

CSE 590 - Data Science Mini Project 2

Kavana Anand - 109929741

Slip days - Used : 1, Remaining : 3

Introduction

The project is mainly intended to analyze and visualize the dataset provided by New York Police Department on Motor Vehicle Collisions and list the possible actions that can be taken by government to reduce the accidents. The data contains sequential records of collisions with associated details.

Approach Plan

Before diving into analyses and computation, understanding the nature of data is important. First approach would be to look into the data and state the obvious facts that you can derive from it. Second approach is to clean the dataset and process the data if required and convert to the format which will help to represent the data visually in graphs and charts. Since this a motor vehicle collisions data, next step would be to divide the entire NYC's collisions into sub locations maybe boroughs. This would help to see where most of the accidents are occurring in the city, has it been reduced over the period of time. Then move on to find the highly accident prone locations in the city and locations which are risky for pedestrians, cyclists and motorists. After this proceed to find the reasons for the accidents. Moving ahead, come up with the solutions to avoid these collisions based on the reasons obtained

Results and Discussions

Understanding the data

Dataset given consists of sequential records of collisions occurred from **01 JUL 2012 to 25 SEP 2015**. Each record will provide details such as date and location of the accident. Dataset gives multiple representations of location. Location can be identified either from latitude and longitude pair or from zipcode or from street names or more generally from borough given in the dataset. Further record specifies number of people injured and killed from a particular accident. More specifically it gives number of pedestrians, cyclist and motorist injured or killed in the accident. Lastly record mentions the reasons that caused accident and how it is related to the motor vehicles/ drivers involved in the accident.

Cleaning and Processing Data

It is necessary to clean the data and retain all the necessary information in the required format. From the visual inspection, it can be noticed that many records have their 'BOROUGH' field empty, however has latitude and longitude values. Google API is used to find the borough name for the given lat and long and this field is filled.

Further to calculate total people affected from the accident, extra columns are added, values for which is calculated from the existing values. For example, for each record -

'NumberOfCyclistAffected' = 'NumberOfCyclistInjured' + 'NumberOfCyclistKilled' and so on.

Further data processing is required to analyse the collisions occurred at particular location. This requires to group all the records with the same 'LOCATION' value and get the values of rest of the field appropriately.

Dividing the dataset among boroughs

To understand the problem and come up with solution, it is useful to divide the problem. In this case, dividing based on the BOROUGH will be useful to analyse where and why most of the accidents are occurring.

Following pie charts represents the distribution among the 5 boroughs of the New York City.

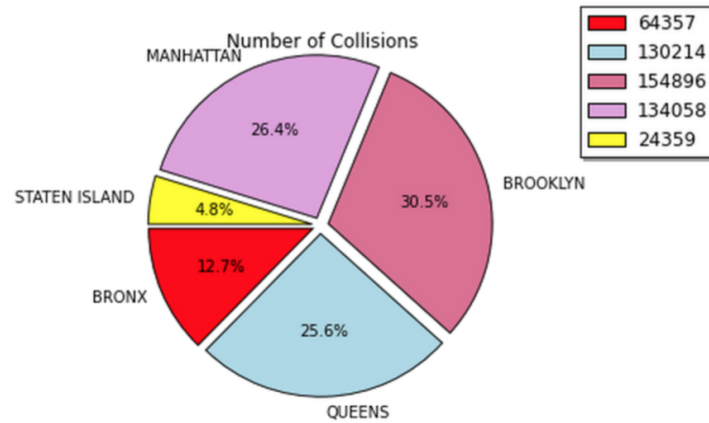


Figure 1 : Distribution of total number of collisions occurred from Summer 2012 to Summer 2015

