CS 432/536: Data Mining

Course Project - Group 6

Motor Vehicle Accidents in New York City

Deliverable 1: Exploratory Data Analysis

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 analyze the prevalence of vehicular accidents in NYC – their causes, effects and trends throughout the city. There will also be some recommendations about key insights for variables to consider that could shape future analysis.

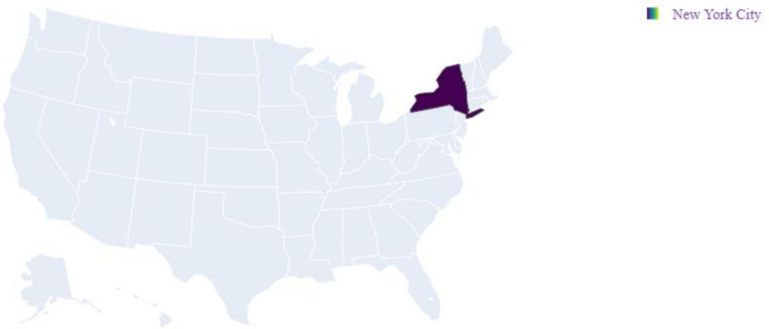


Figure 1: World Map and New York City

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:

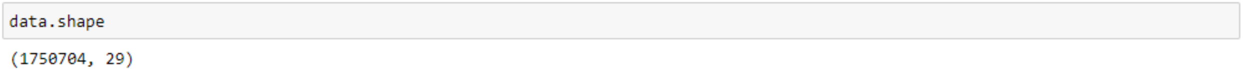
- What is the area-wise trend of accidents in NYC? What inferences can we draw from accident prone boroughs?
- What are the trends of accidents in NYC over different periods of time? Can we make any predictions about the future trends?
- How did the severity of accidents vary throughout NYC?
- What are 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.

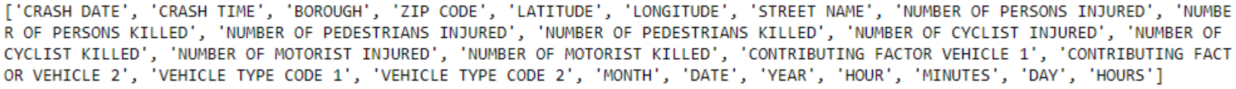
After importing relevant libraries that we needed for our analysis, we loaded our data into a data frame called data. To get an idea about the amount of data and diversity of features we are dealing with, we will first check the shape of data i.e., the number of rows and columns it.



```
data.shape
```

```
(1750704, 29)
```

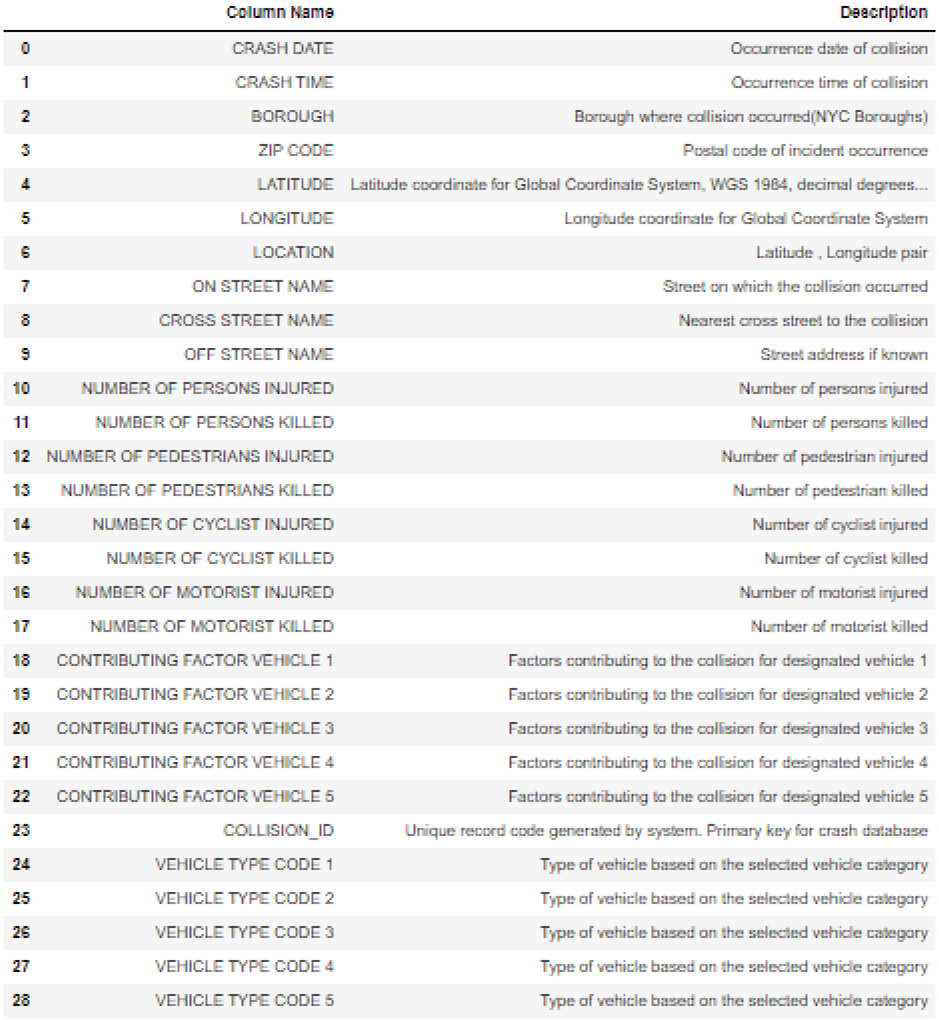
The data consists of more than 1.7 million records of accidents throughout New York and each accident is described by 29 features. It is important to enlist the features here:



```
['CRASH DATE', 'CRASH TIME', 'BOROUGH', 'ZIP CODE', 'LATITUDE', 'LONGITUDE', 'STREET NAME', 'NUMBER OF PERSONS INJURED', 'NUMBE
R 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 FACT
OR VEHICLE 2', 'VEHICLE TYPE CODE 1', 'VEHICLE TYPE CODE 2', 'MONTH', 'DATE', 'YEAR', 'HOUR', 'MINUTES', 'DAY', 'HOURS']
```

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
0	CRASH DATE	Occurrence date of collision
1	CRASH TIME	Occurrence time of collision
2	BOROUGH	Borough where collision occurred(NYC Boroughs)
3	ZIP CODE	Postal code of incident occurrence
4	LATITUDE	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees...
5	LONGITUDE	Longitude coordinate for Global Coordinate System
6	LOCATION	Latitude , Longitude pair
7	ON STREET NAME	Street on which the collision occurred
8	CROSS STREET NAME	Nearest cross street to the collision
9	OFF STREET NAME	Street address if known
10	NUMBER OF PERSONS INJURED	Number of persons injured
11	NUMBER OF PERSONS KILLED	Number of persons killed
12	NUMBER OF PEDESTRIANS INJURED	Number of pedestrian injured
13	NUMBER OF PEDESTRIANS KILLED	Number of pedestrian killed
14	NUMBER OF CYCLIST INJURED	Number of cyclist injured
15	NUMBER OF CYCLIST KILLED	Number of cyclist killed
16	NUMBER OF MOTORIST INJURED	Number of motorist injured
17	NUMBER OF MOTORIST KILLED	Number of motorist killed
18	CONTRIBUTING FACTOR VEHICLE 1	Factors contributing to the collision for designated vehicle 1
19	CONTRIBUTING FACTOR VEHICLE 2	Factors contributing to the collision for designated vehicle 2
20	CONTRIBUTING FACTOR VEHICLE 3	Factors contributing to the collision for designated vehicle 3
21	CONTRIBUTING FACTOR VEHICLE 4	Factors contributing to the collision for designated vehicle 4
22	CONTRIBUTING FACTOR VEHICLE 5	Factors contributing to the collision for designated vehicle 5
23	COLLISION_ID	Unique record code generated by system. Primary key for crash database
24	VEHICLE TYPE CODE 1	Type of vehicle based on the selected vehicle category
25	VEHICLE TYPE CODE 2	Type of vehicle based on the selected vehicle category
26	VEHICLE TYPE CODE 3	Type of vehicle based on the selected vehicle category
27	VEHICLE TYPE CODE 4	Type of vehicle based on the selected vehicle category
28	VEHICLE TYPE CODE 5	Type of vehicle based on the selected vehicle category

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.

