Aditya Kalyan Jayanti
aj8582@rit.edu

Rohit Ravishankar
rr9105@rit.edu

Project 2: New York City Motor Vehicle Collisions
CSCI - 720 Big Data Analytics

1. Abstract

The purpose of this report is to showcase the work performed by us on the New York City Motor Vehicle Collisions dataset. The activities are aimed at exploring the trends and unidentified knowledge in the data. The knowledge could be used by the government to mitigate the number of accidents and make the city safer.

2. Introduction

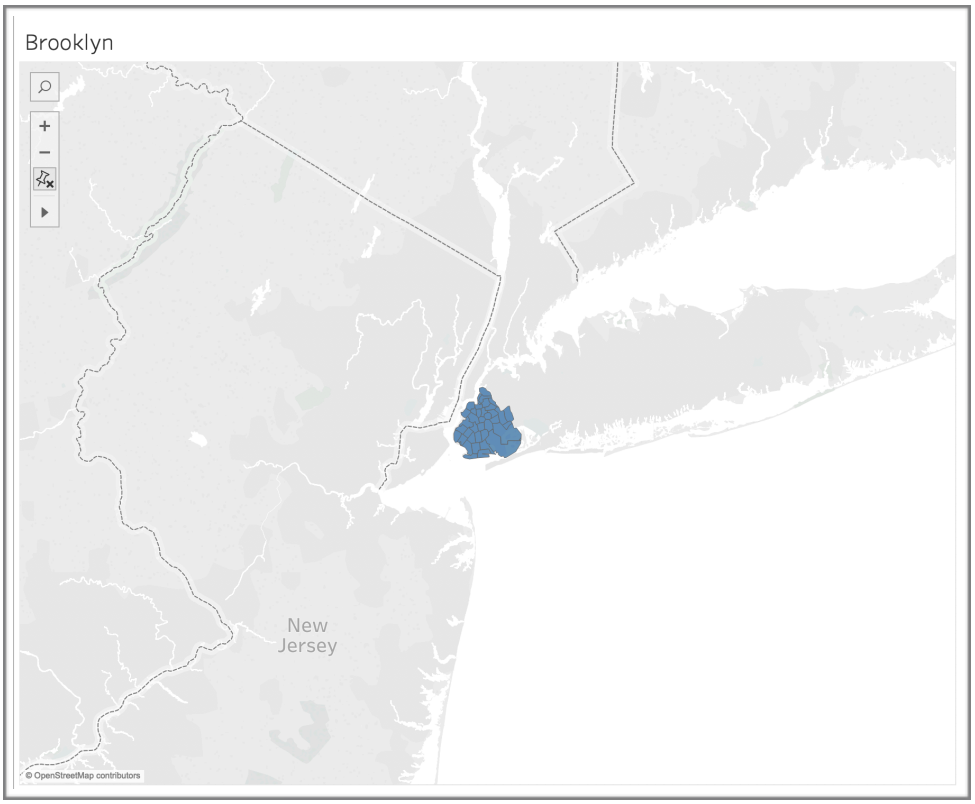
The goal of data mining is to process data in order to uncover trends and knowledge in raw data. The knowledge is present in form of correlations between attributes which cannot be observed by directly looking at raw data. This knowledge can be used to make informed and effective decisions.

Road safety is a critical issue which impacts a large part of the general population. American Automobile Association estimates that crashes cost society approximately $300 billion annually. This suggests that an effective plan needs to be in place for reducing the motor vehicle accidents to save lives and reduce loses. Data mining techniques can be employed to study past incidents and come up with effective measures to achieve this goal.

In the following project, the data set is acquired form the data provided by the city of New York. The data is often updated by the NY Police Department. Effective data mining on the data set can be used to reveal knowledge about them Boroughs prone to accidents, the days with the highest number of accidents, the intersections with the highest number of accidents etc. This

data could be effectively used to install stop signs, increase the time on a signal, creating crosswalks etc. to reduce the number of accidents.

The borough we have chose for performing our data analysis is Brooklyn. We chose Brooklyn based on our online research, which stated that most of the accidents in NYC take place in Brooklyn due it being the most populous borough of NYC.



3. Data

The data is freely provided by the New York State and is regularly updated by the NYPD. This data has information about all the motor vehicle collisions, that took place from July 2012 till date. The different attributes include:-

- Date and Time
- Borough of collision
- 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 unto 5 vehicles
- Vehicle type codes of unto 5 vehicles

This is data was procured possibly on 7th of June, 2018 considering that there is information available on the dataset until 6th of June, 2018 and that the dataset is regularly updated. The dataset has 1,379,936 records. The project focuses on the last 2 years of data.

4. Previous Work

A lot of work has been done in this area by different data science enthusiasts. Almost always, research starts by analyzing the trend over the years of the number of collisions, number of deaths, number of injuries etc. This data could be further used to find correlations and uncover patterns in the data. Some of the previous work done in this area has been reviewed along with interesting discussions.

