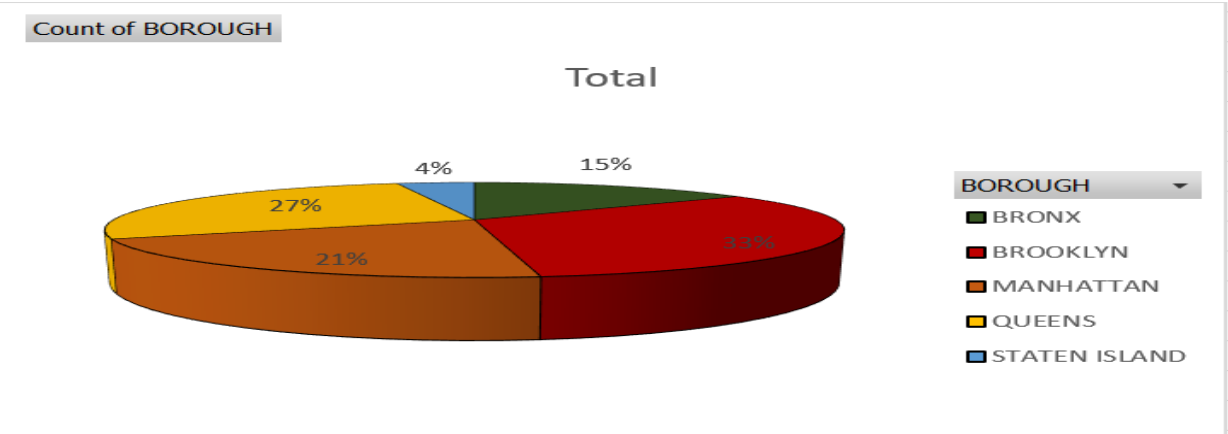
Data Mining Mini Project

Title: **New York Collision Analysis**



Authors:

Utsav Patel: uxp2146 Sailee Rumao: sxr9810

Data source: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95[1].

Tools Used:

Python, Tableau, Excel

Abstract:

New York City is one of the major metropolitan cities in the United States which is why it is densely populated. Vehicle collisions are quite common and is a major concern in such cities. As a result, New York State Police Department (NYPD) has provided the public with the database of all accidents that occur on daily basis in New York city. In this study, we focus on finding common characteristics in the accidents by studying them month-wise,(June-July) year-

wise(2017-2018), day-wise, by street-location, comparing with the contributing factor that led to the accident and understanding all the possible combinations of the attributes to see relationship with the accidents in one of the boroughs of New York city 'Queens'. The study also includes clustering in order to identify the regions with most accidents. The analysis is concluded with general findings, discussion and recommendations to prevent collisions, and for decision-making, to plan for safety measures.

Introduction:

New York city is one of the busiest cities in the world and driving here can be a headache due to the vast amount of traffic and the number of intersections. Thus, a lot of accidents occur here on a daily basis and, safety measures should be taken to reduce the number of such crashes. New York State Police Department (NYPD) has made the data containing these collisions publicly available so that any one can identify the streets with maximum accidents, the time at which these accidents are frequent, and infer why these things happen. For our project, we decided to focus on the months of June and July for the years 2017 and 2018 because during this period of time, the city sees a lot of visitors.

New York city comprises of five boroughs namely, Queens, Brooklyn, Bronx, Manhattan, and Staten Island. For our study, we chose Queens since it is the largest borough among all the other boroughs and the second largest in population after Brooklyn[2]. The first step that we performed in our data mining project was to understand the data provided to us. Each of the 29 attributes were studied in detail to identify the attributes to focus on. The next step was to clean our data by removing the records which do not give us any significant information. After the cleaning process, we generated some features like Time Interval, Street Location, Day, etc. which are explained in detail in the sections below. Using these features and some attributes, we have

visualized our data to observe trends and patterns that can help us to take measures to avoid such accidents. We also implemented clustering algorithms like KMeans and DB-SCAN on our data to identify the locations where the density of such accidents is high.