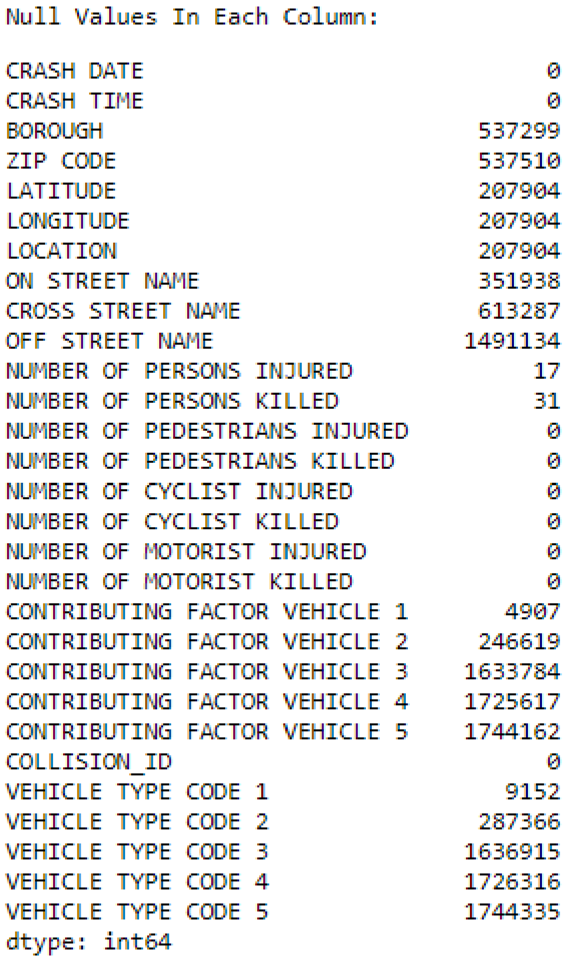
Data Cleaning

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

Data cleaning is a pivotal step in the data science cycle. 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.



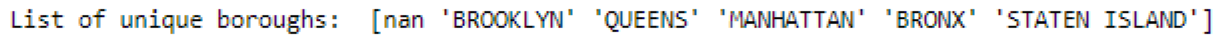
```
Null Values In Each Column:

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
```

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. We will have to drop the records where borough, longitude, latitude is NAN together. (since we know nothing about the location) 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.

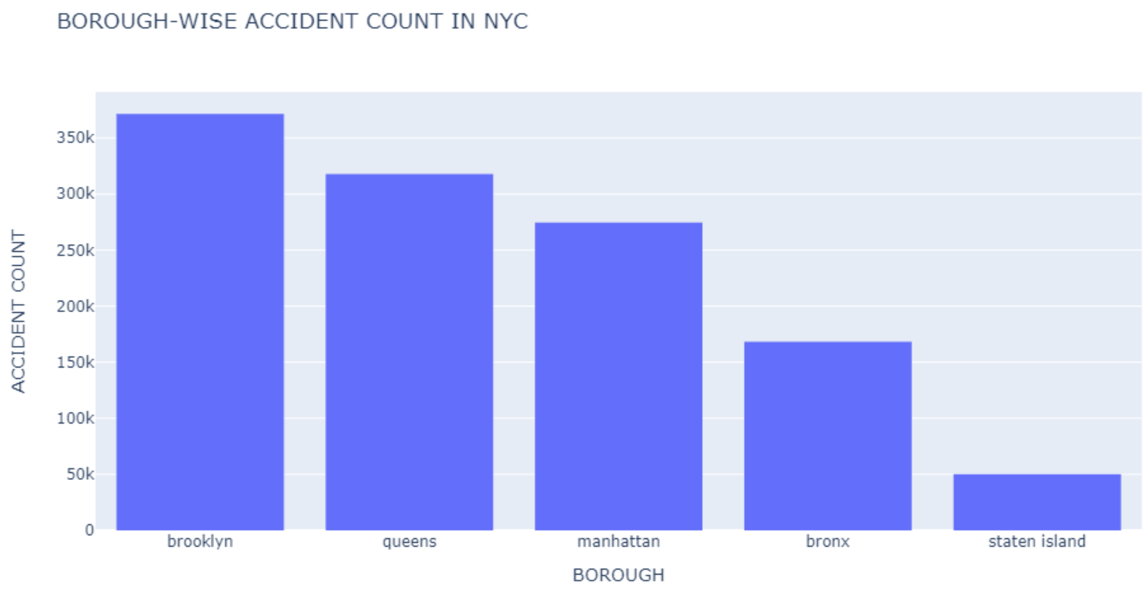
Null Values in Borough

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.



We tried to find the boroughs using latitudes and longitudes, but it results in unpredictable results i.e., boroughs which had higher accident count first moved to lower places resulting in potentially biased side. Therefore, instead of filling the data with uncertain values, we just drop the null records.

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 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 we drop NAN rows of persons injured and number of persons killed.

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

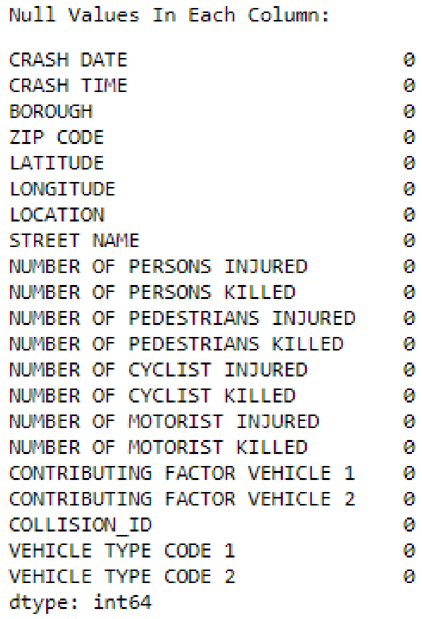
Vehicle Type 1 and 2 (along with their contributing factors) seemed more usable with less Nan and faulty values. Therefore, we will be dropping Vehicle 3,4,5, and their contributing factors because we have decided to use Vehicle 1 and 2 for our analysis.

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'.