4.1 Analysis of NYPD data set

There has been a lot of work which has been already done on the NYPD data set. In [5], the author discusses various trends and compares them for 5 borough. Based on the analysis, the author points out that there are more number of accidents on Friday than any other day of the week. Further, the author talks about how there is a reduction in the number of accidents in the year of 2016 because of the successful vision zero campaign. We further see that Manhattan has the highest number of pedestrian based accidents and is possibly due to the number of people who congregate and walk around in Manhattan. An interesting point we further notice are the causes of crashes. We see that the maximum number of crashes are due to fatigued drivers and drivers driving under the influence of alcohol.

The author further suggests measures for the government to reduce motor vehicle crashes. On another source[7], NYC hourly weather data is discussed along with the NYPD data to understand the correlations. The author analyzed the total number of accidents across different weather conditions. It was found that snow and humidity have a greater impact than temperature and visibility.

On [7], the author uses NYPD data to discuss how, weekends and alcohol have a high positive correlation and how people tend to ignore traffic lights at this time, and snow depth and slippery pavements are correlated.

4.2 Analysis of NHTSA data set

There is a similar data set akin to the one used on this project. It has been provided by National Highway Traffic Safety Administration. This organization's goal is to reduce crashes, fatalities and injuries caused due to motor accidents. The organization has published a report detailing the percentage of increase in crashes and how they have gone since 1995. They discuss percentage change in category e.g < 16 year olds, Male, Female, 65+ year olds. In this they identify maximum increase of 12.4% of the cases involve drivers under the age of 16. They also provide reports such as the cause for the accidents etc.

The author[4] further talks about how males cause greater number of accidents than females. Another interesting point that is made by the author is that most crashes occur on the weekdays between 3:00pm and 9:00pm and on weekends, between midnight and 2:59am.

5. Data Preparation

5.1 How clean is the data?

The data was not clean. There were many missing values for many different attributes considered.

5.2 Did you quantize the data into regions?

No, we did not quantize the data into regions as the data is already quantized into different boroughs.

5.3 Are there any issues with the data?

Yes, the data has issues in terms of many missing values for attributes.

5.4 Is the data from the 2 years comparable? Or are there any issues between 2017 and 2018?

Yes, it is comparable for the months of June and July that we specifically compared. Given the size of dataset, we would need more time to analyze all the months of data. No, there are no issues between 2017 and 2018.

6. Exploratory Data Analysis

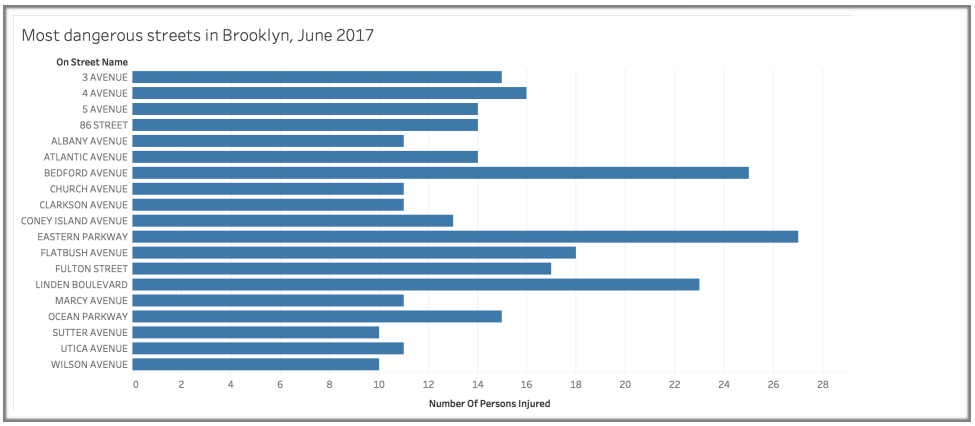
Exploratory Data Analysis is an approach in which various analytical techniques are used, or various graphs are developed from a data set to maximize the insight into the data. EDA can also be used to uncover underlying structure of the data, detect outliers etc.

