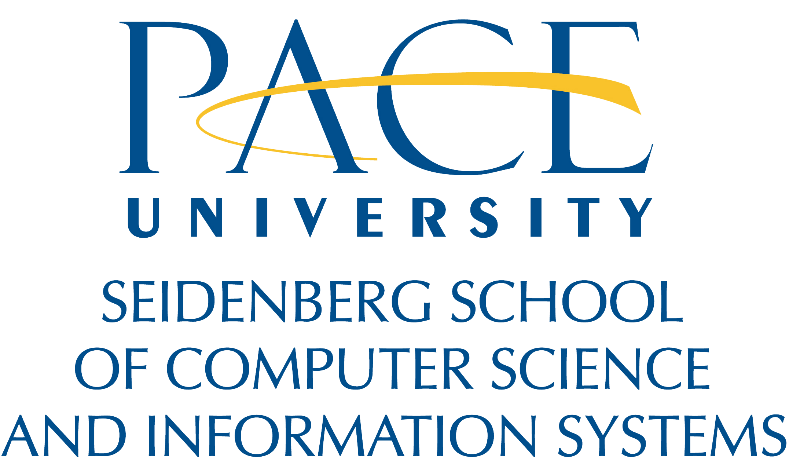
**Data Warehousing, Mining & Visualization**

**Project Report**

**NYPD MOTOR VEHICLE COLLISIONS**



Submitted To: Submitted By:

**PROF. Yegin Genc** Bidhata Subba

**Introduction**

The Dataset taken into consideration is NYPD Motor Vehicle Collisions. The dataset has details of collisions that are reported in which people are injured or killed be it motorist, drivers, pedestrians, cyclist or passengers of the vehicle. We have taken into consideration a business problem where in NYPD wants to decrease the collisions and make the city safe.

**Data Set**

The data set was retrieved from: <https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95>. It consists of various details of collisions in NYC from 2012 to 2016, for example: Date, Time, Borough, Zip code, Latitude, Longitude, Location, On Street name, Cross Street Name, Off Street name, number of persons injured, number of persons killed, number of pedestrians injured, number of pedestrians killed, number of cyclist injured, number of cyclist killed, number of motorist injured, number of motorist killed, contributing factor vehicle 1, contributing factor vehicle 2, contributing factor vehicle 3, contributing factor vehicle 4, contributing factor vehicle 5, Unique Key, Vehicle Type Code 1, Vehicle Type Code 2, Vehicle Type Code 3, Vehicle Type Code 4 and Vehicle Type Code 5.

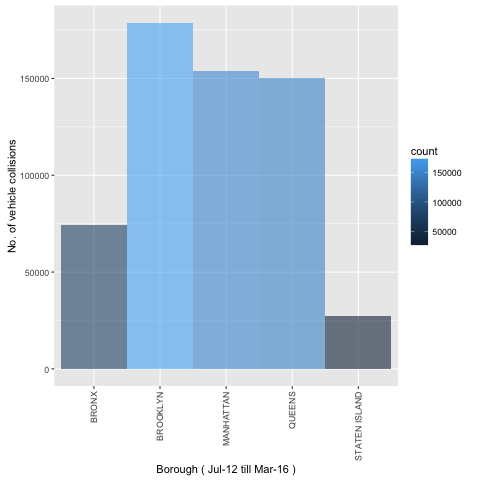
**Descriptive Analytics**

The Date and Time columns describe the date and time details. The Borough, Latitude, Zip Code, Longitude and Location columns provide the location details. The other few columns describe the number of pedestrians, bikers and motorists either killed or injured. The next few last columns describe the contributing factors that led to the collision and the type of vehicles that were involved in the collision. The Unique key column is a unique value given to each collision. The contributing factors that have been considered are drink and drive, using mobile phone, drowsiness, technical problems in vehicle and medication. We noticed that there can be other contributing factors as well that were not considered such as weather.