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. 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
```

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.

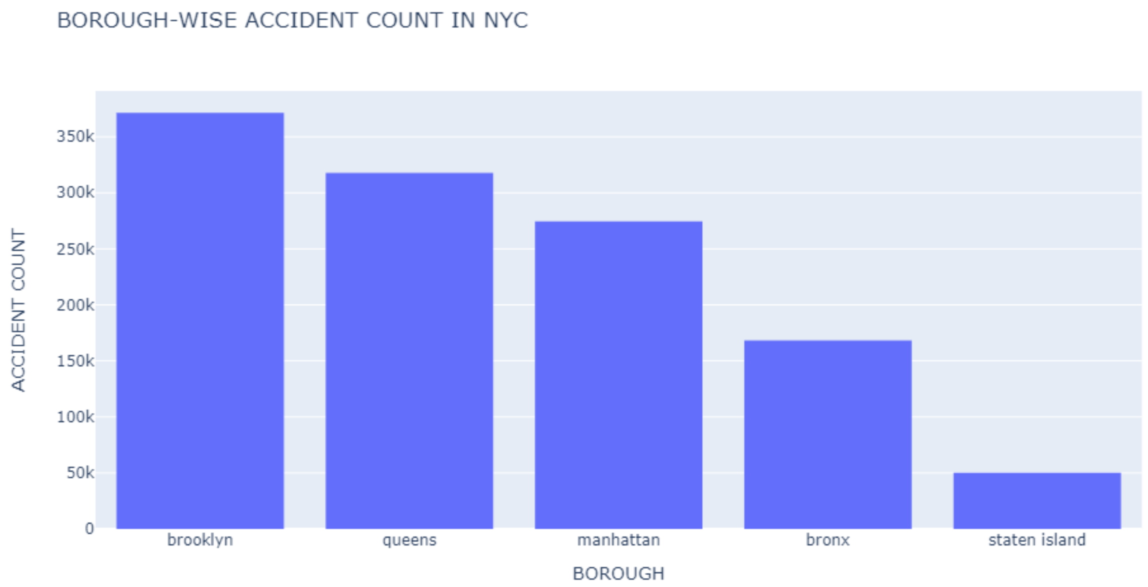
We then had a look at the final cleaned data. It was non-redundant, non-null and usable! This marked the end of data preprocessing stage after which we carried out detailed analysis on the cleaned data.

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 (as mentioned earlier), 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.

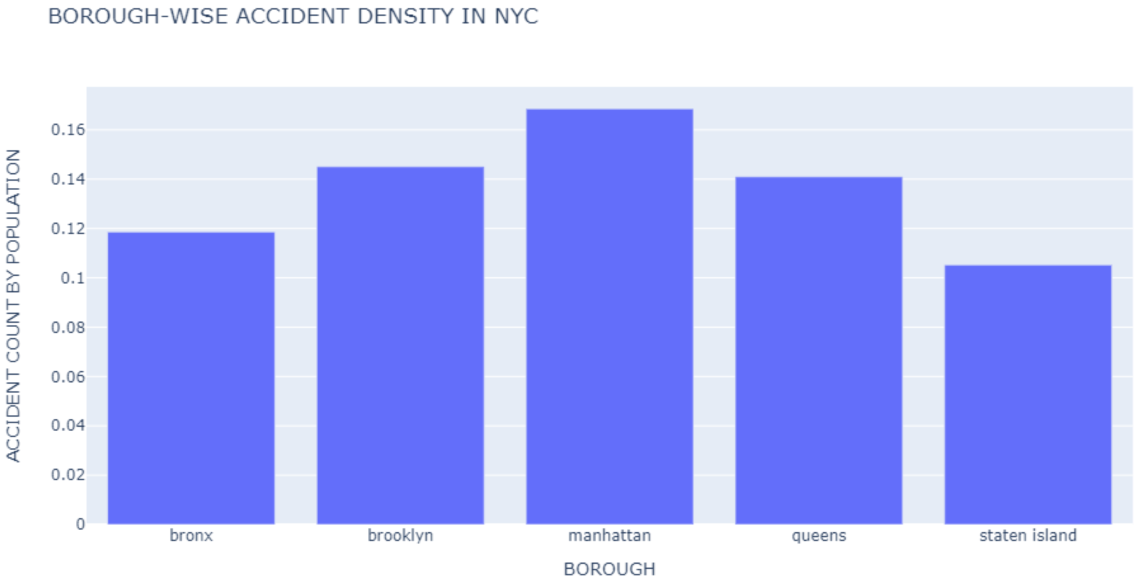
1. What is the area-wise trend of accidents in New York and what inferences can we draw from accident prone boroughs?

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 possible 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.



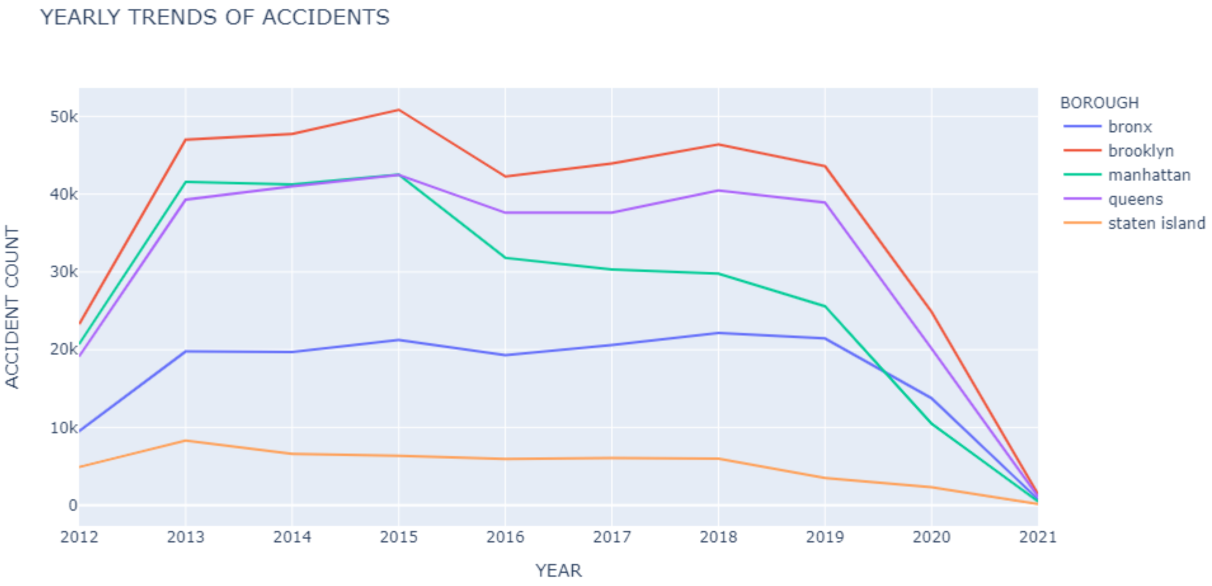
For this plot, we referred to the United States Census Bureau's census for the year 2019. We have assumed that the current population sizes of the five boroughs follow a similar trend as that of 2019. In 2019, Brooklyn had the highest population followed by Queens, and then Manhattan. Staten Island had 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.

