Analysis Of New York City's Motor Vehicle Accidents

- **Vinod Babu Palani**

1. Introduction

1.1. About the Dataset

The data was obtained NYPD **(**New York Police Department) officials at this web address: https://data.ny.gov/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95. It is an open source that can be accessed from the web address. It is updated regularly and the most recent data included in the data analysis is August 2017. The data three years of datasets, i.e. 2015, 2016 and the current 2017. The cut of date for the 2017 dataset is end of August 2017.

The data has 595,553 rows, each representing an accident that occurred in the city of NY, and 21 variables (columns). Different libraries are used to explore, munging the data and to plot and geographically represent the data.

The data was transformed by reducing the original variables and adding new variables, and dropping NA values from the dataset. The transformed dataset has 380425 rows and 12 variables.

1.2. Motives

There are several motives for choosing and working on this dataset. Key motives are listed below.

1. This is a real-time data that represents daily occurrences in actual daily life of the City of New York. As such the data gives better real-life sample than working on synthesized data (randomly generated data).
2. The dataset provides opportunities to analyze the data from different perspectives by considering the causes of the accidents, the damages resulted from the accidents and the human, physical (road and vehicle) conditions and environmental (weather and terrain) situations at the time of the incident.
3. The dataset has also real value as the data analysis can be used to impact actual policies of different sectors of the City of New York.

1.3. Data Attributes

Below are key data attributes from dataset which is used for data analysis and visualization.

- Date
- Time
- Borough
- Zip code
- Latitude
- Longitude
- Location
- Number_of_Injured
- Number_of_Killed
- Contributing_Factor_Vehicle
- Unique Key
- Vehicle_Type_code
-

2. Goals

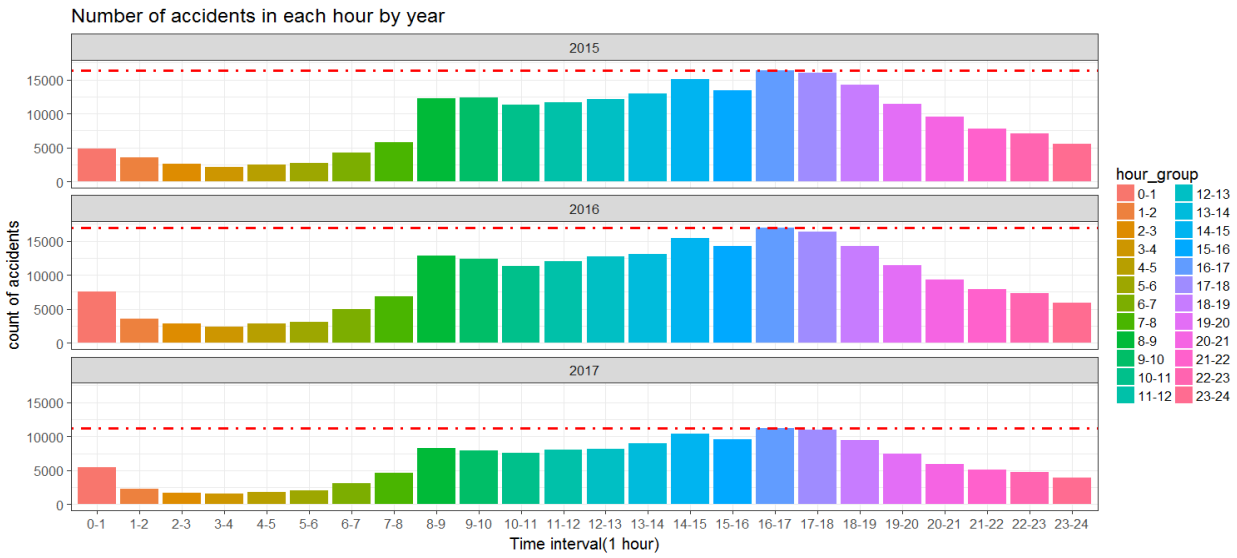
By manipulating and visualizing data from different perspectives, below goals are achieved.

