CS 432/536: Data Mining

Course Project - Group 6

Motor Vehicle Accidents in New York City

Group Member	**Roll Number**
Hareem Raza	22100277
Mehnoor Maqsood	22100191
Mohid Yousaf	22100201
Mursal Junaid Rehman	22100158

Table of Contents

Introduction

In a developed city like New York, the competitive and fast-paced environment offers its own consequences. One of them is increased vehicular traffic and consequently, more accidents and crashes. While some accidents are inevitable, road safety and policies backed up by analysis can lead to apt suggestions about how to curb accidents or at least avoid them. This will be the core motivation of this project. In this report, we set out to ascertain the prevalence of vehicular accidents in NYC – their causes, effects and trends throughout the city. There will also be some recommendations and key insights for variables to consider that could shape future analysis.

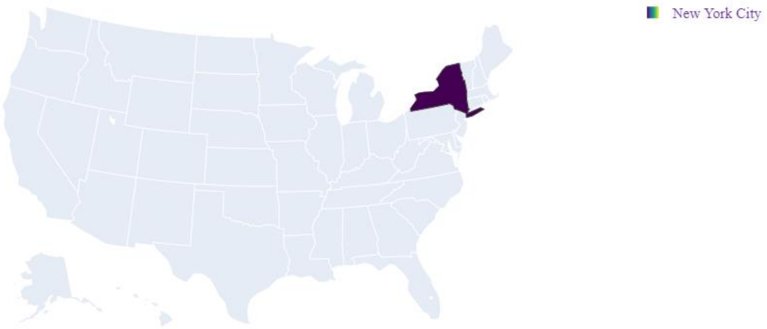


Figure 1: World Map and New York City (Plotted in Python)

Goals

Our primary goal in this project is to use NYC Accident data to not only to observe patterns and trends in the collisions but also to derive meaningful implications and suggestions about ways to reduce the count and intensity of collisions. Subsequent effective policy making can be backed up by our analysis. We have divided our analysis into some sub-questions which include the following:

- What is the area-wise trend of accidents in NYC? What inferences can we draw from accident prone boroughs?
- What are the trends of accidents in NYC over different periods of time? Can we make any predictions about the future trends?
- How did the severity of accidents vary throughout NYC?
- What are the most common causes of vehicular accidents in New York?
- Are there any correlations between different attributes relevant to New York accidents and how do these correlations affect our analysis?

We will be answering them one by one, building one upon the other, and finally connecting the dots to reach a conclusion. We will use the results to pose suggestions and recommendations for NYC administration to alleviate the accident rate. But before that, we will start by having a quick look at the data we have.

Dataset

Our dataset, titled "Motor-Vehicle Accidents in New York City" contains vehicle collisions and crashes data. It is based on a single CSV file where each row represents a vehicular collision incident and each column contains respective information about it.

After importing relevant libraries that we needed for our analysis, we loaded our data into a data frame called data. To get an idea about the amount of data and diversity of features we are dealing with, we will first check the shape of data i.e., the number of rows and columns it which turned out to be `(1750704, 29)`.

The data consists of more than 1.7 million records of accidents throughout New York and each accident is described by 29 features. It is important to enlist the features here:

Features in the dataset
'CRASH DATE', 'CRASH 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', 'COLLISION_ID', 'VEHICLE TYPE CODE 1', 'VEHICLE TYPE CODE 2', 'VEHICLE TYPE CODE 3', 'VEHICLE TYPE CODE 4', 'VEHICLE TYPE CODE 5'

Although the feature names were quite self-explanatory, we made another CSV file (*column_data.csv*) ourselves which had a description of each column. This will be useful to understand the features better and can be referred to throughout the analysis. Let's have a look at the file and see what each feature represents:

(Please turn over to the next page)

Column Name	Description
CRASH DATE	Occurrence date of collision
CRASH TIME	Occurrence time of collision
BOROUGH	NYC Borough where collision occurred
ZIP CODE	Postal code of incident occurrence
LATITUDE	Latitude coordinate for Global Coordinate System
LONGITUDE	Longitude coordinate for Global Coordinate System
ON STREET NAME	Street on which the collision occurred
CROSS STREET NAME	Nearest cross street to the collision
OFF STREET NAME	Street address if known
NUMBER OF PERSONS INJURED	Number of persons injured
NUMBER OF PERSONS KILLED	Number of persons killed
NUMBER OF PEDESTRIANS INJURED	Number of pedestrians killed
NUMBER OF PEDESTRIANS KILLED	Number of pedestrians injured
NUMBER OF CYCLIST INJURED	Number of cyclists injured
NUMBER OF CYCLIST KILLED	Number of cyclists killed
NUMBER OF MOTORIST INJURED	Number of motorists injured
NUMBER OF MOTORIST KILLED	Number of motorists killed
CONTRIBUTING FACTOR VEHICLE (1-5)	Factors contributing to the collision for designated vehicle
COLLISION ID	Unique record code generated by system
VEHICLE TYPE CODE (1-5)	Type of vehicle based on the selected vehicle category

Table 1: Names and descriptions of features of NYC dataset

We then peaked into the actual values the records hold. For this, we took a random *sample* of five accidents from the entire data. All this gave us a basic idea about our data and the records of accidents it contains. Before making a subsequent analysis of the dataset, we need to pre-process it in order to pass the best possible data for exploratory data analysis.

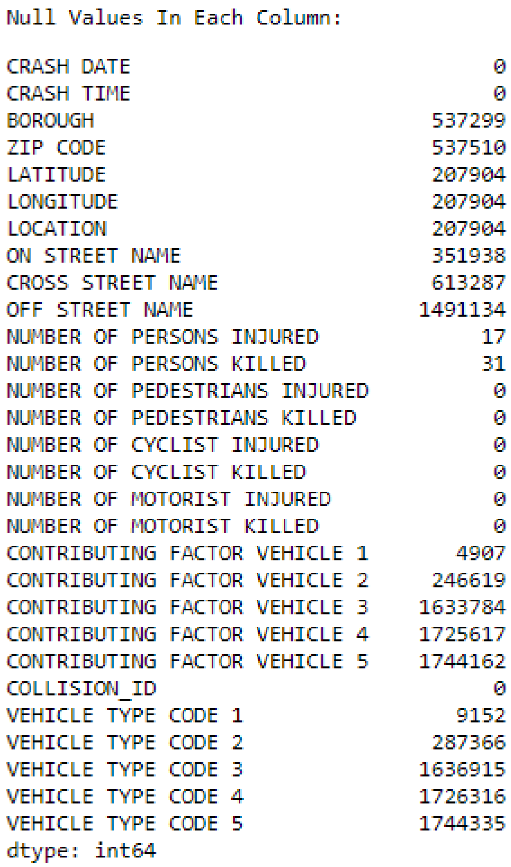
Data Cleaning

Ever heard of the famous phrase “garbage in, garbage out?”

Before we can derive meaningful insights and inferences from our data, we need to validate its correctness and ensure that it is in a standardized and useable format. Thus, we will divide the cleaning process into various stages before moving to the analysis. So, let’s charge into it. We will take things one at a time. Each pre-processing part is divided into a different section as shown below.

Finding Null Values

Let’s check if there are any missing values in the data that we need to account for.



```
Null Values In Each Column:

CRASH DATE                             0
CRASH TIME                             0
BOROUGH                           537299
ZIP CODE                          537510
LATITUDE                          207904
LONGITUDE                         207904
LOCATION                          207904
ON STREET NAME                    351938
CROSS STREET NAME                 613287
OFF STREET NAME                  1491134
NUMBER OF PERSONS INJURED             17
NUMBER OF PERSONS KILLED              31
NUMBER OF PEDESTRIANS INJURED          0
NUMBER OF PEDESTRIANS KILLED           0
NUMBER OF CYCLIST INJURED              0
NUMBER OF CYCLIST KILLED               0
NUMBER OF MOTORIST INJURED             0
NUMBER OF MOTORIST KILLED              0
CONTRIBUTING FACTOR VEHICLE 1       4907
CONTRIBUTING FACTOR VEHICLE 2     246619
CONTRIBUTING FACTOR VEHICLE 3    1633784
CONTRIBUTING FACTOR VEHICLE 4    1725617
CONTRIBUTING FACTOR VEHICLE 5    1744162
COLLISION_ID                           0
VEHICLE TYPE CODE 1                 9152
VEHICLE TYPE CODE 2               287366
VEHICLE TYPE CODE 3              1636915
VEHICLE TYPE CODE 4              1726316
VEHICLE TYPE CODE 5              1744335
dtype: int64
```

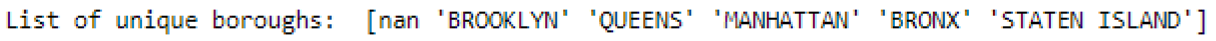
There are numerous null values in our data. Let's cater to them it feature by feature. Our goal will be to fill as many null values as possible and trying to minimize the introduction of any biases or discrepancies with in the data.

Null Values in Latitude and Longitude

Since the data is based on location wise analysis of the accidents, it is important to know at least one or more details about the location of the accident. In case there are records where borough, longitude, latitude is NAN together, we know nothing about the location and we have to drop the records. We also drop the columns where latitude or longitude is not known at all. (This is because we can find unknown town from latitudes and longitudes, but we cannot find the exact latitude and longitude from a given town). To avoid misinformation, we will remove all the records with null values here. Fortunately, such instances were quite low as compared to the total number of records we had.

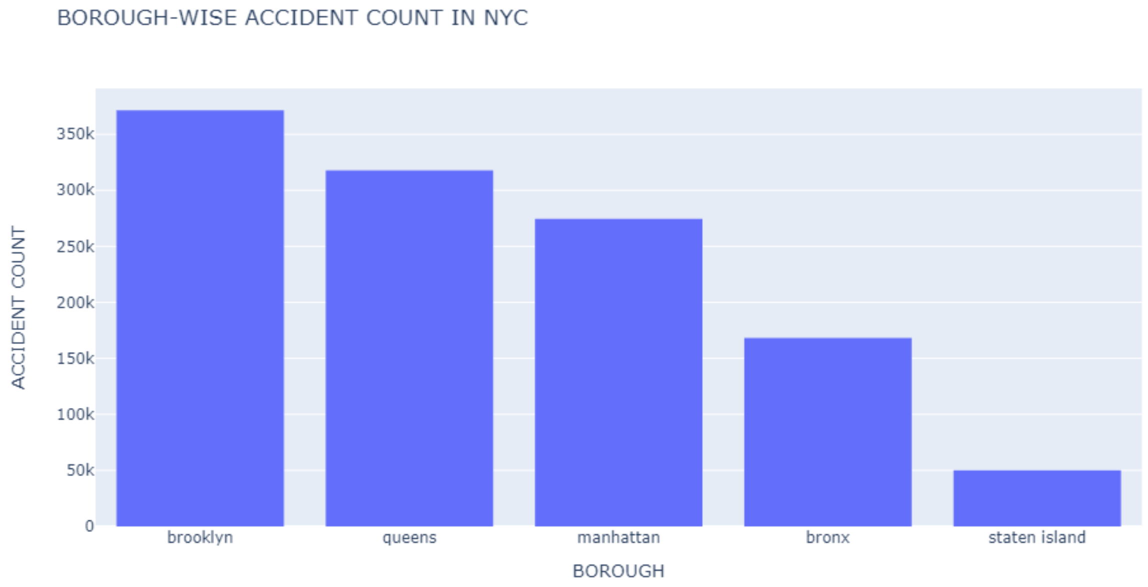
Null Values in Borough

In some of the records in ‘BOROUGH’ column, we have latitudes and longitudes given but not the borough information. Let’s try to find the missing boroughs from the latitude and longitudes. For that, let’s have a look at the list of boroughs in the data:



```
List of unique boroughs:  [nan 'BROOKLYN' 'QUEENS' 'MANHATTAN' 'BRONX' 'STATEN ISLAND']
```

Before we change the Nan values, let’s have a look of the accident count by borough. This will enable us to check whether the latitude and longitude help us predict the town or should we look for another approach.



Initially, we tried to find the boroughs by manually using latitudes and longitudes, but it resulted in unpredictable results and contrary to the original trend i.e., boroughs which had higher accident count first moved to lower places resulting in potentially biased side. Therefore, we tried another approach called *Reverse Geocoding*, our data size is very large, so we have applied various approaches to cater the Null values in Borough by using *myGeocoder* as our user agent and time out is set as 10min, and we could see in the output that null values were converted into respective boroughs by the assistance of point pair of longitudes and latitudes in reverse geocoding.

Null Values in Street Name

We have three street names given - ON STREET NAME, OFF STREET NAME and CROSS STREET NAME. Since some of them are null and some are not, we decide to keep a single feature ON STREET NAME and merge the rest in order of decreasing priority. This forms a new column STREET NAME which gives us an idea about the street (without further subdivisions in categories).

In case there is absolutely no information about street names, we fill it with unknown/unspecified. We chose not to drop the rows solely based on streets, because we do have valuable information about the latitudes, longitudes, and boroughs.

Null Values in Persons Injured and Persons Killed

The NULL values in these columns are negligible but filling them with mode, mean, median or any value means introducing potential discrepancy in a sensitive content like 'death' and 'injuries'. Thus, these columns are supposed to be accurate for our interpretation, so what we are doing is adding respective injuries to find the total injuries and similarly for deaths i.e., we used the following formula:

```
Persons Injured = Pedestrians Injured + Cyclists Injured + Motorists Injured

Persons Killed = Pedestrians Killed + Cyclists Killed + Motorists Killed
```

Null Values in Vehicle 3,4,5 and Contributing Factors

Vehicle Type 1 and 2 (along with their contributing factors) seemed more usable with less Nan and faulty values i.e., in most cases a collision was between two vehicles. Therefore, considering their low usability for our analysis, we will be dropping the features Vehicle 3,4,5, and their contributing factors.

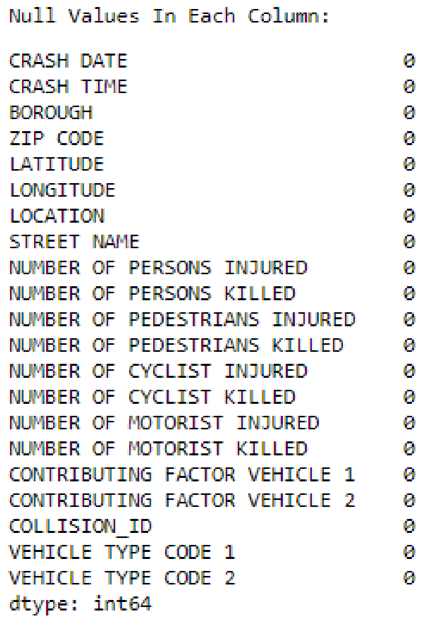
Null Values in Vehicle 1,2 Contributing Factors

Despite being useable overall, there were still a few Null values in Vehicle 1 and 2 and their respective contributing factors. We will not drop them. Instead, we fill null values of CONTRIBUTING FACTOR VEHICLE 1, CONTRIBUTING FACTOR VEHICLE 2 with 'unspecified' which would indicate that the contributing factor for the respective collision went unrecorded or is unknown in general.

Null Values in Zip Code

In cases where zip code was missing, we used a different strategy and filled the Nan values in the missing zip codes with the mode of that particular borough’s zip code. No records were dropped and manually analysis and comparisons with World Zip Codes database revealed that the zip codes were more or less aptly filled.

After numerous attempts and trying various techniques, we successfully catered for Null values in each feature. Let’s check the final count of Null values (after cleaning).



```
Null Values In Each Column:

CRASH DATE                       0
CRASH TIME                       0
BOROUGH                          0
ZIP CODE                         0
LATITUDE                         0
LONGITUDE                        0
LOCATION                         0
STREET NAME                      0
NUMBER OF PERSONS INJURED        0
NUMBER OF PERSONS KILLED         0
NUMBER OF PEDESTRIANS INJURED    0
NUMBER OF PEDESTRIANS KILLED     0
NUMBER OF CYCLIST INJURED        0
NUMBER OF CYCLIST KILLED         0
NUMBER OF MOTORIST INJURED       0
NUMBER OF MOTORIST KILLED        0
CONTRIBUTING FACTOR VEHICLE 1    0
CONTRIBUTING FACTOR VEHICLE 2    0
COLLISION_ID                     0
VEHICLE TYPE CODE 1              0
VEHICLE TYPE CODE 2              0
dtype: int64
```

A total of 1574180 cleaned and apt records out of 1750704 are now available for our analysis.

Redundant Features

When we were analyzing the features one by one, we saw that there was one redundant feature. The location feature is an ordered pair of latitude and longitude. Since this information is redundant, we can drop it and use the other columns when needed.

Standardizing Data

As the last part of data pre-processing, we decided to standardize the data a bit for our ease. This included the following:

Setting Index

In the column descriptions that we saw earlier, we noticed that *collision_id* is used as a primary key for crash database. Since it is unique, we decided to use it as the index for our data as well. We will set our index to collision id.

Lower Case Values

Although this does not impact our analysis in a significant way, having the columns in a similar format makes the data easily interpretable and usable. Therefore, we decided to convert all the column strings to lowercase.

Columns for Date

The date is given in DD/MM/YY format. Separating the time periods into separate features can help in future time-based analysis.

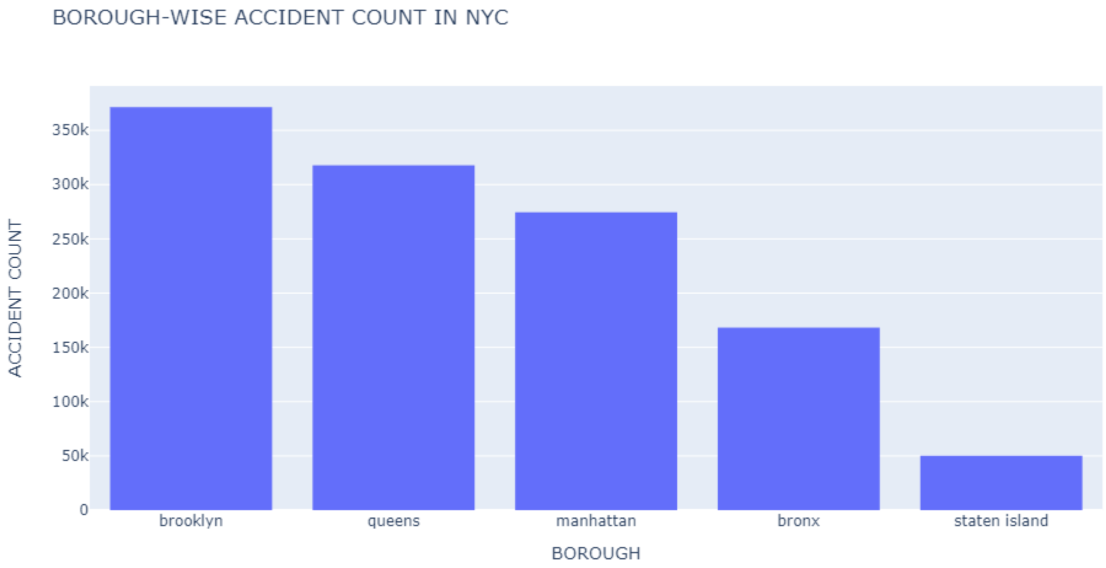
We then had a look at the final cleaned data. It was non-redundant, non-null and usable! This marked the end of data preprocessing stage after which we carried out detailed analysis on the cleaned data.