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.

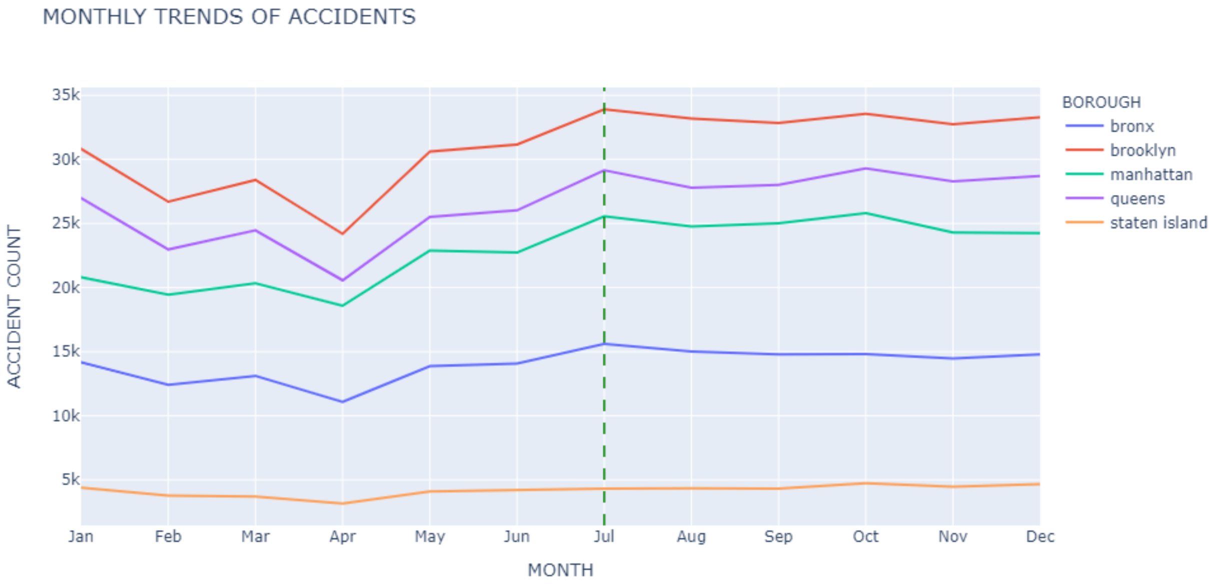


- Firstly, **Staten Island** has had the least number of accidents over the years. This is because its population is the smallest among the five boroughs. Moreover, the accidents in Staten Island also seem to decrease after the year 2013.
- **Brooklyn**, as seen above, has the highest number of accidents 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.



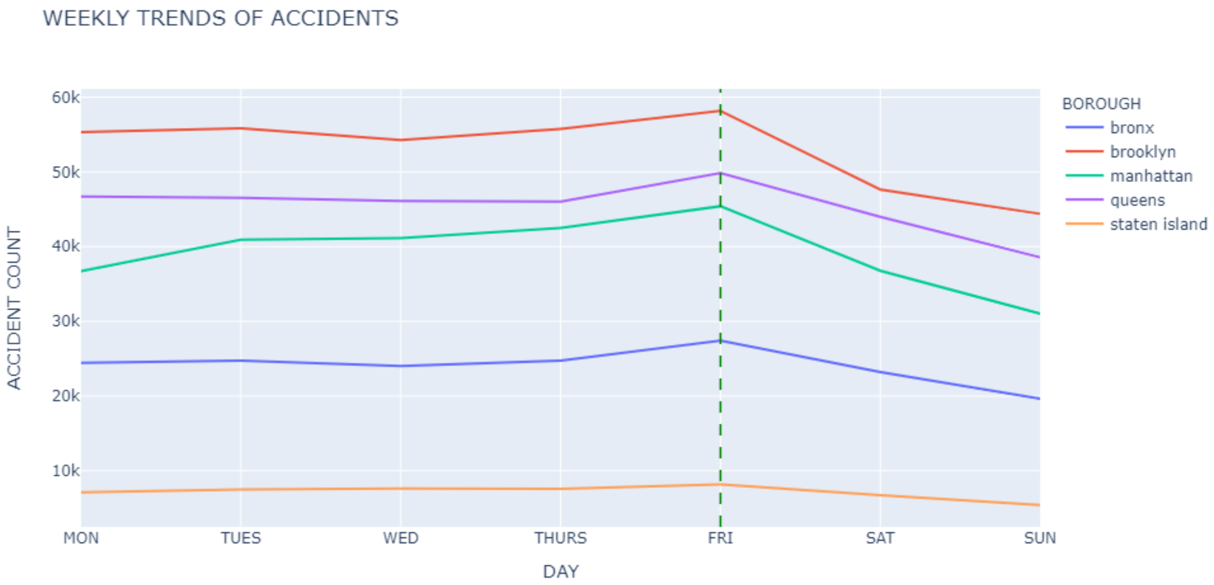
Here are a few intriguing observations that we found:

- 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 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. 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

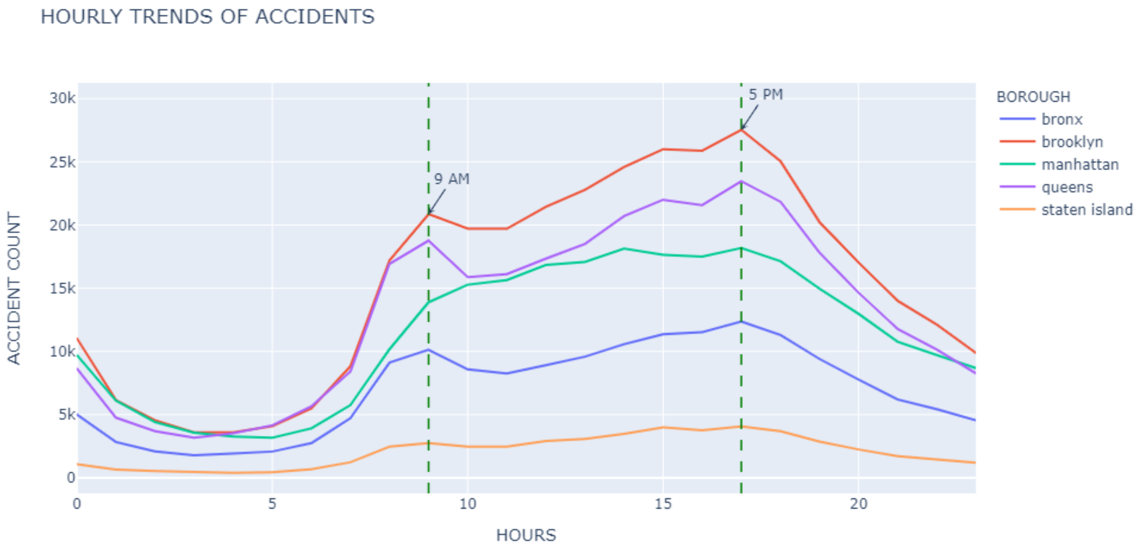
In order to determine the changes in the count of accidents on different days of the week, we again drew a similar plot which shows the trends of borough-wise accidents throughout the week:



- 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.

Hourly trends of accidents