Data Preparation:

The data set we downloaded online had 1.39 million rows and 29 columns. Some of these columns/attributes are Date, Time, Borough, Location, On Street Name, Contributing Factor, Number of Persons Injured, etc. The attribute Location and Borough had many missing values and thus, we removed all those records as they won't make any sense without the location of the crash. We tried to impute the missing values of Boroughs by reverse geocoding the location coordinates using an online API, but we got a timeout error since it only works for a few number of records. For this project, we are going to analyze the months of June and July for the years 2017 and 2018 and since, we chose Queens as the borough, the size of our data reduced from 1.39 million records to just 13,623. In order to perform better analysis, we generated some features like Time Interval, Street Location, Day, Month, Year, etc. For the feature Time Interval, we made 7 categories based on domain knowledge. The categories are as follows: -

5 AM TO 10 AM – People Travelling to Work
10 AM TO 12 PM – Non-traffic hours
12 PM TO 3 PM – Lunch Time
3PM TO 5 PM – Non-Traffic Hours.
5 PM TO 8 PM – People returning from Work
8 PM TO 10 PM – Dinner Time
10 PM TO 5 AM – Sleeping Time

We observed for the street locations that when the On Street name and Off Street name is missing, the Cross Street name data is given. Thus, we merged all these three attributes to get our

feature Street Location. We extracted days, month and years from the Date attribute.

Data Analysis and Visualization:

1. Number of Accidents in June-July in 2017-2018

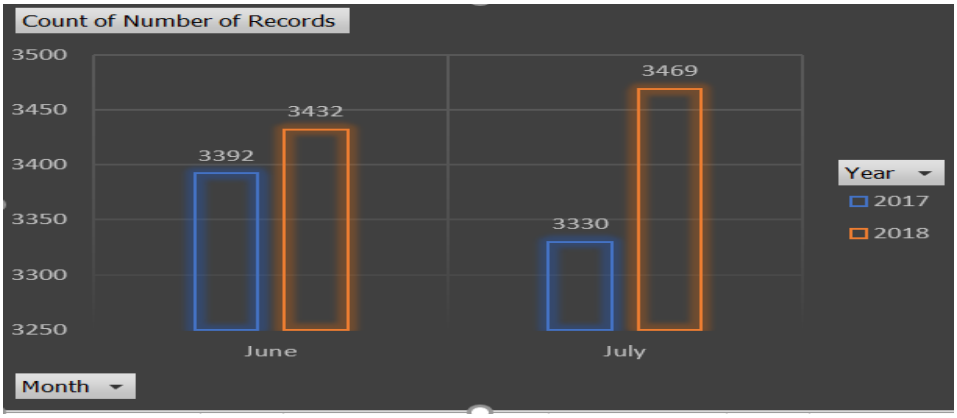


Fig (1.)

We can clearly see from the above graph that the number of accidents in July are more than the number of accidents in June. One inference we can make from this is that, the schools are closed in the months of July and August and a lot of people come to visit the city.

2. Number of Accidents by street location in each month

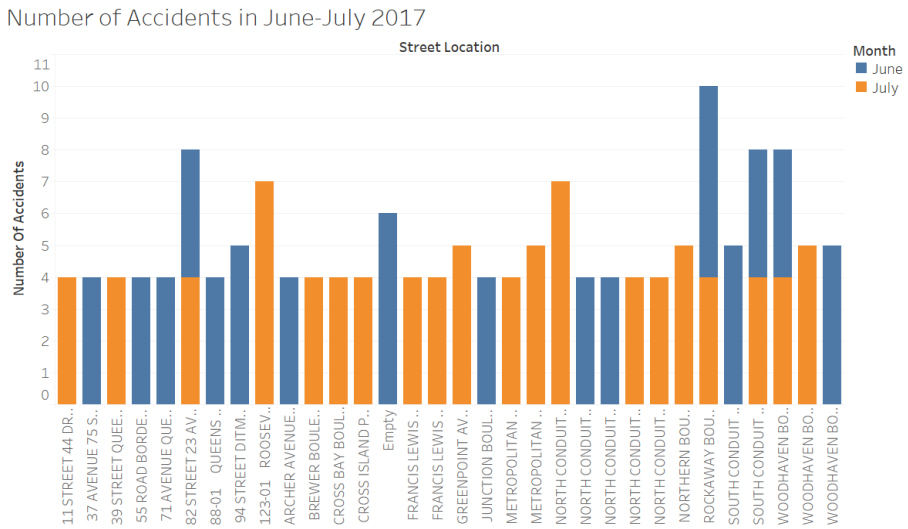


Fig (2.a)

From this graph, we can see that the common streets where accidents took place both in June and July 2017 are 82 Street 23 Avenue, South Conduit Avenue Farmers Boulevard, Woodhaven Boulevard Union Turnpike, and Rockaway Boulevard Brewer Boulevard. We can clearly see from Fig (2.b) that Rockaway Boulevard Brewer Boulevard is the street with the maximum crashes.

