Motor Vehicle Accidents in New York City

04.25.2021

—

Anusheh Zohair Musatafeez, Hajira Zaman, Danyal Haroon & Ahmed Hassaan Mirza

Data Mining

Table of Contents

1. Overview

Our dataset consisted of land vehicle accident data from New York City from 2012-2020 comprising 1.5 million values. The data was collected from the five different boroughs of the city and contained date and time information, location information, information about the frequency and categories of victims and information on culprit vehicles and the contributing factor behind the accident.

This dataset can be used to learn more about which type of vehicles cause the most accidents, what times of the day are accident rates higher, which type of accidents result in more fatalities as well as many other meaningful insights. This information can be used by policy-makers to introduce new regulations to help reduce the number and severity of accidents in New York City.

2. Data Cleaning and Exploratory Data Analysis (EDA)

2.1 Data Preprocessing

1. After loading our data into a pandas dataframe, we **set *Collision ID* as the index** because it is the primary key of our data.
2. Next, we checked for **NULL values** in every 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
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

3. We **filled NULL values** in these columns with appropriate filling techniques corresponding to the nature of the data.

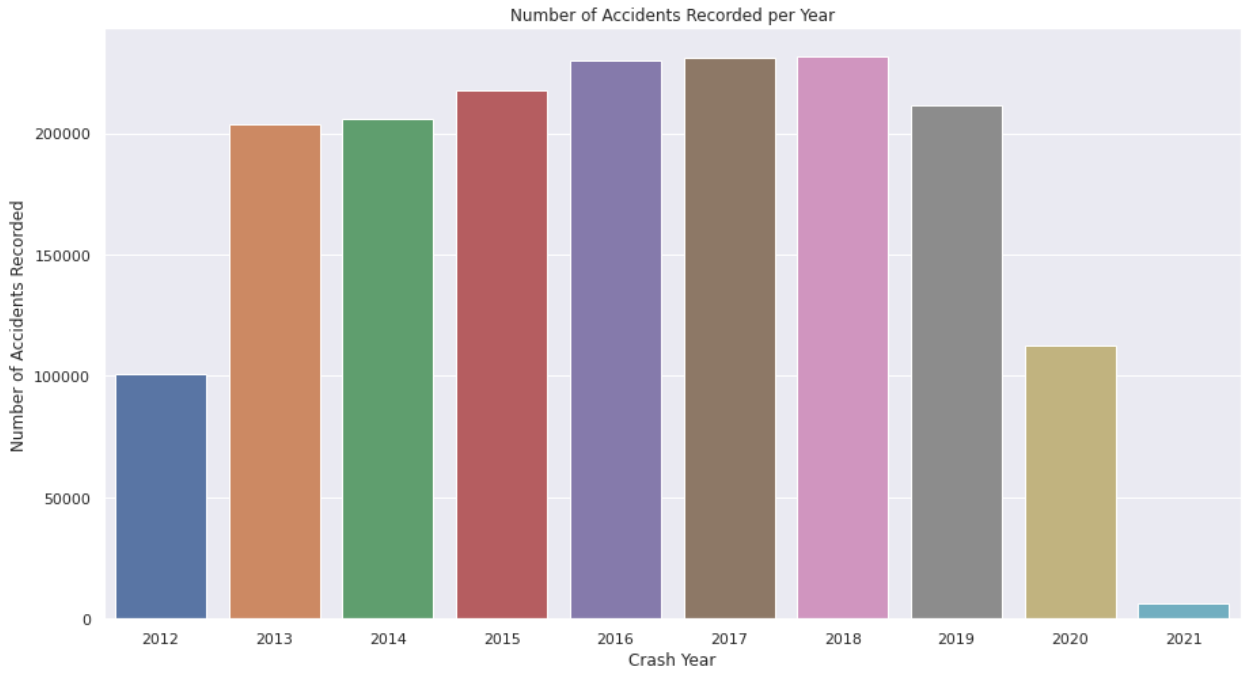
   We found the missing number of persons killed and injured by adding up the motorists, pedestrians and cyclists.

4. We also changed all cases of vehicle records to lowercase to remove discrepancies.
5. Rows with missing values remaining were deleted from the data frame.
6. We **deleted duplicate records**.
7. This left us with 1,541,487 records, which is 88% of our original number of records.

2.2 Answering Interesting Questions through EDA and Visualization

1. **Number of Accidents Recorded in each Year**

   We counted the number of accidents each year to see that there were approximately 200,000 accidents per year other than 2012 and 2021 (complete years not covered in the dataset), 2020 (Covid pandemic struck). The accidents drastically decreased in the two years of the pandemic.

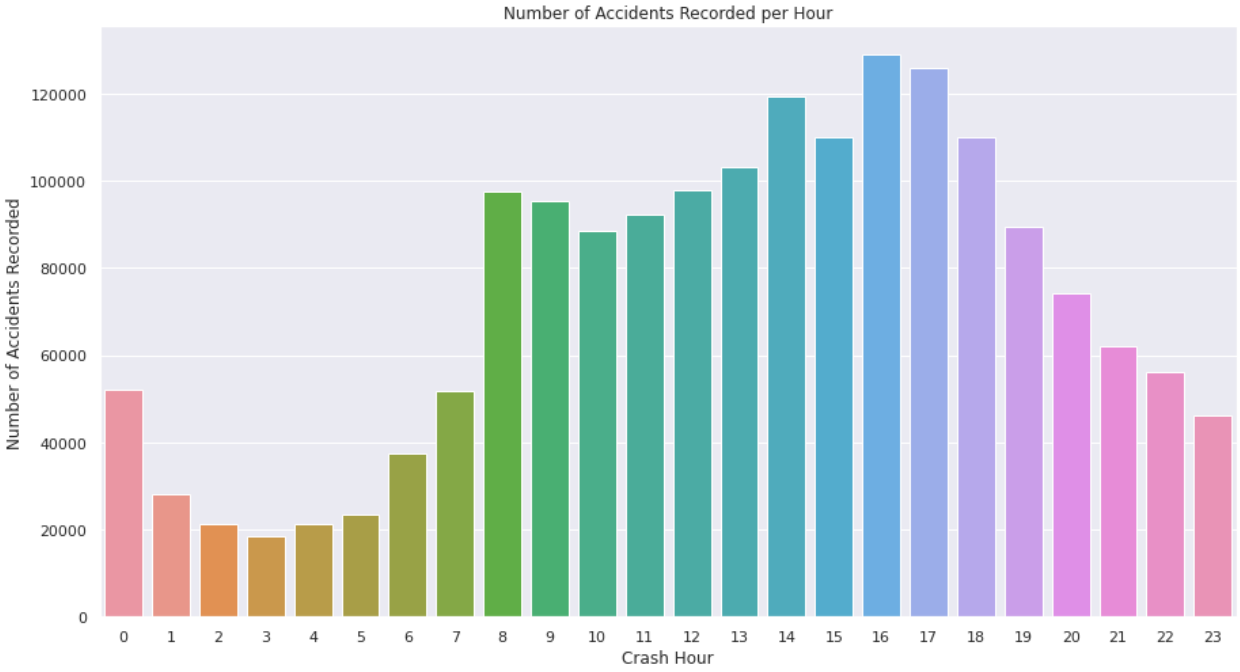


2. **Average Number of Accidents per Year from 2012 - 2020**
   Excluding 2021, we found the average number of accidents per year to be **218,082.**

3. **Visualising Number of Accidents Recorded at each Hour**

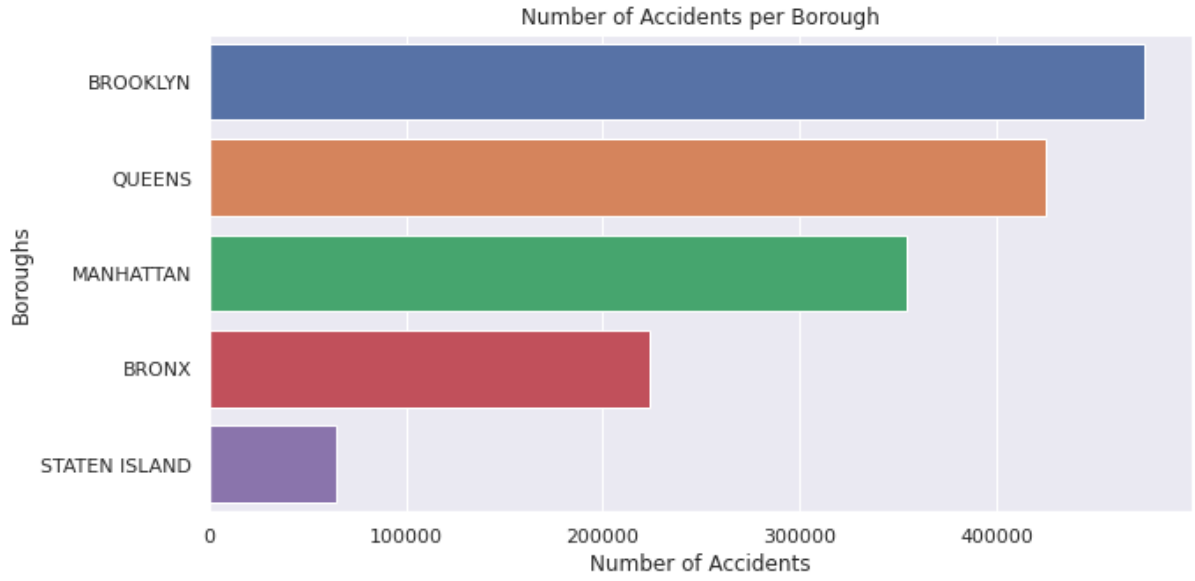
We converted the timestamp data given to datetime to then study the amount of accidents across times of the day.

The data shows that accidents are most likely to happen between 2pm and 7pm. Similarly we can see that accidents are least likely to happen between 1am and 7am.