- Time series of accidents and fatalities
- Geographical representation of fatal accident zones

2.1. Time series of accidents and fatalities

Hour of the day when the maximum number of accidents happen

On plotting the NYPD collision data from 2015 to Aug 2017 on the time scale with 1-hour interval, we observe below pattern.

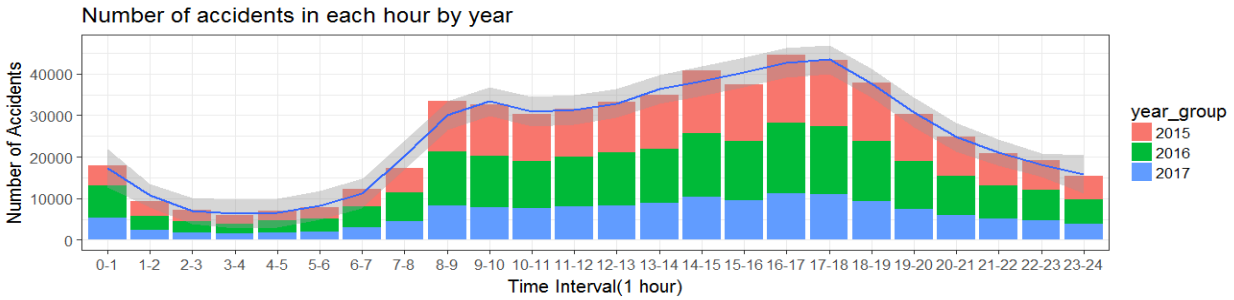


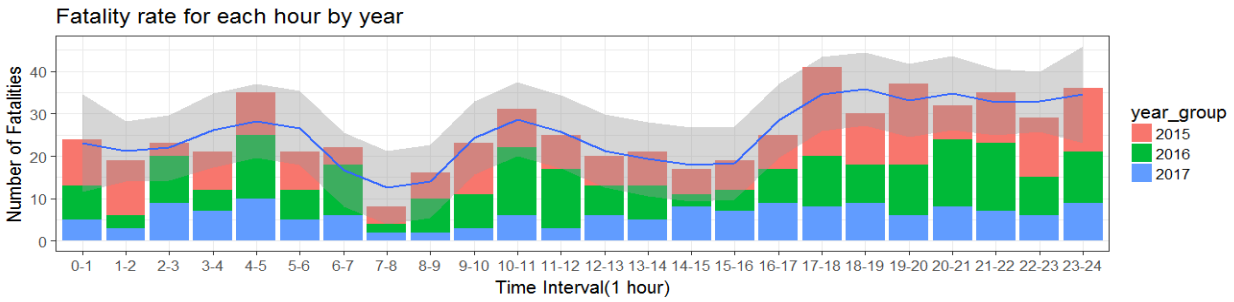
Below information is inferred from the graph,

- Maximum number of accidents happen between 4 PM and 6 PM of the day which is peak time of the day when most of the people return home from their work.
- Least number of accidents happen at late night and early morning when the number of commuters are less.

Time Series Analysis of Fatalities percentage and Accidents percentage

- Before understanding the time series of fatalities and accidents, it is needed to understand the spread of fatalities and accidents over the time. Below are the graphs for the same.