We can infer that this place has the highest number of accidents **since it is is near a bridge** which we observed on Google Earth**.**

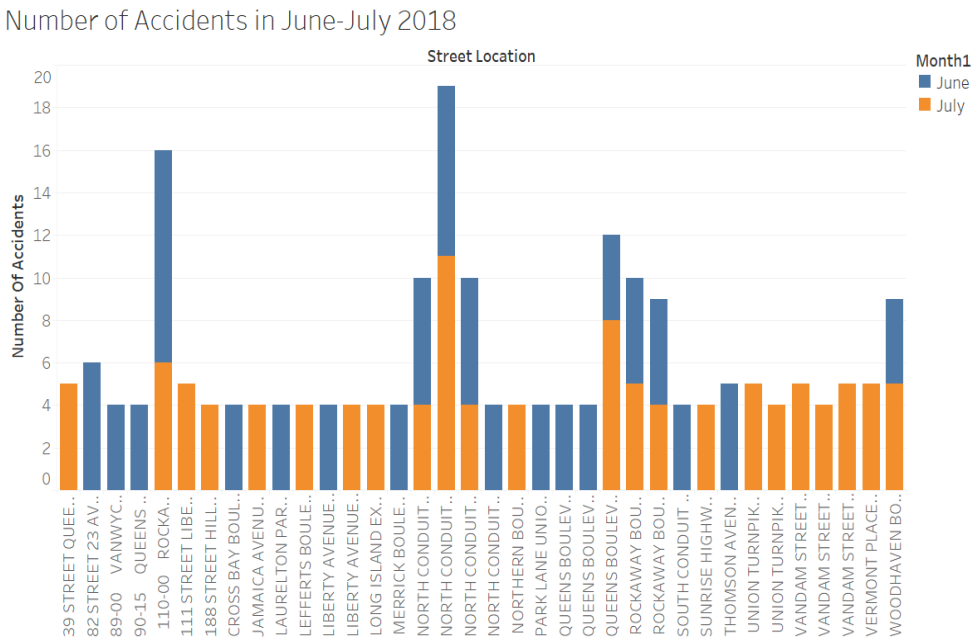


Fig (2.b)

In 2018, we see that highest number of accidents occur in NORTH CONDUIT AVENUE SPRINGFIELD BOULEVARD.

3. Number of Accidents by Contributing factor in June-July 2017-2018

The two major contributing factors have been Driver's inattention and unspecified which is constant for both the months throughout 2017-2018 as can be seen in the graph below along with the statistics of other contributing factors.