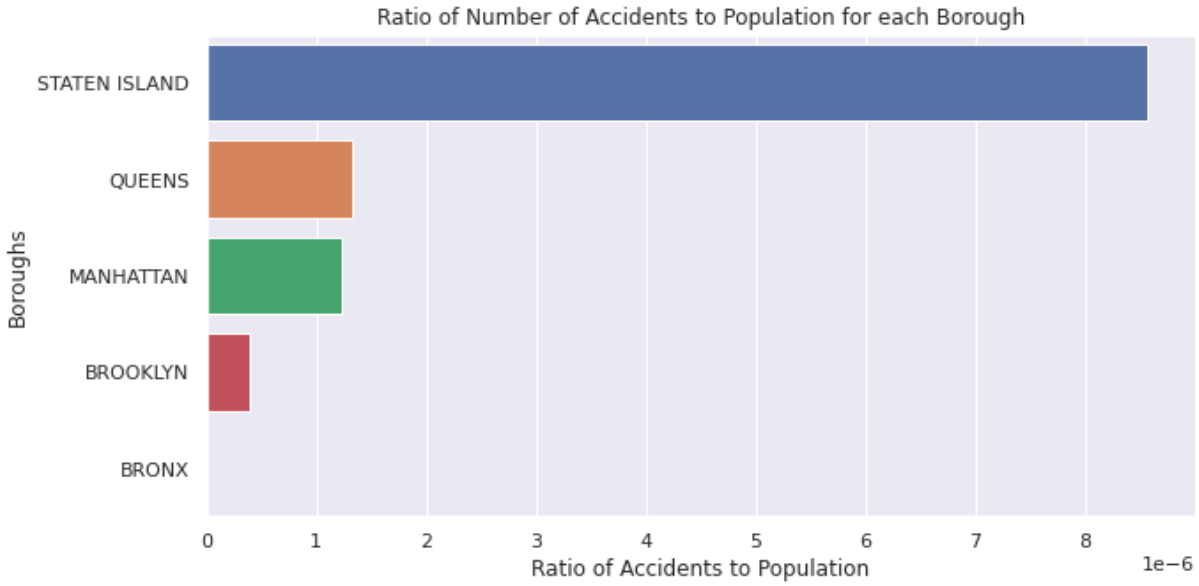


4. Visualising Number of Accidents Recorded at each Borough

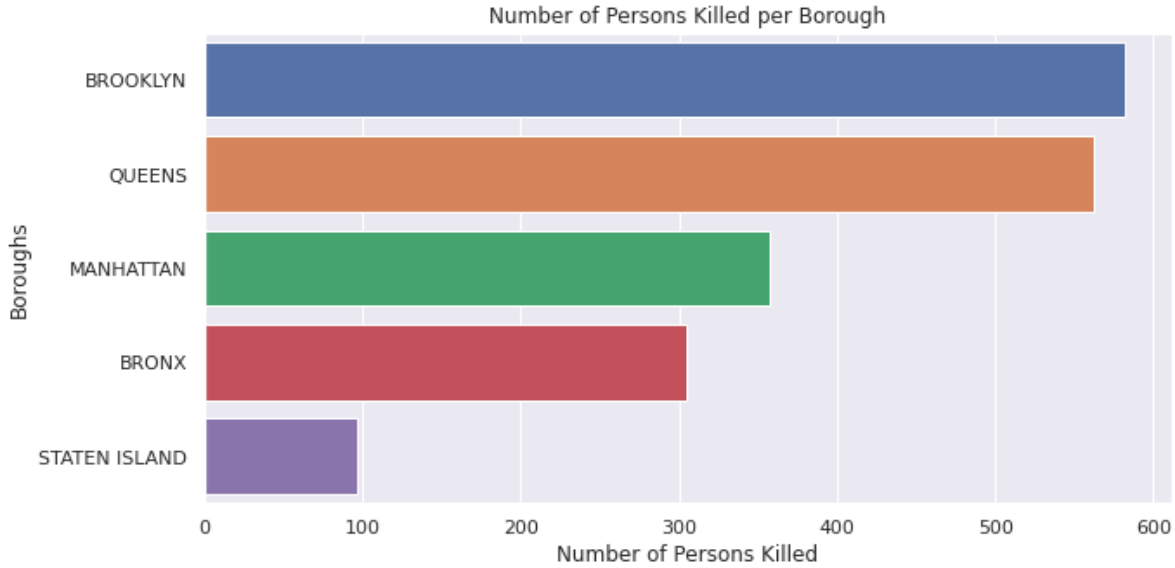
We studied which borough is known to have the most accidents. Staten Island was a clear winner. Location wise distribution is done in more detail later on.



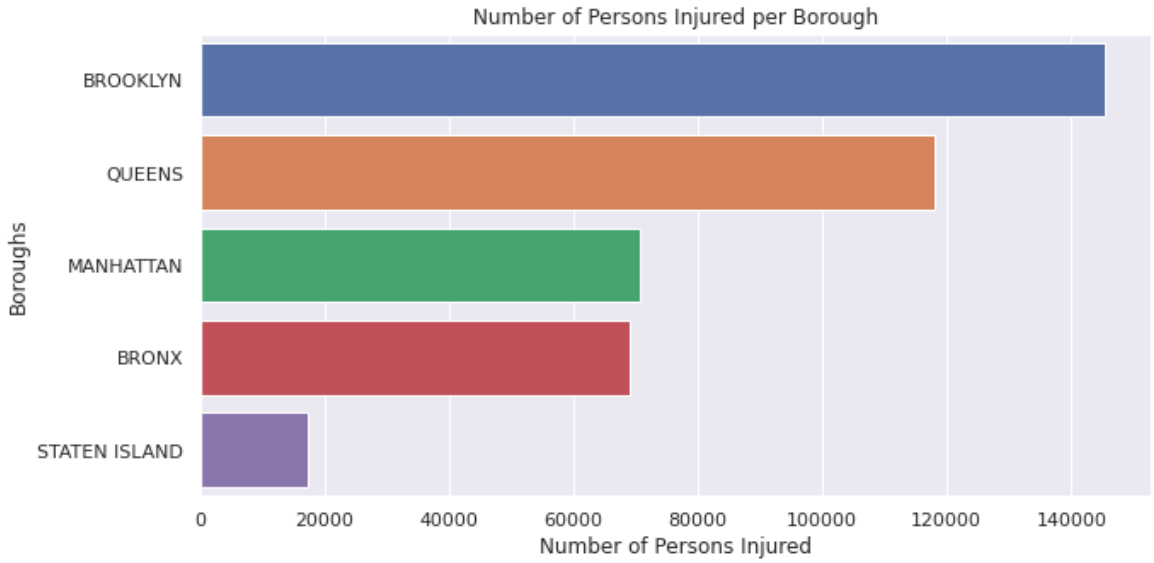
5. Ratio of Accidents to Population for Each Borough



6. Number of Persons Killed Per Borough



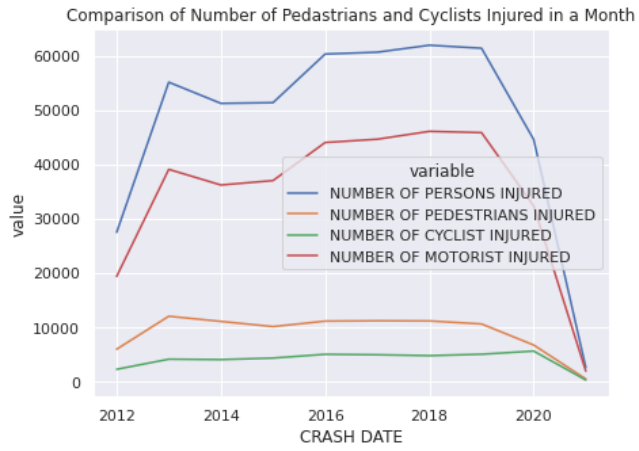
7. Number of Persons Injured Per Borough

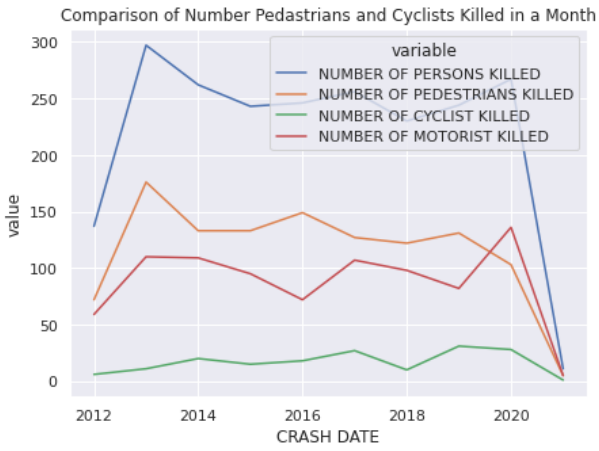


8. Average Number of Accidents Per Borough:

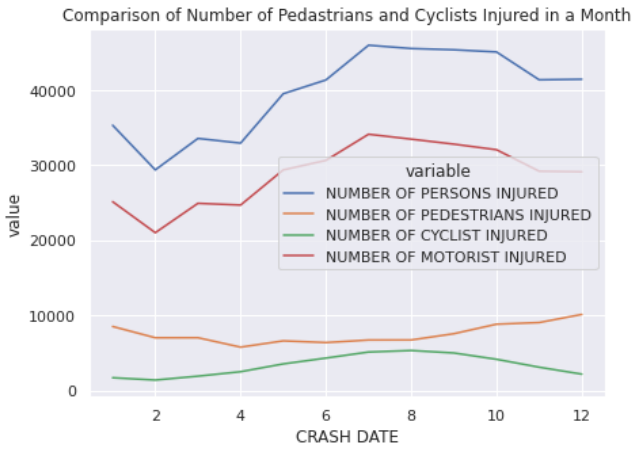
The average number of accidents per borough was **348,932.**

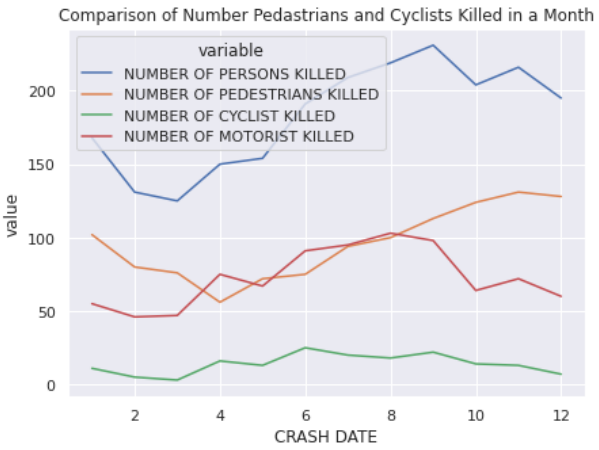
9. Plotting Trend of Accidents Across Years





10. Plotting trend of accidents across months



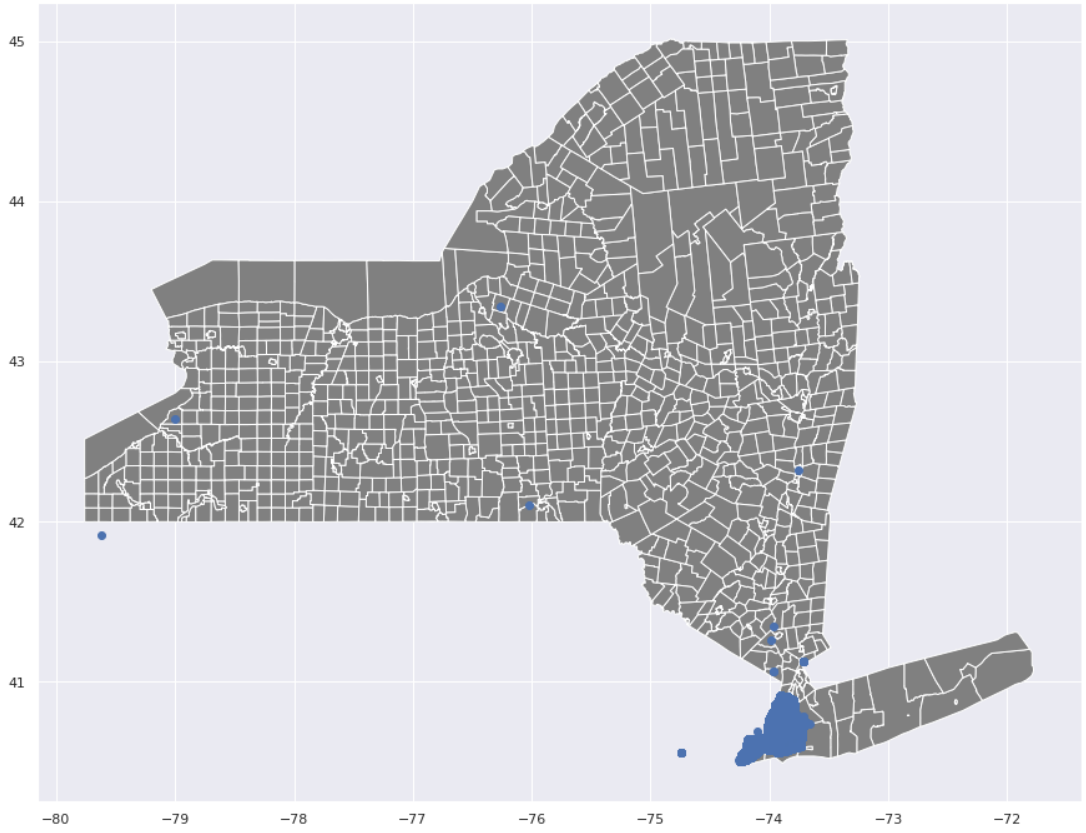


2.3 Location Based Analysis

11. First of all, we tried plotting the latitude longitudes of the accident locations on a new york city map shape found from a government website, as well as a scatterplot with Borough as hue.