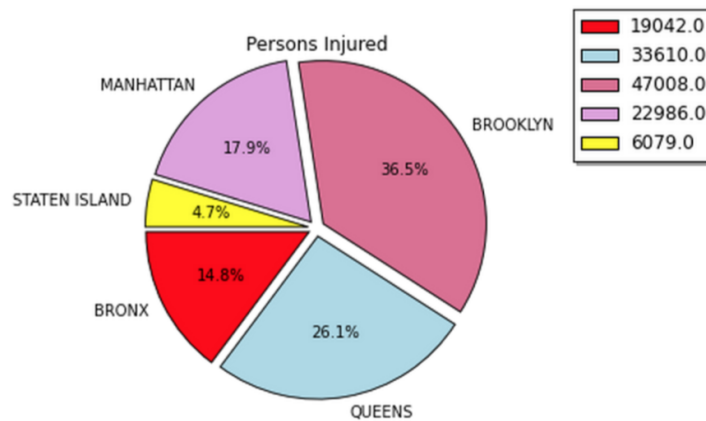


Figure 2 : Distribution of number of people injured from collisions occurred from Summer 2012 to Summer 2015

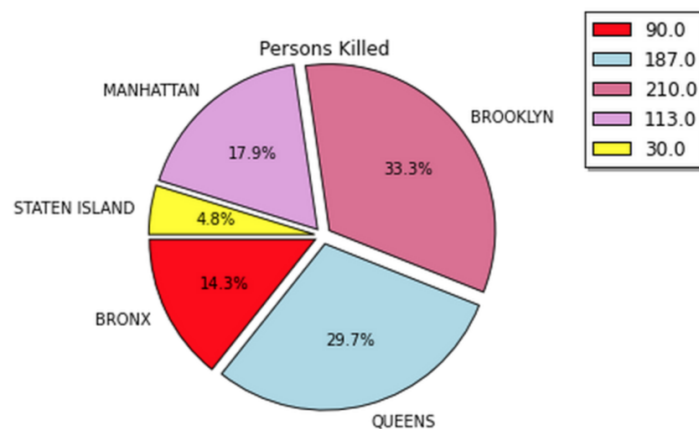


Figure 3 : Distribution of number of people killed from collisions occurred from Summer 2012 to Summer 2015

From above visualization we can see that Brooklyn has highest rate of accidents and death rate with Staten Island having lowest.

Analysing the collision trend over the period

In the motor vehicle analyses, its necessary to check if possible precautions have already been taken care to reduce/avoid the accidents at certain locations. This can be observed from following figure.

Following figure shows the number of collision occurred over the given period. For more specific explanation and understanding number of collisions has been divided into the boroughs.