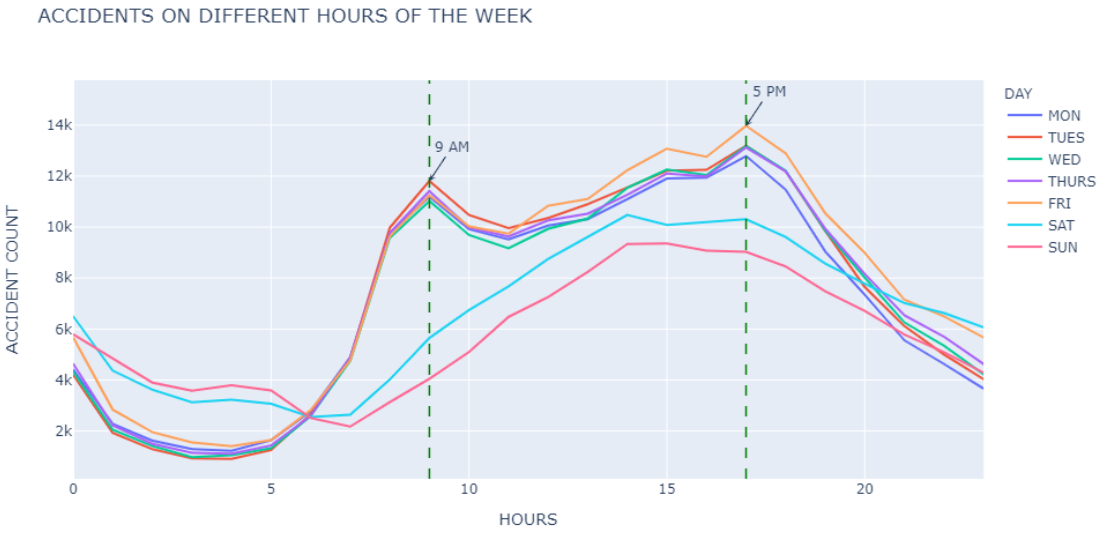
We all know how some hours of the day are rush hours. We also know that they are somewhere around the morning (school and office timings) and somewhere around early evening. To confirm this speculation, the plot below shows the hourly trends of borough-wise accidents throughout the day.



- 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.

- 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?

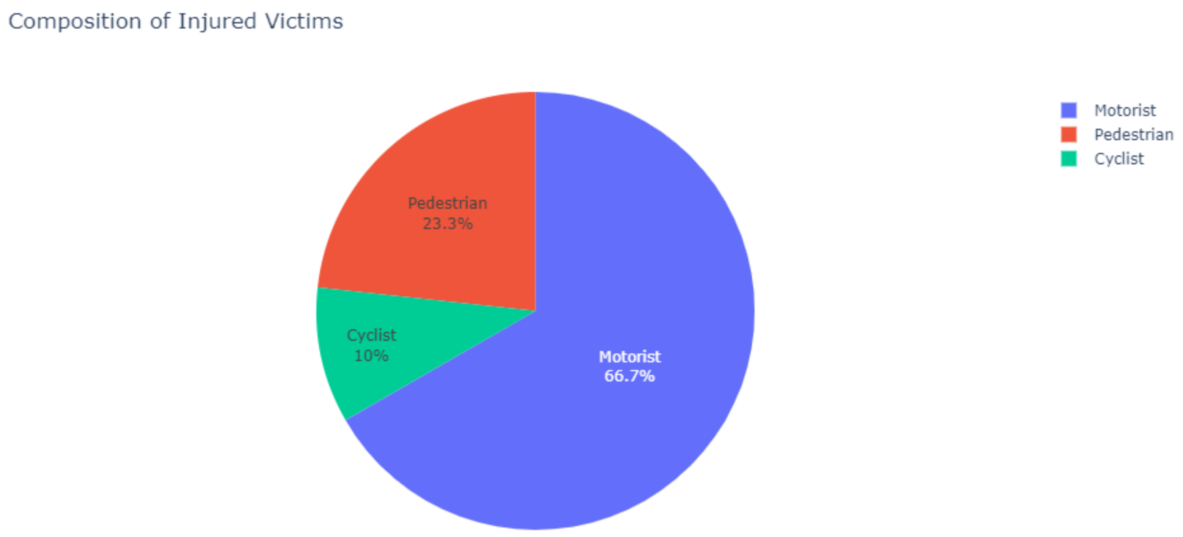
We were given multiple columns hinting towards the severity of accidents. They not only include the count of people injured and killed in each accident but also the type of victim (pedestrian, cyclist, motorist). Let's first look at the type of people injured and killed in accidents.

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

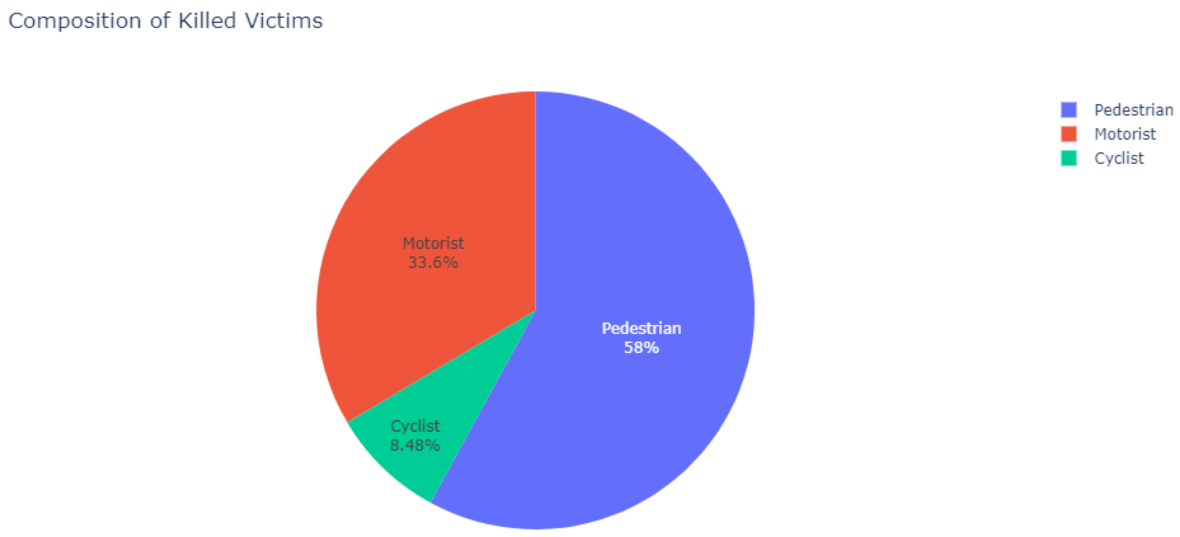
The pie plot above shows the composition of different victim types who were injured in the accidents in New York over the years 2012 to 2021 and also enables us to compare the ratios:



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.

- ### Composition of victims who were killed

