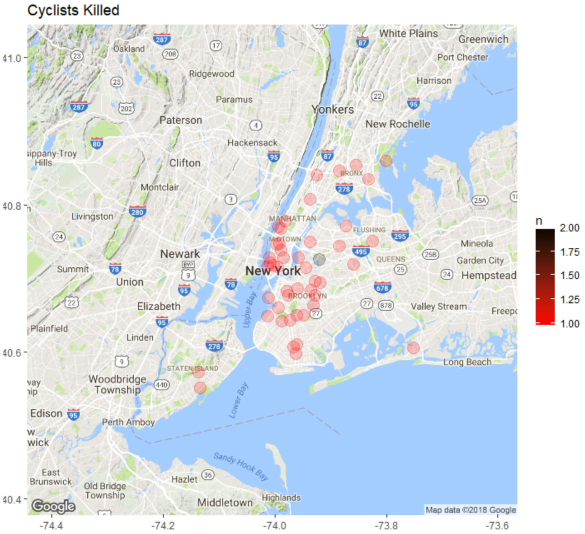

Cyclists Killed

Figure 15: Heat-map for number of cyclists killed

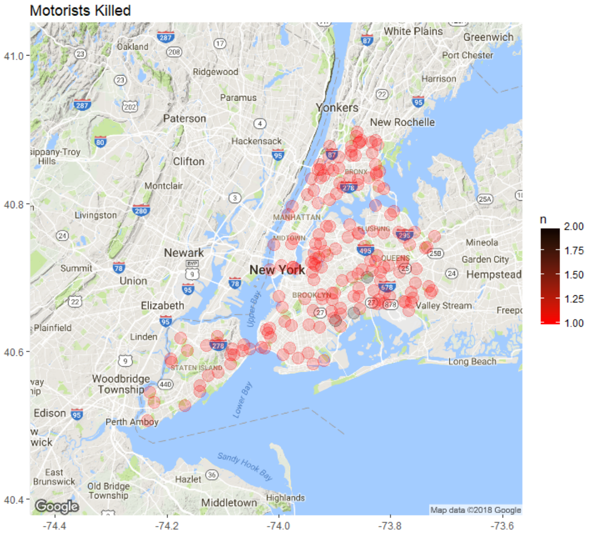

Motorists Killed

Figure 16: Heat-map for number of motorists killed

5.9 Number of accidents during daylight savings

Hypothesis: *Accidents increase around day light savings changes*
Figure 17 shows the number of accidents by day during daylight saving changes in November and March. There is no observable peak in accidents as compared to

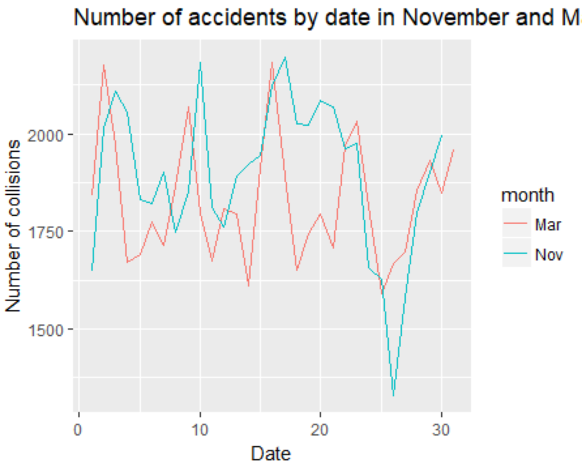


Figure 17: Accidents during Daylight savings changes

whole month around 4th November and 11th March.

6 Algorithms

6.1 Data

A subset of data that was used for EDA was extracted to perform classifications on it. This data was filtered as follows:

- Accidents happened on top 20 most dangerous streets of NYC
- Borough and on street name should not be null
- group across day, hour, month and on street name.

We get a dataset of $27,180$ rows from above. This dataset gives a number of accidents that occur on a street, given a month, day and hour. This number *'n'* varies between 1 and 14. From this number a class label was derived that depicts the danger of an accident or no. If normalized value of *'n'* is less than 0.14, then no accident otherwise, there is an accident.
This data is a trivial one which can be used to predict if on a given day if someone uses a street, will there be an accident. Figure 18 shows distribution of different attributes across the two classes. Weka was used for this analysis.