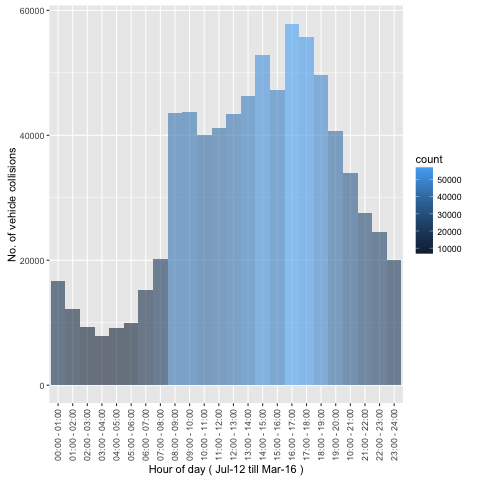
The total number of rows in the data set is around 770 thousand and like every other data set, this data set also had some missing values. For example, around 184 thousand rows had missing Borough & Zip code values; around 120 thousand rows had missing Latitude, Longitude and Location values. The total number of records that don’t have Borough, Zip code, Latitude, Longitude, and Location was around 121 thousand, and these records were completely ignored while analyzing the data based on the mentioned columns.

The data set was analyzed using R over the period of July 2012 till March 2016 & various facts were observed.

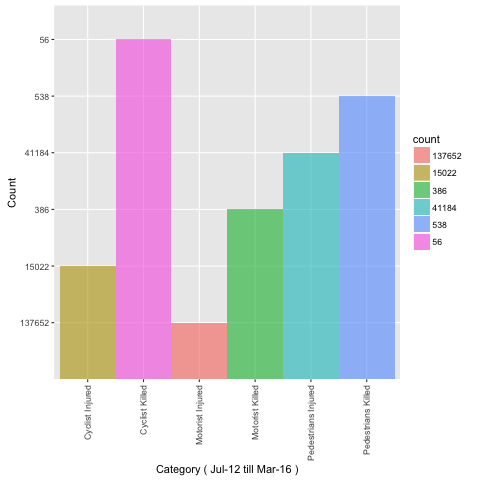
It was observed that the number of accidents is highest at Brooklyn Borough. The graph below furthermore emphasis the same point:



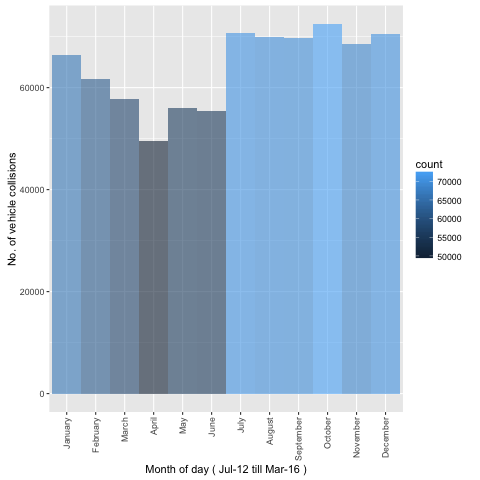
Further analysis led to the observation that the numbers of accidents are highest between 4PM and 6PM. The graph below descripts the same.



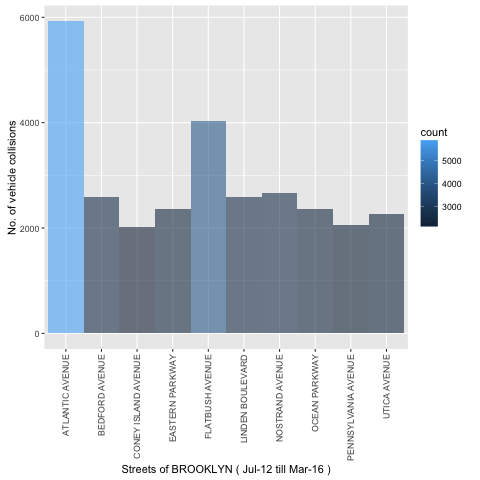
The analysis also shows that in the collisions, the cyclists are the ones that are killed the most. The graph below shows the same.



The date wise analysis led to the observation that the month of October has the most number of collisions. Also, it can be seen that the number of collisions tends to increase from June. It can also be said that the collisions increase during festive period that is during Thanks Giving and New Year’s. The graph shows the same.



Since we know the time and month when the collisions had occurred so further the analysis showed us the exact street where most number of collisions occurred. Our analysis concluded that in Brooklyn mostly collisions occurred on Atlantic Avenue. The graph below describes the same.



With the help of R, the data set can be further analyzed to gain more insights. This will be a great help for NYPD because by knowing these factors, they can take preventive measures in the right direction. For instance, they can deploy more officers at Atlantic Avenue from 4PM to 6PM. They can start campaigns from June as the data analysis shows that the collisions tend to increase from June.