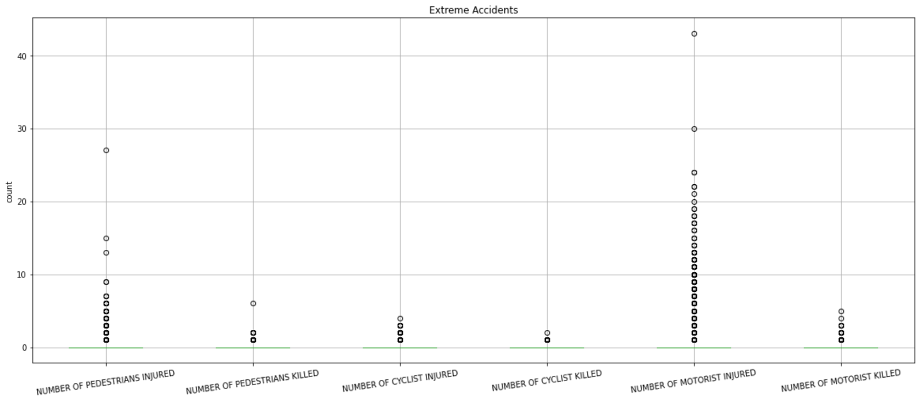
In case of the victims who were killed in the respective accidents, the pie plot below shows the composition of different victim types who were killed in the accidents in New York over the years 2012 to 2021.



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 data has two major outliers, one 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.

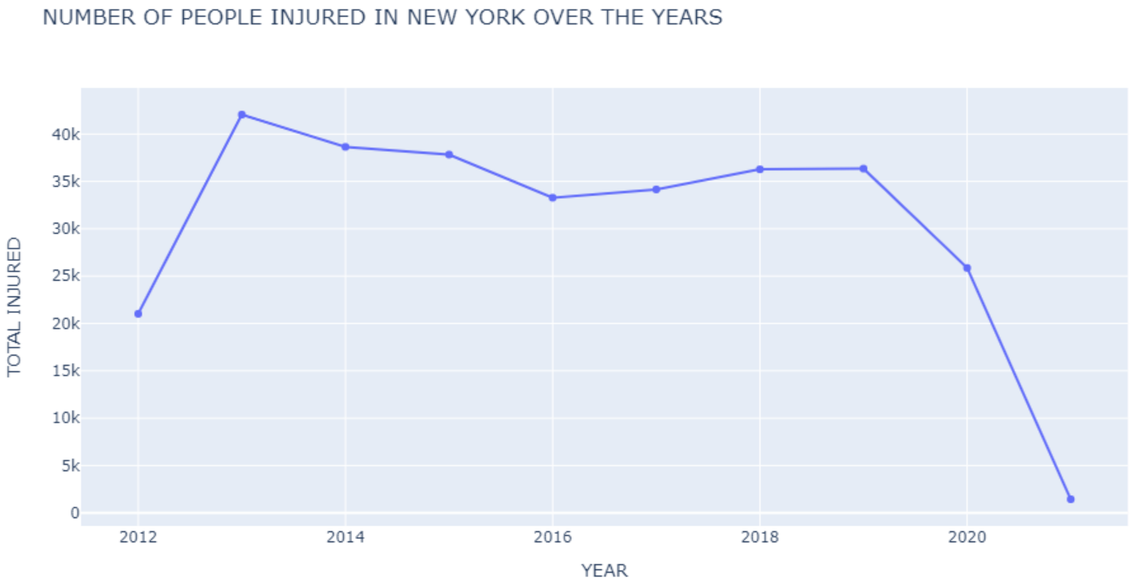
The second one happened in Brooklyn on September 9,2013 when a car crashed into a bus. 40+ people were injured.

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

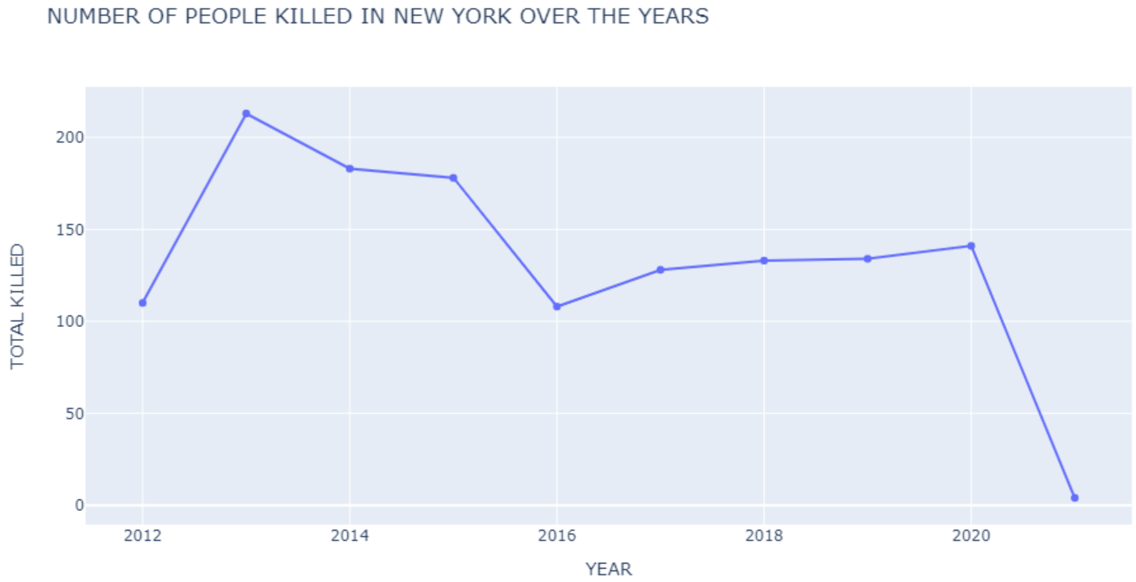
The plot below shows the number of people injured in New York accidents over the years 2012 to 2021.



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 lets 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.

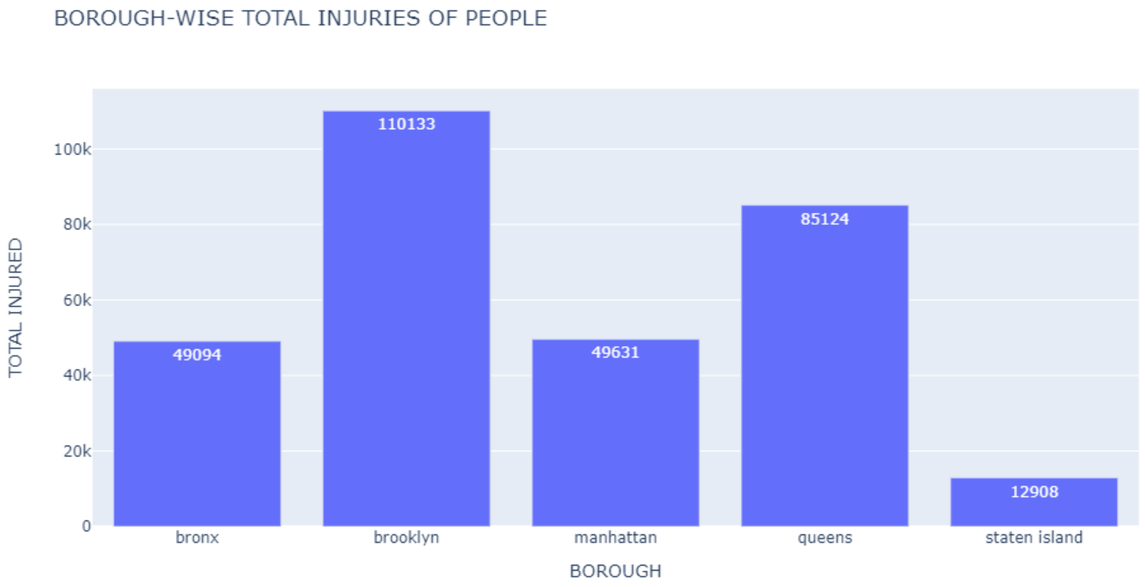


Some important observations from the plots above are listed below:

- 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.

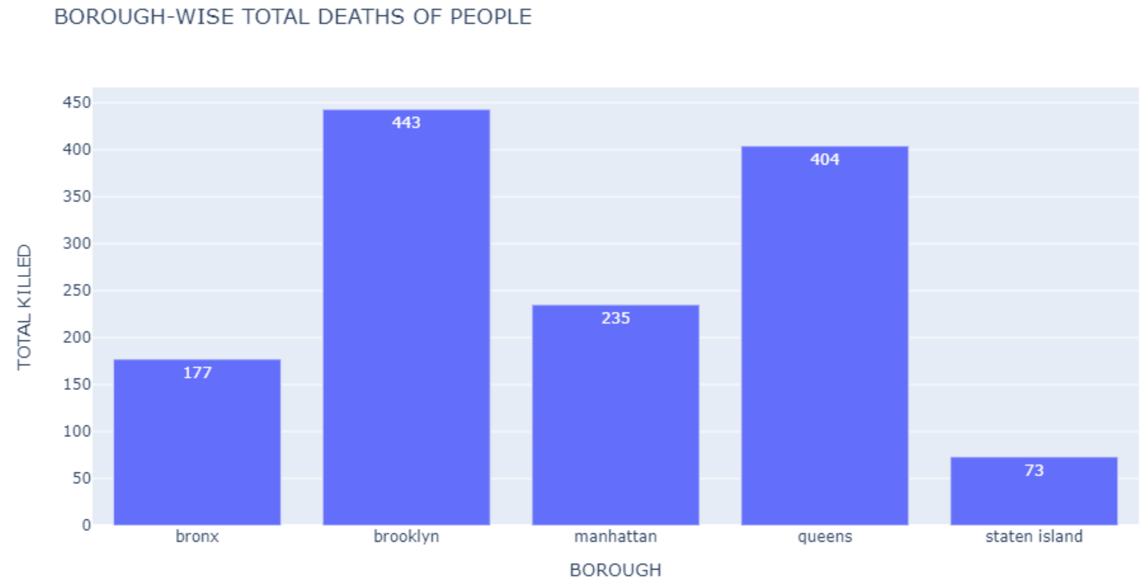
Borough-wise overall injuries