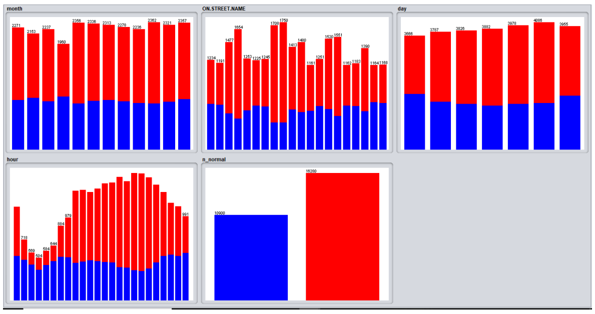


Figure 18: Distribution of attributes.

6.2 One-Rule model

A one rule classifier generates a threshold for every attribute in the dataset to provide classification and then chooses the attribute for which the error-rate was minimum. The data that is generated above is used to model a one-rule classifier in weka [20] along with 10 fold cross-validation.
The accuracy of the one rule model is 65.74%. The performance of this simple model is good given the data and the problem. The confusion matrix is in Table 1

Table 1: Confusion matix for OneRule

Prediction	0	1
Actual		
0	3730	7170
1	2143	14137

The attribute used by one rule model is hour. This tells us that amongst all the other attributes, hour has the most amount of correlation with our class label. The rules are in Table 2

Table 2: OneRule output for hour

Hour	Prediction
0	1
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	0

The false-positive rate of this model is high. However the true positive rate is also very high. The miss rate is very low, which is desirable. This performance is pretty good considering the simplicity of the model and the data provided.

6.3 Decision Tree

Decision tree is another model that uses the information content of attributes for classification. A decision tree can be thought of a cascaded series of if-else-if blocks. Figure 19 shows the layout of the decision tree built by Weka. There are 213 leaf nodes in the tree.

The accuracy of this model is 67.68%, which is a slight improvement over the one rule classifier and immense increase in the complexity. The confusion matrix for this model is given in Table 3

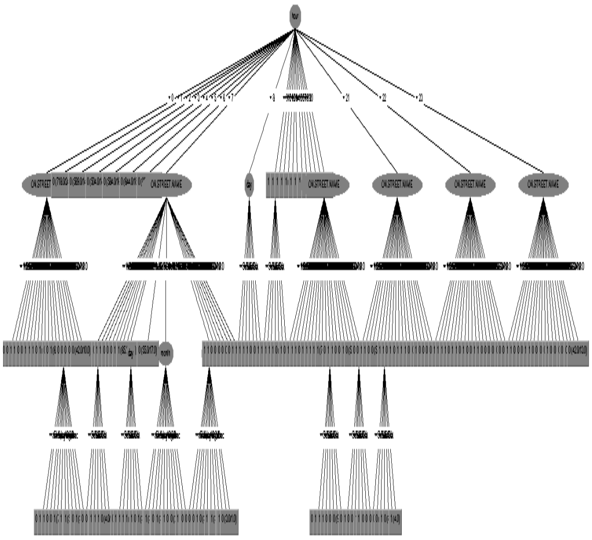


Figure 19: Decision tree structure

Table 3: Confusion matrix for Decision Tree

Prediction	0	1
Actual		
0	4824	6018
1	2767	13513

The False positive rate is a bit better than one rule while the true positive rate has been reduced slightly compared to one rule.

6.4 Naive Bayes

Naive Bayes is very different from the above two classifiers. Naive Bayes is a probabilistic classifier that predicts the class labels using Bayes theorem. It uses probability of a data point belonging into a class and then predicts the class with highest probability. Naive Bayes Classifier model was built in Weka using 10-fold cross validation. The accuracy of the model is 68.1%. This is slightly better than the previous two models. The confusion matrix of this model is given in Table 4

Table 4: Confusion matrix for Naive Bayes

Prediction	0	1
Actual		
0	4657	6243
1	2429	13851

7 Discussion

The project leads to discovery of some interesting trends in the data. It also helps in validating certain hypotheses that are made about the data just from the problem statement. This project also helps in understanding many data mining concepts. It shows that Exploring the data and understanding it is very important in order to be able to perform any data mining tasks.

It also shows how a very simple model like a one rule classifier will do a good enough job as compared with other more complex models. It also presents that trade-off between slight accuracy in performance and increase in complexity. Like in our case, one rule does the job well enough with a very simple approach, while more complex models perform just slightly better. Table 5 compares the accuracy of different models.