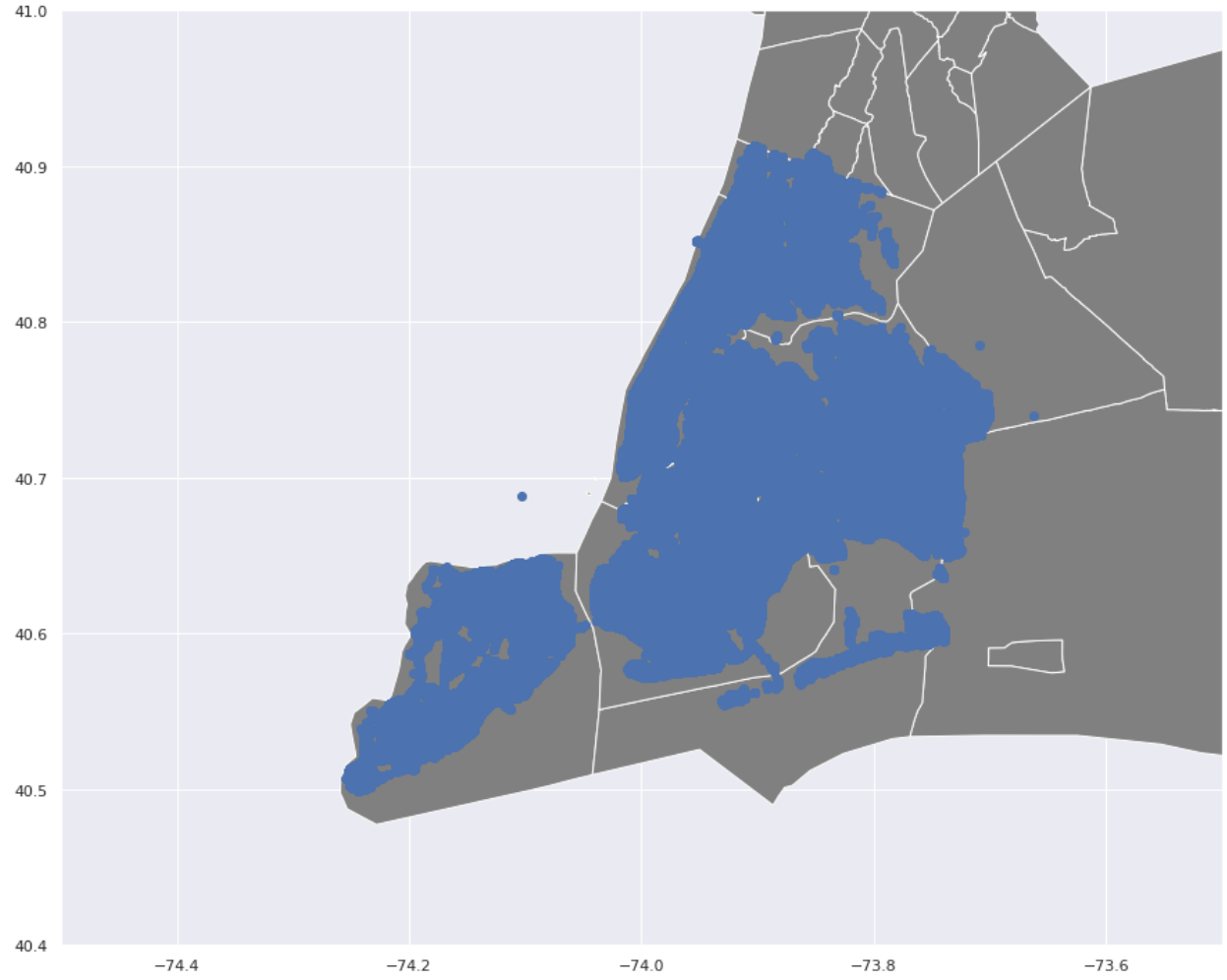
12. Plotting accidents on New York Map

We got a government County Map of New York and showed the accident locations on the Map using latitude and longitude. Most of the accidents are seen to happen in the bottom right corner of the map which is New York City.

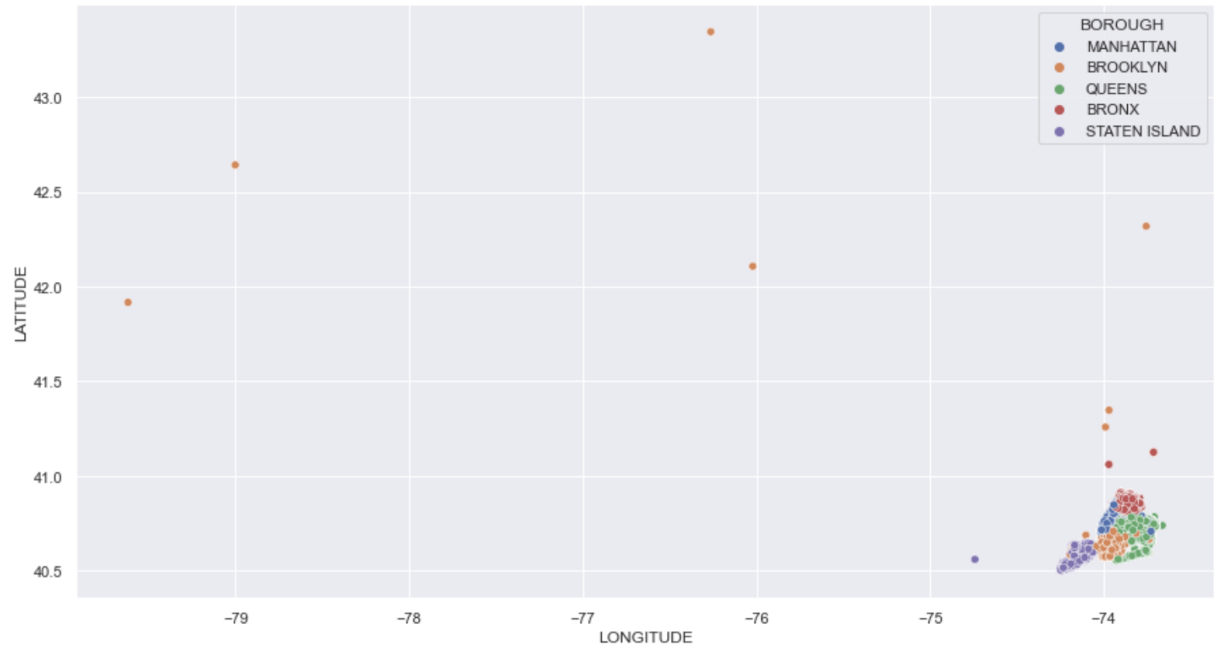


13. Zooming up on accidents in New York City

Zooming up on New York City shows the scatterplot of accident locations.


41.0
40.9
40.8
40.7
40.6
40.5
40.4
−74.4
−74.2
−74.0
−73.8
−73.6

This meant we had to drop all points which had NaNs as latitudes and longitudes.



As is visible, the smaller corner of the graph is the actual relevant area, NYC and a lot of latitude longitudes are outliers and do not even lie in the range of latitude longitude range for New York City which a quick Google search showed, was between Latitudes (-74.3 and -73.6) and Longitudes (40 and 41).

I. Filtering Out Obvious Outliers

We removed the obvious outliers which did not lie in the specified longitude and latitude range, leading to this graph.

```
g_df = df.dropna(subset=['LATITUDE', 'LONGITUDE'])

g_df = g_df[(-74.3 < g_df['LONGITUDE']) & (g_df['LONGITUDE'] < -73.6)]
g_df = g_df[(40.0 < g_df['LATITUDE']) & (g_df['LATITUDE'] < 41.0)]
sns.scatterplot(data=g_df, x="LONGITUDE", y="LATITUDE", hue="BOROUGH")
```

As is very evident in this, there are still outliers (or misclassified points) in this data, i.e for example there are some orange and blue points in the green region. These are points that belong to a different Borough than the ones they are labelled as according to their Latitude Longitudes.

Since we don't have any reason to believe that Latitude Longitude data is more reliable than Borough data, the best solution is to discard these points.

II. Filtering out other Outliers by similarity threshold

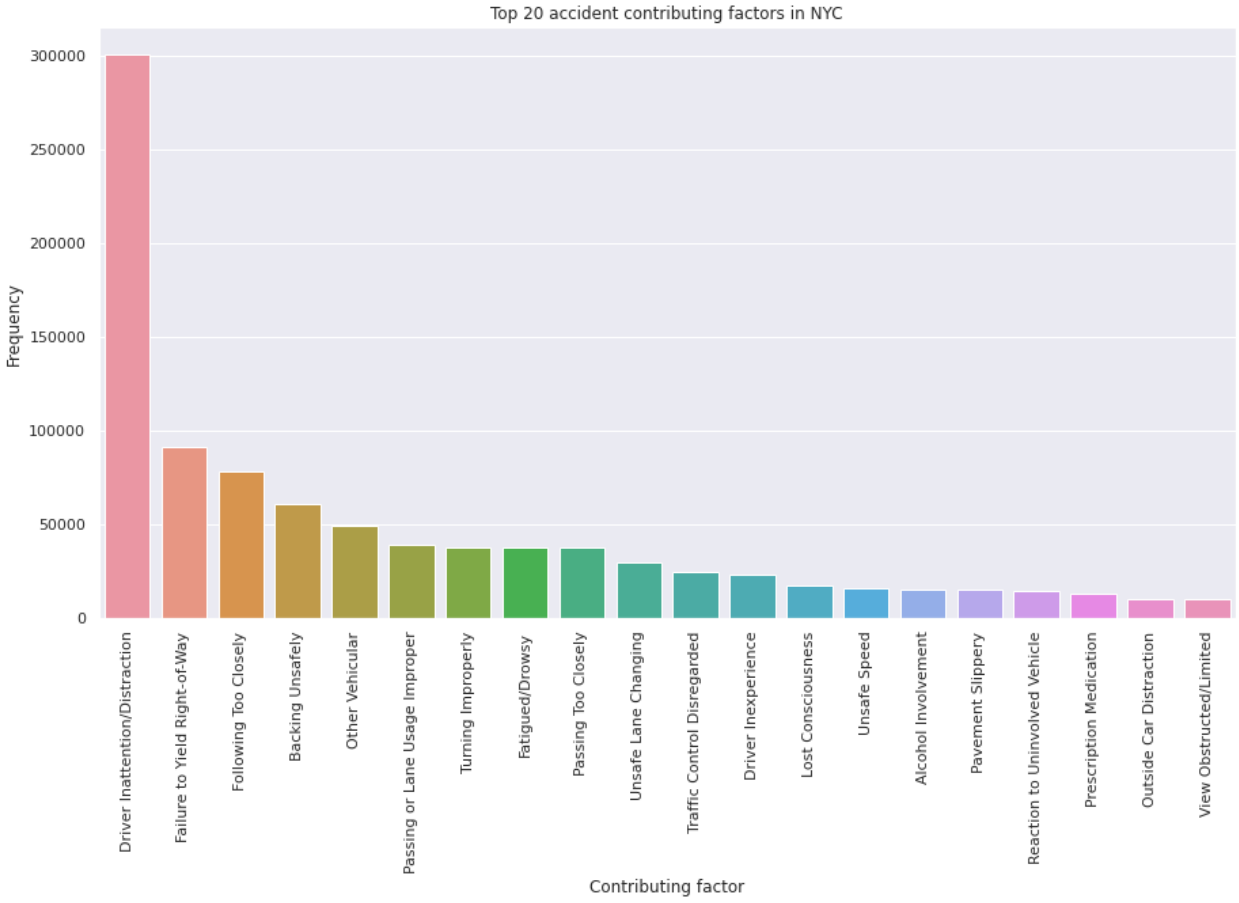
Used a clustering technique to filter out falsely labeled borough data points by looking at a small cluster of around each point and if the region contained a higher than 95% of points from a borough that was not the same as the points' borough, then drop the point from the data.

```
radius = 0.005
g_df['Number of Neighbours'] = ""
g_df['Number of Similar Neighbours'] = ""

def getNeighbourhood(row):
    lat = row["LATITUDE"]
    long = row["LONGITUDE"]
    borough = row["BOROUGH"]
    small_df = g_df[(lat-radius < g_df['LATITUDE'])    &
                    (g_df['LATITUDE'] < lat+radius)    &
                    (long-radius < g_df['LONGITUDE']) &
                    (g_df['LONGITUDE'] < long+radius)
                   ]
    numNeighbours = small_df['LOCATION'].count()
    numSimilarNeighbours = small_df[small_df["BOROUGH"] == borough]['LOCATION'].count()
    return numNeighbours, numSimilarNeighbours

g_df["Number of Neighbours"], g_df["Number of Similar Neighbours"] = zip(*g_df.apply(getNeighbourhood, axis=1))
g_df["% Similarity w/t Neighbour"] = g_df["Number of Neighbours"]*100/g_df["Number of Similar Neighbours"]
g_df = g_df[g_df["% Similarity w/t Neighbour"] > 95]
```