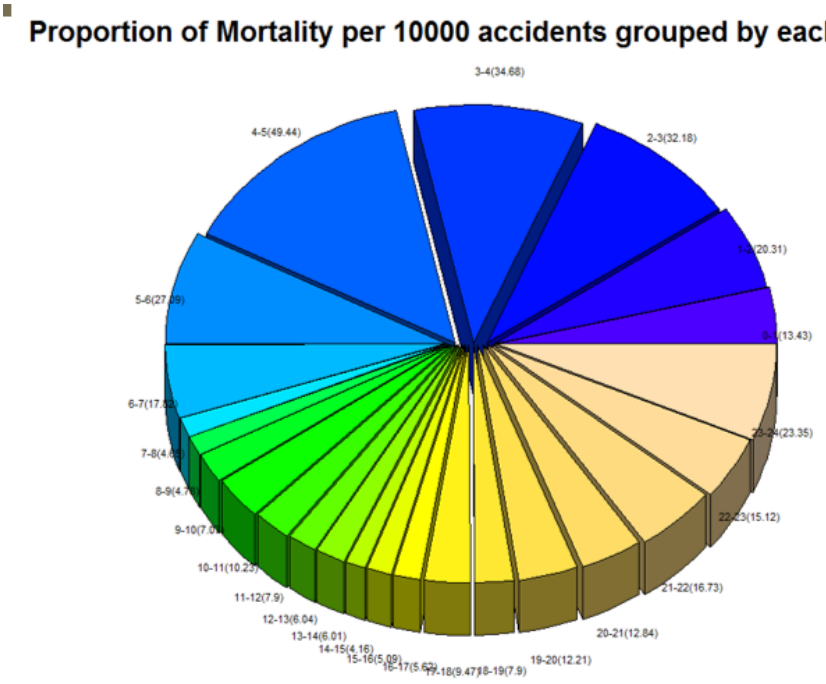




- Above graph indicates that the trend of Number of Fatalities and Number of accidents happening over time interval is not the same and follows different patterns.
- Below is the Time Series plot of Accidents percentage and Fatalities percentage to understand the relationship between two better.

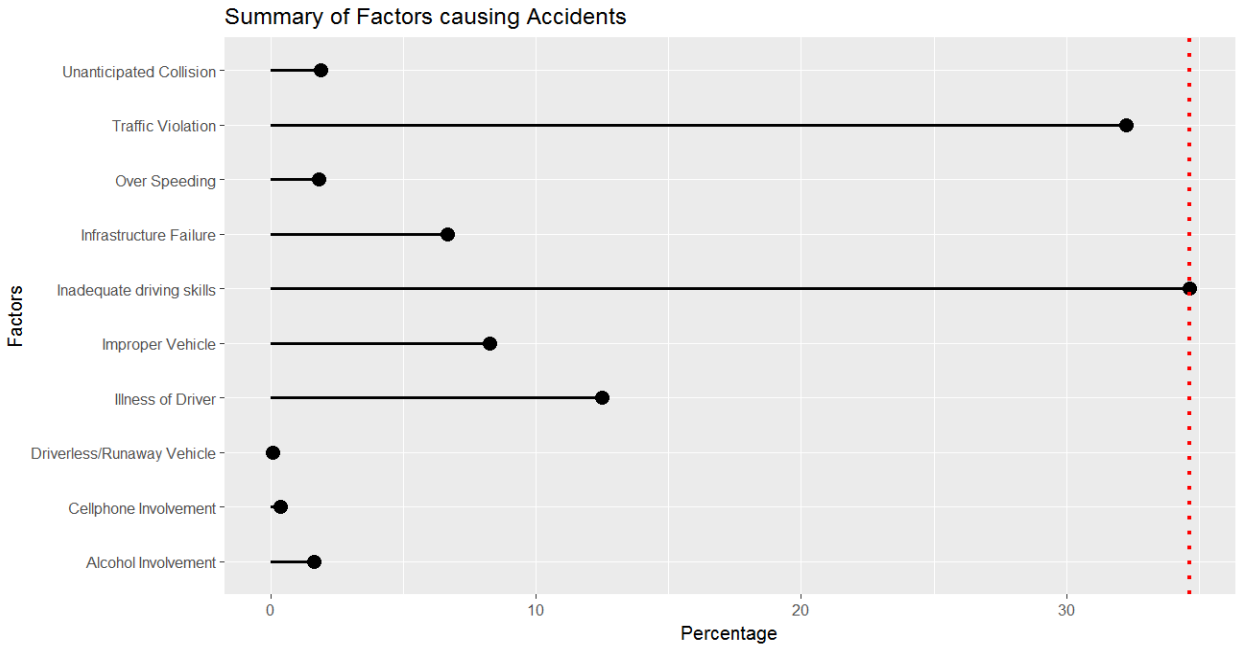
- Below are insights obtained from the Time series graph
    - Accidents percentage is more during broad daylight (7 AM – 6 PM) and is starting to fall late evening and early morning hours Whereas fatalities percentage is more during late evening and early morning hours (6 PM and 7 AM) than in the broad daylight.
    - Likelihood of a person dying in an accident happening during late evening and early morning hours is much more compared to that of the daylight hours.
    - Most of the fatal accidents happen during late evening and early morning hours whereas most of the minor accidents happen during daylight hours.
    - Given an accident happened during late evening and early morning hours, it is likely to be more fatal and severe compared to accidents that might happen during broad daylight.