Exploratory Data Analysis

It's finally time to dig deep into the data, summarize and analyze it and try to find answers to the questions we posed earlier. We have divided our EDA into 5 parts, each part answering a single question which will then lead on to the next one. In the end, we will combine them all to look at the bigger picture of our findings.

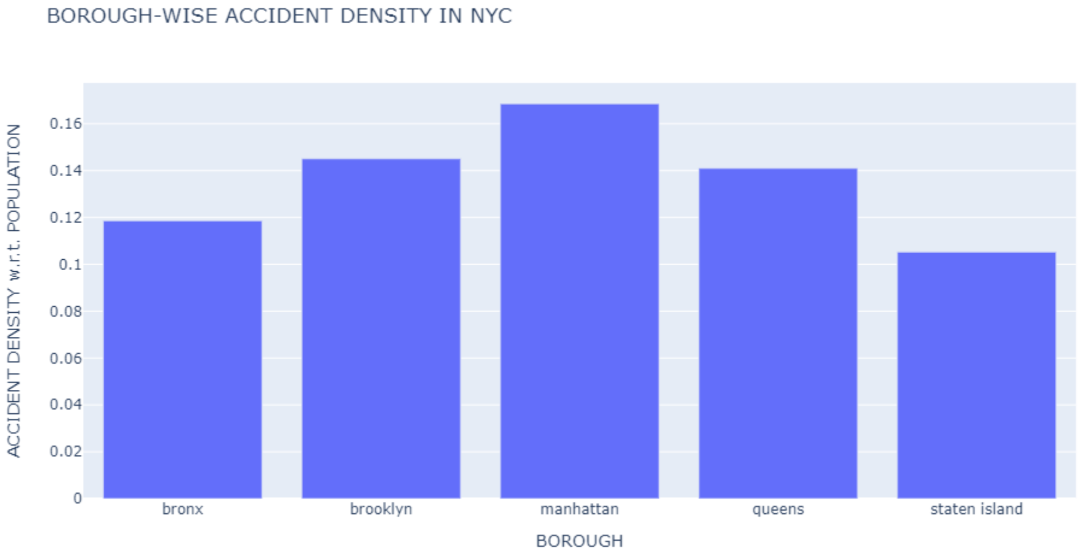
1. What is the area-wise trend of accidents in New York and what inferences can we draw from accident prone boroughs?

We grouped the accidents according to the Boroughs in which they took place to find out which Borough has the highest number of accidents happened over the years 2012 to 2021:



From this plot, it can be seen that Brooklyn has had the highest number of accidents happened over time followed by Queens, Manhattan, Bronx, and Staten Island. But this plot does not give any significant information regarding the most accident-prone Borough as it does not take the areas or populations of Boroughs into account. It is plausible that Brooklyn's area or population is significantly greater than the other boroughs therefore its accident count is also high.

In order to find out the most accident-prone borough, we need to also take the populations of these Boroughs into account.



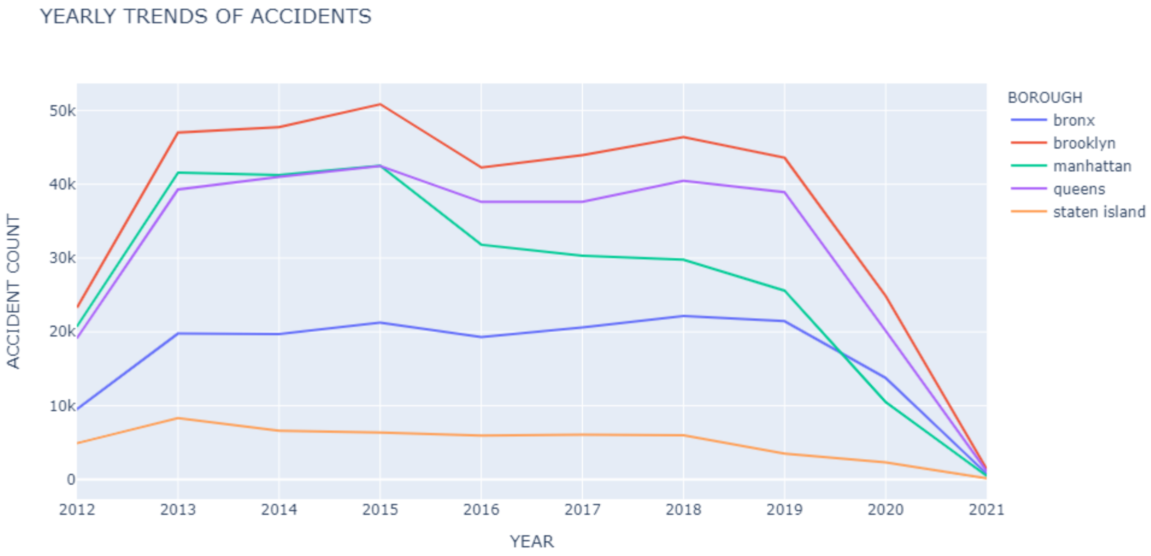
For this plot, we referred to the United States Census Bureau's census for the year 2019. We have assumed that the current population sizes of the five boroughs follow a similar trend as that of 2019. In 2019, Brooklyn had the highest population followed by Queens, and then Manhattan. Staten Island had the lowest population. If we consider the population of the five boroughs, it can be seen that Manhattan has the highest accident density. Therefore, although Brooklyn has had the highest number of accidents over time, Manhattan is the most accident-prone borough as it has had the highest number of accidents happen within a smaller population size compared to Brooklyn. But given the number and density of accident counts, Brooklyn, Manhattan and Queens are quite accident-prone.

2. What are the trends of accidents in New York over time?

Next, we analyzed the yearly, monthly, weekly, and hourly trends of accidents in New York over the years 2012 to 2021. To do so, we draw primary line plots across the years (one line for each borough) and then drew inferences

Yearly trends of accidents

The plot below shows the yearly trends of Borough-wise accidents over the years 2012 to 2021.

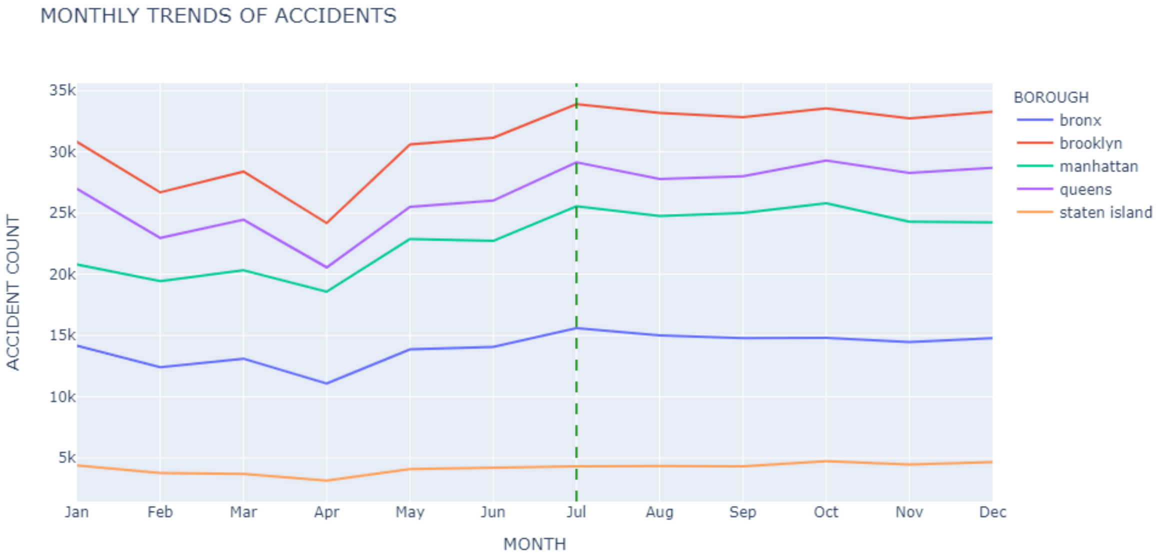


- Firstly, **Staten Island** has had the least number of accidents over the years. This is because its population is the smallest among the five boroughs. Moreover, the accidents in Staten Island also seem to decrease after the year 2013.
- **Brooklyn**, as seen above, has the highest number of accidents among the five boroughs owing to its large population size.
- Accidents in 2012 were the lowest for all the five boroughs. One possible explanation for this can be that the data collection methodologies were not so advanced as they are today and therefore insufficient information regarding the accidents was collected.
- The accidents sharply rose from 2012 to 2013 where more accidents might have started getting officially reported. The trend does not vary much from 2013 to 2015, however the accident count drops in the year 2016.

- The drop in accidents in 2016 may be due to the **"Vision Zero"** traffic control program launched by Mayor de Blasio on January 15, 2014. It aimed at eliminating traffic deaths and injuries in New York City by pressing charges against traffic violators, by reducing the speed limit from 35 to 20 mph and some other measures.
- After 2019, the accidents start to drop in all the five boroughs. This can be due to the **Coronavirus Pandemic** which started around December, 2019. The closure of educational institutions and works led to lesser vehicular traffic on the roads therefore, it resulted in a sharp drop in accidents.
- As 2021 is still going on, the data regarding accidents in this year is not complete therefore it shows the lowest number of accidents.

Monthly trends of accidents

Just like the plot for years, the plot below shows the monthly trends of borough-wise accidents.



Here are a few intriguing observations that we found:

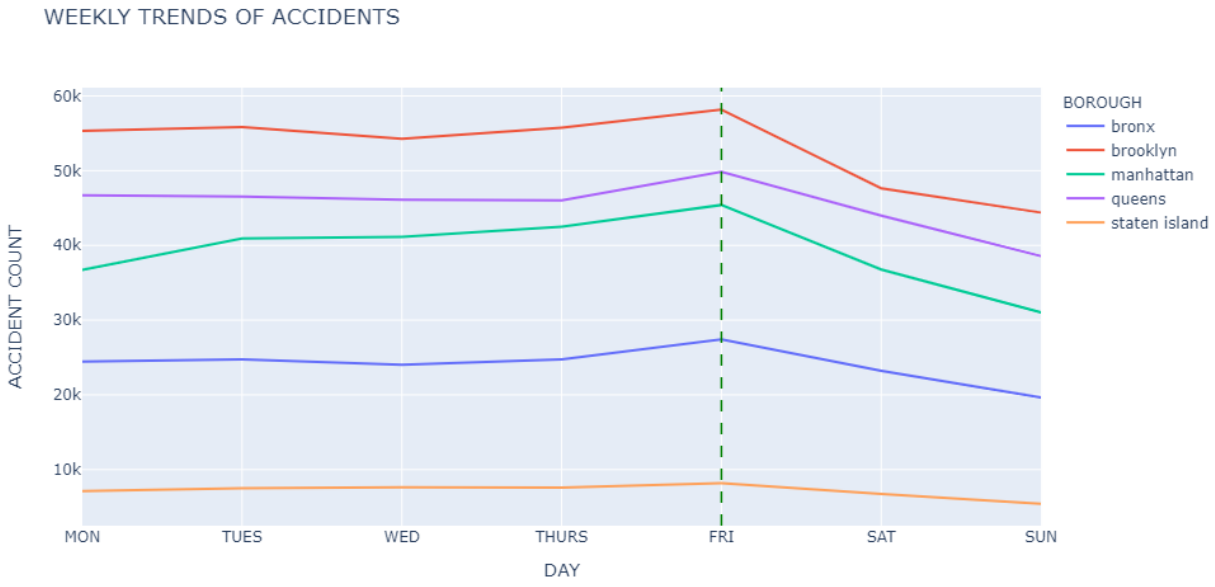
- From January to March, the trend does not change much. Accidents slightly decrease in February and increase in March.
- The lowest number of accidents are observed to be in **April**. This may be due to the transition in the weather from Winter to Summer. Lesser snowfall in New York and increased sunlight may improve visibility for the drivers on the road.
- After April, the accidents significantly increase and they peak in **July**. Excessive heat in the summer season may frustrate the drivers making it difficult for them to focus on the road. Moreover, the vehicular traffic can also increase as people would more likely be hanging out to enjoy their summer holidays. This could include the residents of the city as well as the people

or tourists who visit NYC for summers. These can be a few possible explanations for the high accidents in July.

- Although people are more likely to stay inside their homes during the Winter season, therefore one can expect for the accidents to drop from November onwards. But the trend remains nearly the same which can be explained by the poor weather conditions increasing the risks of accidents.

Weekly trends of accidents

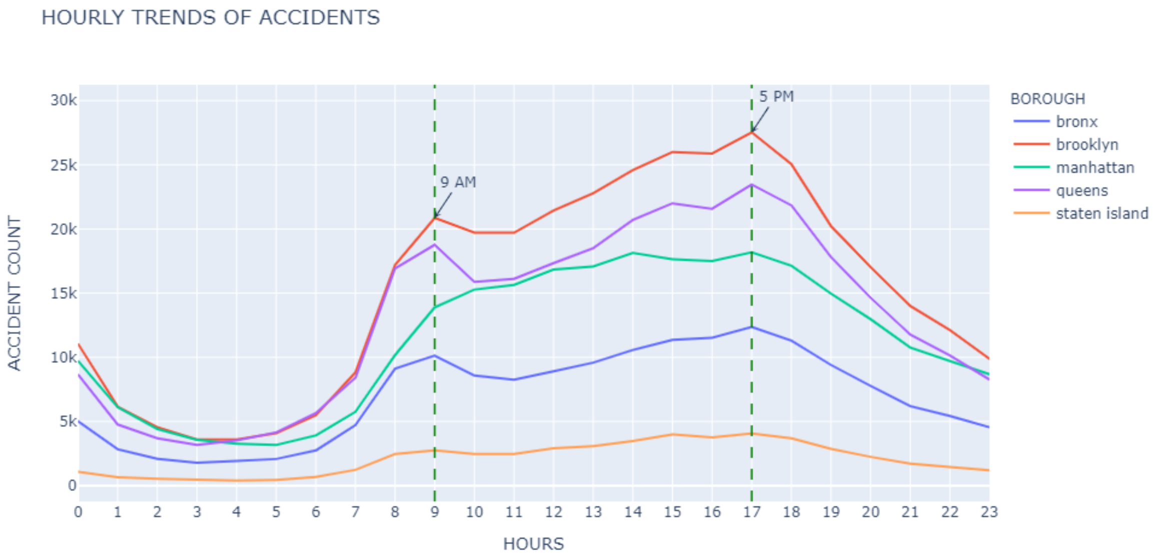
In order to determine the changes in the count of accidents on different days of the week, we again drew a similar plot which shows the trends of borough-wise accidents throughout the week:



- The number of accidents is higher during the weekdays, peaking on Friday. As Friday is the last working day of the week, people are more likely to rush to their homes thus increasing the risks of accidents.
- Accidents are significantly lesser over the weekends which may be due to the presence of fewer vehicles on the road. This could also be attributed to the alertness and robustness of accident tracking methods during the weekday. But this is just an assumption and no supporting evidence could be found.

Hourly trends of accidents

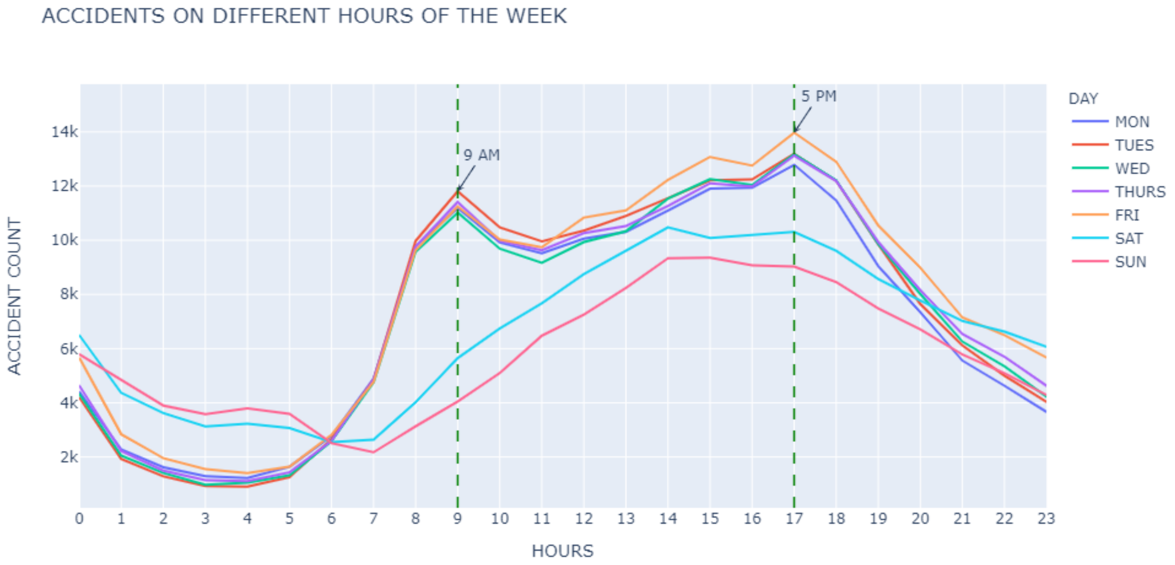
We all know how some hours of the day are rush hours. We also know that they are somewhere around the morning (school and office timings) and somewhere around early evening. To confirm this speculation, the plot below shows the hourly trends of borough-wise accidents throughout the day.



- The number of accidents seem to be the highest around 9AM and 5PM. Clearly, **9AM** and **5PM** are considered rush hours as more people are heading to their work and schools around 9 AM and returning to their homes around 5PM. Busy roads and high vehicular traffic can be associated with the peak office hours here.
- From 6PM to 5AM, the accidents tend to decrease because of fewer vehicles on the road.

Hourly trends of accidents with respect to days of the week

Now another interesting thing was analyzing the hours during which the accidents mostly happen in a week.



The plot above shows the hourly trends of Borough-wise accidents throughout the week.

- The most important point to note here was the clear difference in the collision trends in the weekdays and weekends interpreted by the legend on the right.
- Similar to the trends we have seen earlier, the accidents during the **working weekdays** peak around **9AM and 5PM** with these hours being the rush hours. Although during the weekends, accidents are also high around these hours but they are comparatively lesser than the weekdays.
- Around **4AM**, the accidents are higher in number on the **weekends** than the working weekdays. This may be because people tend to stay out till late at night to enjoy their weekends.
- After 5PM, the accidents decrease on all days of the week.