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:

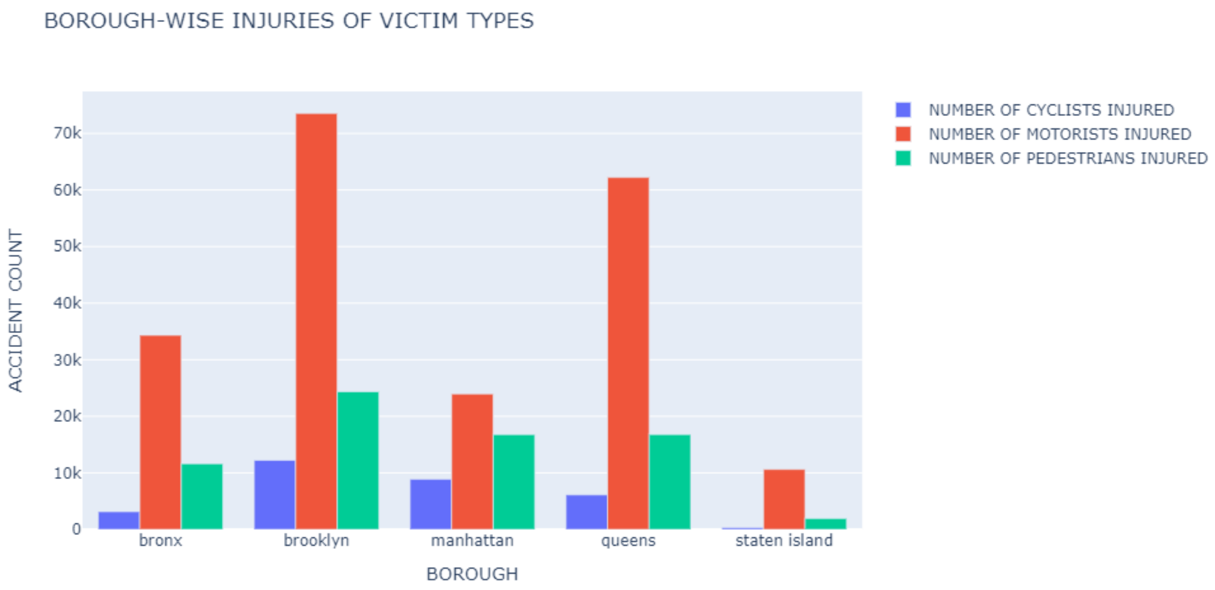


Some common observations which are in accordance with the previous plots are as follows:

- 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

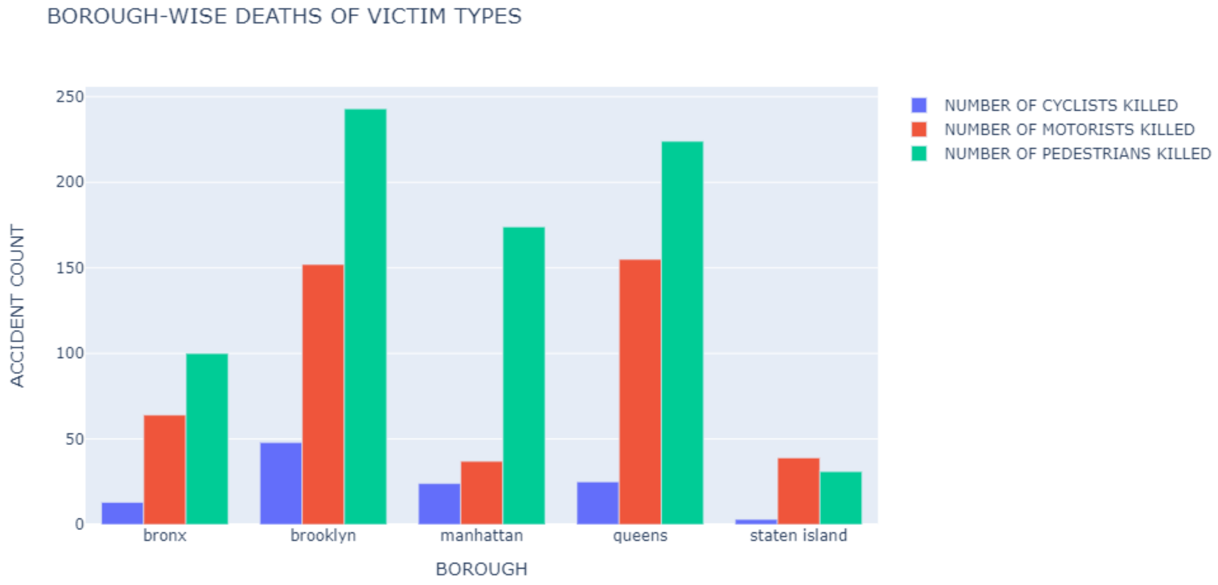


In the general trend of injuries in New York, we observed that the injured people mostly consisted of **motorists**. A similar trend can be seen in the five boroughs of New York. Following observations were made from the plot above:

- 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.

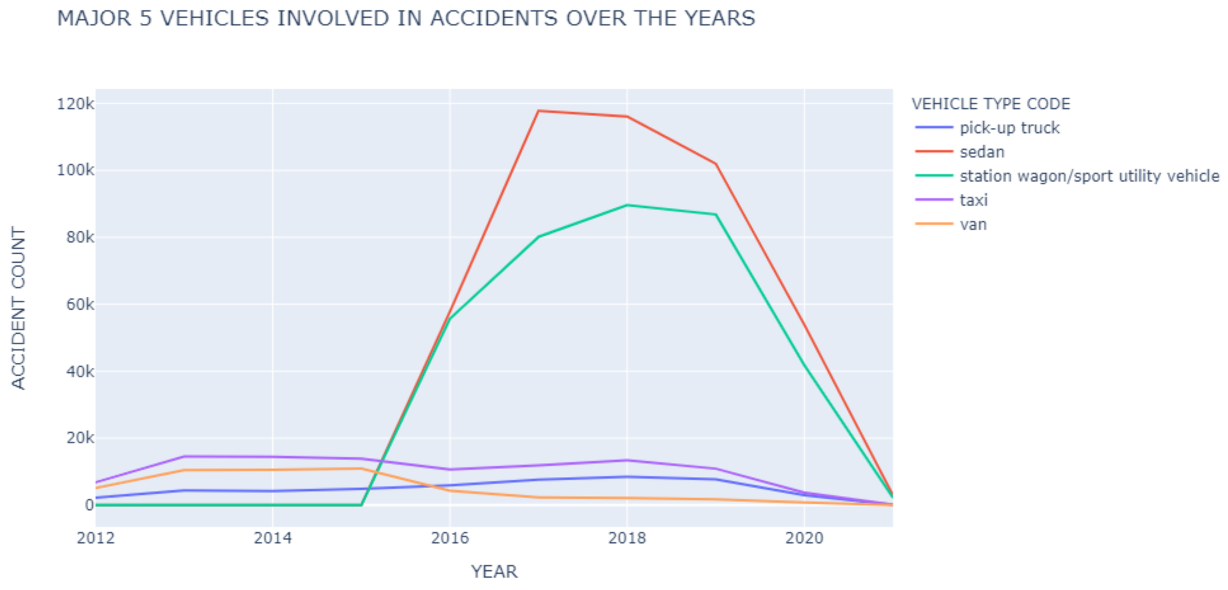
- Brooklyn again has the highest number of pedestrians killed.
- Contrary to other boroughs, Staten Island has slightly more number of motorists killed than pedestrians.

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.

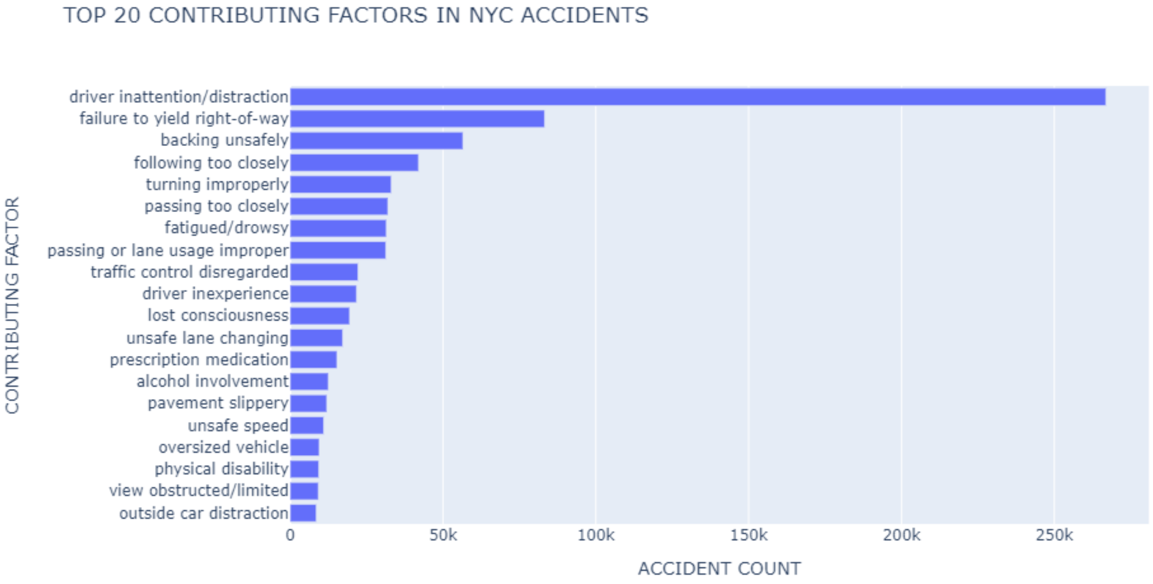


The plot above shows how the trend of the top 5 vehicles' involvement in accidents changed over the years.

- 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:

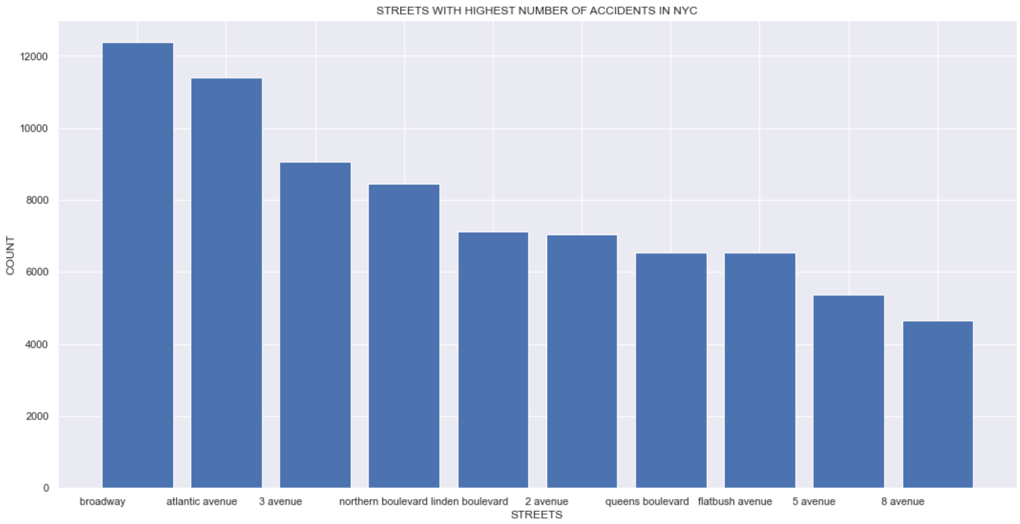