This took care of the outliers as can be seen. There aren't any points in obviously wrong regions.

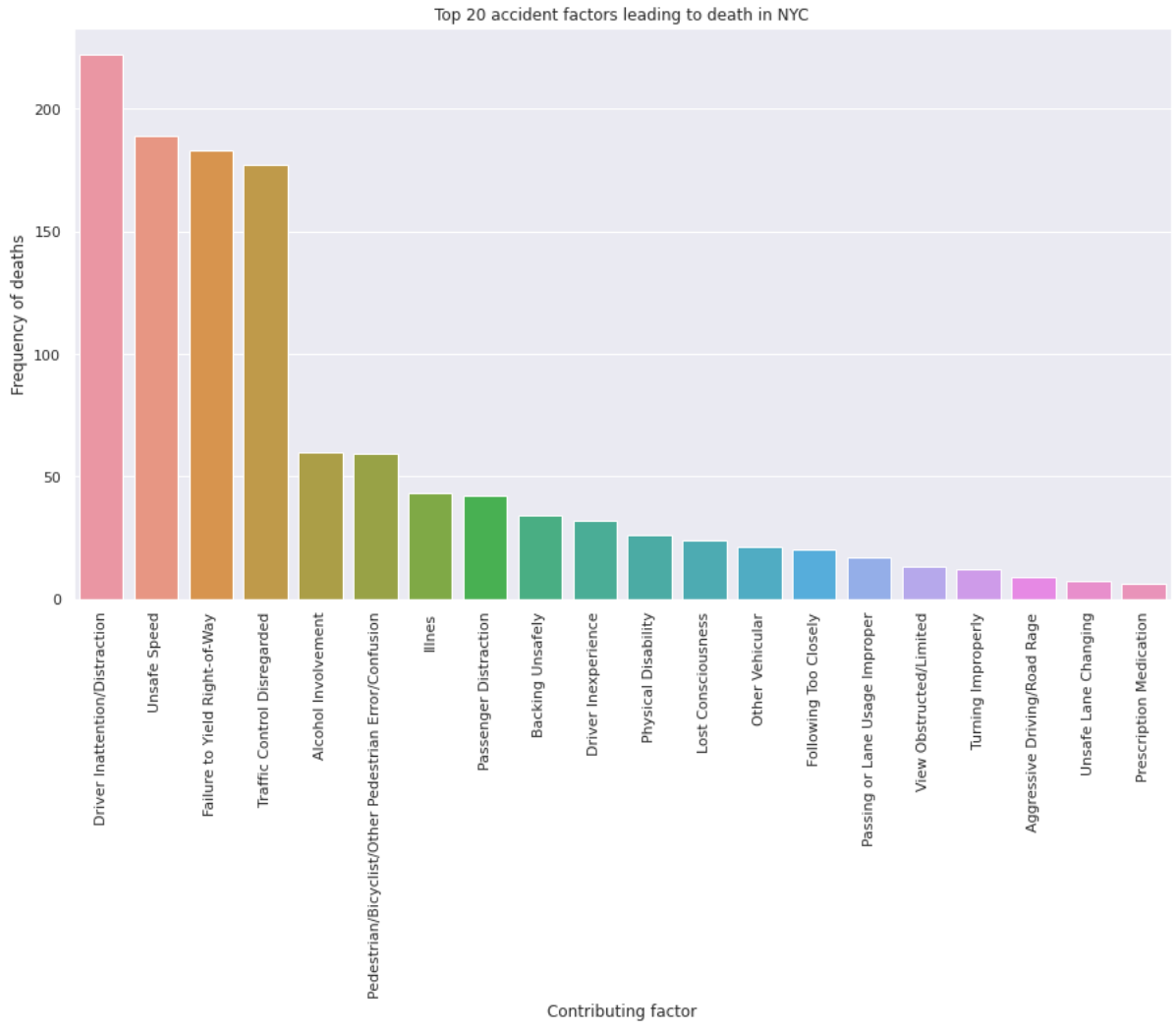
14. Most common contributing factors to accidents in NYC

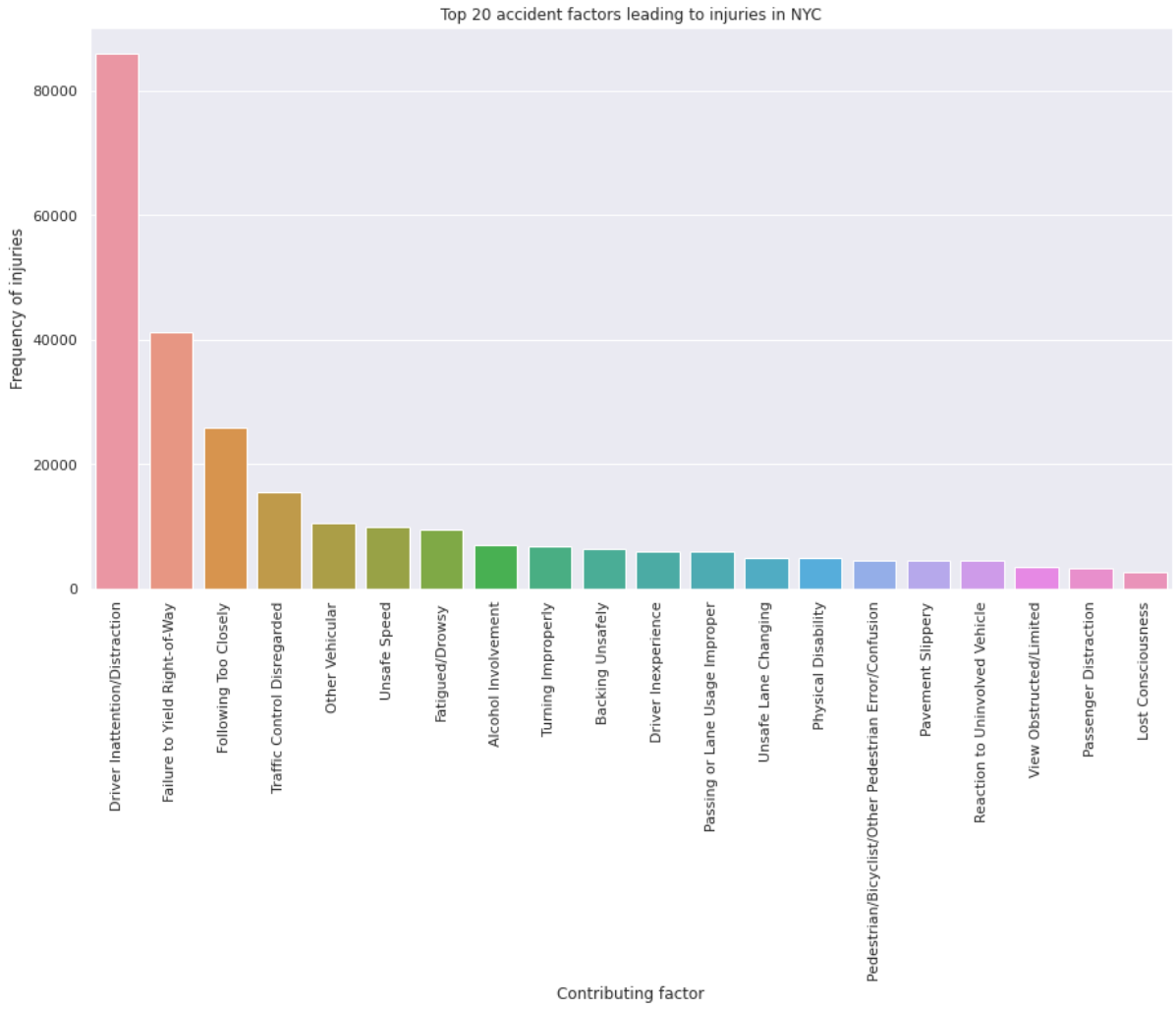
The most common contributing factors had to do with the driver e.g. driver distraction was the most harmful factor alongside other faults on their side. The results are as followed:



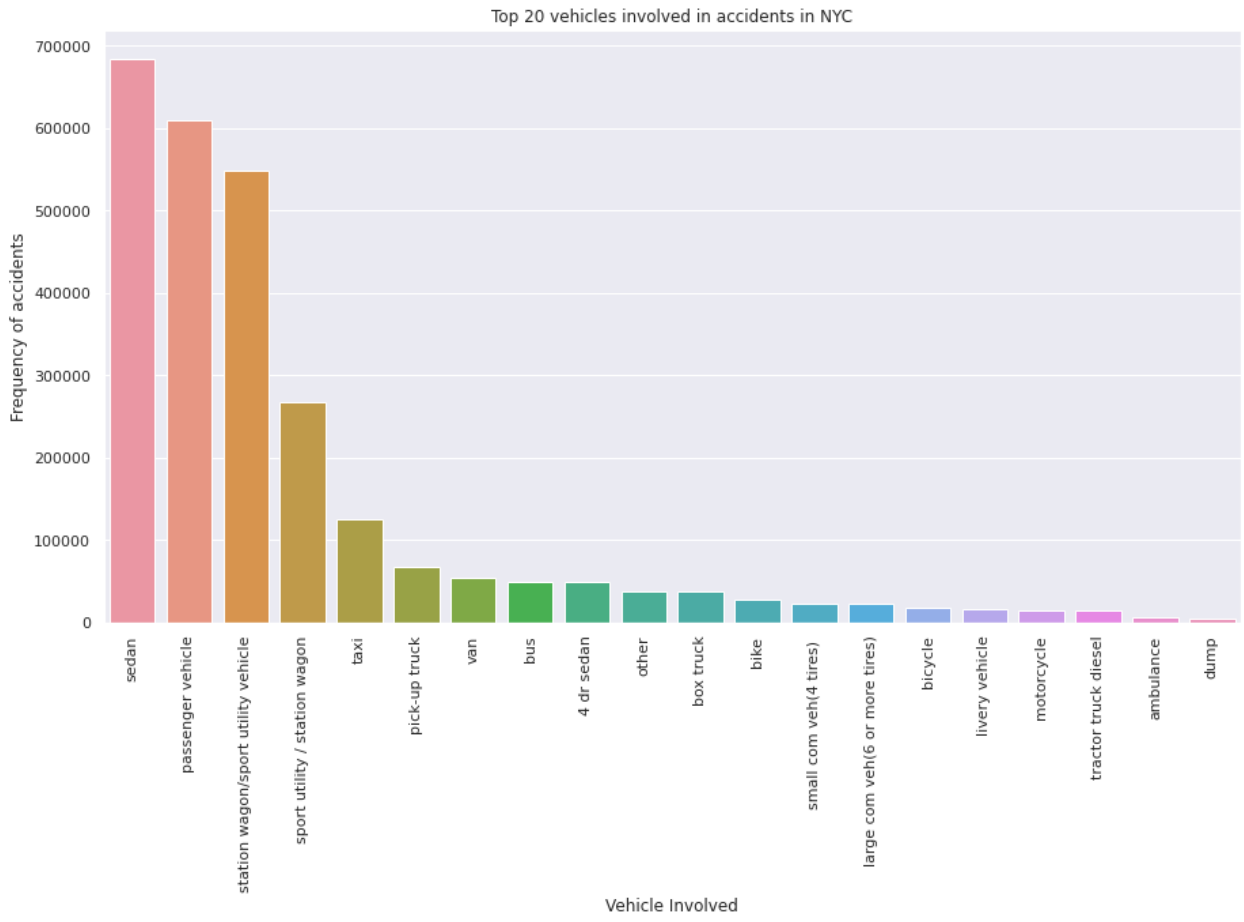
15. Contributing factors leading to the most damage (deaths and injuries)

The number one factor which led to deaths in accidents in NYC was Driver Inattention/Distraction because of which over 200 deaths occured. This was followed by unsafe speed and failure to yield right of way.




Top 20 accident factors leading to injuries in NYC
Frequency of injuries
0
20000
40000
60000
80000
Driver Inattention/Distraction
Failure to Yield Right-of-Way
Following Too Closely
Traffic Control Disregarded
Other Vehicular
Unsafe Speed
Fatigued/Drowsy
Alcohol Involvement
Turning Improperly
Backing Unsafely
Driver Inexperience
Passing or Lane Usage Improper
Unsafe Lane Changing
Physical Disability
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion
Pavement Slippery
Reaction to Uninvolved Vehicle
View Obstructed/Limited
Passenger Distraction
Lost Consciousness
Contributing factor

16. Most common vehicles involved in accidents in NYC



17. Judging Severity of Accident by location based on Number of People Injured or Killed

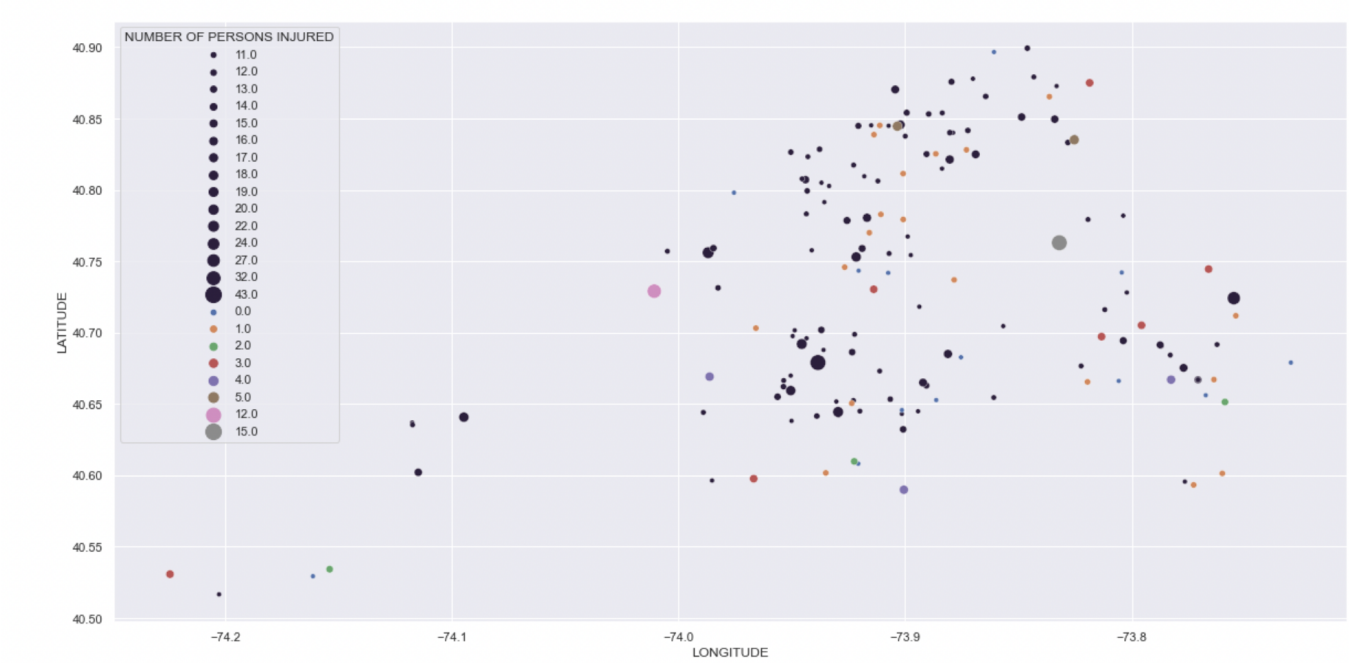
Now that the outliers from location data have been removed, we can go ahead and do more location based analysis.

Here we look at the graph of people injured or killed in accidents by location. We have filtered out the accidents with one death and fewer than 10 injured members to filter out less severe accidents.

The black circles/dots on the graph show the injured people and the colored ones show the people that were killed.

The size of the circles/dots show us the number of people impacted in the accident and the position shows the latitude longitude location on a map.

We can see a fairly even distribution on severity of accidents across the map mostly, except for Staten Island Borough which shows very few severe accidents and Queens Borough which shows relatively sparse severe accidents.



2. Clustering and Pattern mining:

In this section we will run a clustering algorithm on different attributes of the data set to make clusters with some similar properties. We will also do pattern mining to find out interesting patterns in data that will help us determine any causal relationship between different attributes. We will do clustering and pattern mining on a collective data set and then will do it borough wise to find out more detailed information.

2.1 Clustering

Feature selection

Before doing clustering , we need to decide which attributes are suitable for clustering and what useful information we can get from them. One case would be knowing about the relationship between location and death number to figure out which locations have more deaths that can help in deploying emergency departments near that location. Similarly other cases could be finding the relationship between cause of accident and type of vehicle that could help improve the vehicle for that particular type of cause and so on. We will be doing clustering analysis on 2 of the following feature set

- Cause of accident and type of vehicle

- Clustering to identify severity of accidents (borough-wise)

2.1.2 Data preparation

For Cause of accident and type of vehicle, we took most 5 common attribute values and clustered that data. For both types of clustering we used the K-Modes method.

Clustering with cause of the accident and vehicle type

First we find k using the elbow curve method. The plot across various Ks is followed as:

We did the following steps for whole clustering process:

- Preprocessing the data.
- Selecting required features for clustering (features selection)
- Reducing values to top 20 most occurring
- Used KModes clustering method (since all attributes were categorical)
- Select Appropriate K using elbow method
- Plot the clusters against each attributes

First of all , after preprocessing data, we selected four attributes (Vehicle Type1,Vehicle Type2,Contributing Factor Vehicle Type1,Contributing Factor Vehicle Type2). We dropped columns for other factors and vehicle types because the majority of their values were nul and hence do not contain any useful data for clustering. Since there were many values for these attributes, we filtered out the top 20 most occurring values from the remaining data set. Final view of data set is followed as

	CONTRIBUTING FACTOR VEHICLE 1	CONTRIBUTING FACTOR VEHICLE 2	VEHICLE TYPE CODE 1	VEHICLE TYPE CODE 2
0	Aggressive Driving/Road Rage	Aggressive Driving/Road Rage	Taxi	Station Wagon/Sport Utility Vehicle
1	Unspecified	Unspecified	Station Wagon/Sport Utility Vehicle	Unspecified
2	Failure to Yield Right-of-Way	Passing or Lane Usage Improper	Station Wagon/Sport Utility Vehicle	Bike
3	Unspecified	Unspecified	Sedan	Unspecified
4	Failure to Yield Right-of-Way	Unspecified	Station Wagon/Sport Utility Vehicle	Sedan

Then we choose the K-Mode method for clustering. To choose the appropriate K value, we ran clustering for 10 k values from 1 to 10 and stored costs in a list. Then we plot cluster numbers and costs to find the appropriate k using the elbow method. We found 4 to be the good value for k.

Plotting cost against clusters to figure out K

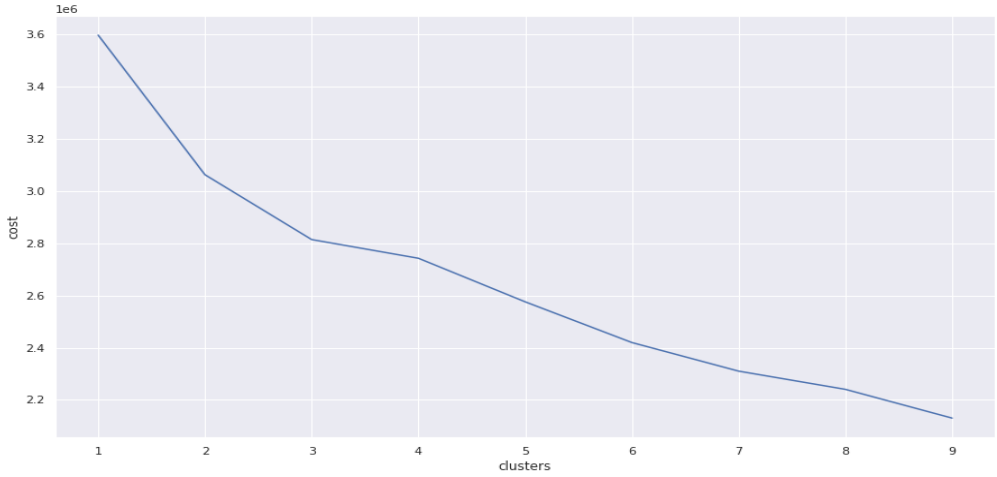


Fig1 Trend of Costs as we increase the number of clusters.

In the final step, we plot all the clusters against each attribute to analyze and observe the distribution of different values across different clusters. Following are the coloured bar plots for each attribute where plot height is showing quantity and color is showing the cluster number.