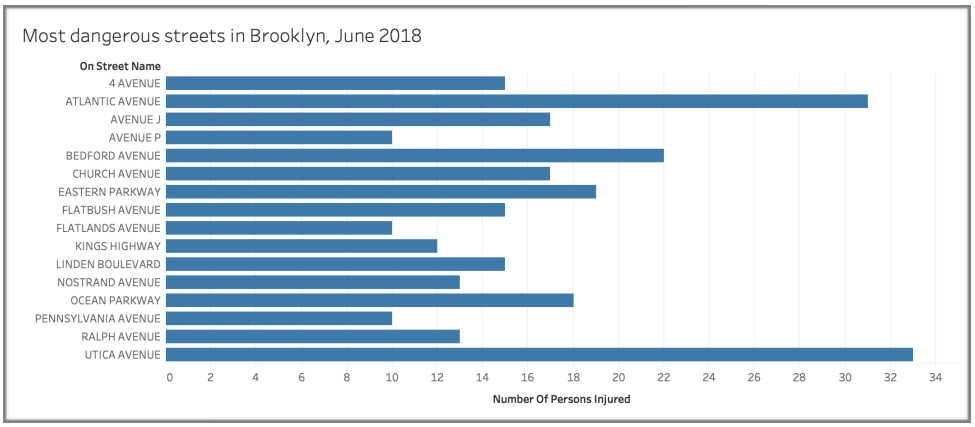
The tools we have used to perform EDA include, R, Rattle and Tableau. We started our analysis with Python, however, we gravitated towards R due to to ease of access and support for EDA. When we tried reading the dataset into Python using *pandas,* the reading was extremely slow and our laptops were heating up. Also, while attempting to access attribute within the dataset, PyCharm stopped responding. On the other hand, we did not experience the same difficulties with R.

Once we had decided to use R, we created a new column using *lubridate* package in R, to change the date format on the dataset from integer to a date-time format of YYYY-MM-DD. We had already analyzed Python for the individual values, and we weren't able to see a stark difference in the number of injuries and kills for specific data such as motorists, cyclists and pedestrians and hence, we decided to make use of the aggregated values. We made use of Tableau to get nice looking and interactive plots.

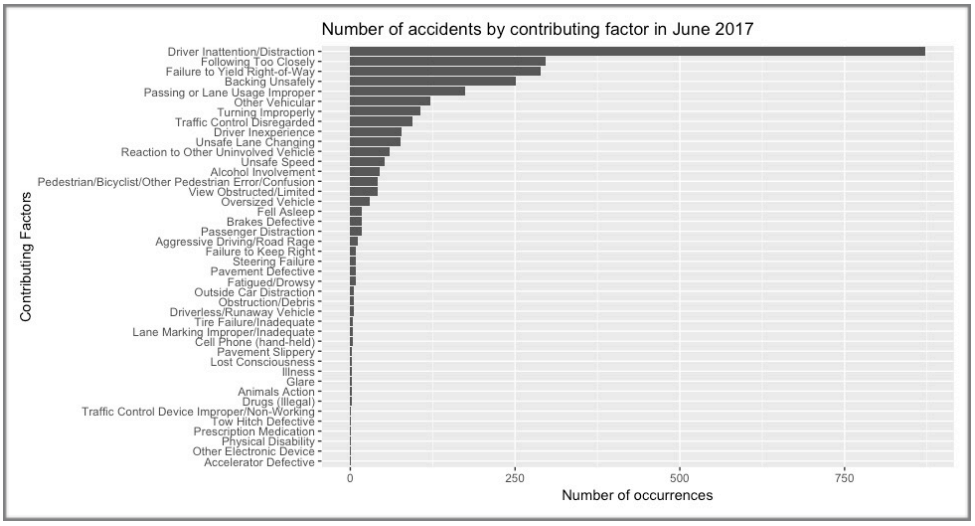
6.1 How was June of 2018 different than June of 2017?

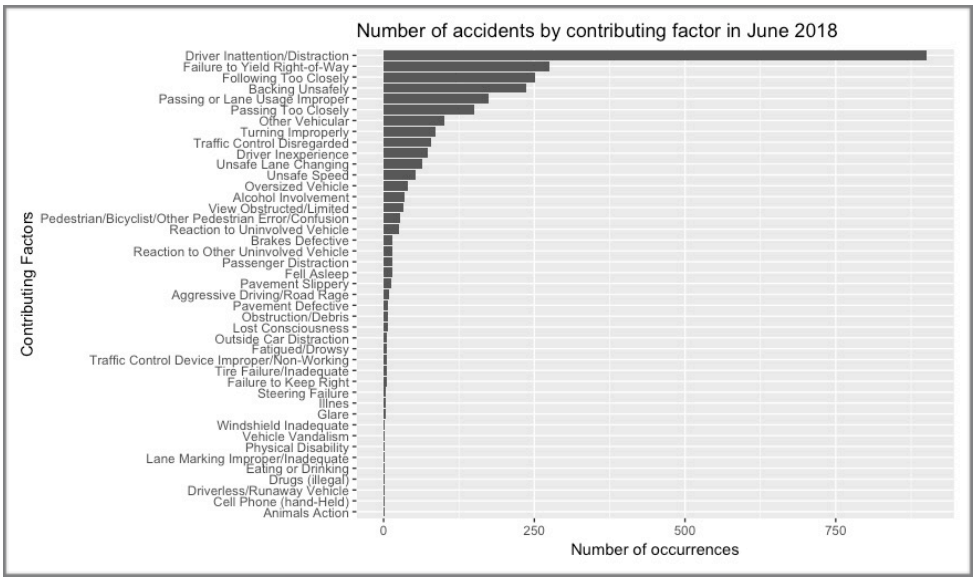
In June 2018, *'Utica Avenue'* was the most dangerous street in all of Brooklyn with maximum number of traffic based injuries happening, while in June 2017, 'Eastern Parkway' was the most dangerous street. Another interesting point to note is that, Bedford Avenue has a high number of injuries across both, June 2017 and June 2018.