3. How did the severity of accidents vary throughout NYC?

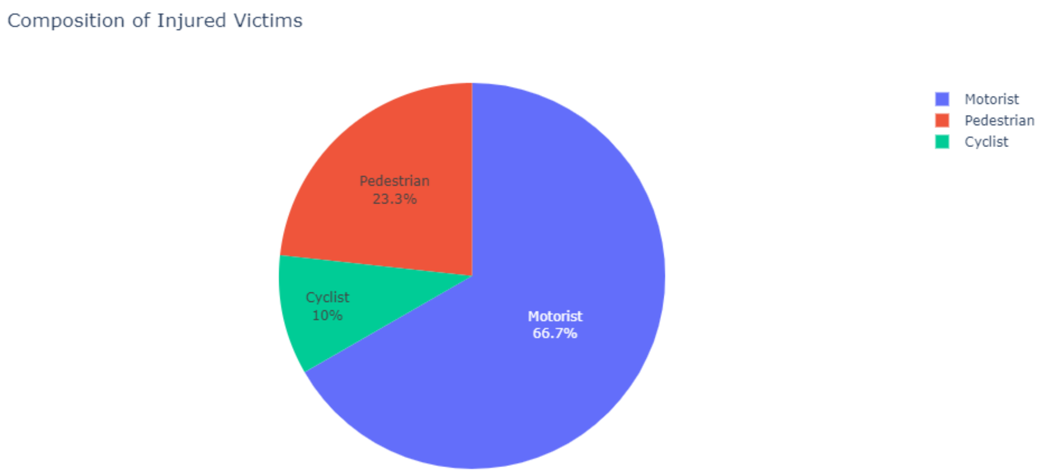
We were given multiple columns hinting towards the severity of accidents. They not only include the count of people injured and killed in each accident but also the type of victim (pedestrian, cyclist, motorist). Let's first look at the type of people injured and killed in accidents.

Composition of victims

While we have discussed the trend of accidents including the area and time in which they occurred, we will now have a look at the victims of the accidents. Were they pedestrians, motorists or someone else? Who was most adversely affected by the accidents? Let's have a look.

• Composition of victims who were injured

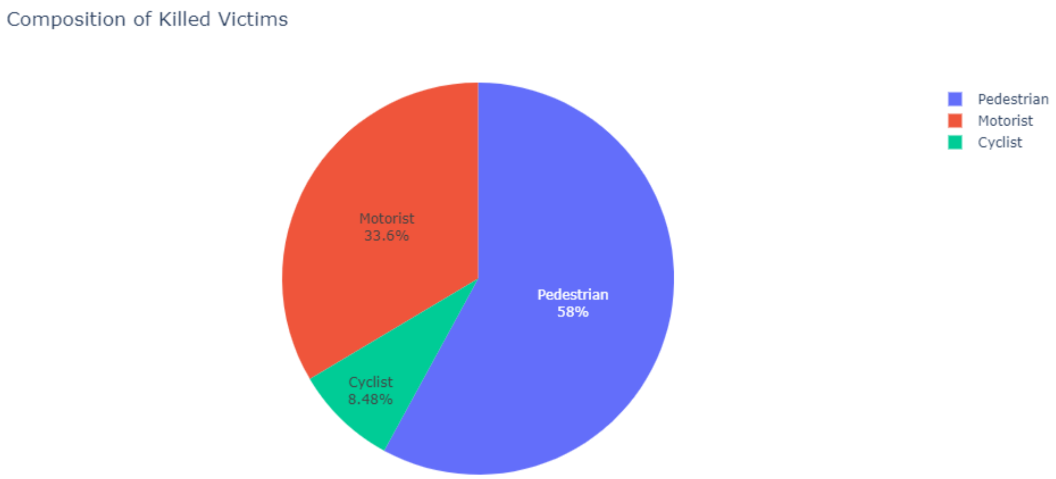
The pie plot above shows the composition of different victim types who were injured in the accidents in New York over the years 2012 to 2021 and also enables us to compare the ratios:



We see that majority of the injuries were faced by motorists followed by pedestrians and cyclists. This is because care accidents are usually severe and higher in number resulting in injuries to the passengers. Also, in a busy city like New York, the proportion of motorists is definitely higher as well.

• Composition of victims who were killed

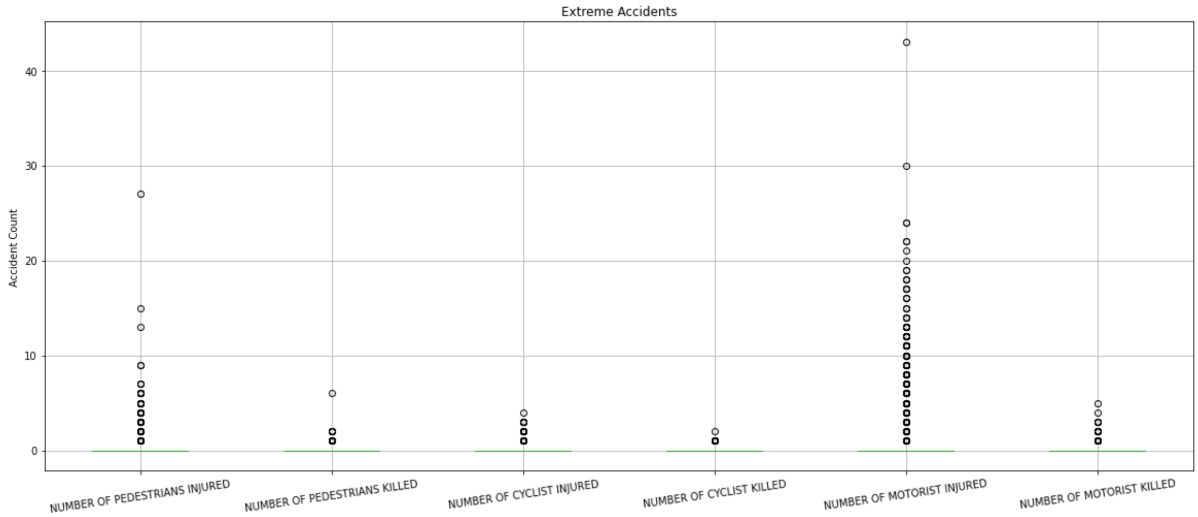
In case of the victims who were killed in the respective accidents, the pie plot below shows the composition of different victim types who were killed in the accidents in New York over the years 2012 to 2021.



From both of these pie charts, it can be observed that **motorists** were majorly injured in the accidents that occurred in New York over the years 2012 to 2021 while the deaths were dominated by **pedestrians**. A possible reason may be that in case of an accident with a car/vehicle, the cases of survival are more as the collision is not direct. In both cases, cyclists were involved in the smallest number of accident injuries or killings.

Extreme Accidents

To visualize extreme accidents, we used a boxplot as it gives the best representation of the outliers or potential anomalies present in the data. The aim here was to visualize the outliers and not the five-number summary or relevant statistics, so we visualized them keeping the outliers as focus as follows:



The data has two major outliers:

- The first one is where 25 pedestrians were injured and the second where 40+ motorists were injured. The first incident happened on May 18, 2017 when a Car rammed into pedestrian in Times Square and killed one person, injuring more than 20 others.
- The second one happened in Brooklyn on September 9,2013 when a car crashed into a bus. 40+ people were injured.

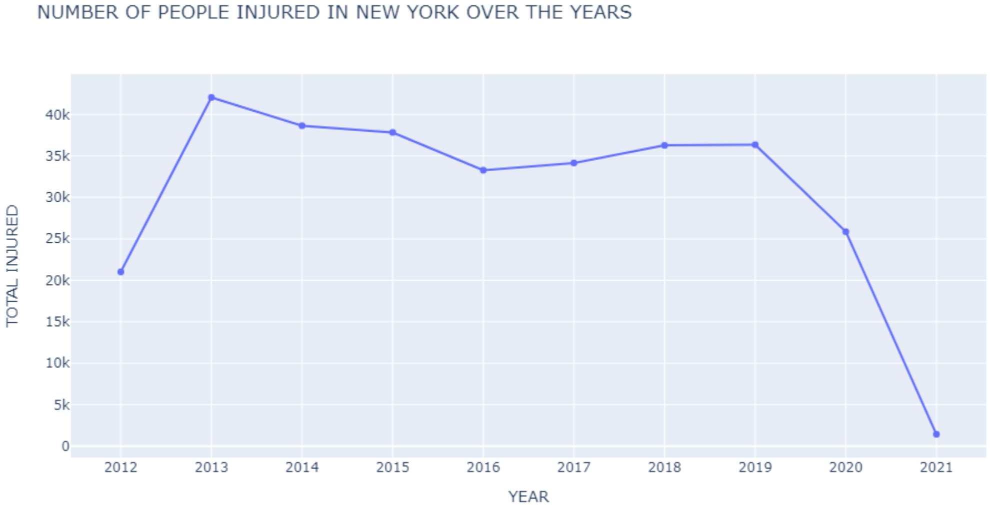
Manual analysis of the aforementioned dates and comparison with NYTimes news revealed the same.

Trends of injuries and deaths in New York

After establishing that majority of the injured victims consisted of motorists, and death victims consisted of pedestrians, we will now analyze how the general and borough-wise trends of victim injuries and deaths changed over the years.

General trend of injuries over the years

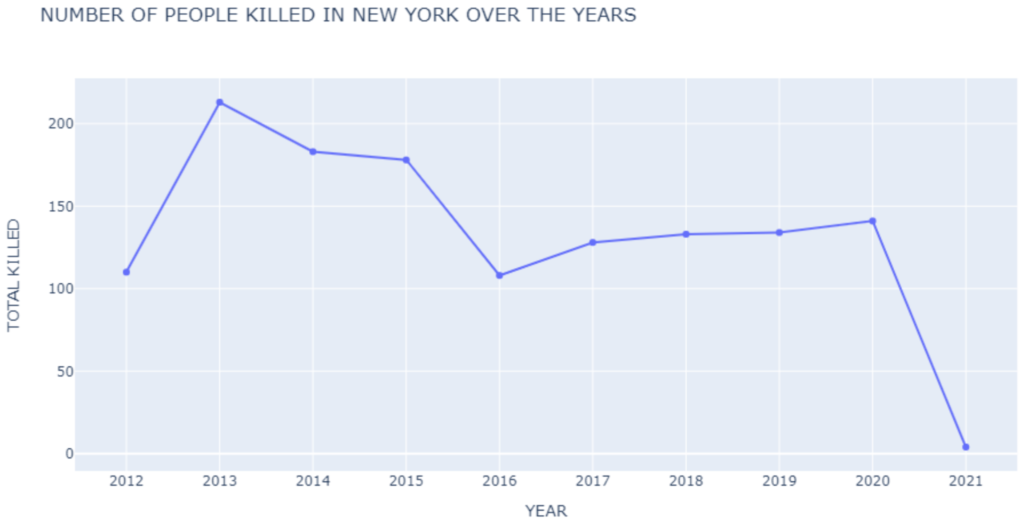
The plot below shows the number of people injured in New York accidents over the years 2012 to 2021.



We see quite an interesting shape of the line plot. The highest count was somewhere around late 2012 and the lowest was in 2020 and 2021. We will discuss them later but first let's have a look at the trend of deaths as well.

General trend of deaths over the years

The plot below shows the number of people who died in New York accidents over the years 2012 to 2021.

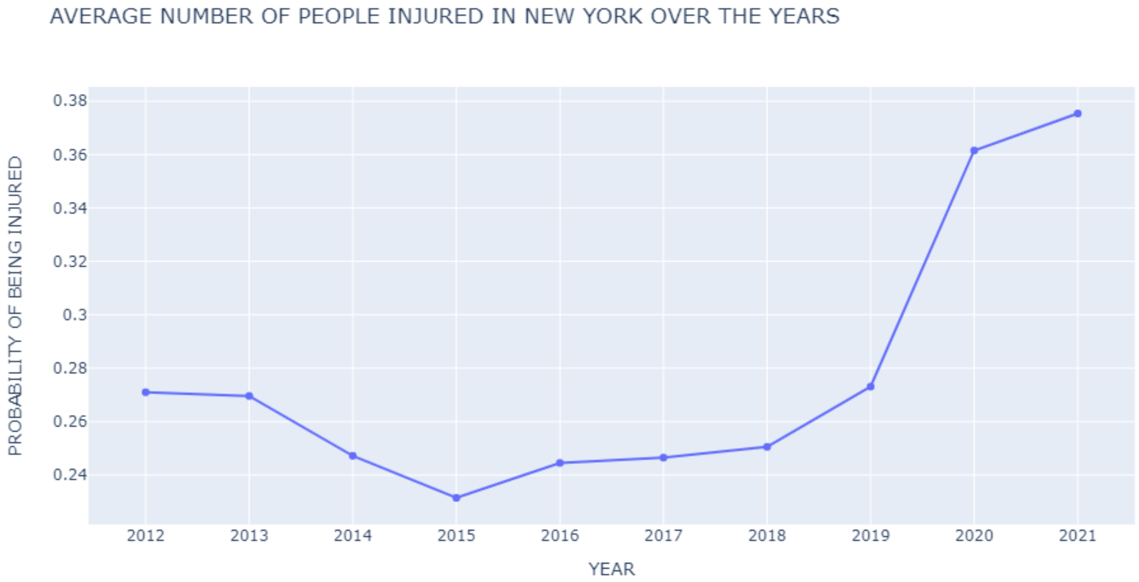


Some important observations from the plots above are listed below:

- 2012 has the lowest number of people injured and killed which, as stated earlier, may be due to insufficient data collected regarding the accidents at that time.
- The highest number of deaths and injuries occurred in the year 2013. This sharp rise from 2012 can be due to improvement in the data collection strategies.
- Following 2013, the years 2014 to 2016 show a sharp drop in injuries and deaths which may be the result of the "Vision Zero" program launched in the year 2014 which aimed at reducing deaths and serious injuries.
- From 2016 onwards, the injuries and deaths slightly increased until the year 2020. There are no sharp peaks after 2016 which means that the "Vision Zero" program might have helped in decreasing injuries and deaths.
- 2020 has the significantly lesser injuries which can be due to closure of educational institutions (COVID-19) and work offices leading to lesser vehicular traffic on the roads. However, the death rate did not decrease in 2020, which means that most of the accidents that had happened in 2020 led to the death of victims.

- **Probability of people injured per accident over the years**

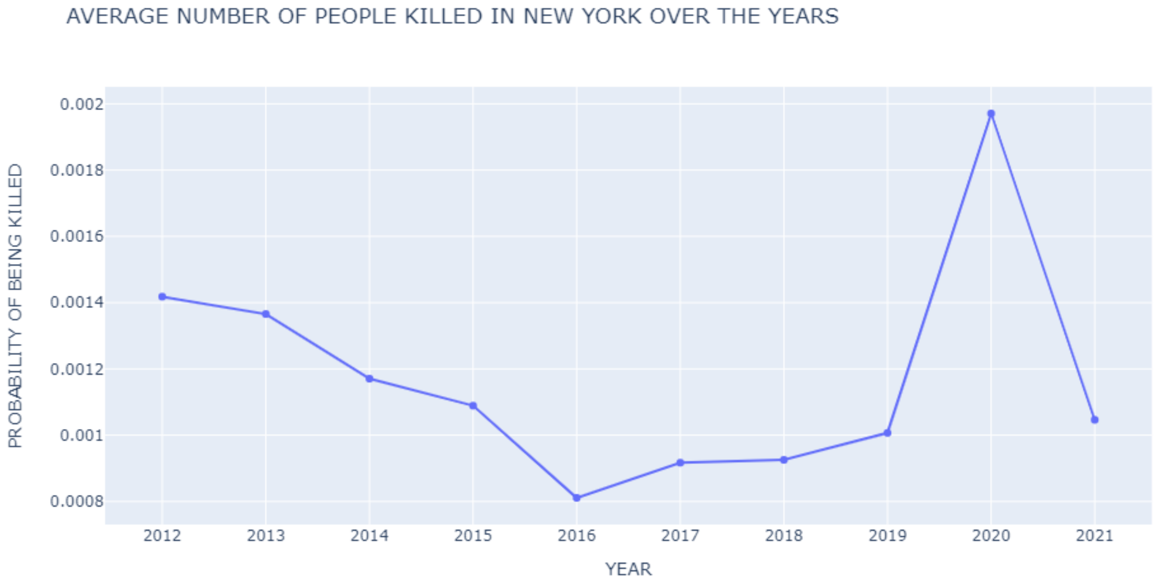
The plot below shows the probability of injury in case of accidents over the years 2012 to 2021:



The y axis represents the probability of of a person being injured in a particular accident over a year. For example, .27 for 2012 means that in 2012, there were 0.27 of people being injured per accident. We see an interesting plot here, with the graph decrease from 2012 to 2015-16. However, we see a sharp increase to 2020 to 2021. This is in contrary to what we posed earlier – 2020 and 2021 had lower number of accidents. While the number of accidents in 2020 and 2021 were admittedly less, they were higher in their severity causing higher probability of injuries within people. This is alarming as the risk associated with accidents in increasing in recent times.

- **Probability of people killed per accident over the years:**

The plot below shows the probability of killings in case of accidents over the years 2012 to 2021:

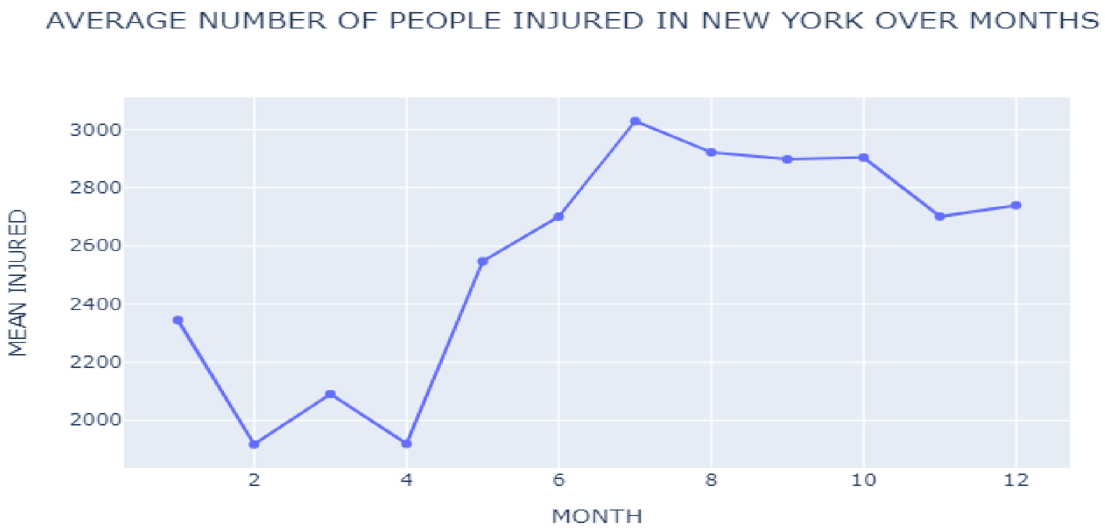


Some important observations from the plots above:

- 2012 has 0.275 people injured per accident and .0014 deaths per accident.
- Following 2013, the years 2014 to 2016 show a sharp drop in probability of injuries and deaths which may be the result of the "Vision Zero" program launched in the year 2014 which aimed at reducing deaths and serious injuries.
- After this it starts to gradually rise until hitting an all-time high in 2020 and 2021.
- This is interesting to note as the total injured decreases in 2020 as shown in previous plots however average injuries and deaths increase. This may be because of incomplete data for 2020 or some biases.