Plot against Vehicle1

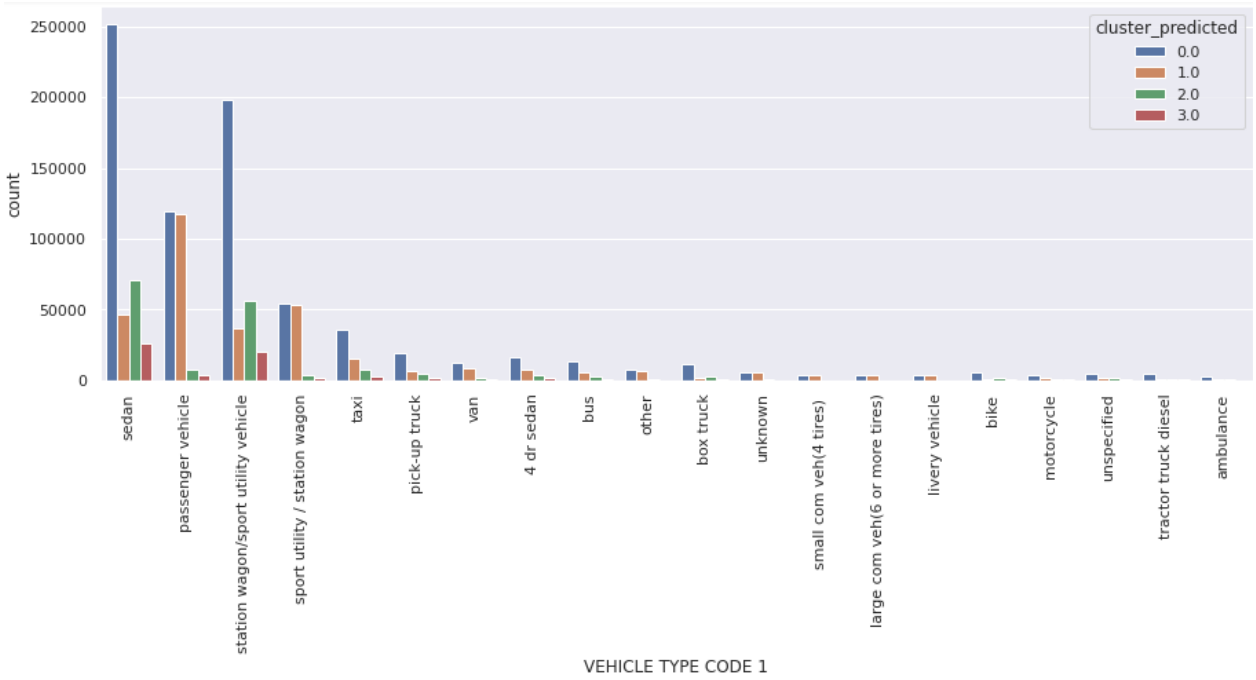


Fig1 Bar graph showing frequency of each vehicle type in each cluster

Plot against Vehicle2

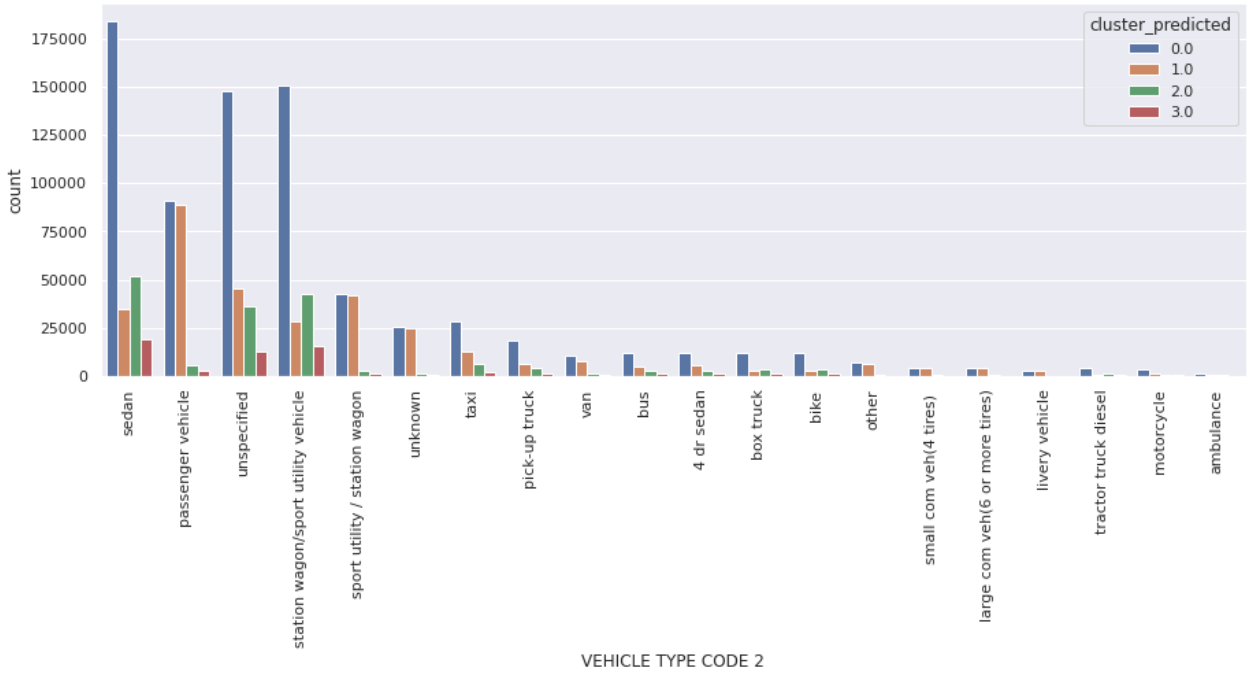


Fig1 Bar graph showing frequency of each vehicle type in each cluster

Plot against Factor1

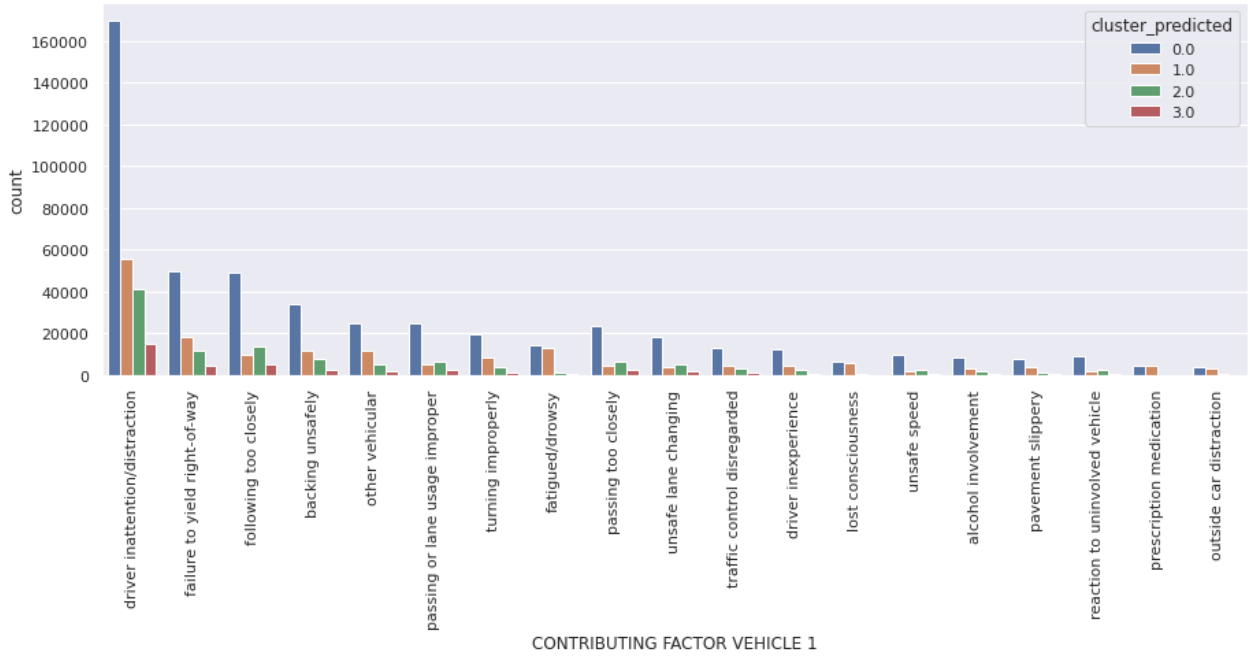


Fig1 Bar graph showing frequency of each contributing factor in each cluster

Plot against Factor2

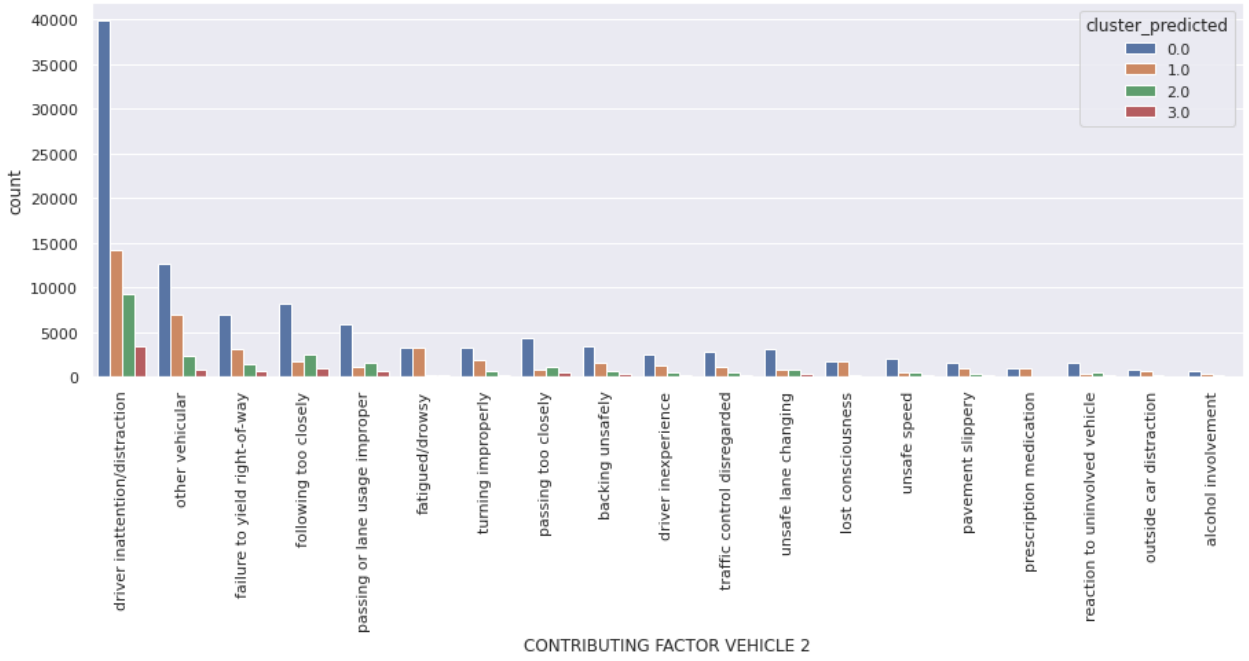


Fig1 Bar graph showing frequency of contributing factor in each cluster

Interpretation

Clusters have following properties: Cluster 1 contains a high high number of accidents caused by sedan and high number of causes of driver inattention or distraction, Cluster 2 contains type of accidents in which the most contributing factor is driver distraction but the vehicle type in this case is sedan. In the last two clusters, the most contributing factor is again driver in attention but the most frequently occurring vehicles are sports wagons and sedans. These are the four categories of accidents that are clustered on vehicle type and contributing factor. First type shows that sedan vehicles accidents mostly happen due to driver inattention. This is an important insight which can be drawn from cluster 1. We can take necessary actions such as increasing training for sedan drivers and imposing more laws restrictions such as low speed, upgradation checks e.t.c to reduce this type of accidents.