The plot above shows the top most common causes of accidents in New York. We have filtered the "unspecified" and "other vehicular" causes as they are vague and do not give any meaningful information to help our analysis.

- 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

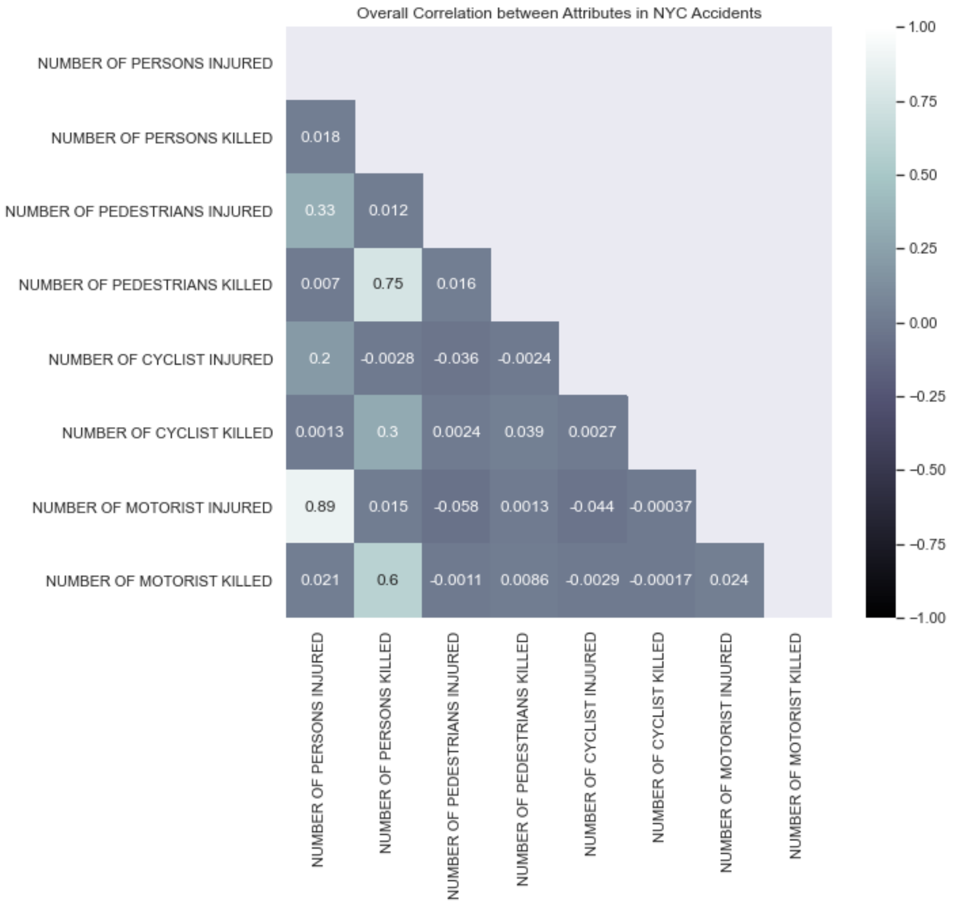
We would like to see which streets are the most dangerous in New York City for both accident count and fatalities. To investigate we formed a pivot table and sorted values by accident count. We then used the python group by function for street name and number of persons killed. This was sorted in ascending 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 correlated with each other. Persons killed and injured is basically a sum of all other attributes and thus showing positive strong correlations. However, 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.

Conclusion

By looking at such a comprehensive data set, preprocessing it and then finally conducting exploratory data analysis 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.

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.

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/.