- **Average number of injuries over months:**

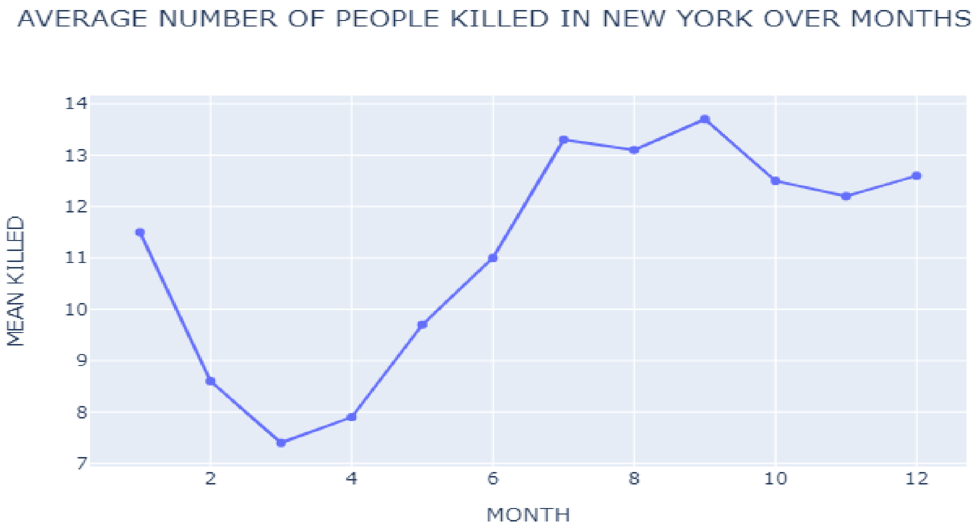
The plot below shows the average monthly injuries in NYC over the years 2012 to 2021:



This plot shows us a low in average injuries in February or the 2nd month while it gradually increases till hitting a peak during the summer. There is a gradual decrease as we come toward the 11th and 12th months of November and December, representing winter.

- **Average number of deaths per months:**

The plot below shows the average monthly deaths in NYC over the years 2012 to 2021:

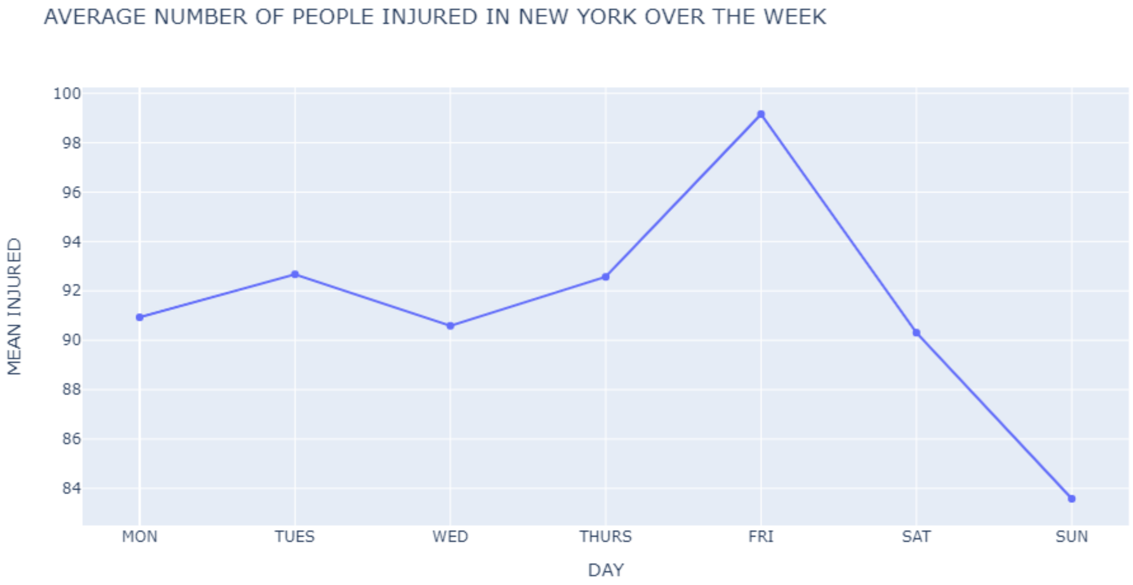


Important observations regarding the two plots:

- Average injuries and deaths are less in the winter than summer due to the harsh weather in winter where less people go out and so there is less traffic on the road
- The graphs show a decreasing trend from January to February, with increase till August and September, after which there is another decreasing trend.
- Highest number of injuries and deaths in summer during August.

- **Average number of injuries over the week:**

The plot below shows the average weekly injuries in NYC over the years 2012 to 2021:

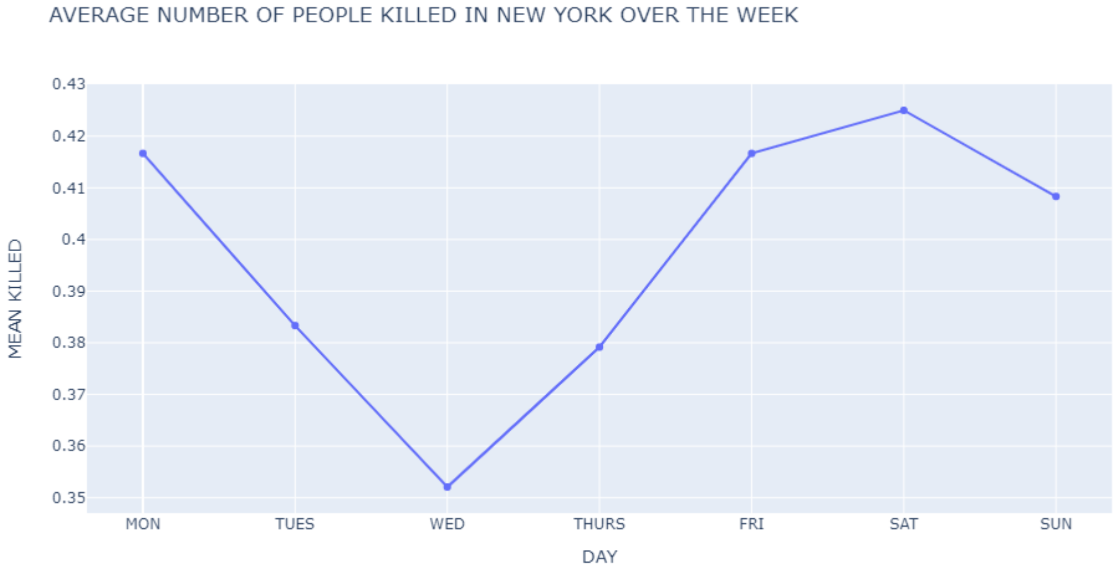


Observations from the plot above are as follows:

- Average injuries over the week almost remain consistent from Monday to Thursday with slight fluctuations.
- Average injuries are the highest on Friday. As it is the last working day of the week, vehicular traffic is likely to be high and it can lead to more accidents and injuries.
- Average injuries decrease over the weekends because of lesser vehicular traffic on the roads.

- **Average number of deaths over the week:**

The plot below shows the average monthly deaths in NYC over the years 2012 to 2021:

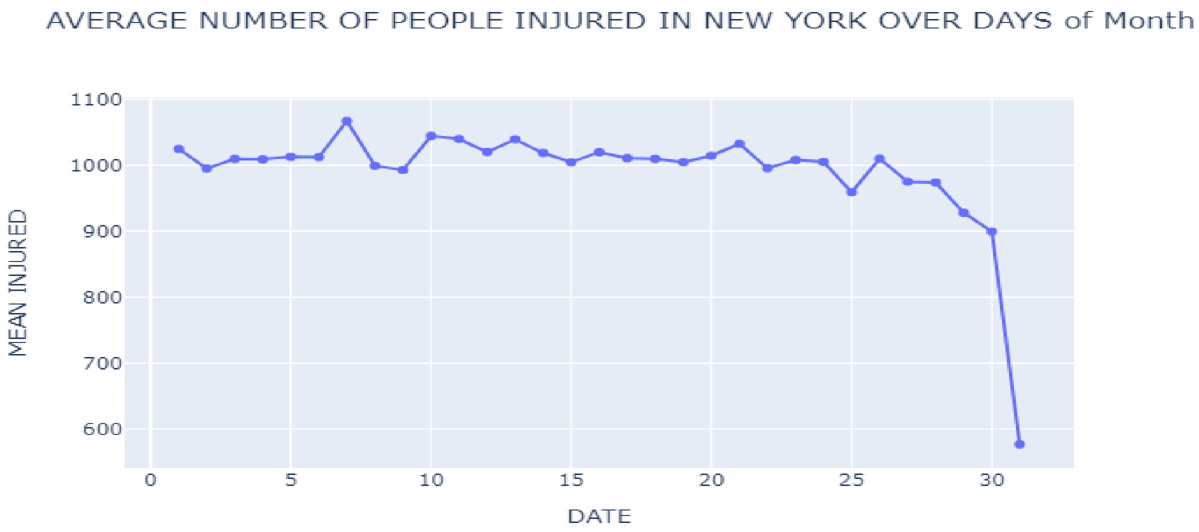


Observations from the plot above are as follows:

- Average deaths on any particular day of the week are quite low as evident from the y-labels being quite low.
- Average deaths decrease from Monday to Wednesday.
- Wednesday has the least number of deaths on average over the years 2012 to 2021.
- Average deaths increase after Wednesday and are the highest on Saturday.
- The plot above implies that although injuries and accidents are low over the weekend, but most of these accidents lead to deaths.

- **Average number of injuries per days:**

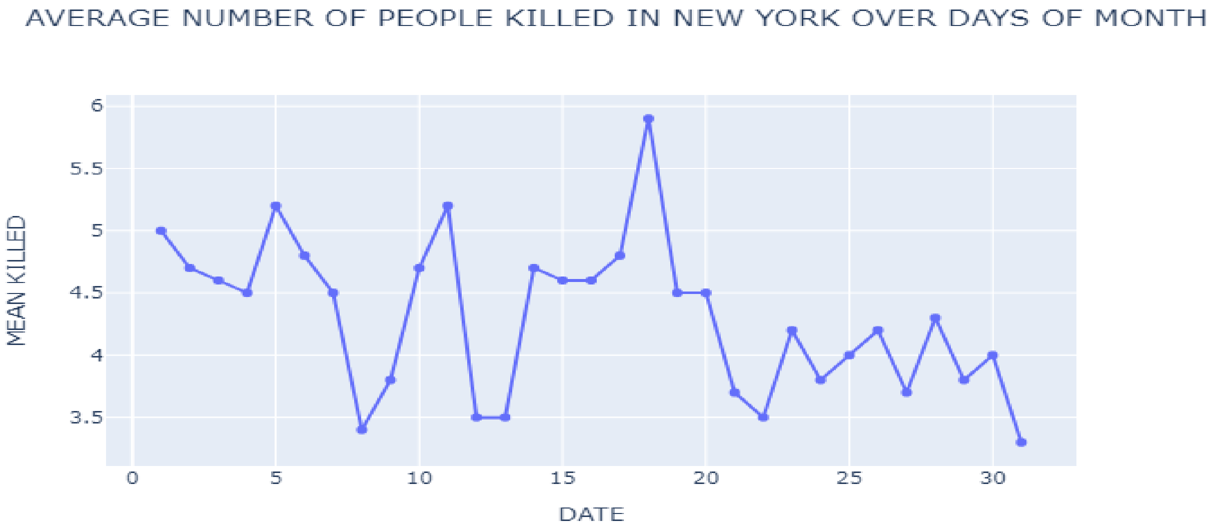
The plot below shows the average daily injuries in NYC over the years 2012 to 2021:



We see a very random plot with values withing a small range. It is difficult to spot any trends. There is a sharp decrease on 31 because less months have 31 days.

- **Average number of deaths per days:**

The plot below shows the average daily deaths in NYC over the years 2012 to 2021:

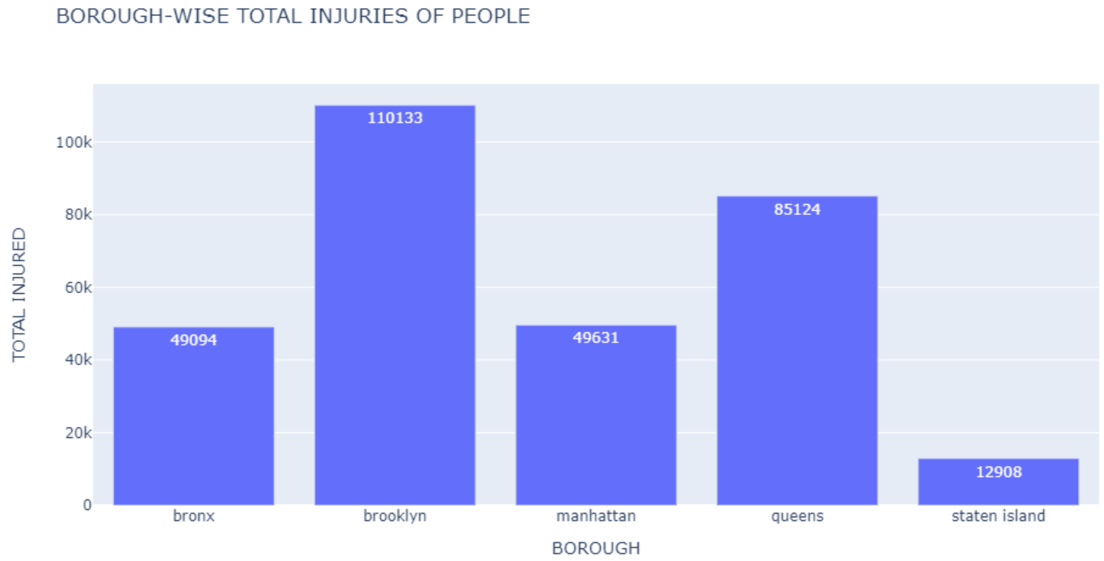


Important observations regarding the plots:

- It is difficult to spot any trends or patterns as the graph is very random.
- However, there are more injuries and deaths during the start and end of the month as opposed to the middle.

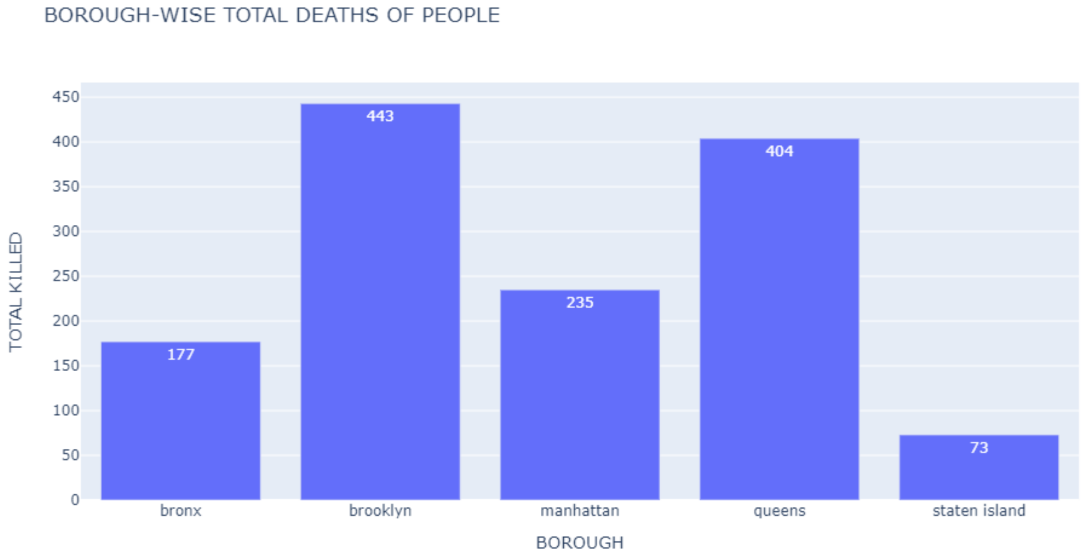
Borough-wise overall injuries

The plot below shows the borough-wise overall injuries of people in New York over the years 2012 to 2021.



Borough-wise overall deaths

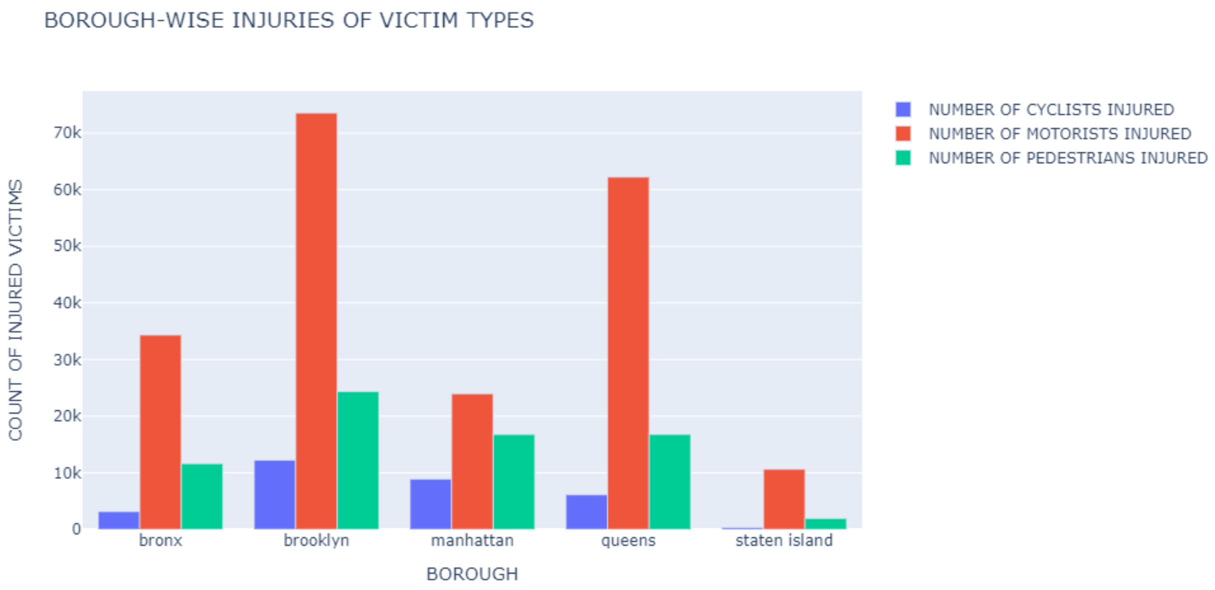
The plot below shows the borough-wise overall deaths of people in New York over the years 2012 to 2021:



Some common observations which are in accordance with the previous plots are as follows:

- Brooklyn, as expected, has the highest deaths and injuries among the five boroughs, followed by Queens.
- The number of people who died in the accidents are far less than the injured ones. This means that a vast majority of the injured people end up recovering.