Clustering to find Accident Severity (Manhattan Borough)

Here we applied the K-Modes clustering algorithm on the data for Manhattan Borough. For K-Modes we selected the attributes *Crash Time, Number of Persons Injured, Number of Persons Killed* and *Contributing Factor Vehicle 1*.

Since there were many values in the *contributing factor*, we kept the 5 most frequent and changed the rest to 'other'.

Crash Time was not a categorical attribute; to make it categorical, we will created the following bins (or time slots):

- Night: 00:00 hrs to 05:59 hrs
- Morning: 06:00 hrs to 11:59 hrs
- Afternoon: 12:00 hrs to 18:59 hrs
- Evening: 19:00 hrs to 23:59 hrs

We also created new categories for Number of Persons Killed:

- 0
- 1-2 (inclusive)
- 3-5 (inclusive)

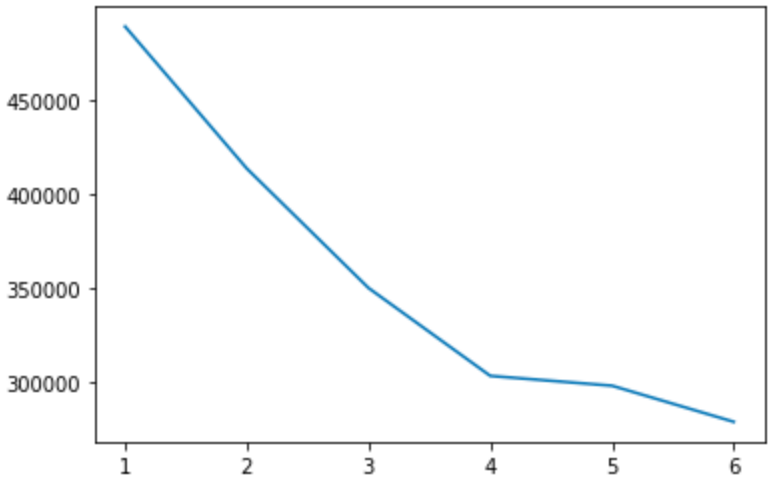
This will help in making the clusters.

COLLISION_ID	CONTRIBUTING FACTOR VEHICLE 1	TIME OF DAY	PERSONS KILLED	PERSONS INJURED
3934661	other	Afternoon	0	0
3947774	Unspecified	Afternoon	0	0
3942156	Driver Inattention/Distraction	Night	0	1-2
3943793	Driver Inattention/Distraction	Evening	0	1-2
3941432	Failure to Yield Right-of-Way	Afternoon	0	0

We encoded our attribute values using the sklearn library.

COLLISION_ID	CONTRIBUTING FACTOR VEHICLE 1	TIME OF DAY	PERSONS KILLED	PERSONS INJURED
3934661	24	0	0	0
3947774	22	0	0	0
3942156	2	3	0	1
3943793	2	1	0	1
3941432	4	0	0	0

Then we ran the K-Modes algorithm using the Cao method for selection of initial K objects. To find the ideal value of K, we plotted the costs from K = 1 to 6.



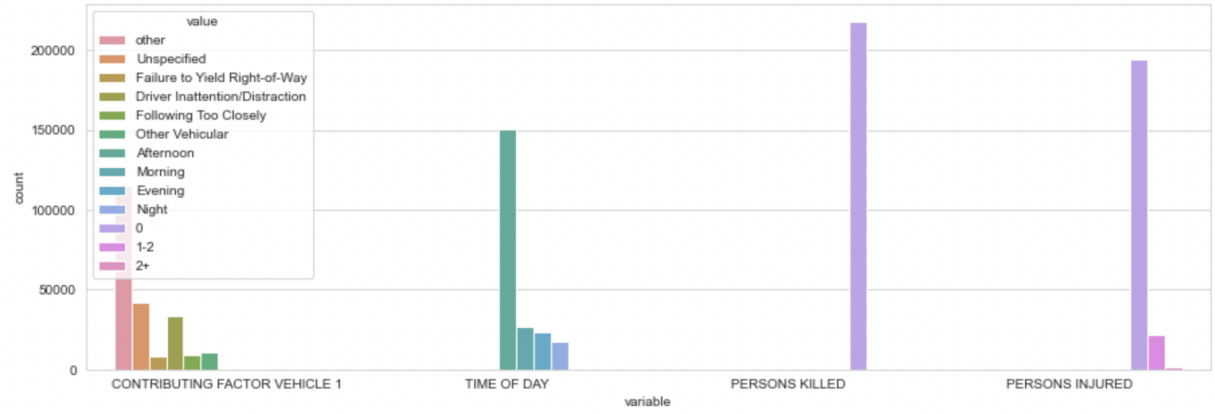
We saw that K=4 is the elbow point and chose to make 4 clusters.

```
Number of records in cluster 1: 217999
Number of records in cluster 2: 57520
Number of records in cluster 3: 46446
Number of records in cluster 4: 32080
```

Cluster 1

This cluster represents accidents that happen in the afternoon and show that the contribution factors are predominantly unspecified but many are also caused by driver inattention.

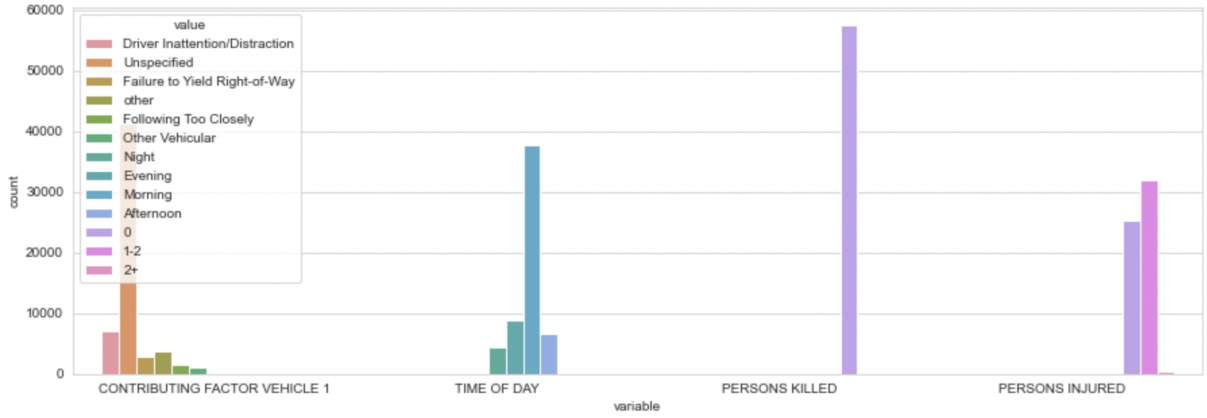
The severity of accidents is not very high, as is shown by the fact that persons killed and injured are predominantly zero.



Cluster 2

This cluster represents accidents that happen in the morning and show that the contribution factors are predominantly unspecified but many are also caused by driver inattention.

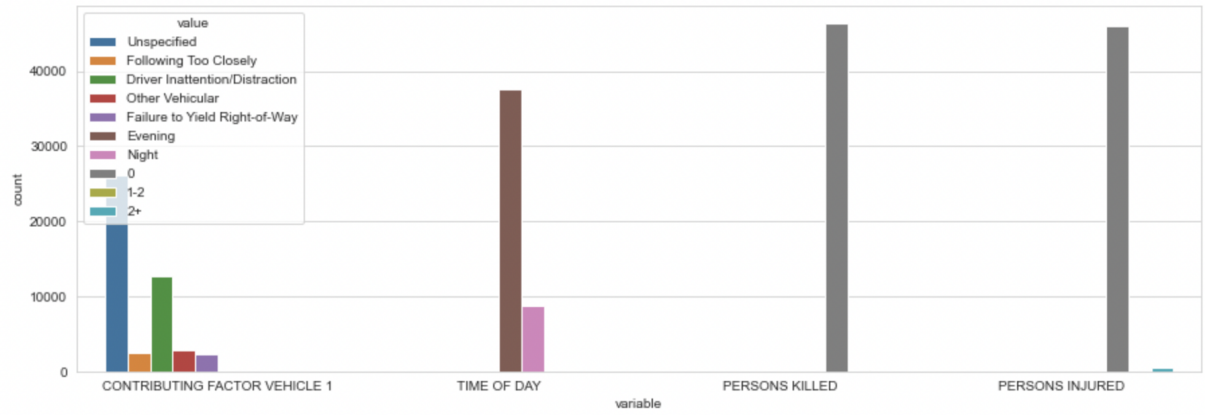
The severity of accidents here is not extremely high, shown by the fact that predominantly zero people are killed, however there are a high number of accidents where 2+ persons are injured in this cluster.



Cluster 3

This cluster represents accidents that happen in the evening and show that the contribution factors are predominantly unspecified but many are also caused by driver inattention.

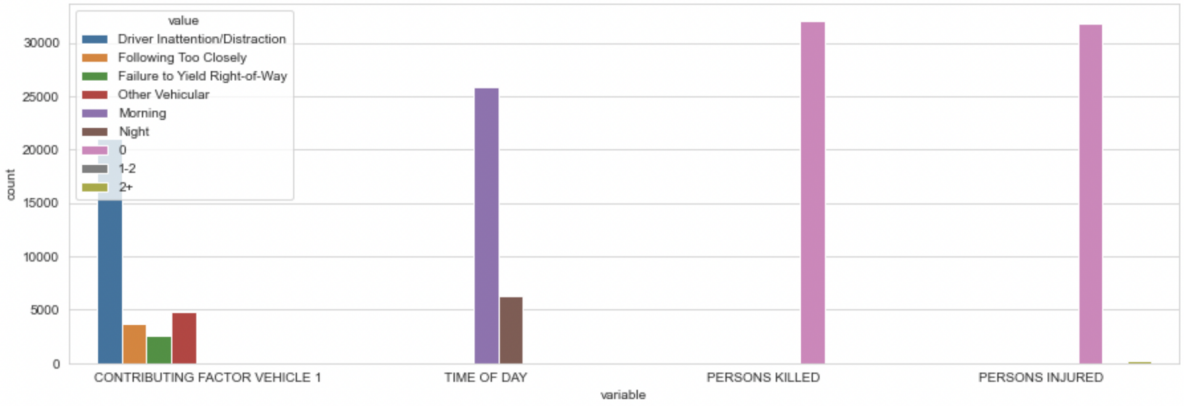
The severity of accidents here is not extremely high, shown by the fact that predominantly zero people are killed and zero people are injured.



Cluster 4

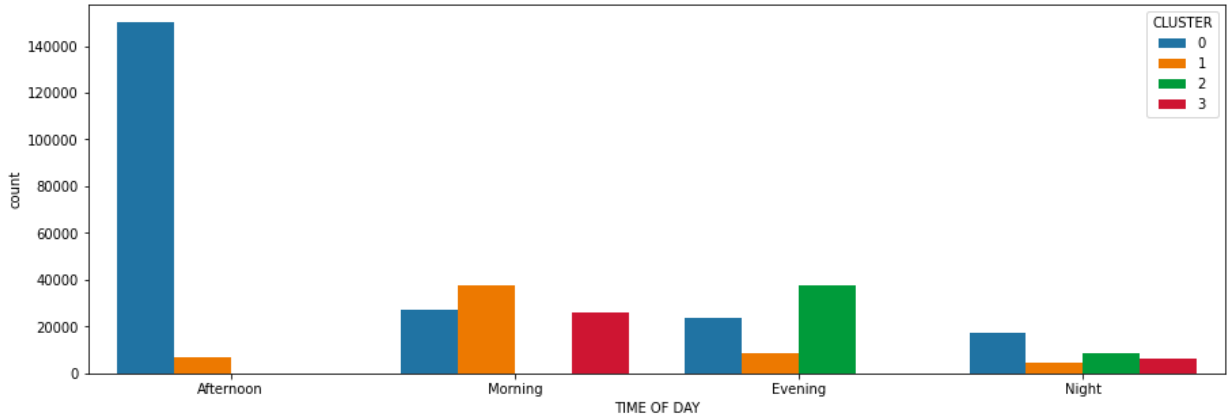
This cluster represents accidents that happen in the morning and show that the contribution factors are predominantly caused by driver inattention.