Table 5: Accuracy of different models

Model	Accuracy
One-rule	65.74%
Decision Tree	67.68%
Naive Bayes	68.1%

8 Conclusion and Future work

Performing Exploratory Data Analysis helped in getting answers for many questions and hypotheses and also provided many insights into the data that helped in creating a data set that could be used to create classification models. Some hypotheses were validated true while some were not, using the visualizations.

One could improve the model and try to explore the effect of using weather data along with the street name, and time information. It is a natural hypothesis that weather would affect visibility and that could cause increase in accidents.

Along with weather, traffic information could be used to make the model more robust and practical. The trained systems could then be use in GPS systems to provide routes that have the least chances of an accident. This will be a very advanced system that could predict the chances of an accident if the driver took a particular route given the traffic information on that route, weather of the day, day of the week and time. This model will make travel safe in the city by alarming drivers by possibilities of accidents.

References

[1] NYPD Motor Vehicle Collisions, *https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95* Accessed: 04/13/2018.

[2] AAA Study finds costs associated with traffic crashes are more than three times greater than congestion costs, *https://newsroom.aaa.com/2011/11/aaa-study-finds-costs-associated-with-traffic-crashes-are-more-than-three-times-greater-than-congestion-costs/* Accessed: 04/14/2018.

[3] Traffic Safety Facts - Research Note, *https://crashstats.nhtsa.dot.gov/Api/Public /ViewPublication/812318* Accessed: 04/14/2018.

[4] Susan Li. What I Learned From Analyzing and Visualizing Traffic Accidents Data, *https://towardsdatascience.com/what-i-learned-from-analyzing-and-visualizing-traffic-accidents-data-7cd080a15c15* Accessed: 04/14/2018.

[5] Hua Yang. New York City Motor Vehicle Collision Data Visualization, *https://nycdatascience.com/blog/student-works/new-york-city-motor-vehicle-collision-data-visualization/* Accessed: 04/14/2018.

[6] Vision Zero, *http://www1.nyc.gov/site/visionzero/ index.page* Accessed: 04/14/2018.

[7] Hua Yang. New York City Weather and Vehicle Collision Data Analysis, *https://nycdatascience.com/blog/student-works/new-york-city-weather-and-vehicle-collision-data-analysis/* Accessed: 04/14/2018.

[8] Weather Underground, *https://www.wunderground.com/history/ airport/KNYC* Accessed: 04/14/2018.

[9] Kaggle, *https://www.kaggle.com/nypd/vehicle-collisions/data* Accessed: 04/14/2018.

[10] Adhokshaja Pradeep. Exploratory Data Analysis, *https://www.kaggle.com/adhok93/exploratory-data-analysis* Accessed: 04/14/2018.

[11] Milan Bala. NYC Collision Analysis, *https://www.kaggle.com/milanbala/nyc-collision-analysis?scriptVersionId=878022* Accessed: 04/14/2018.

[12] Paul. Date/Time/Location Crash Correlation, *https://www.kaggle.com/thestats/date-time-location-crash-correlation?scriptVersionId=777746* Accessed: 04/14/2018.

[13] John Bencina. Analyzing NYPD Motor Vehicle Collisions with Python (Part 2), *http://www.jbencina.com/blog/2017/06/14 /analyzing-nypd-motor-vehicle-collisions-python-part-2/* Accessed: 04/14/2018.

[14] Google BigQuery, *https://cloud.google.com/bigquery/public-data/nypd-mv-collisions* Accessed: 04/14/2018.

[15] Lecture Notes provided by Instructor *https://mycourses.rit.edu/d2l/le/content /686207/Home* Accessed: 04/14/2018.

[16] Lubridate documentation *https://www.rdocumentation.org/packages /lubridate/versions/1.7.4* Accessed: 04/28/2018.

[17] DPLYR documentation *https://www.rdocumentation.org/packages /dplyr/versions/0.5.0* Accessed: 04/28/2018.

[18] GGPLOT2 documentation *https://www.rdocumentation.org/packages /ggplot2/versions/2.2.1* Accessed: 04/28/2018.

[19] GGMAP documentation *https://www.rdocumentation.org/packages/ggmap/versions/2.6.1* Accessed: 04/28/2018.

[20] WEKA documentation *https://www.cs.waikato.ac.nz/ml/weka/documentation.html* Accessed: 04/28/2018.