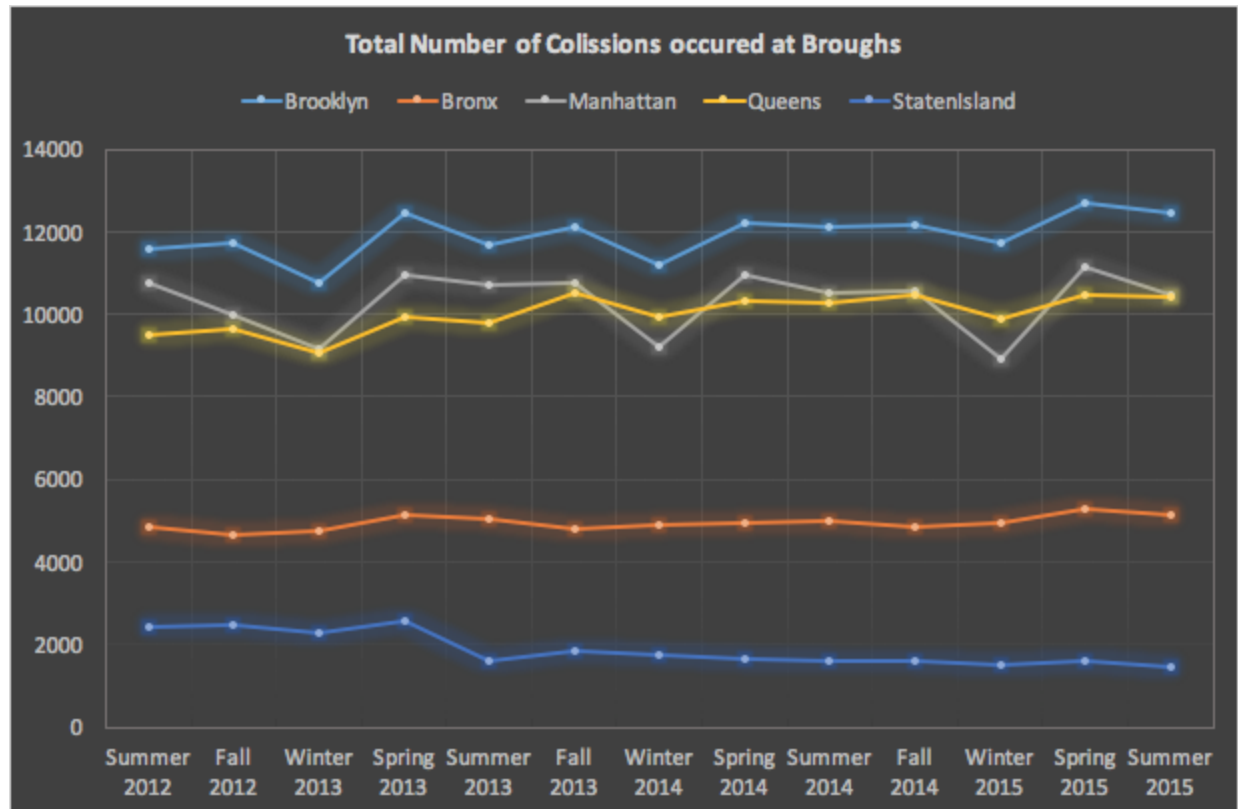


Figure 4: Total Number of Collisions occurred at different boroughs

From the figure we can see that, total number of collisions has been reduced only in Staten Island. However, it can be seen that in other boroughs the number is almost the same. This means that necessary precautions have not been taken to reduce/avoid the accidents. So to come up with the possible solutions or precautions we need to analyze deeply to find out the reasons behind the accidents. This will be discussed in next section.

One interesting finding from the above figure is that the number of collisions is less during winter season. This could probably be because of weather. Since it snows in New York city during this time probably not many motor vehicles would be running or drivers could be extra cautious because of slippery roads and pavements.

Finding out the reasons

Before finding out reasons for accidents occurring at all locations it is necessary to find out the locations where most number of accidents are occurring or most number of people are affected either injured or killed. From further analysis and processing of data it is seen that the maximum number of accidents are occurring at following locations.

STREETS/ INTERSECTION	BOROUGH	NO OF COLLISIONS
FLATBUSH AVENUE EXTENSION, TILLARY STREET	BROOKLYN	512
EAST 59 STREET, 2 AVENUE, SECOND AVENUE, QUEENSBORO BRIDGE APPROACH	MANHATTAN	423
8 AVENUE, WEST 42 STREET	MANHATTAN	408
PENNSYLVANIA AVENUE, ATLANTIC AVENUE	BROOKLYN	379
PENNSYLVANIA AVENUE, LINDEN BOULEVARD	BROOKLYN	374
WEST 34 STREET, 7 AVENUE	MANHATTAN	320
BOWERY, DELANCEY STREET, KENMARE STREET, DELANCY STREET	MANHATTAN	318
EAST 57 STREET, 3 AVENUE, THIRD AVENUE	MANHATTAN	314
WOODHAVEN BOULEVARD, QUEENS BOULEVARD	QUEENS	312
9 AVENUE, WEST 42 STREET	MANHATTAN	308

Analyzing the records that contains above 10 locations, it can be seen that the top reasons causing accidents at this locations are as represented in following chart -