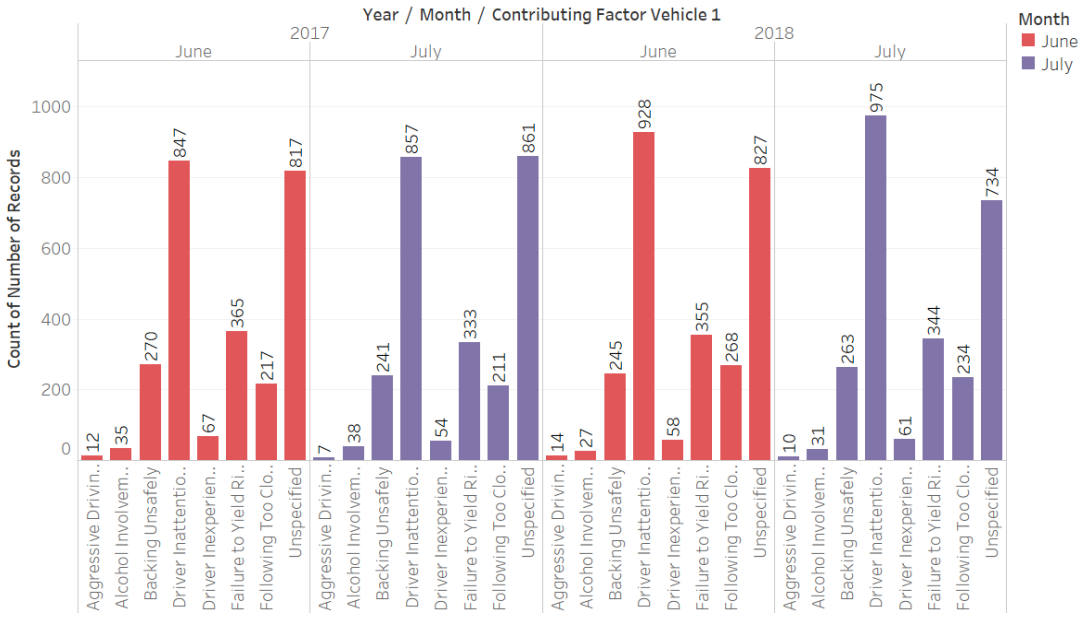


Fig (3.a)

We further studied the Contributing factors by the time-interval for both the months and figured that accidents due to alcohol involvement happen mainly during 10 pm – 5 am and a few of them between 5 am- 10 am as can be seen in the graph. This time frame is considered to be sleeping time for most people but from this study it can also be labelled as party time for a few people.

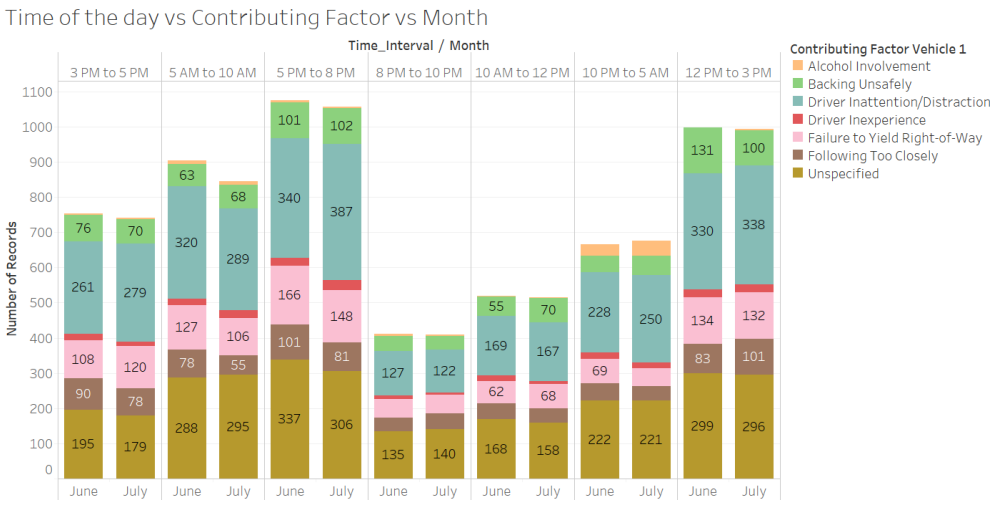


Fig (3.b)

To see if there is any relationship between Alcohol and week days, we plotted the following graph.