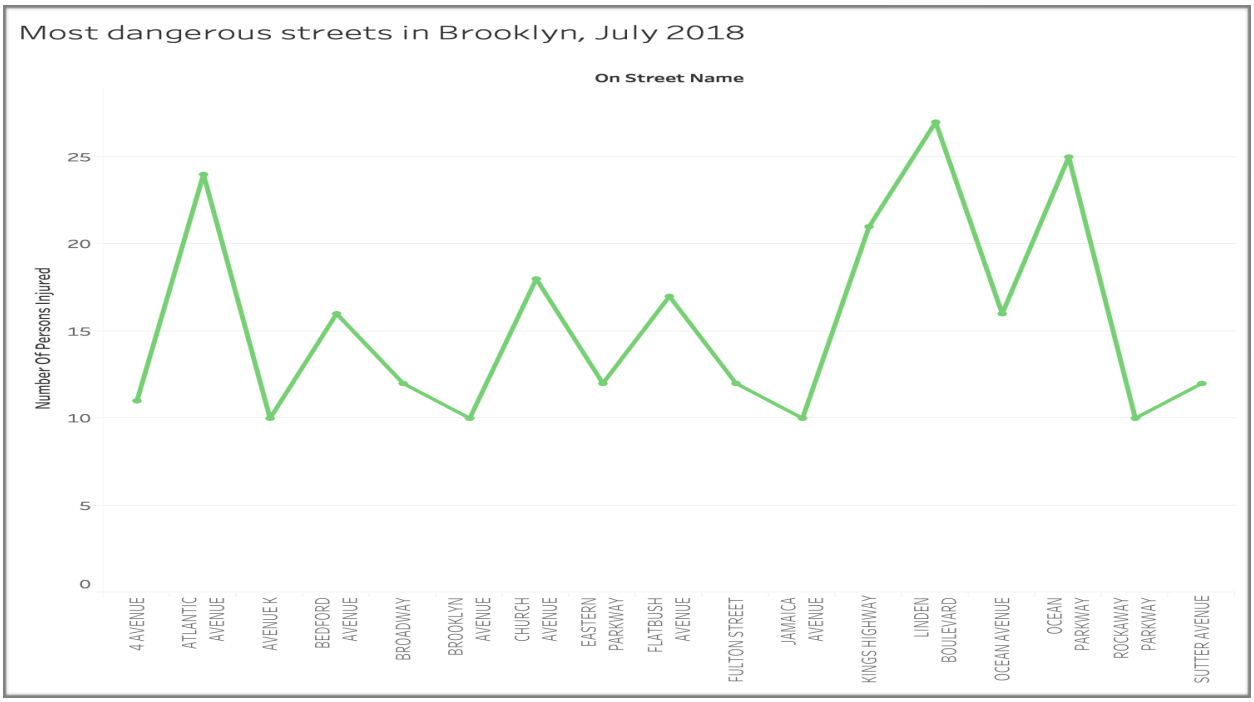
Further, we decided to analyze the probable causes for the accident, and not surprisingly, we see that the most common reason in both the cases was 'Driver Inattention/Distraction' being the reason for the accident and the top ten causes, in both the cases, remain the same.