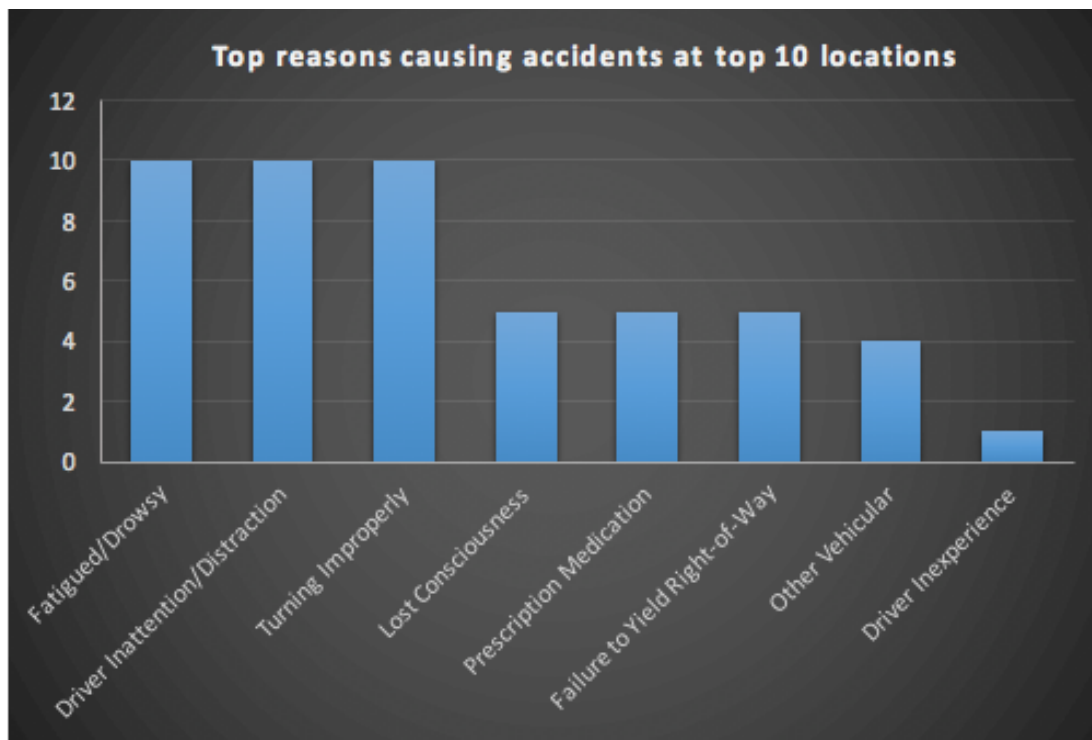


Figure 5: Top reasons causing accidents

On similar lines it is also necessary to find out the risky locations for pedestrians, cyclists and motorists and the reasons for the same. Similar to the previous analyses, top 10 locations where pedestrians, cyclists and motorists have been injured or killed are taken into consideration.

Analysing the records at these locations, following three charts are used to represent the top reasons for injuring or killing pedestrians, cyclists and motorists respectively.