Borough-wise injuries of different victim types

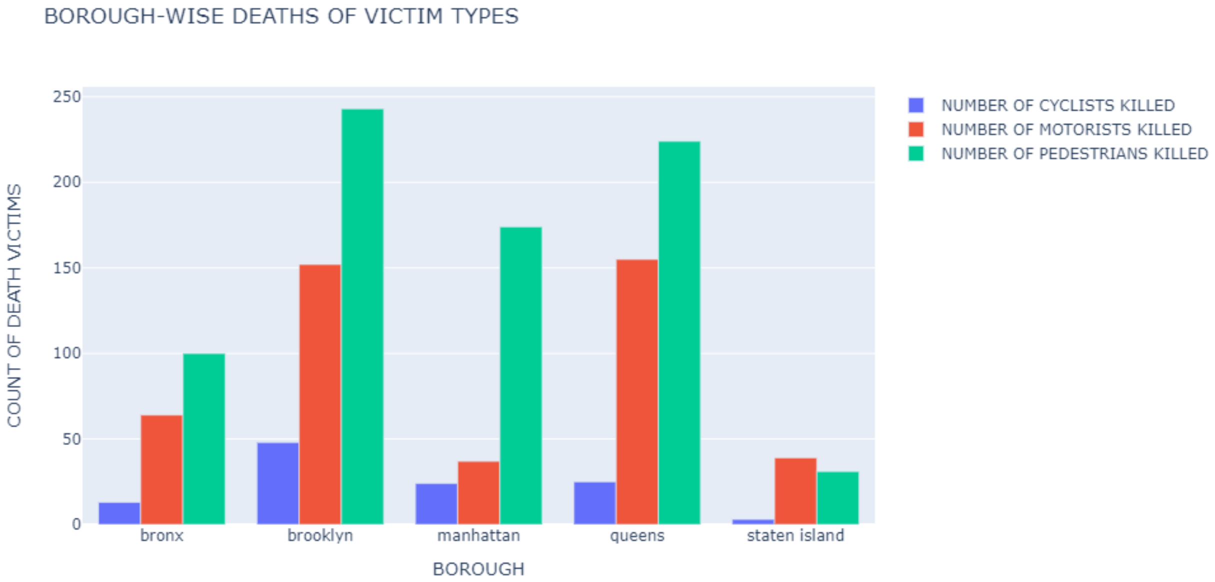


In the general trend of injuries in New York, we observed that the injured people mostly consisted of **motorists**. A similar trend can be seen in the five boroughs of New York. Following observations were made from the plot above:

- In all the five boroughs, motorists were injured greatly over the years 2012 to 2021 followed by pedestrians, and then cyclists.
- Number of motorists injured is the highest in Brooklyn.
- Number of cyclists injured is the lowest among all the five boroughs. Staten Island has the lowest number of cyclists injured.

Borough-wise deaths of different victim types

In the general trend of deaths that took place in New York due to accidents over the years, we observed that pedestrians made up the major portion of death victims



Similar to it, using the plot above, we see that borough-wise deaths consist mainly of **pedestrians** followed by motorists and then cyclists.

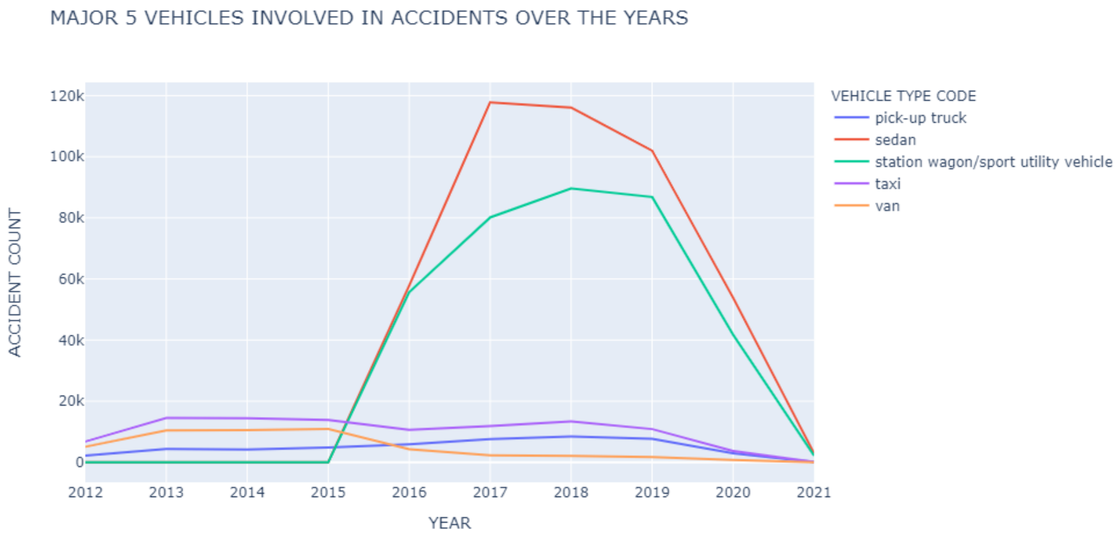
- Brooklyn again has the highest number of pedestrians killed followed by Queens and Manhattan.
- Contrary to other boroughs, Staten Island has slightly a greater number of motorists killed than pedestrians.

4. What are the most common causes of vehicular accidents in New York?

For each accident in the dataset given to us, the vehicles involved and the causes of accidents were also mentioned. Using this information, we have visualized the top vehicles and causes of accidents in New York below.

Major vehicles involved in accidents

Some of the categories of vehicles involved in accidents were either unknown or redundant i.e., one category was listed multiple times with different names. We have not included these while plotting the top 5 vehicles which contributed to accidents in New York over the years 2012 to 2021.

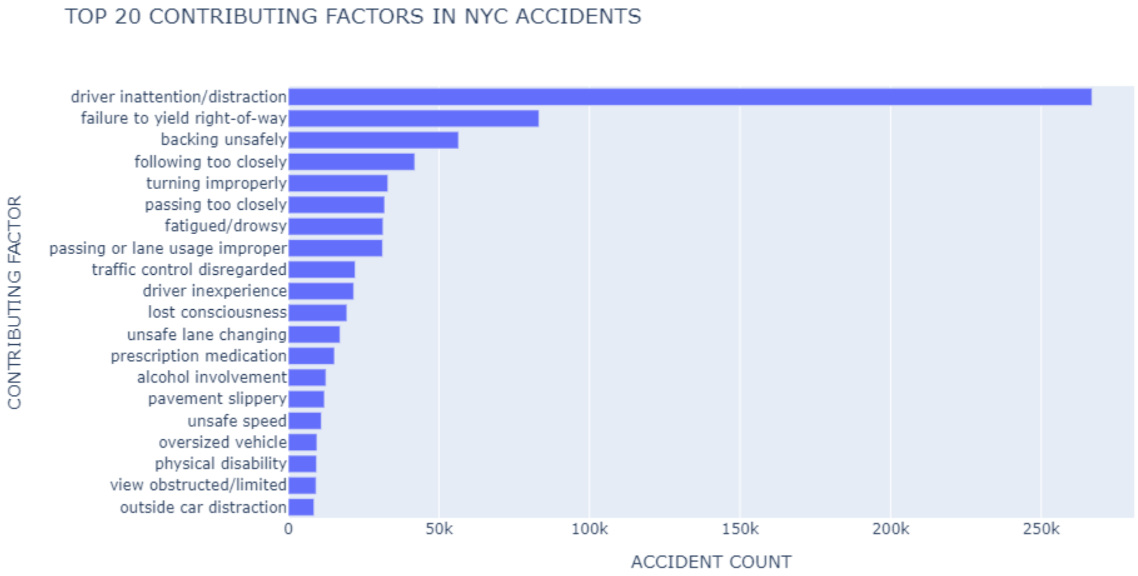


The plot above shows how the trend of the top 5 vehicles' involvement in accidents changed over the years.

- Taxis were more involved in accidents from 2012 to 2015 and station wagons were the least involved during these years.
- 2015 onwards, there is a sharp rise in accidents due to sedans followed by station wagons. This may be because station wagons and sedans are more common in New York, therefore they contribute to more accidents.
- The accidents due to pick-up trucks, taxis, and vans remain nearly constant throughout the years.

Major causes of accidents

It will now be interesting to see what actually caused the accidents. Let's have a look at the top 10 contributing factors:

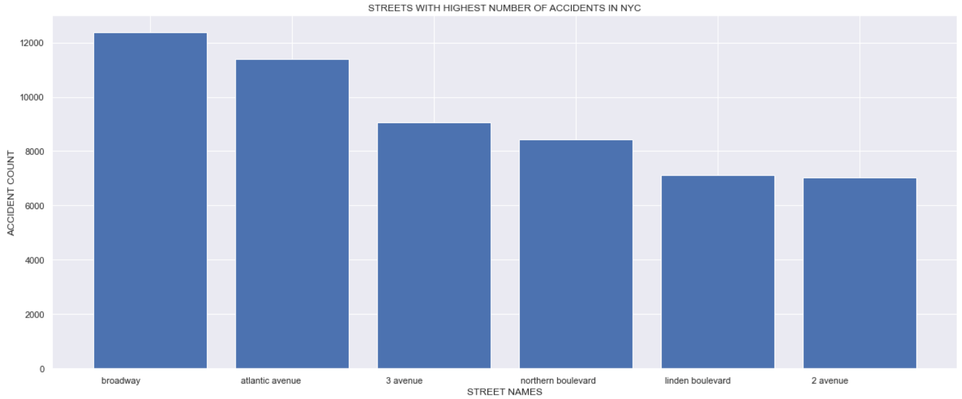


The plot above shows the top most common causes of accidents in New York. We have filtered the “unspecified” and “other vehicular” causes as they are vague and do not give any meaningful information to help our analysis.

- Driver inattention/ distraction is the top cause of accidents in New York; however, it is still a broader category and does not give much detail regarding the accident. Other causes such as “alcohol involvement”, “unsafe speed”, “fatigued”, “outside car distraction” etc. can also be listed as driver inattention or distraction.
- Some causes such as alcohol involvement, unsafe speed, improper lane usage etc. show that the traffic rules are not properly being followed by people. Therefore, proper enforcement of traffic rules and stricter fines in case of their violation can help in decreasing accidents.

Most Dangerous Streets in New York City

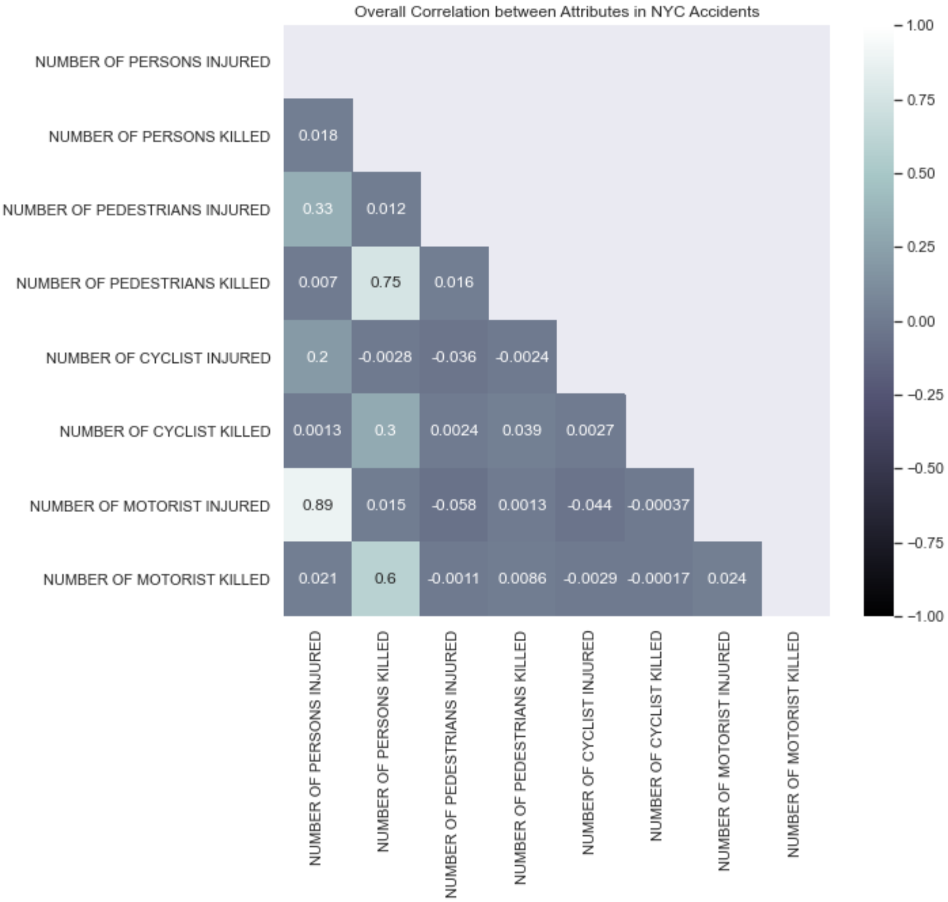
We would like to see which streets are the most dangerous in New York City for both accident count and fatalities. To investigate we formed a pivot table and sorted values by accident count. We then used the python group by function for street name and number of persons killed. This was sorted in ascending order to get the streets with most fatalities.



We found streets with most accidents are the ones where there are the most deaths. Most of these streets are located in Manhattan which is the busiest area in New York. Broadway and 3 avenue where the most accidents and deaths occur are located in the center of New York City and act as a hub for tourists, business and drama.

5. Are there any correlations between different attributes relevant to New York accidents and how do these correlations affect our analysis?

Before calculating correlations and plotting heatmaps, we will be extracting only the quantitative values from the dataset as correlation can’t be run on categorical values. We then calculated a correlation matrix. Since it is difficult to understand the values as it is, we visualized them in the heatmap shown below. In the correlation matrix and the corresponding heatmap, a value near 1 represents a strong positive correlation while one near -1 represents a weak negative correlation. Values near 0 mean the attributes are not related to each other.



- As can be observed in the above matrix the attributes are not as such correlated with each other.
- Persons killed and injured is basically a sum of all other attributes and thus showing positive strong correlations.
- It is interesting to note that persons injured and killed are not correlated to each other. This has also been proved earlier where we saw accidents with injuries are more common than those with fatalities.
- We also ran correlation for each specific borough and the results were largely consistent with the overall trend in NYC. However, there is a slight difference in the correlation value in Manhattan where the correlation between number of pedestrians killed and number of cyclists killed was 16. On further investigation we found this number was impacted by a single accident on the 31st of October 2017 which killed 6 pedestrians and 2 cyclists. It was a terror attack, which on further analysis proved to be the deadliest accident in the dataset.

Frequent Pattern Mining

After cleaning the data and applying exploratory data analysis, we got hold of multiple interesting findings about the collisions in New York City. While we found a lot of trends and patterns through the visualizations and summaries in the EDA, we intend to confirm them or find even more interesting patterns in this section using frequent pattern mining techniques. Before we dive into the main frequent pattern mining section, it is first important to identify and select the right features in order to get meaningful results. Therefore, we divide this section into two parts:

- Feature Selection and Extraction
- Frequent Pattern Mining and Relevant Findings

Feature Selection

Our cleaned data had multiple columns. Since we had already carried out exploratory data analysis, we had an idea that some of the features are more significant in this phase. So, we identified them using a stepwise approach. Let's have a look at the features we have:

Initial Features (In Cleaned Data Set)
'CRASH DATE', 'CRASH TIME', 'BOROUGH', 'ZIP CODE', 'LATITUDE','LONGITUDE', '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', 'VEHICLE TYPE CODE 1', 'VEHICLE TYPE CODE 2', 'MONTH', 'DATE', 'YEAR'

Since we had a picture in mind about the questions we wanted to answer in this phase, we shortlisted the columns which were needed and dropped the rest (we however were open to reconsider our choice if needed). Following features were dropped:

- **Crash Date:** We had already separated out date, month and year from crash date. We wanted to analyze frequent patterns primarily based on months or years instead of specific dates, we dropped the column crash date.
- **Latitude and Longitude:** We intended to use boroughs are a determinant of location of accidents and their frequency. Therefore, we dropped the feature latitude and longitude.
- **Street Name:** Boroughs provided information about the location on a broader scale while zip code provided information on an intermediate scale. Street name, however, had many unique values (~146,000) and were too specific to be included in our frequent pattern mining analysis. Therefore, we dropped this feature as well.

Following table includes the key features we selected for frequent pattern mining. We intended to use this set or combinations of its subsets to determine variety of frequent patterns:

Features Selected
'CRASH TIME', 'BOROUGH', 'ZIP CODE', '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', 'VEHICLE TYPE CODE 1', 'VEHICLE TYPE CODE 2', 'MONTH', 'DATE', 'YEAR'

Feature Extraction

Now that we had selected the features that we needed for our frequent pattern mining analysis ahead, it is time for feature extraction. We wanted to bring the existing features into their most usable form to extract maximum information from them. Most of the features had apt information and format already. The ones which needed feature extraction were as follows:

- **Crash Time:** Crash Time was given in 24-hour clock format. However, the probability of multiple accidents occurring at the exact same time is quite low. Therefore, we decided to bin the crash timings into 4 groups such that the crash timings fall into one of the following: *12AM to 12PM, 12PM to 3PM, 3PM to 6PM or After 6PM (to 12AM)*. This would make it easier for us to understand the general trend of the time of the day in which accidents frequently occurred.
- **Month:** We had already extracted day, month and year from the given crash dates. But to make interpretability of the results easier and more intuitive, we converted months feature into a more readable format i.e., the exact calendar name.

Feature Modification

Our final step before moving on the actual frequent pattern mining part was to modify the features such that they were as readable and easily interpretable in the resulting frequent item sets. Features like Borough, Month, Year, Time, Vehicles and Contributing Factors were fairly straightforward and self-explanatory.

The remaining features were numeric in nature and may not be identifiable in the transactions or frequent patterns. Therefore, key words based on a specific column were added to them. Following is the table of prefixes (keywords) added to the respective columns:

(Please turn over to the next page)

Feature Name[1]	Prefix Added	Final Feature Entry Example
Number of Persons Killed	PK	PK:1

[1] This table can be referred to for easier interpretations of frequent patterns

Number of Persons Injured	PI	PI:1
Number of Pedestrians killed	PedK	PedK:1
Number of Pedestrians injured	PedI	PedI:1
Number of Cyclists Killed	CK	CK:1
Number of Cyclists Injured	CI	CI:1
Number of Motorists Killed	MK	MK:1
Number of Motorists Injured	MI	MI:1
Vehicle Type Code 1	V1	V1: bike
Vehicle Type Code 2	V2	V2: passenger vehicle

Table 2: Prefix Conventions for Pattern Interpretation

The rest of the features required little to no feature extraction. We decided to keep the features as simple as possible so that results could be clear and precise. The features were now in their usable form so we moved on with the frequent pattern mining.

Frequent Pattern Mining

Instead of acquiring a broad one-way approach and finding the overall frequent patterns in the accidents, we decided to divide this section into the following sub questions and answer them one by one:

- What are the frequent patterns collisions of different vehicles over the regions and over time?
- Are there any frequent patterns in the number and type of victims in collisions?
- How do the frequent patterns of collisions vary in different regions of New York City?

The next section intends to answer these questions, build them upon each other and reach a conclusion.

1. What are the frequent patterns of collisions of different vehicles and over time?

In earlier phases (particularly exploratory data analysis), we found out that accidents were more pertinent in a particular time of the day than the other. Similarly, some months and years had greater number of accidents than the rest. A similar trend was observed in with respect to the regions as well. Therefore, we decided to select all region and time-based features to form a dataset, formed their transactions and then ran frequent pattern mining algorithm on them.