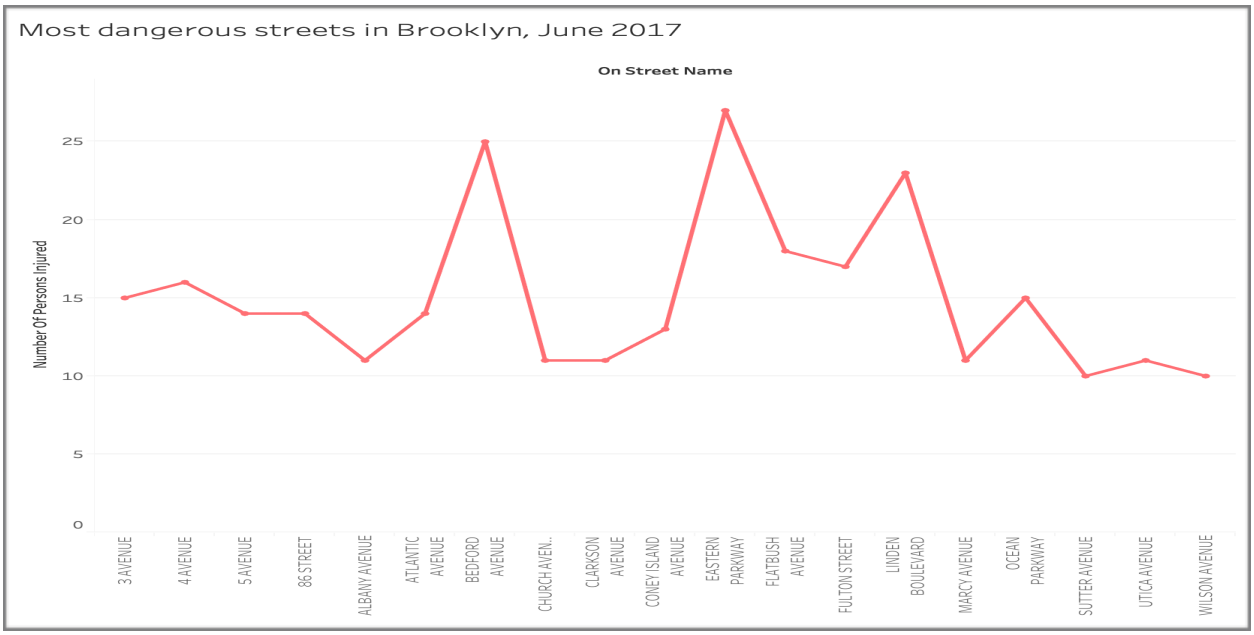




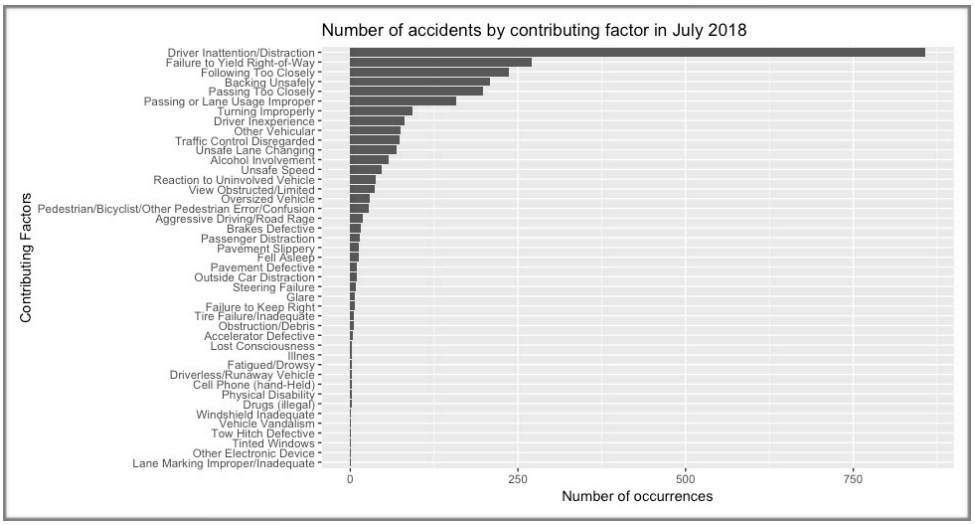
6.2 How was July of 2018 different than June of 2017?

In July 2018, *'Linden Blvd.'* was the most dangerous street in all of Brooklyn with maximum number of traffic based injuries happening, while in June 2017, 'Eastern Parkway' was the most dangerous street.