- To quantify the claim made in above time series analysis, proportion of mortality per 10000 accidents for each hour is calculated and plotted as below.
    - It can be inferred that amount of lives claimed per 10000 accidents is highest during 4 AM and 5 AM which is 56.51 and is on the higher side during late night and early morning hours (Red highlighted values) compared to other hours.



hour_group	prop_per_10000
12AM-1AM	15.11
1AM-2AM	21.38
2AM-3AM	36.38
3AM-4AM	34.68
4AM-5AM	56.51
5AM-6AM	29.67
6AM-7AM	18.62
7AM-8AM	4.65
8AM-9AM	4.78
9AM-10AM	7.07
10AM-11AM	10.23
11AM-12PM	8.22
12PM-1PM	6.04
1PM-2PM	6.58
2PM-3PM	4.16
3PM-4PM	5.09
4PM-5PM	5.62
5PM-6PM	9.47
6PM-7PM	8.96
7PM-8PM	12.21
8PM-9PM	12.84
9PM-10PM	16.73
10PM-11PM	15.12
11PM-12PM	24

Summary of Factors causing accidents

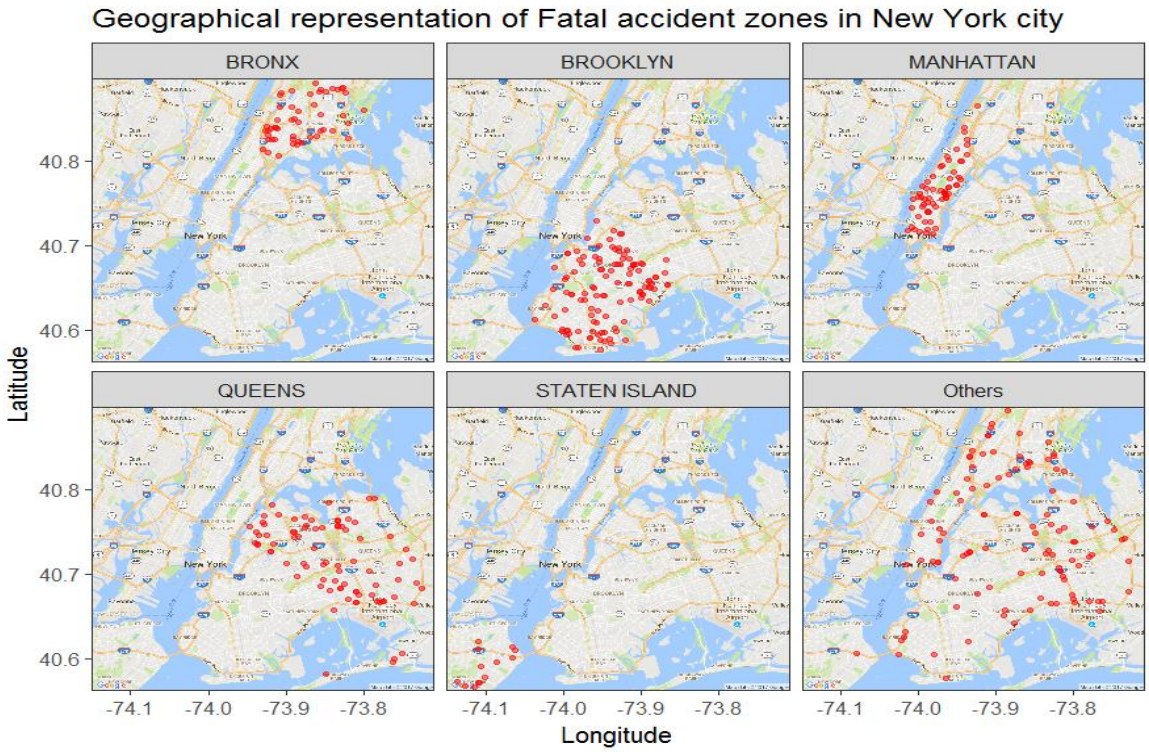


- Inadequate driving skills and Traffic Violation has been the primary reasons for causing accidents
- Motion chart representing the variation of top 5 factors over the years (2015-2017) is made using google visualization package. Below is the code for the same.

```
plot4 <- gvisMotionChart(
data = filter(data_motion_chart,Contributing_Factor_Vehicle %in%
c('Illness of Driver','Improper Vehicle','Inadequate driving
skills','Infrastructure Failure','Traffic Violation')),
idvar = 'Contributing_Factor_Vehicle',
timevar = 'year_group',
yvar = 'Count of Accidents',
options = list(height = 420,width = 500,showSidePanel=FALSE))
plot(plot4)
```

- From the motion chart, it is understood that Inadequate driving skills and Traffic Violation is rising over the years (2015-2017).

2.2. Geographical Representation of Fatal Accidents zone in New York city Boroughs



3. Conclusion:

Below are the very vital conclusions from analyzing the dataset

- Maximum number of accidents happen between 4 PM and 6 PM of the day which is peak time of the day when most of the people return home from their work.
- Least number of accidents happen at late night and early morning when the number of commuters are less.