As a starting point, we ran both Apriori Algorithm and FP Growth Algorithm to check their performance but contrary to our expectations, Apriori Algorithm did not fail to work efficiently on substantial amounts of data. A plausible reason can be that the library version of Apriori is enhanced and efficient. Therefore, we opted for Apriori Algorithm for the rest of the frequent pattern mining.

To get an idea, we observed the frequent patterns under different minimum support and minimum length and observed a clear distinction in the number of patterns obtained. As we increased the support count or length, the number of frequent patterns decreased. But reducing them significantly could lead to low-quality frequent patterns. Therefore, it was important to choose a correct threshold. Following is a table that we made (after observing the effect of change of threshold on the number of frequent item sets obtained):

Minimum Support	Minimum Length	Number of Frequent Patterns
0.05	5	None
0.05	4	>20
0.1	4	1
0.1	3	>15
0.2	4	None
0.2	3	5

Table 3: Number of frequent patterns returned by FPM algorithm under different thresholds

Since frequent patterns extracted using 0.1 minimum support and minimum length 3 were quite reasonable in number and quality, we based our analysis primarily on it (however we alluded to frequent patterns extracted using 0.05 minimum support and minimum length 4 off as well). Now let's have an in-depth discussion of the frequent patterns from different dimensions:

Vehicles Involved

With more than 600 unique vehicle types in the dataset, we did not expect to find a few of them in involved in collisions a high majority. To check this notion, we now decided to observe if there were any specific vehicle types in the frequent item sets. Following were the interesting frequent item sets with respect to the vehicles involved:

- `['V2: passenger vehicle', 'PI:0.0', 'V1: passenger vehicle', 'PK:0.0']`
- `['V1: station wagon/sport utility vehicle', 'PI:0.0', 'PK:0.0']`
- `['V2: station wagon/sport utility vehicle', 'PI:0.0', 'PK:0.0']`
- `['V2: sedan', 'PI:0.0', 'V1: sedan', 'PK:0.0']`

Contrary to our expectations, vehicles like sedan, passenger vehicle and station wagon/sport utility vehicle were in a higher number as compared to the rest and were present in the frequent item sets most probably because these are the most common vehicles in New York. Collisions of sedan with sedan and passenger vehicle with passenger vehicle were also quite common. Again, little to no injuries or killings were caused by in most cases.

We separated out the severe accidents, i.e., the ones in which at least one person was injured or killed and found the following interesting pattern with respect to vehicles:

- ['V1: passenger vehicle', 'V2: passenger vehicle', 'PI:1.0', 'PK:0.0']
    - **Rule:** V1: passenger vehicle -> V2: passenger vehicle
    - **Support:** 0.14212093055268774
    - **Confidence:** 0.5330343169681032

The confidence of the rule (V1: passenger vehicle -> V2: passenger vehicle) is 0.53 which indicates that in the cases of severe accidents where one of the vehicles was passenger vehicle, the other one was also a passenger vehicle 53% of the times. This also validates our claim that passenger vehicles' collisions with each other are not only frequent, but also severe in most cases.

Time

Analysis of the frequent patterns based on time (time of the day, month, and year) revealed something similar. Following are some interesting frequent patterns:

- ['V2: passenger vehicle', 'Before 12PM', 'PK:0.0']
- ['brooklyn', '3PM to 6PM', 'PI:0.0', 'PK:0.0']
- ['Before 12PM', 'PI:0.0', 'PK:0.0']
- ['12PM to 3PM', 'PI:0.0', 'PK:0.0']
- ['3PM to 6PM', 'PI:0.0', 'PK:0.0']
- ['PI:0.0', 'After 6PM', 'PK:0.0']
- ['2013', 'PI:0.0', 'PK:0.0']
- ['2014', 'PI:0.0', 'PK:0.0']
- ['PI:0.0', 'PK:0.0', '2015']

Overall, collisions were frequent in all the timings of the day. Earlier, in exploratory data analysis, we analyzed that the highest number of accidents occurred around 5PM. But here we see that accidents occurred throughout the day in a busy city like New York, it is just that the number of accidents were a little higher on timings before 12PM and between 3PM to 6PM (as discussed earlier as well). All the years had significant numbers of collisions (i.e., they were all found to be in 1-itemsets and 2-itemsets) but interestingly, in the 3-itemsets shown above the collisions in 2013, 2014, and 2015 were prominent and had little to no injuries and killings. Intuitively, this can be attributed to the increased risk of deaths and injuries in the recent years as the city progresses, number of vehicles increases and number and damage of collisions increases.

We also observed an interesting finding that in most cases where the reasons of the accidents were unidentified by the authorities (86%), the timings were around 3PM to 6PM

- ['unspecified', '3PM to 6PM']
    - **Rule:** unspecified -> 3PM to 6PM

- **Support:** 0.8649810325458719
- **Confidence:** 0.8649810325458719

This may be attributed to this timeframe being the busiest due to which the contributing factor of the accidents went unnoticed.

2. Are there any frequent patterns in the type of victims in collisions?

This time around our focus was to determine the type and number of people frequently killed and injured in collisions. We incorporated each type of victim (motorists, cyclist, pedestrians) into the features of data set we used earlier. We then followed a similar trend of minimum support counts and lengths as shown in Table 3 and ran our frequent pattern mining algorithm. The most intriguing frequent patterns were as follows:

- `['V2: passenger vehicle', 'PI:0.0', 'V1: passenger vehicle', 'PK:0.0']`
- `['PI:0.0', 'V1: passenger vehicle', 'PK:0.0']`
- `['V2: passenger vehicle', 'PI:0.0', 'PK:0.0']`
- `['MK:0', 'PedK:0', 'CK:0', 'CI:0', 'PI:0.0', 'MI:0']`
- `['Before 12PM', 'PI:0.0', 'PK:0.0']`

As expected, the number of people killed and people injured was always minimum in most accidents. Intuitively, extreme accidents are quite less in number and probability of occurrence as compared to minor accidents. Therefore, the number of people killed and injured as observed in the frequent patterns were 0 in most cases. Motorists killed and injured, cyclists killed and injured and pedestrians killed and injured also turned out to be 0 in the frequent patterns. This hints that less severe accidents i.e., the ones in which the victims were unharmed were more as compared to the ones in which there were major injuries and deaths.

To zoom into the severe accidents I.e., the ones in which at least a person was injured or killed, we shortlisted the accident records and ran similar frequent pattern on them. This time around, we intended to have an even more in-depth analysis. Following were the most interesting pattern:

- `['V1: passenger vehicle', 'PI:1.0', 'PK:0.0']`
  - **Rule:** `V1: passenger vehicle -> PI:1.0`
  - **Support:** `0.20769667561428867`
  - **Confidence:** `0.7789806553621271`

The support for passenger vehicle is 0.2 I.e., a passenger vehicle was involved in one-fifth of the severe accidents. The confidence for the rule is 0.77 which means that out of all the severe accidents involving passenger vehicles, 77% of them caused an injury!

Another intuitive, yet interesting pattern was as follows:

- `['PI:1.0', 'PK:0.0']`
  - **Rule:** `PI:1.0 -> PK:0.0`
  - **Support:** `0.7903718425218529`
  - **Confidence:** `0.7903718425218529`

Even in severe accidents, although there were people injured, the number of killings remained low. The above rule confirms it as well. With a confidence of 79%, it shows that even in the accidents in which there were some injuries, there were no people killed 80% of the time.

3. How do the frequent patterns of collisions vary in different regions (i.e., different boroughs) of New York City?

The dataset itself was large with a granularity of one collision per record and despite trying to dissect the data and performing frequent pattern mining on them, we wanted to look into collision region wise. The intent here was to determine if the collisions in different boroughs had similar patterns within them or not. We divided the original data frame into 5 parts where each part corresponded to the collisions within a single borough and performed frequent pattern mining on them. Following are the findings for each borough:

Brooklyn

Frequent patterns from Brooklyn indicated a clear trend I.e., in most of the collisions recorded the reason of the collision was unspecified and the number of people killed and injured (including motorists, pedestrians and cyclists) were typically low. The longest and most interesting frequent itemset here was the following:

- ['brooklyn', 'PedI:0', 'PedK:0', 'unspecified', 'PI:0.0', 'MK:0', 'PK:0.0', 'CK:0', 'MI:0', 'CI:0']

It clearly shows the trend that we mentioned in this and the prior sections. Killings and injuries are low. A rationale for contributing factor of the collision being ‘unspecified’ can indicate towards in limitation of the dataset i.e., the exact reasons for most of the accidents were not recorded aptly.

Again, we observed the severe accidents in Brooklyn and found a very interesting pattern:

- ['brooklyn', 'V2: passenger vehicle', 'PI:1.0']
  - **Rule:** brooklyn -> V2: passenger vehicle
  - **Support:** 0.13827517379035034
  - **Confidence:** 0.5006853582554517

With a confidence of 50%, it was quite interesting to note that 50% of the accidents that occurred in Brooklyn had a passenger vehicle involved, once again hinting towards the gravity of the issue.

Staten Island

Staten Island, having the least number of collisions in the dataset, was absent from the overall frequent patterns. Therefore, its individual pattern mining was quite interesting. In Staten Island, the exact trend as Brooklyn was observed. The longest and most interesting frequent itemset just like Brooklyn was the following:

['MK:0','PedK:0','CK:0','PedI:0','unspecified','PK:0.0','CI:0', 'PI:0.0', 'staten island','MI:0]

Again, the number of deaths and injuries were negligible and the reason for most of the collisions was unspecified.

But in case of severe accidents, unexpectedly, there was a unique trend:

- ['staten island', 'Before 12PM', 'PI:1.0', 'PK:0.0']

  - **Rule:** `staten island -> Before 12PM`
  - **Support:** `0.24457509353287013`
  - **Confidence:** `0.7695930036999664`

In Staten Island, most severe accidents occurred before 12pm causing some injuries but little to no deaths. This was in contrary to what we had observed earlier.

Queens

For Queen, yet again, a similar pattern was observed where the longest frequent itemset was as follows:

```
['queens','PedI:0','PedK:0','unspecified','PI:0.0','MK:0','PK:0.0','CK:0','MI:0',
'CI:0']
```

Just like the collisions in Brooklyn and Staten Island, the number of deaths and injuries caused by most accidents was 0 and the reason for most of them was unspecified.

In Queens, yet again, a majority of accidents and injuries pertained to passenger vehicles:

- `['queens', 'V2: passenger vehicle', 'PI:1.0']`
  - **Rule:** `V2: passenger vehicle -> PI:1.0`
  - **Support:** `0.38292918573894863`
  - **Confidence:** `0.8297697086179084`

Manhattan

In Manhattan, following was the longest and most interesting frequent itemset:

```
['PedI:0','PedK:0','PI:0.0','MK:0','PK:0.0','CK:0','MI:0','manhattan','CI:0']
```

Unlike the frequent item-sets of the boroughs we saw up till now, the reason of collisions in Manhattan was not ‘unspecified’ in most cases. This means that there were a variety of reasons for the accidents most of which were known.

In case of severe accidents in Manhattan, 87% of the times, there was just one injury as indicated from the confidence and support below.

- `['manhattan', 'PI:1.0', 'PK:0.0']`
  - **Rule:** `manhattan -> PI:1.0`
  - **Support:** `0.8707684883580649`
  - **Confidence:** `0.8707684883580649`

Bronx

The last borough, Bronx, had the following longest and most interesting frequent itemset:

```
['MK:0','PedK:0','CK:0','PedI:0','bronx','PK:0.0','CI:0','PI:0.0','MI:0']
```

In Bronx, yet again the number of people killed and injured frequently was 0. Like Manhattan and unlike the other boroughs, the reasons for accidents were well-recorded and ‘unspecified’ was not frequently used as a contributing factor.

Now let's look at an interesting frequent pattern that we found within the severe accidents at bronx:

- `['bronx', 'V2: passenger vehicle', 'PI:1.0', 'PK:0.0']`
    - **Rule:** `bronx -> V2: passenger vehicle`
    - **Support:** `0.4201762977473066`
    - **Confidence:** `0.543450722067393`

Again, a major part of severe accidents in Bronx, just like Brooklyn and Queens consists of passenger vehicles while the injuries were 1 and killings were minimum.

While there is a clear trend of frequently occurring collisions and their severity with respect to the injuries and death in each borough, the most interesting finding here was that the contributing factors for collisions were aptly recorded in Bronx and Manhattan. A plausible rationale for it can be the small size of boroughs which results in data collection and recording to be easier and more specific. Other than that, overall, the number of deaths and injuries in all boroughs as well as collectively was quite low.[2]

[2] This section only mentions the most interesting frequent patterns. The lists of all frequent item sets and their support count are in the Jupyter Notebook. A copy of the item sets can also be accessed here

Clustering

For this part our goal was to find interesting clusters regarding all the different variables present in our dataset. We used the preprocessed data as our base to start. In order to find the most viable clustering technique for our dataset, we tried different methods. DBSCAN and Agglomerative Clustering were not suitable for the large amount of data that we had because of high computational complexity (which was also confirmed by a timeout in Jupyter notebook). Splitting the data borough or year wise was not of much help either. Therefore, we opted for k-means[3] which worked effectively despite the large size of our dataset containing around 1.7 million rows. We divided our analysis based on clustering in the following sub questions:

- How did the severity of collisions of persons vary over time and how can they be grouped together based on specific trends?
- How were the vehicles involved in accidents in NYC over the years 2012 to 2021?
- How were the accidents spread throughout NYC based on region?

Let's perform some preliminary steps first.

Feature Selection

Before applying k-means clustering it was important to choose the right features (the ones which would help us answer the above questions), it was necessary to select the right features and make sure the features are in numerical and usable format.

- Borough and Vehicle type codes were of categorical datatype therefore we encoded them into numerical values to them using dictionaries.
- Although the variety of collisions included the injuries and deaths of motorists, cyclists, and pedestrians, we have focused only on the number of ‘persons’ killed or injured which include all three victim types collectively into a single feature.

After careful selection of the features which could prove beneficial for clustering in order to answer our research questions, the features we used in K-Means clusters are as follows:

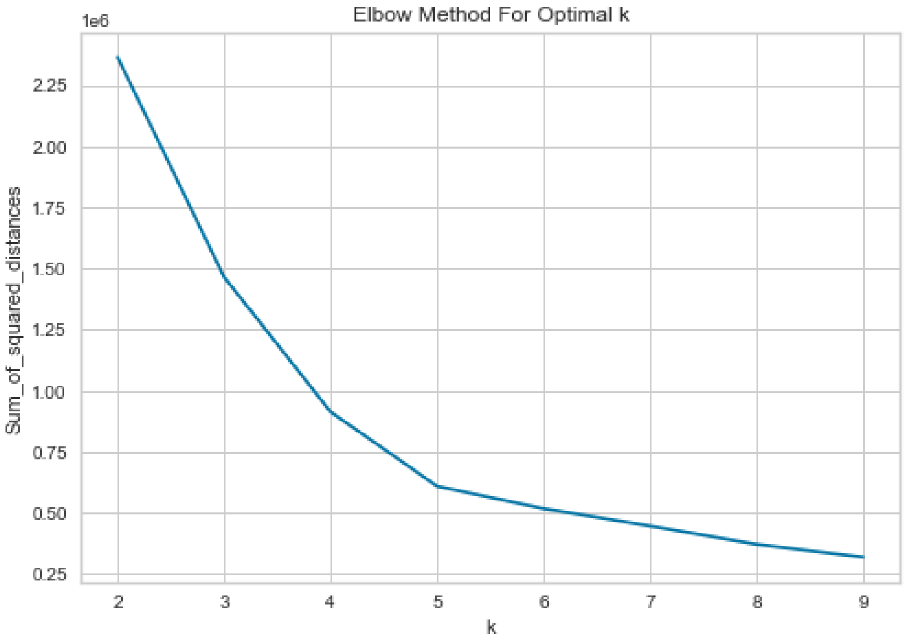
No.	Features
1	BOROUGH
2	LONGITUDE
3	LATITUDE
4	NUMBER OF PERSONS INJURED
5	NUMBER OF PERSONS KILLED
6	VEHICLE TYPE CODE 1
7	VEHICLE TYPE CODE 2
8	YEAR
9	MONTH
10	HOUR

[3] Using k-means seemed the most viable option here. However, it required significant feature selection and engineering which is discussed in the later sections.

Table 4: Features selected for clustering

Optimal Number of Clusters

In order to make sure k-means clustering gave the most accurate and meaningful clusters, it was necessary to start off by finding out the optimal number of clusters. For that we used the "Elbow Method". We performed k means clustering on the numerical features mentioned above for different values of k ranging from 2 to 9 and calculated their distortion scores i.e., sum of squared distances for each point from its assigned cluster's center. The point after which sum of squared distances seem to decrease (elbow) represents the optimal number of clusters. From the plot given below, this elbow appears to be at 5, therefore the optimal number of clusters for most our dataset is 5. In other cases, it turned out to be 4. In order to confirm this, elbow method was applied on various features, pair by pair as well and the plots showed similar results to the one shown below[4]:



[4] The optimal number of clusters for most cases were k=5. In other instances, the number of optimal clusters will be mentioned alongside.

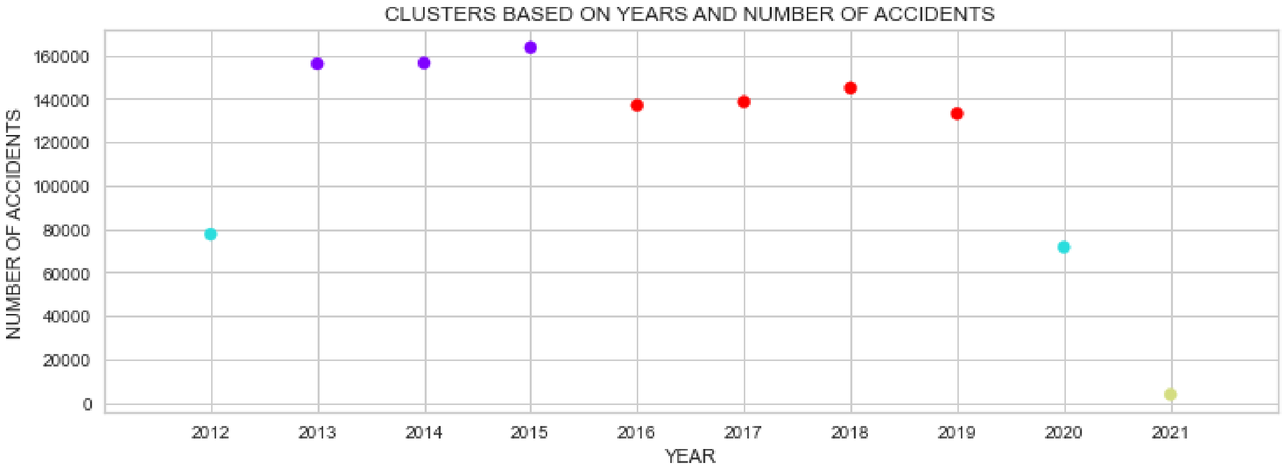
1. How did the severity of collisions of persons vary over time?