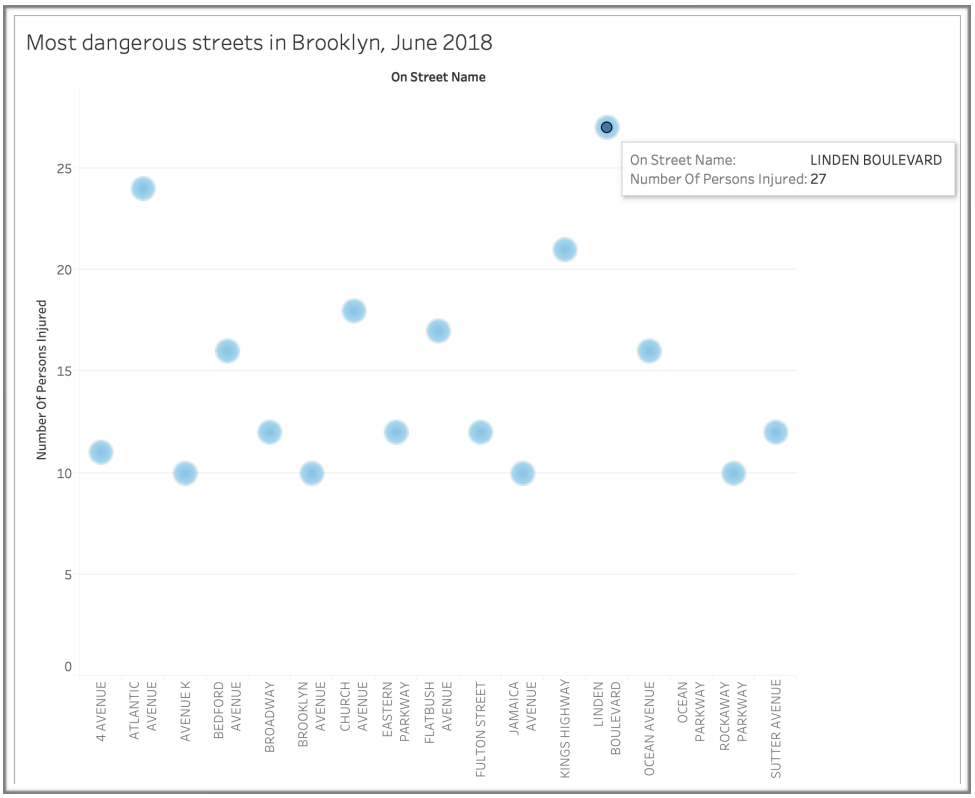


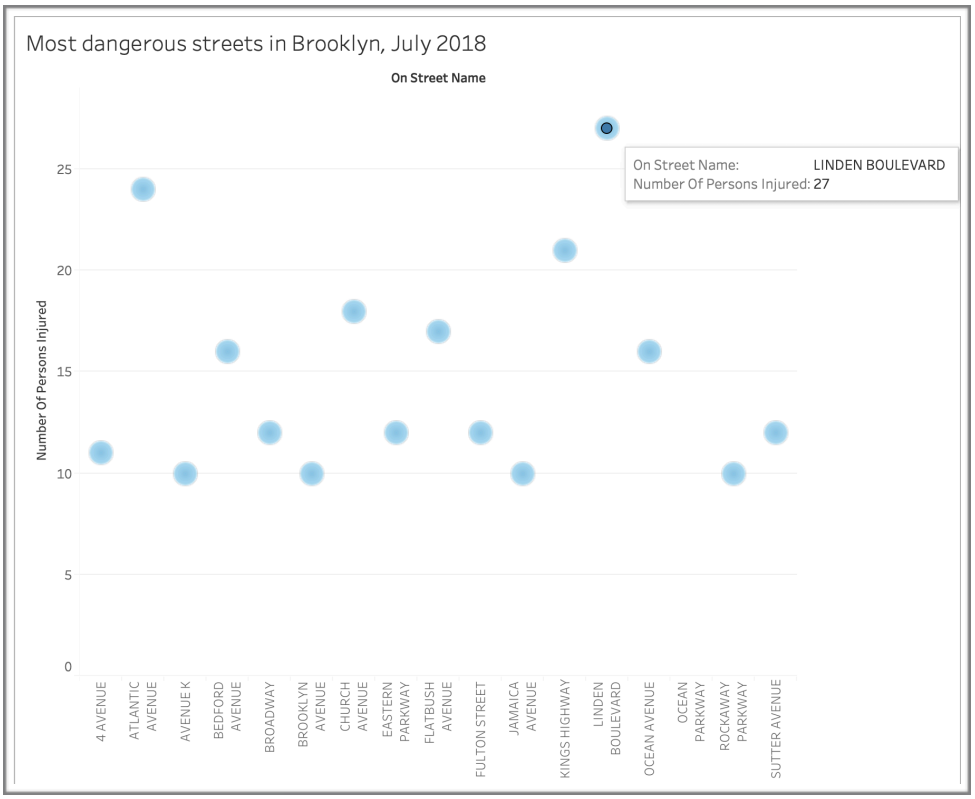
Further, we see a similar trend as in the previous question, *'Driver Inattention/Distraction'* is the largest cause for accidents in the 2 months in consideration.



6.3 How was July of 2018 different than June of 2018?

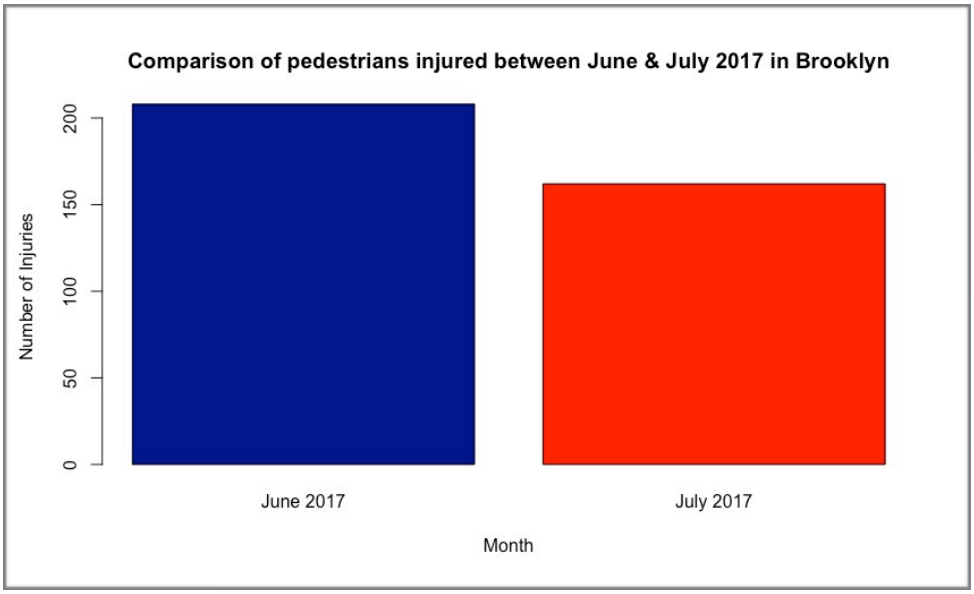
In July 2018, *'Linden Blvd.'* was the most dangerous street in all of Brooklyn with maximum number of traffic based injuries happening, while in June 2018 the most dangerous street remains the same.