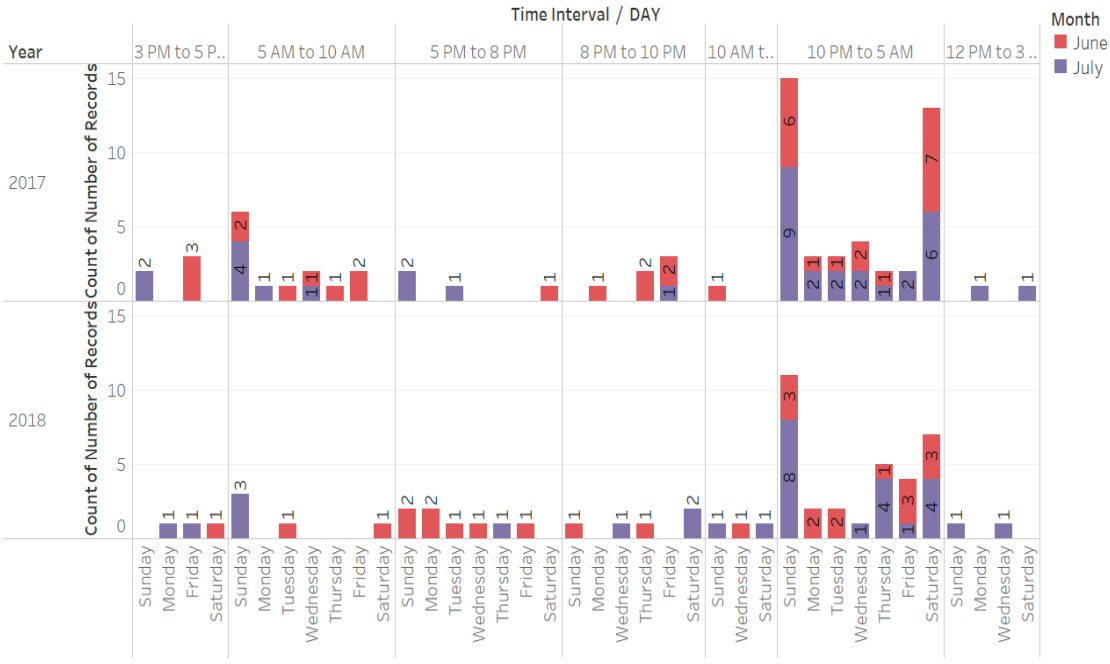


Fig (3.c)

From this graph we can clearly see that maximum number of accidents due to alcohol in the above mentioned time-frame happen on weekends (Saturdays & Sundays).

4. Clustering approaches

K-Means.

In K-Means clustering, we cluster our data into k groups and observe the trends in each of these clusters. In our project, we created clusters on our data set using the latitude and longitude attributes of our data. Initially we performed K-means on the entire data set and then we performed it on just the Queens data set. The results are as follows: -

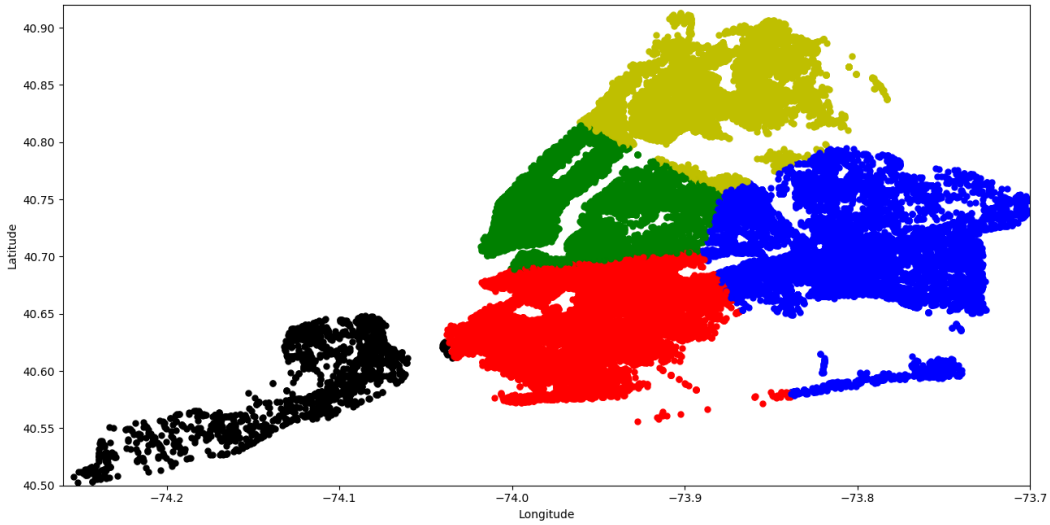


Fig (4.a)

This graph does not tell us much about the locations where the accidents are frequent. Since, accidents occur almost everywhere in the city, this is not of much help.