- Accidents percentage is more during broad daylight (7 AM – 6 PM) and is starting to fall late evening and early morning hours Whereas fatalities percentage is more during late evening and early morning hours (6 PM and 7 AM) than in the broad daylight.
- Likelihood of a person dying in an accident happening during late evening and early morning hours is much more compared to that of the daylight hours.
- Most of the fatal accidents happen during late evening and early morning hours whereas most of the minor accidents happen during daylight hours.
- Amount of lives claimed per 10000 accidents is highest during late evening and early morning hours compared to other hours
- Inadequate driving skills and Traffic Violation has been the primary reasons for causing accidents and is been rising over the years

- There is a relationship that can be derived from the accidents caused due to inattention and distraction more in summer and spring due to higher number of accidents in summer and spring and not the same observed in winter and fall seasons. Thus, conclusion can be arrived that drivers are cautious in harsh weather conditions and lethargic in comfortable weather conditions
- Passenger vehicles are involved in more collisions and need to be educated on the common contributing factors involved in accidents and how to avoid them.
- Brooklyn has the has number of fatalities and injuries followed by Queens and Manhattan and Bronx.

4. Next Steps for the Data Scientists

Given the analysis made so far, Dataset can be further explored in upcoming courses to find out below vital information.

- Estimating accident prone routes
- Routes where over-speeding is very likely.
- Accidents due to traffic infrastructure failure
- Identifying cause of fatal accidents
- Predict the next accident by regression to the time series
- Accident rates in Highways and city routes
- Roads that needs speed tracking device and road signs to be installed
- Borough that needs better traffic regulations