Most dangerous streets in Brooklyn, June 2018
On Street Name
On Street Name: LINDEN BOULEVARD
Number Of Persons Injured: 27
Number Of Persons Injured
0
5
10
15
20
25
4 AVENUE
ATLANTIC AVENUE
AVENUE K
BEDFORD AVENUE
BROADWAY
BROOKLYN AVENUE
CHURCH AVENUE
EASTERN PARKWAY
FLATBUSH AVENUE
FULTON STREET
JAMAICA AVENUE
KINGS HIGHWAY
LINDEN BOULEVARD
OCEAN AVENUE
OCEAN PARKWAY
ROCKAWAY PARKWAY
SUTTER AVENUE


Most dangerous streets in Brooklyn, July 2018
On Street Name
On Street Name: LINDEN BOULEVARD
Number Of Persons Injured: 27
Number Of Persons Injured
0
5
10
15
20
25
4 AVENUE
ATLANTIC AVENUE
AVENUE K
BEDFORD AVENUE
BROADWAY
BROOKLYN AVENUE
CHURCH AVENUE
EASTERN PARKWAY
FLATBUSH AVENUE
FULTON STREET
JAMAICA AVENUE
KINGS HIGHWAY
LINDEN BOULEVARD
OCEAN AVENUE
OCEAN PARKWAY
ROCKAWAY PARKWAY
SUTTER AVENUE

6.4 How was July of 2017 different than June of 2017?

We noticed that the number of people injured and the number of people killed were different for June and July 2017. Hence, we decided to dig in further and found that the number of pedestrians injured differed across the months. We can hypothesize that the number of pedestrians injured are lesser in July 2017 than June 2017, because schools were closed in July 2017 and hence, there were lesser number of parents and children walking on the sidewalks and crosswalks.