In this part, we have performed k-means clustering on various combinations of injuries and deaths of Persons grouped by time factors such as year, months, and hours of the day.

Number of Accidents over the Years

In order to analyze the severity of collisions over the years, we ran K-Means algorithm on "NUMBER OF ACCIDENTS", "NUMBER OF PERSONS INJURED", and "NUMBER OF PERSONS KILLED" grouped by "YEAR". The number of clusters used were 4.

Scatterplot for "NUMBER OF ACCIDENTS" over the "YEARS is shown below:

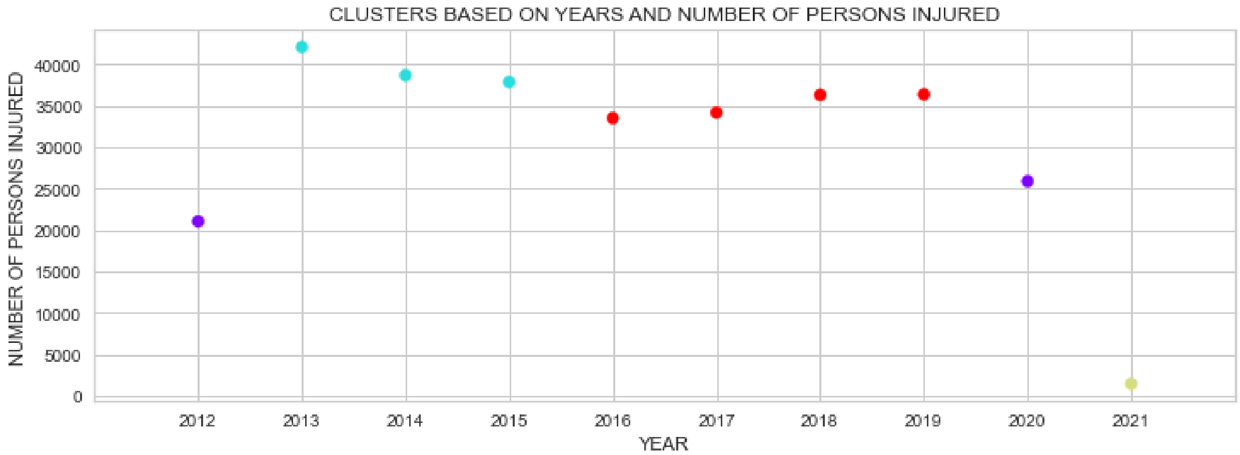


The clusters observed in scatterplot are as follows:

- **Purple Cluster** includes the highest number of accidents which occurred between the years **2013** to **2015**.
- **Blue Cluster** includes lesser number of accidents occurred in years **2012** and **2020**. In the EDA, we also observed that 2012 had lesser accident count as compared to other years possibly because of lesser data available regarding accidents. Small number of accidents in 2020 can be explained by lockdowns due to pandemic leading to lesser vehicular traffic.
- **Red Cluster** includes significantly higher number of accidents than the blue cluster, but lesser than the purple cluster. These accidents are spread over the years **2016** to **2019**.
- **Green Cluster** shows the least number of deaths because it includes the accidents happened in the current year.

Persons Injured over Years

Scatterplot for "NUMBER OF PERSONS INJURED" over the "YEARS" is shown below:



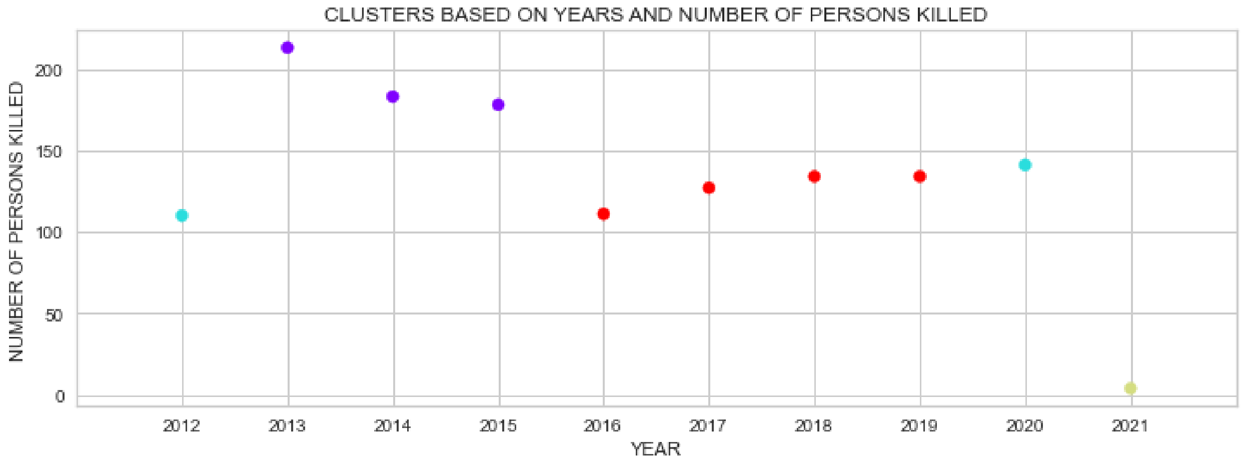
The clusters observed in scatterplot are as follows:

- **Blue Cluster** shows large number of injuries occurred over the years **2013** to **2015**
- **Red Cluster** shows comparatively lesser injuries than the blue cluster which may be due to the implementation of **Vision Zero Program** launched in 2014, and **COVID-19** pandemic which started around 2019. It includes years from **2016** to **2019**.
- **Purple Cluster** shows less injuries. It includes the years **2012** and **2020**.
- **Green Cluster** shows least number of injuries as it includes the current year **2021** whose complete data is not available yet.

Persons Killed over Years

Scatterplot for “NUMBER OF PERSONS INJURED” over the “YEARS” is shown below:

(Please turn over to the next page)



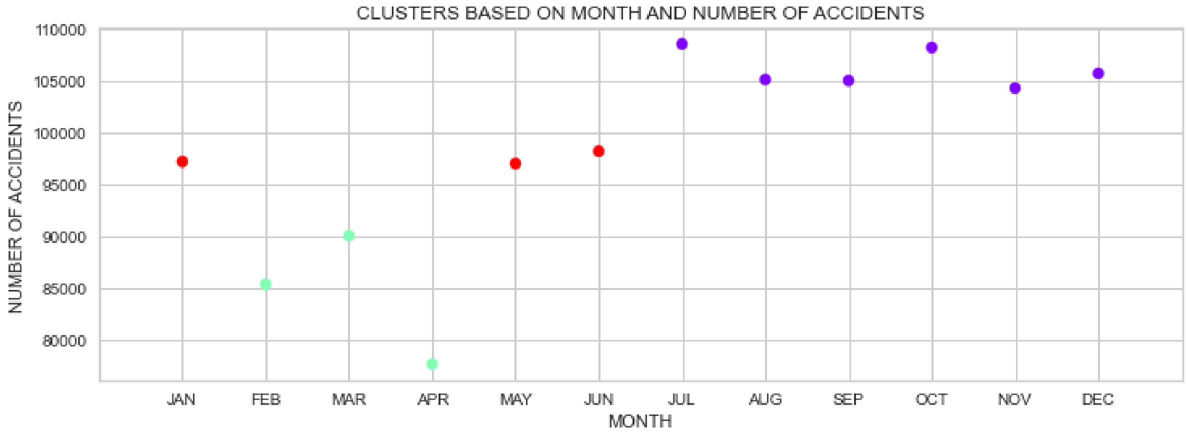
The clusters observed in scatterplot are as follows:

- **Purple Cluster** shows highest number of deaths occurred between years **2013** to **2015**.
- **Blue Cluster** shows lesser deaths than the purple cluster however it is within the range of red cluster. It includes the years **2012** and **2020**. The cluster point at 2020 shows **higher deaths** than the injuries which means that although 2020 had lesser number of accidents and injuries due to lesser vehicular traffic, but most of the accidents led to death of victims.
- **Red Cluster** shows less deaths and it includes the years from **2016** to **2019**.
- **Green Cluster** shows the least number of deaths as it includes the current year **2021**.

Number of Accidents over the Months

In order to analyze the severity of collisions over the months, we ran K-Means algorithm on “NUMBER OF ACCIDENTS”, “NUMBER OF PERSONS INJURED”, and “NUMBER OF PERSONS KILLED” grouped by “MONTH”. The number of clusters used were 3.

Scatterplot for “NUMBER OF ACCIDENTS” over the “MONTHS” is shown below:

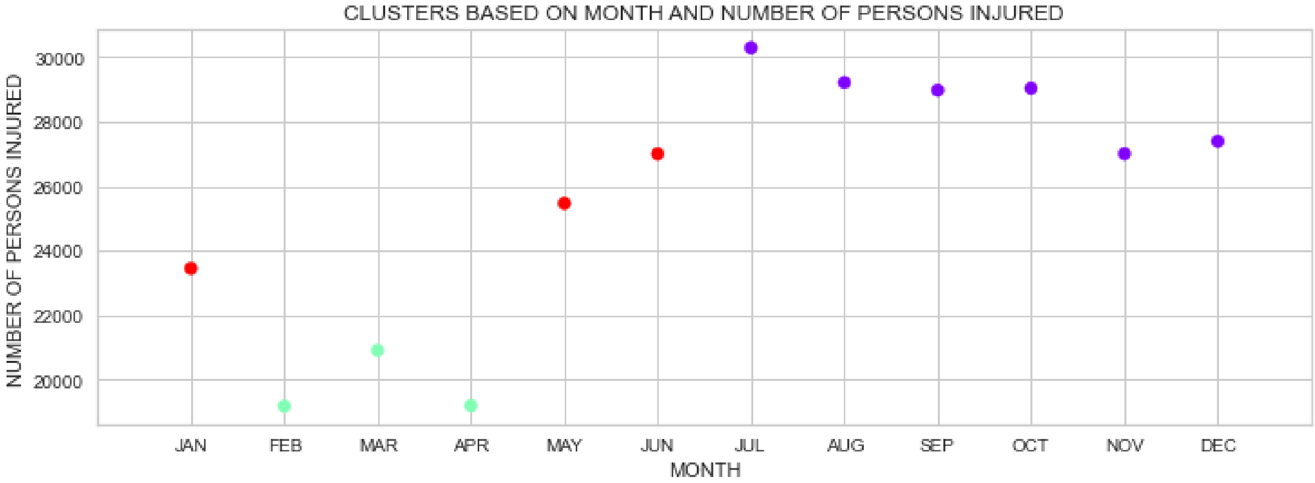


The clusters observed in scatterplot are as follows:

- **Purple Cluster** shows highest numbers of accidents happened within the months from **July** to **December**. The earlier months are extremely hot and September onwards, the weather starts transitioning to winter. Poor weather conditions such as high humidity in the summer months, and snowfall in the winter months can be the cause of accidents.
- **Red Cluster** includes months with median number of accidents which include **January, May**, and **June.** The weather in January may be too cold for people to go out, thus leading to lesser vehicular traffic on the road.
- **Green Cluster** shows the least number of accidents. This maybe because of stable weather conditions from **February** to **April**.

Persons Injured over the Months

Scatterplot for "NUMBER OF PERSONS INJURED" over the "MONTHS" is shown below:

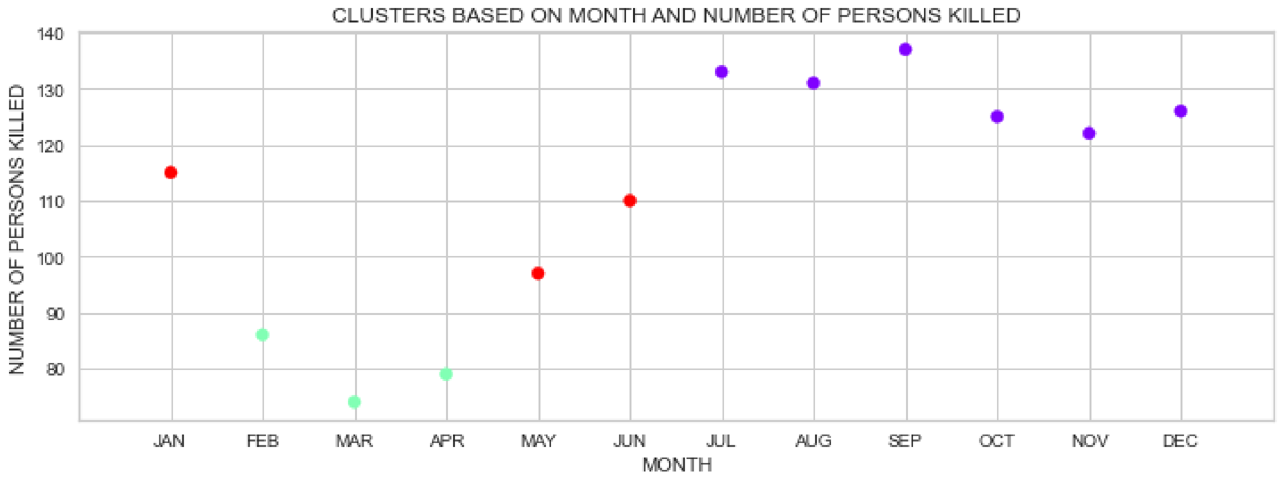


The clusters observed in scatterplot are as follows:

- **Purple Cluster** shows the highest number of injuries occurred in the months from July to December. These are the months with highest number of accidents.
- **Red Cluster** covers median number of people injured within the months January, May, and June.
- **Green Cluster** includes the months with least number of injuries, which are February, March, and April.

Persons Killed over the Months

Scatterplot for "NUMBER OF PERSONS KILLED" over the "MONTHS" is shown below:



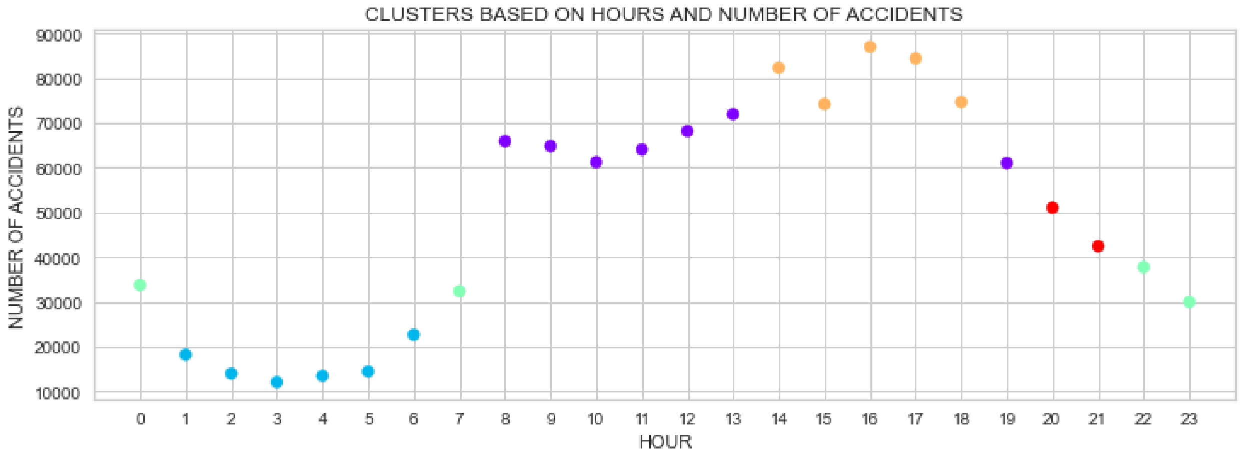
The y-axis scale for persons killed is lesser than that of persons injured. This is because accidents involving fatalities are quite less than those with injuries. The clusters observed in scatterplot are as follows:

- **Purple Cluster** consists of months from July to December. Most deaths happened within these months.
- **Red Cluster** consists of months with lesser numbers of deaths. These include January, May, and June.
- **Green Cluster** consists of months with least number of deaths. It includes the months of February, March, and April.

Accidents occurred on Different Hours of the Day

Next, we ran K-Means algorithm on "NUMBER OF PERSONS INJURED", "NUMBER OF PERSONS KILLED", and "ACCIDENT COUNT" grouped by hours of the day.

Scatterplot of clusters for "HOUR" and "NUMBER OF ACCIDENTS" is shown below:



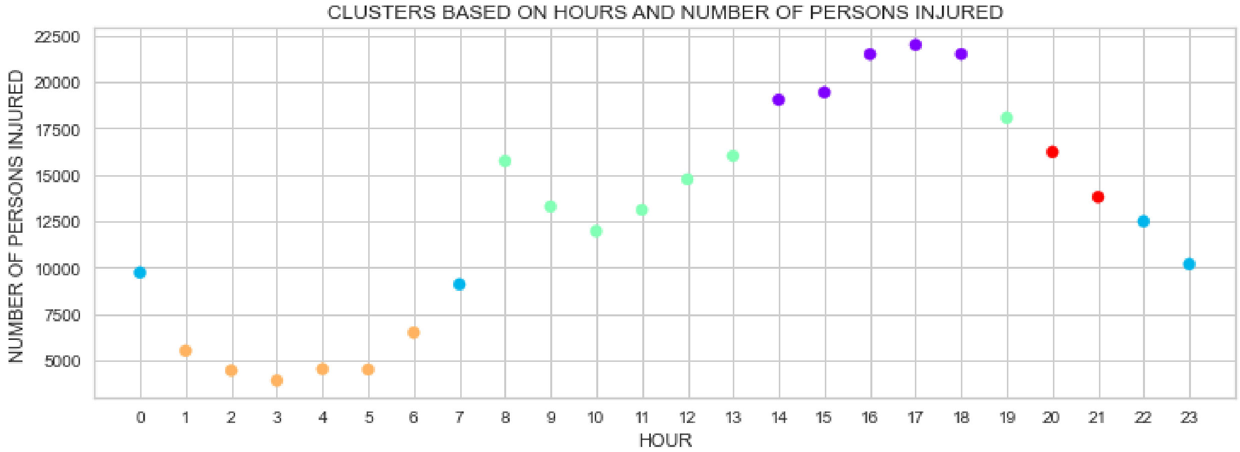
The clusters observed in scatterplot are as follows:

- **Yellow Cluster** shows hours with highest number of accidents. These include hours between **2PM** to **6PM**.
- **Purple Cluster** includes hours from **8AM** to **1PM** having the second highest number of accidents.
- **Red Cluster** covers accidents happened around **8PM** and **9PM**. These are lesser than the previous two clusters.
- **Green Cluster** has accidents within the range of 30k to 40k.
- **Blue Cluster** has the least number of accidents because of lesser vehicular traffic on roads early in the morning around **1AM** to **6AM.**
- These clusters represent lesser accidents late at night and early in the morning.