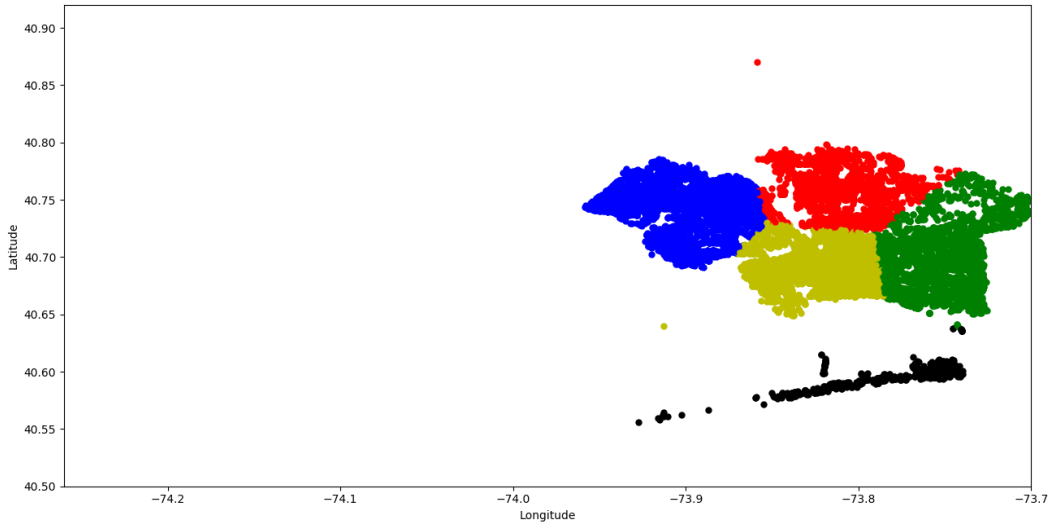


Fig (4.b)

This figure results from implementing K-Means on just the Queens data set. Even this figure doesn't tell us much about specific locations where accidents are quite frequent.

DB-Scan

In DB-Scan clustering, we find highly dense clusters which is the case for our data. There are some intersections in Queens, where the number of accidents are quite frequent. Thus, DB-scan helps us identify all such locations and eliminates all the outliers. For this algorithm, we used an epsilon value of 0.005 and min_samples as 10.

The results are as follows :-

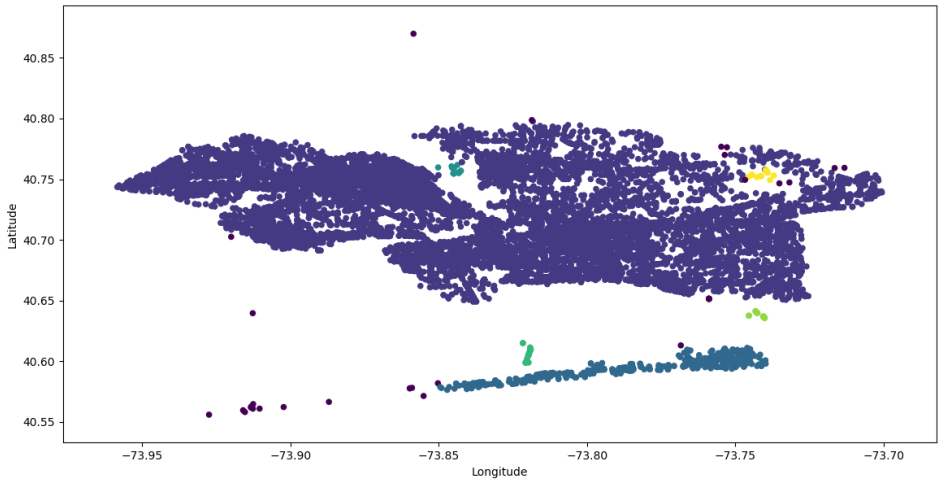


Fig (3.c)

From this figure, we can clearly see that the yellow regions, green regions and the blue regions are identified which was not visible in K-Means where the accidents occur more. After getting these clusters, we saw that the locations that we obtained from them were matching the street locations mentioned above with the most number of accidents.

Discussion and Conclusion

The study was good overall and the experience was great as we dealt with a real data set. This study was beneficial as the next time we, our friends or family visit the city, we can be cautious. The only clustering algorithm that we didn't implement in this course was DB-Scan. We applied that in this project and it turned out to be the best clustering approach as it helped identify the locations where accidents are more frequent. These locations that we obtained from the

clustering approach were observed in detail with respect to the contributing factors, time-interval, weekends, etc. We observed that for the location, Rockaway Boulevard Brewer Boulevard, the two major reasons for the accidents to happen are driver inattention and following too closely. Thus, we can say that just including slow zones in such areas will not solve the problem. The infrastructure of these places should also be taken into reconsideration. Another thing that we observed was that, most accidents occur on weekends due to alcohol consumption. Thus, stricter drink and drive laws should be enforced.

We also learned how to better visualize our data. Most of the visualization plots were rendered in Tableau. For this term project, we analyzed only the months of June and July for this year and the past year. In the future, we can extend this study for all the boroughs and years to find better trends and patterns.

References:

[1] **https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95** **[Accessed December 7, 2018]**

[2] **https://en.wikipedia.org/wiki/Queens** **[Accessed December 10, 2018]**

[3] **https://blog.alookanalytics.com/2017/02/14/advanced-analytics-with-python-and-tableau/** **[Accessed December 7, 2018]**