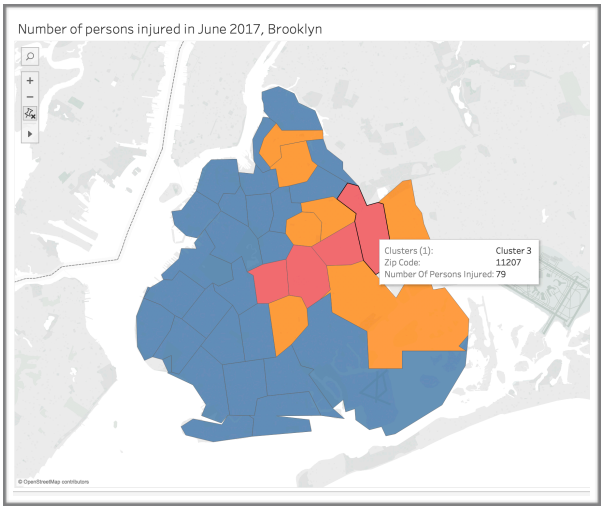


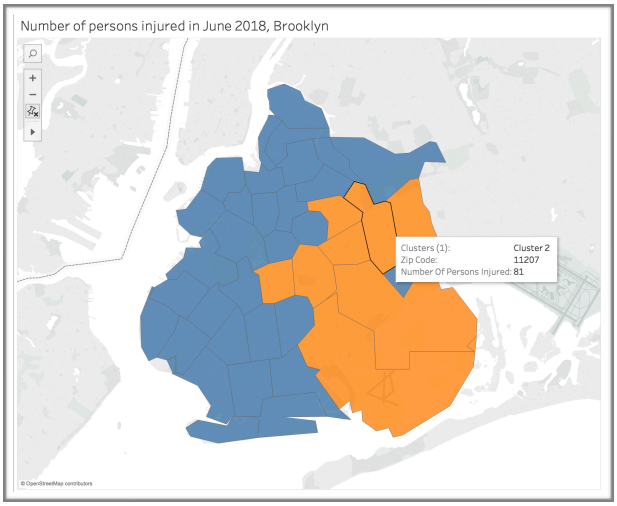
6.5 Final Viewpoints

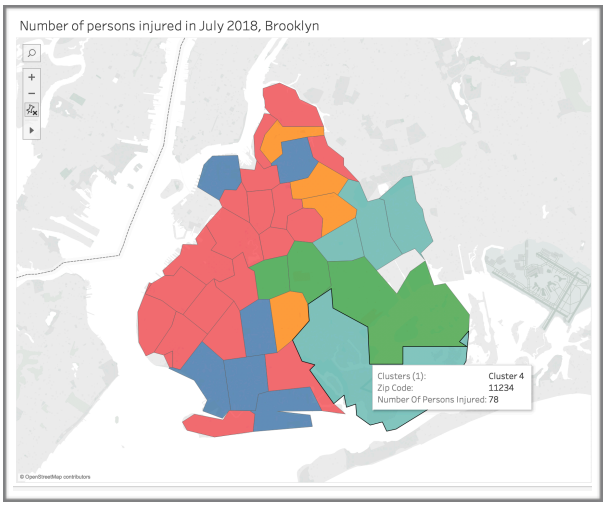
In our view, the most distinct cluster was June 2018, because it had the least number of clusters when compared to the other months in question. More often than not, the most number of people injured are in the zip code of 11207 possibly implying that it has an unsafe intersection or blind crosswalk etc. This is clearly visible from the heat map as depicted below.

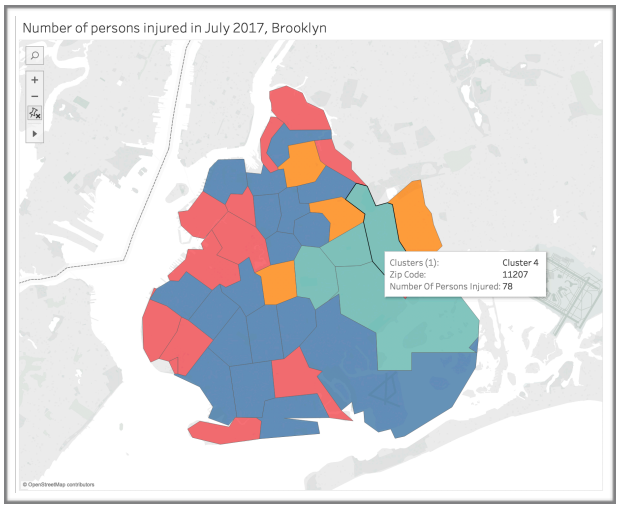
From the heat maps below, the average values of people injured is visible as well. In July 2017, at zip code 11207 there were 10 pedestrians who were injured and this is comparable to June 2018, where 11 pedestrians were injured. While the numbers are similar, pedestrians need to be careful in the zip code of 11207.

Also, the number of motorists killed/injured in 11207 is close to 0 in all the 4 heat maps, and hence, we can say that it is possibly an area with a lot of vehicular traffic and a lot of pedestrians walking alongside.









7. Conclusion

The biggest challenge we faced during the course of this project was that, the tools we had previously used and were useful did not scale well for this problem. Also, we had to learn new skills to make use of R and Tableau in order to visualize our data.

It was interesting to note that a powerful tool such as Python, can have drawbacks and it is very important to choose the right tools for any problem. Another interesting part of this project was that we learnt how to use Tableau, a Business Intelligence tool, to provide powerful interactive visualizations which can be used by people with possibly no knowledge of coding to improve the safety of New York City.

Something we would like to share about this project was that we enjoyed learning and working with a new tool which is commercially used at large companies as well. We wanted to, in fact, provide a lot more visualizations but were constrained due to a lack of time and limitation of the report size.

Performing Exploratory Data Analysis, we could answer several questions and finally conclude on many aspects about the safety of people living in Brooklyn, NYC. One could improve the model and try to explore the effect of using weather data along with the street name, and time information. Along with weather, traffic information could be used to make the model more robust and practical.

Something really interesting about data mining that we learnt through the course of this project, was we put a lot of concepts we learnt in class into use, such as

- Occam's razor - given a choice of models with the same accuracy, choose the simplest model. We had the option of picking Weka, R, Rattle, Tableau or Python. We chose Tableau due it being easy.
- No Free Lunch Theorem - One model does not fit all. We usually worked with Python through the course of this semester for the homeworks and projects. However, for the project we had to make use of other tools.
- "Try them all" - We tried getting visualizations in Python, R/Rattle and Weka before we settled for Tableau as it provided us with the best visualizations and helped us find the best insights.

These were some of our most important learnings.

References

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

[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/

[3] Traffic Safety Facts - Research Note, https://crashstats.nhtsa.dot.gov/API/Public/ViewPublication/812318/

[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

[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/

[6] Vision Zero, http://www1.nyc.gov/site/visionzero/ index.page

[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/