Persons Injured on Different Hours of the Day

Clusters formed by running K-Means algorithm on “HOUR” and “NUMBER OF PERSONS INJURED” are shown in the scatterplot below:

(Please turn over to the next page)

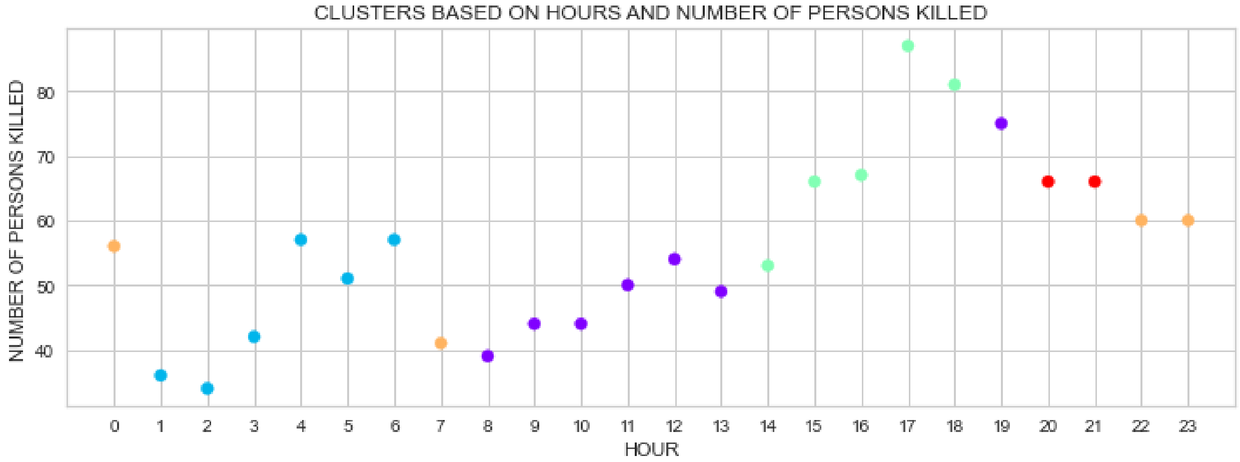


The clusters observed in scatterplot are as follows:

- **Purple Cluster** represents high number of injuries around the hours **2PM** to **6PM**.
- **Green Cluster** includes hours from **8AM** to **1PM**, and **7PM**. Injuries are lesser than the purple cluster.
- **Red Cluster** includes hours **8PM** to **9PM**. Injuries seem to decrease from **6PM** onwards.
- **Blue Cluster** is spread out around hours early in the morning and late at night. This cluster also represents lesser injuries as people are usually inside around this time.
- **Yellow Cluster** represents the least number of injuries due to lesser accidents early in the morning.

Persons Killed on Different Hours of the Day

Clusters formed by running K-Means algorithm on “HOUR” and “NUMBER OF PERSONS KILLED” are shown in the scatterplot below:



This plot also has less points than injuries vs hours because accidents involving fatalities are quite less than those with injuries. The clusters are as follows:

- **Blue Cluster** represents early morning collision deaths.
- **Yellow Cluster** is spread out across the plot representing some deaths in the early morning and some late at night.
- **Purple Cluster** represents morning to early afternoon collision deaths.
- **Green Cluster** represents death from early afternoon to evening.
- **Red Cluster** represents night collision deaths.

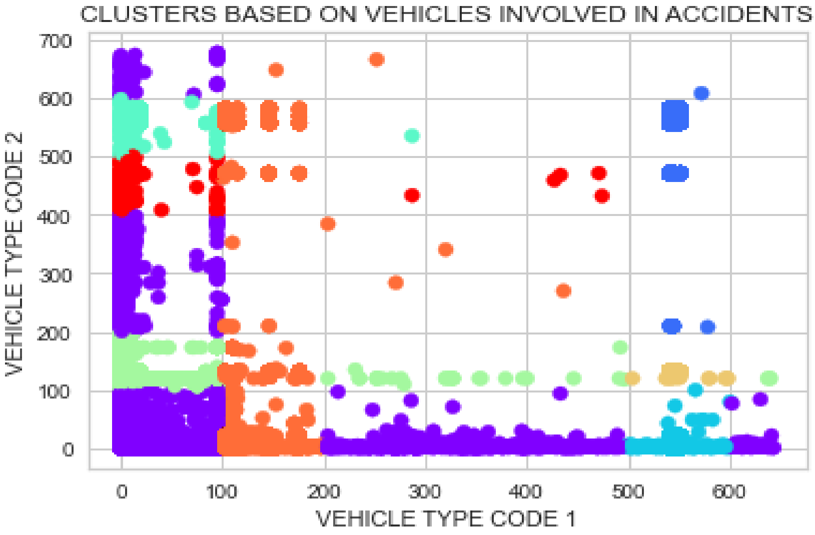
2. How were the vehicles involved in accidents in NYC over the years 2012 to 2021?

In this section, we have performed K-means clustering on various combinations of features including vehicle types, injuries and deaths of persons, and boroughs.

Vehicles Involved in Accidents

To analyze which vehicles had frequently collided with each other in NYC accidents throughout the years 2012 to 2021 through clustering, we had to consider Vehicle Type 1 and Vehicle Type 2 as the main features. Since vehicle types were categorical in nature, k-means could not be applied without some preprocessing. Therefore, we employed **binning** and **one-hot encoding** for vehicle type code 1 and 2. We made seven bins by grouping values (0-100,100-200,200-300,300-400,400-500,500-600,600-700) based on the decreasing popularity of vehicles. These were then one hot encoded. As the above plot shows we see some interesting clusters. Most of the clusters are based on the specific vehicle types.

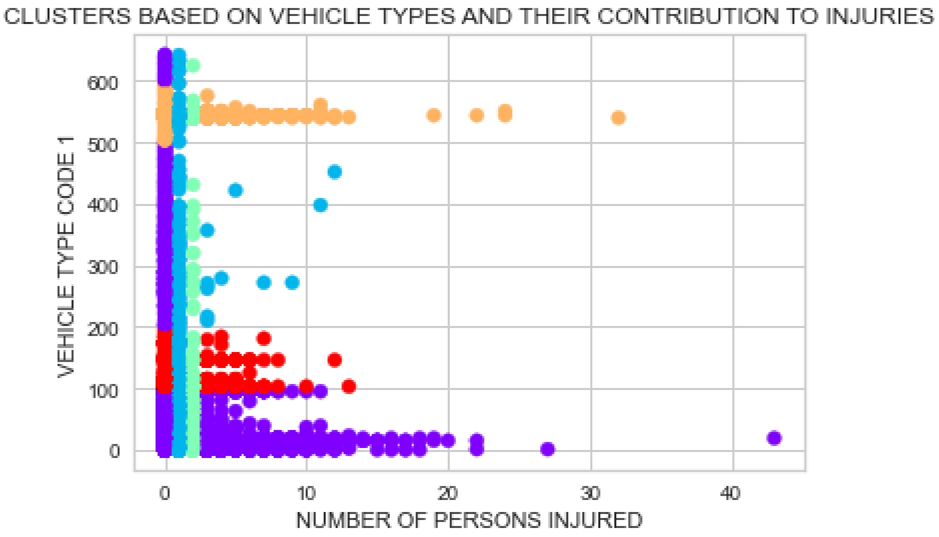
Overall, common vehicles like sedan, passenger vehicles, taxi, bikes had a lower code while less common vehicle types had higher codes. Clusters formed by running K-Means algorithm on the types of the two vehicles colliding are shown in the plot below:



From the plot, it can be seen that the clusters are mostly spread along y axis where vehicle type 2 has smaller codes and along x axis where vehicle type 1 has smaller codes. We listed the codes of vehicles which were most frequently involved in accidents above and most of them belong to the lesser range of codes (Bins 1 to 3) and the scatterplot shows a similar picture. Both vehicle types having smaller codes frequently collided with all other vehicles. Only the small com and large com vehicles have codes around 500 which can be observed in the blue cluster around 540.

Persons Injured by Vehicles

Clusters formed by running K-Means algorithm on the types of vehicles and the number of people injured are shown in the scatterplot below:

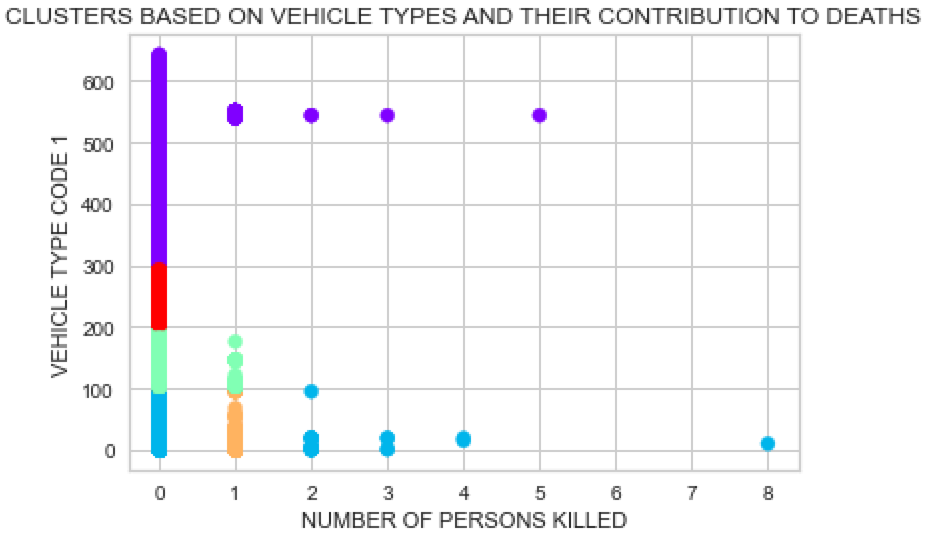


We observed in the previous section that vehicles having codes around 0 to 30 were most involved in accidents. Here

- **Purple Cluster** is the sparsest and it contains vehicle codes around 0. Range of injuries is 0-25. There seems to be one instance where a vehicle belonging to this cluster injured around 45 people.
- **Red Cluster** having vehicle codes ranging from 100 to around 200, and injuring 0 to around 15 people.
- **Orange Cluster** contains vehicles having codes between 500 to 600. It also contains vehicles that have injured a range of people from 0 to around 12 with a few outliers will 32. We saw two vehicles above “Large com veh” and “small com veh” belonging to this range and the behavior of the blue cluster at this point may be explained by these vehicles to some extent.
- **Blue Cluster** is denser than the previous clusters. Vehicles in this cluster mostly seem to have injured 1 person.
- **Blue Cluster** contains only accidents with two injuries

Persons Killed by Vehicles

Clusters formed by running K-Means algorithm on the types of vehicles and the number of people killed are shown in the scatterplot below:

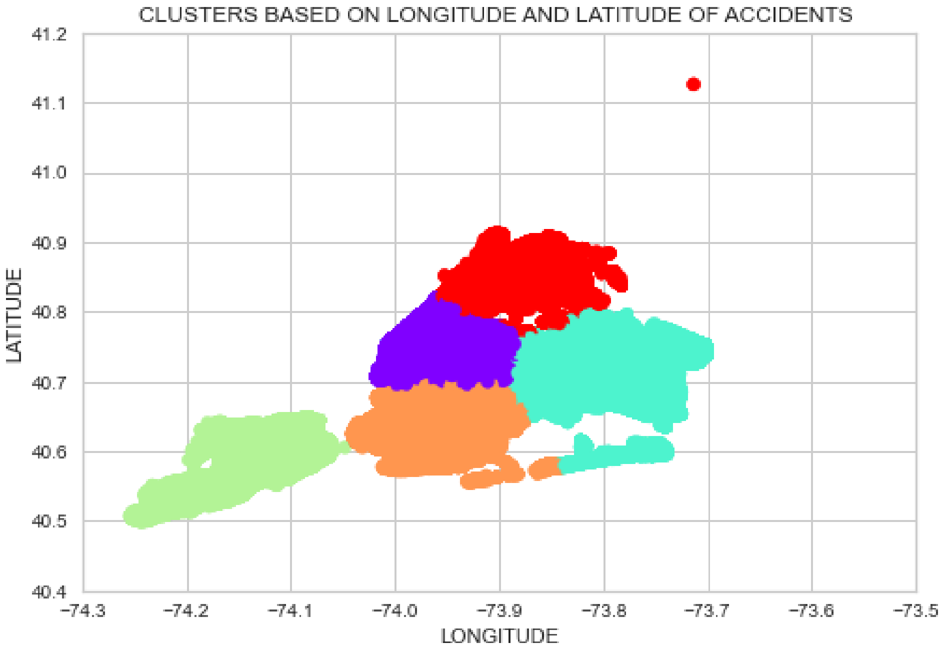


This scatterplot above of the clusters of people killed by different vehicles is sparser than the previous plot because number of persons killed were far less than those injured: The five clusters are as follows:

- **Purple Cluster** containing vehicle type codes from 300 to 600 which killed 0 to 5 people.
- **Blue Cluster** containing vehicles above 0-100 killed 0 to 4 people. The outlier around 8 shows that 1 or more vehicles belonging to this cluster killed 8 people in one accident.
- **Green Cluster** contains vehicles having codes around 100 and 200 and these vehicles mostly killed no people but, in some cases, killed 1 to 2 people.
- **Red Cluster** contains a wider range of people a vast majority of which killed nobody. These vehicles belong to bins with codes between 200 and around 300.
- **Yellow Cluster** is the densest among the 5 clusters. Most of the vehicles belonging to this cluster killed 1 person.

3. How were the accidents spread throughout NYC?

Here, we are analyzing the longitude and latitude of accidents through K-means clustering to figure out the location-wise spread of accidents in New York City,



The five clusters (depicted by one color each) represent the five boroughs of New York City where the accidents took place. Let's have a look at them one by one:

- **Green Cluster** is Manhattan with average coordinates of 40.7, -74.8.
- **Red Cluster** is Bronx with 40.85, -73.85.
- **Purple Cluster** is Queens with coordinates 40.75, -74.95.
- **Orange Cluster** is Staten Island with coordinates 40.55, -74.15.
- **Blue Cluster** is Brooklyn with coordinates of 40.65, -73.9. The most concentrated clusters are Manhattan and Brooklyn where most accidents occur.

Surprisingly, the clusters that we formed on the bases of latitudes and longitudes were somewhat similar to the map of the boroughs in World Atlas which indicates that the collisions were more or less spread out in the throughout each borough.

Overall, clustering enabled us to answer the questions posed earlier one by one and finally reach a conclusion that indeed, there are certain very obvious trends and groupings in the collisions in New York City which further strengthened the claims we made in the earlier phases.

Conclusion and Recommendations

By looking at such a comprehensive data set, preprocessing it and then finally conducting exploratory data analysis, frequent pattern mining and clustering on it, we derived several important findings which may be useful in many ways. We determined the time-based and area-based trends of vehicular accidents in New York and tried to correlate them with plausible reasons. We also found out the main causes of the accidents and their severity. We had a bird's eye view of the type of victims involved in the accidents and were shocked by the difference in percentages in injuries and deaths of people based on their mode of transportation. Finally, we not only drew correlations between the causes and effects of accidents overall but also calculated region-wise correlations separately and drew meaningful comparisons.

All the analysis and findings were backed up by secondary research which then enabled us to chalk down the following policy suggestions to curb accident rates in New York:

- Special attention should be paid to the policy-making in accident-prone boroughs and the boroughs which are bigger in size. These include Brooklyn, Queens and Manhattan.
- People travelling through streets and boroughs where there are higher accident rates should be alerted about the potential risk factors.
- Strict traffic policing and measures should be observed during office hours (I.e., 9am - 5am) on working days and on all hours on public holidays and busy days.
- Since winter weather (in the latter months of the year) was accompanied by an increased number of accidents, clearing streets and vehicular path (to cater for snow and fog) should be mandated by the relevant authorities.
- A combination of disaster relief and traffic policies should be prepared to cater for extra-ordinary circumstances where a large number of people are injured or killed. (A few such instances were observed in our analysis)
- In order to alleviate the number of injuries in pedestrians, sidewalks and crossings should be declared a safe space for them. First aid should be available on all accident-prone or busy sites.
- Passenger vehicles, being the most common, are most prone to accidents, that too severe in nature. Therefore, a separate set of policies focused on passenger vehicles (and similar vehicles) should be devised and enforced.
- Just like contributing factors of accidents are rigorously recorded in Manhattan and Bronx, the factors should be observed and recorded in other boroughs as well so that the root cause of collisions can be identified and curbed.

Limitations

Although the data and its intent were thorough and it provided a multi-dimensional view about vehicular accidents in New York, we found a few loop holes and areas of improvement in it. The analysis can only be as good as the quality of data provided. Once catered for, the quality of analysis and implications can be improved drastically. Following are few limitations that we found in the dataset:

- In addition to numerous missing values in the location (borough, latitude, longitude), there was an approximation in the location I.e., the location was not 100% accurate. Measuring accurate information about the boroughs and coordinates will enable better region and area-based analysis of accidents. Another approach could be to have geo-coded information in the data instead of mere names.
- Missing and unspecified values in fields like contributing factors hinder us from completely understanding the cause of accidents, leading to biases and assumptions to prevail.
- There is a room for several other important features to be included in the data which are not present in the data for now. These may include more detailed victim information, exceptions, disasters etc.
- As evident from the results of frequent pattern mining, most of the contributing factors of collisions were not recorded in the data. In order to devise apt conclusions, it is necessary to have a clear idea of the causes of collisions. Therefore, overall, the quality of records can be improved so that the issue can be curbed from the roots.

Way Forward

The utility of the dataset should not only be limited to time or crash intensity-based analysis. Future work on the data can also incorporate correlation between crime-data and accidents data in New York city (an aspect which is currently missing in it). The correlation between certain months and increase in accident count also hints towards possible weather-based analysis of the accident data. Similar analysis can be extended to accident counts on special events or national holidays. Recording features about the victims (e.g., gender, age, occupation) may also reveal meaningful insights in this regard.

An impressive and promising factor about the dataset that we found online is that the dataset is updated daily at New York Open Data. With such a robust data collection system, there can be even more significant deductions from the data from a research point of view with the recommendation stated above. This, however, would require considerable time and effort.

References

The references mentioned here are in the order in which they were referred to in the report:

Mueller, Benjamin, et al. “Terror Attack Kills 8 and Injures 11 in Manhattan.” *The New York Times*, The New York Times, 31 Oct. 2017, www.nytimes.com/2017/10/31/nyregion/police-shooting-lower-manhattan.html.

“U.S. Census Bureau QuickFacts: New York City, New York; Bronx County (Bronx Borough), New York; Kings County (Brooklyn Borough), New York; New York County (Manhattan Borough), New York; Queens County (Queens Borough), New York; Richmond County (Staten Island Borough), New York.” *Census Bureau QuickFacts*,

Rosenberg, Stack et al. “One Dead and 22 Injured as Car Rams Into Pedestrians in Times Square.” *The New York Times*, The New York Times, 18 May. 2017,

CBS New York. “Brooklyn Bus Accident Leaves At Least 41 Injured.” *CBS New York*, CBS New York, 9 Sept. 2013, newyork.cbslocal.com/2013/09/09/brooklyn-bus-accident-leaves-at-least-6-injured/.

“United States.” *New York, United States Zip Codes*, worldpostalcode.com/united-states/new-york/.

Kiprop, Victor. “The Boroughs of New York City – NYC Boroughs Map.” *WorldAtlas*, WorldAtlas, 17 Dec. 2018, www.worldatlas.com/articles/the-boroughs-of-new-york-city.html.