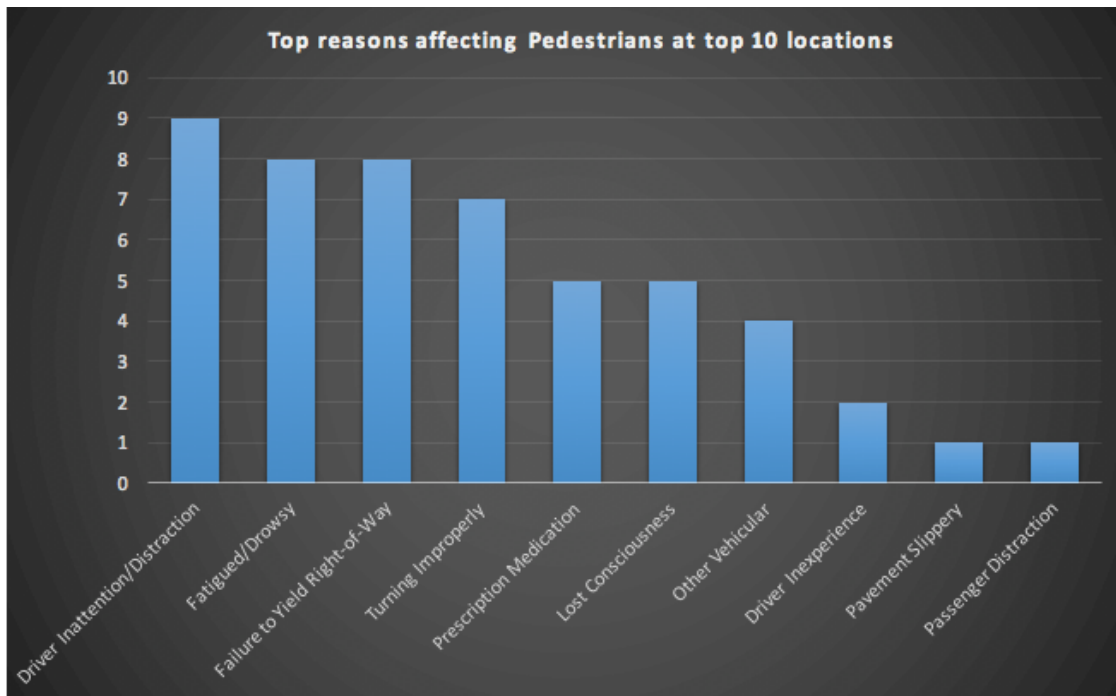


Figure 6: Top reasons harming Pedestrians

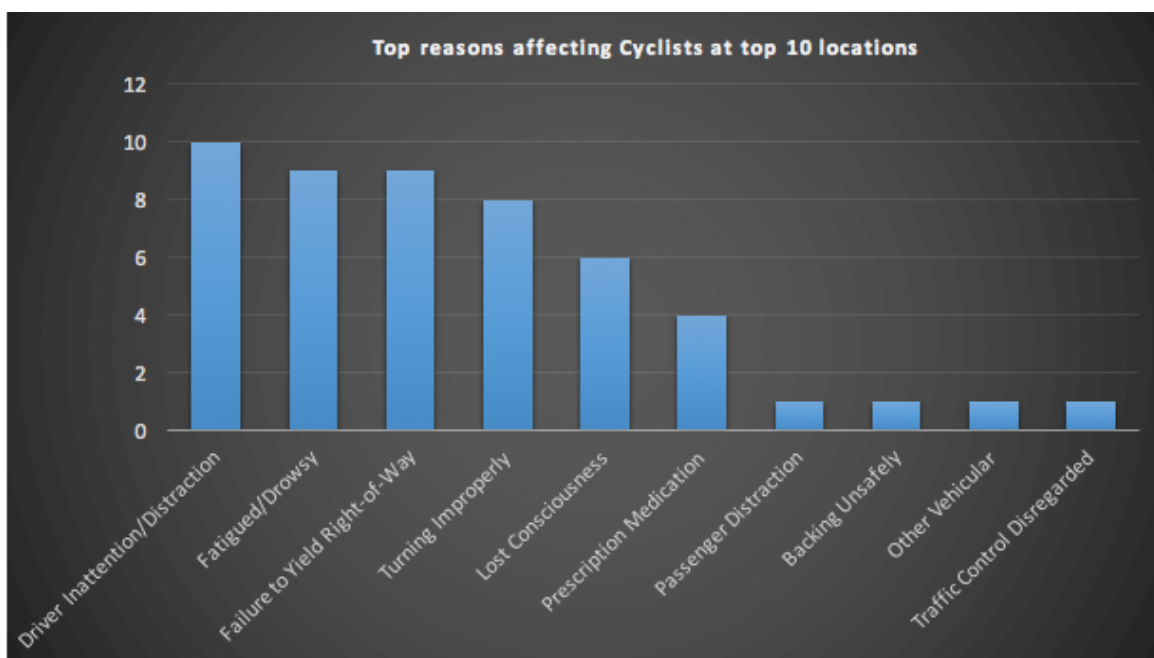


Figure 7: Top reasons harming Cyclists

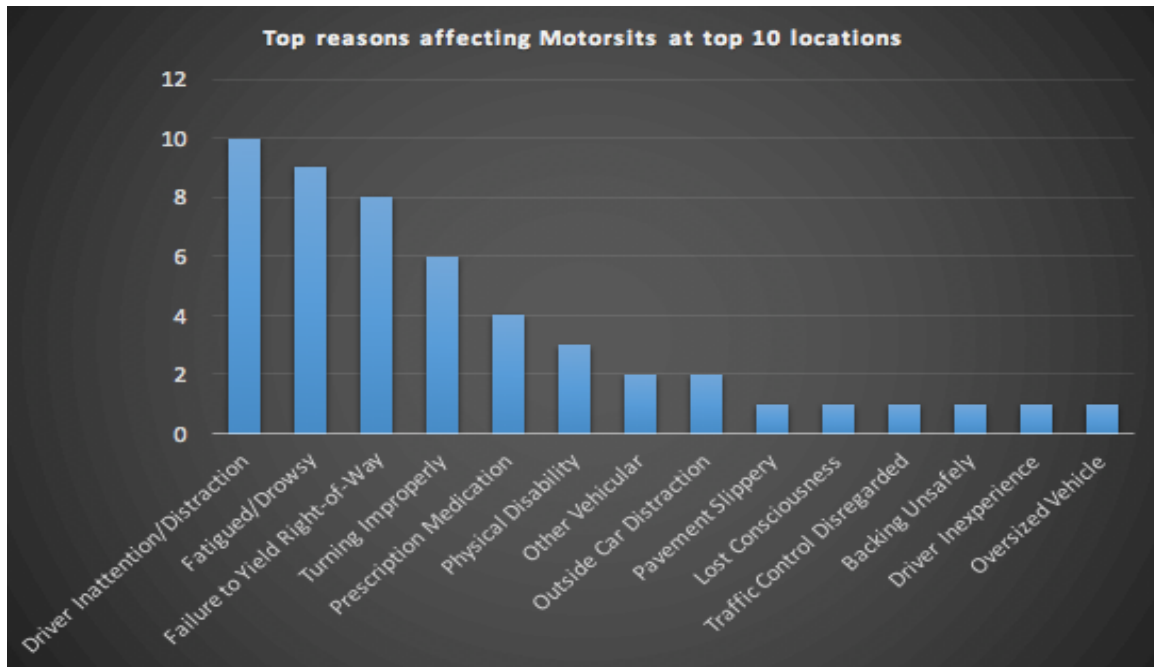


Figure 8: Top reasons harming Motorists

From above analysis, it can be seen that top reasons are -

- Driver Inattention/Distracted
- Driver Inexperience
- Passenger Distraction
- Failure to Yield Right-of-Way
- Turning Improperly
- Traffic control disregarded
- Backing unsafely
- Oversized vehicle
- Fatigued/Drowsy
- Lost Consciousness
- Physical Disability
- Prescription Medication
- Pavement Slippery
- Other Vehicular
- Outside Car Distraction

Precautions/Solutions to reduce the collisions and make NYC a better place

This project aims to provide the solutions to improve NYC. Following are possible precautions that can be taken to avoid above mentioned reasons.

- **Driver Inattention/Distracted, Driver Inexperience and Passenger Distraction**
These are the causes mostly in hands of the drivers and passengers of the motor vehicles. Every driver and passenger in the car must be aware of to be very attentive while driving. Passengers should not talk to drivers and distract them especially when driving through busy intersection.

- **Failure to Yield Right-of-Way, Turning Improperly, Traffic control disregarded, Backing unsafely and Oversized vehicle**

Every driver must be issued license only after all kinds of turns, signals have been tested. Driver must be aware of Blind spots. Driver must be given limited number of warnings combined of total violation after which his license must be revoked. This way he will be alert to follow every rule, big or small and would not dare to ignore any rules. He would stop for pedestrians to cross, he would turn properly and not violate signals. He would be extra cautious with blind spots and make sure no vehicle or people are harmed.

- **Fatigued/Drowsy, Lost Consciousness, Physical Disability and Prescription Medication**

Drivers can be under the influence of medication or illegal drugs. At the highly risky locations, check for the drugs and alcohol must be increased and driver must be fined by the government. Awareness among drivers must be created that drivers who are feeling tired or dizzy, must sit for a while or ask for help from people around and get back to driving only after he feels better.

- **Pavement Slippery, Other Vehicular and Outside Car Distraction**

Every driver before getting into the vehicle, must check the traffic before getting into the road and also status of the road. This will help to avoid unknowingly hitting the other vehicle already on the way. And also driver while crossing small intersections must slow down and then cross when the road is clear and no vehicle is joining the road.

Conclusion

This project concludes that the given dataset was visualized and analysed by plotting various charts. This project proposes analyses of sublocations for collisions to identify the locations prone to accidents the most. Project also specifies that reasons must be well know to come up with the solution. Finally and most importantly, project also provide possible solutions for the top reasons found during the analyses and by putting these solutions into effect collisions can be reduced and NYC can be a better place.