The severity of accidents here is not extremely high, shown by the fact that predominantly zero people are killed and zero people are injured.



Alternate way to view the same Clusters

Plot against Time of Day:

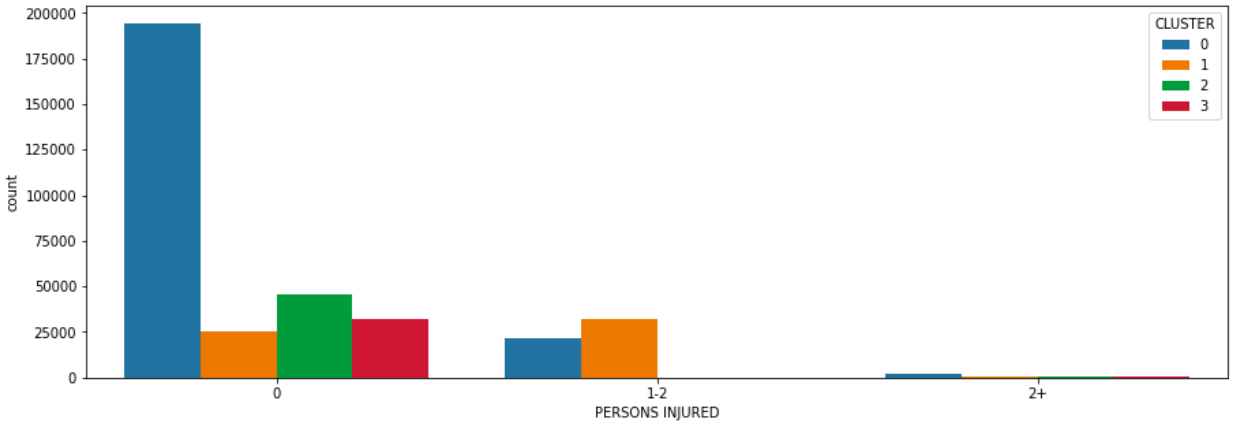


Cluster 1: Most accidents occur in the Afternoon.

Clusters 2 and 4: Most accidents occur in the Morning.

Cluster 3: Most accidents occur in the Evening.

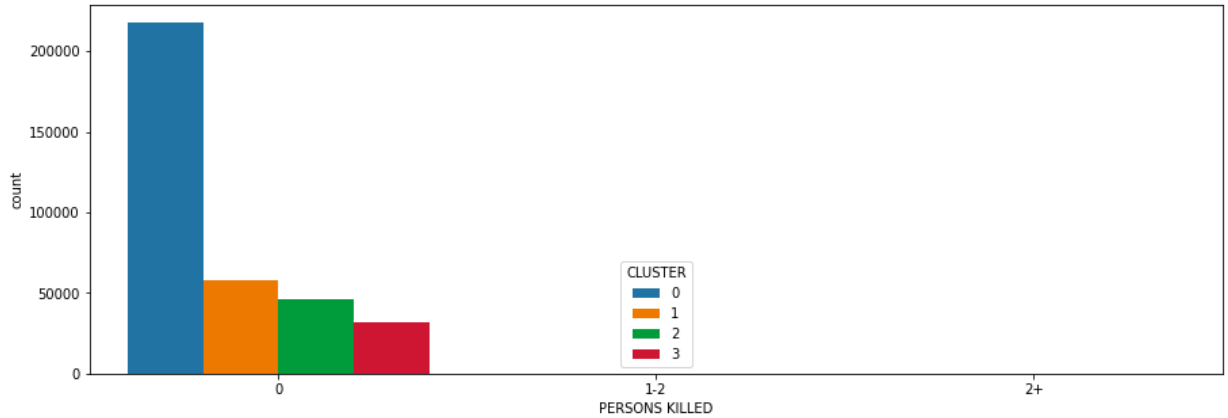
Plot against Persons Injuries:



Cluster 1, 3, and 4: Mostly zero persons injured.

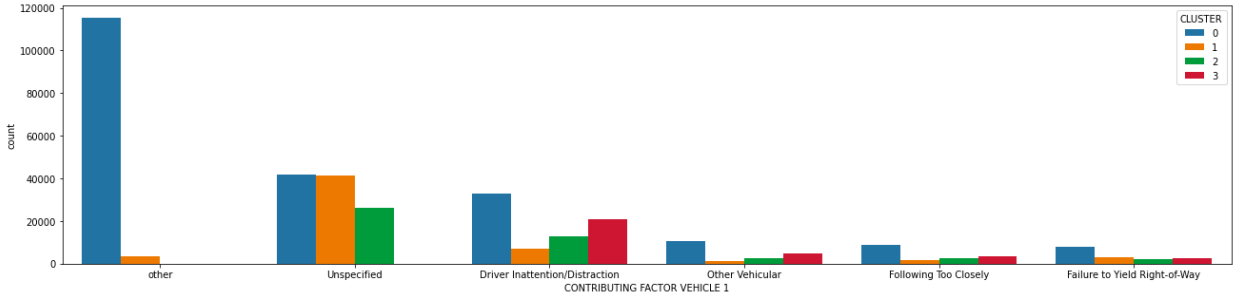
Cluster 2: Mostly 1-2 persons injured.

Plot against Persons Killed:



All clusters: Mostly zero people killed.

Plot against Persons Killed:



Cluster 1, 2 and 3: Major causes include Driver Inattention/Distraction and Unspecified Cause.

Cluster : Major causes include Driver Inattention/Distraction and Other Vehicular.

Conclusions

The clusters reveal that in Manhattan, most accidents occur during daylight hours. Drivers are more likely to be distracted in the morning as compared to the afternoon or evening. Moreover, accidents caused by driver inattention in the morning are more likely to result in injury than in the afternoon or evening.

K-Means Clustering

Data Preparation

Since our dataset is extremely large with over 1,000,000 records, hierarchical clustering can be inefficient since it has a quadratic computational complexity. So we turn to K-means which is a more scalable and efficient technique. However, since K-Means uses Euclidean distance as a similarity matrix, our features cannot contain categorical data.

Before we begin our dimensionality reduction, this is what our columns and their data types looked like:

```
CRASH DATE                        datetime64[ns]
CRASH TIME                        datetime64[ns]
BOROUGH                                  object
ZIP CODE                                 object
LATITUDE                                float64
LONGITUDE                               float64
LOCATION                                 object
ON STREET NAME                           object
CROSS STREET NAME                        object
OFF STREET NAME                          object
NUMBER OF PERSONS INJURED               float64
NUMBER OF PERSONS KILLED                float64
NUMBER OF PEDESTRIANS INJURED             int64
NUMBER OF PEDESTRIANS KILLED              int64
NUMBER OF CYCLIST INJURED                 int64
NUMBER OF CYCLIST KILLED                  int64
NUMBER OF MOTORIST INJURED                int64
NUMBER OF MOTORIST KILLED                 int64
CONTRIBUTING FACTOR VEHICLE 1            object
CONTRIBUTING FACTOR VEHICLE 2            object
CONTRIBUTING FACTOR VEHICLE 3            object
CONTRIBUTING FACTOR VEHICLE 4            object
CONTRIBUTING FACTOR VEHICLE 5            object
VEHICLE TYPE CODE 1                      object
VEHICLE TYPE CODE 2                      object
VEHICLE TYPE CODE 3                      object
VEHICLE TYPE CODE 4                      object
VEHICLE TYPE CODE 5                      object
dtype: object
```

To fix this, we drop columns not relevant to our analysis and map the relevant categorical ones to numeric data.

To do this we:

- Converted Number of Persons killed and injured to int from float for consistency
- Dropped contributing factor and vehicle type columns as they contained too many distinct values to be able to meaningfully map them to numeric values.
- Also dropped columns the Location, ZIP code and Street Names columns.
- Mapped categorical Borough names to numbers from 1-5.

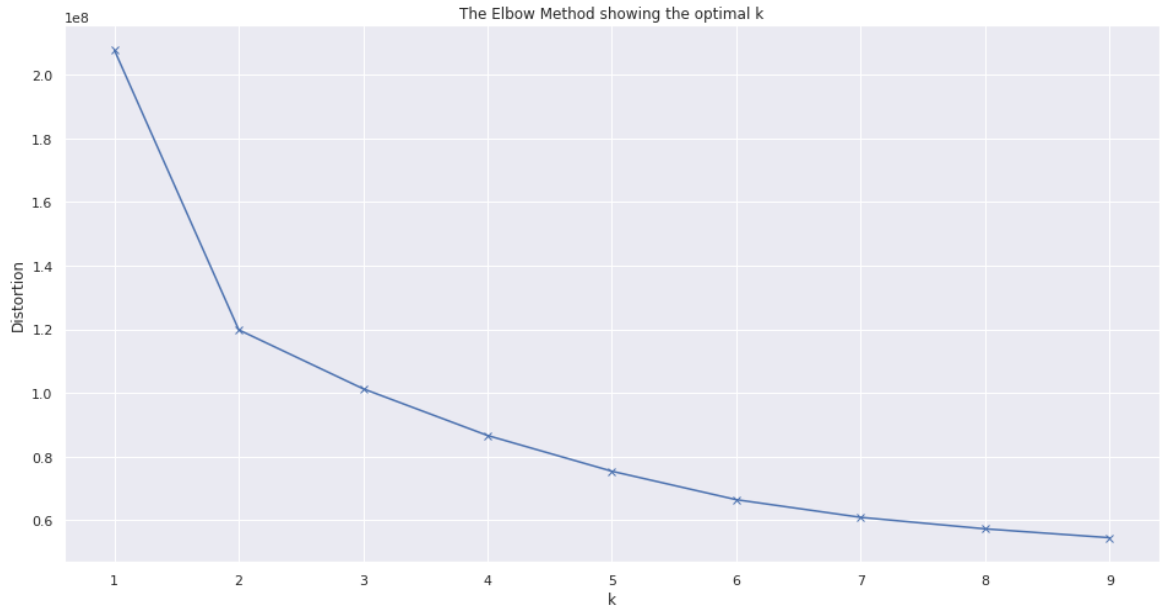
- Convert datetime columns CRASH TIME and CRASH DATE to numeric by adding numerical Year, Month, Day and Hour columns to the cluster dataframe and dropping the original datetime columns.

After the above steps, we finally have a dataset which we could run K-means clustering on:

```
BOROUGH                          int64
LATITUDE                       float64
LONGITUDE                      float64
NUMBER OF PERSONS INJURED        int64
NUMBER OF PERSONS KILLED         int64
NUMBER OF PEDESTRIANS INJURED    int64
NUMBER OF PEDESTRIANS KILLED     int64
NUMBER OF CYCLIST INJURED        int64
NUMBER OF CYCLIST KILLED         int64
NUMBER OF MOTORIST INJURED       int64
NUMBER OF MOTORIST KILLED        int64
Year                             int64
Month                            int64
Day                              int64
Hour                             int64
dtype: object
```

Determining the optimal value for K using the Elbow Method

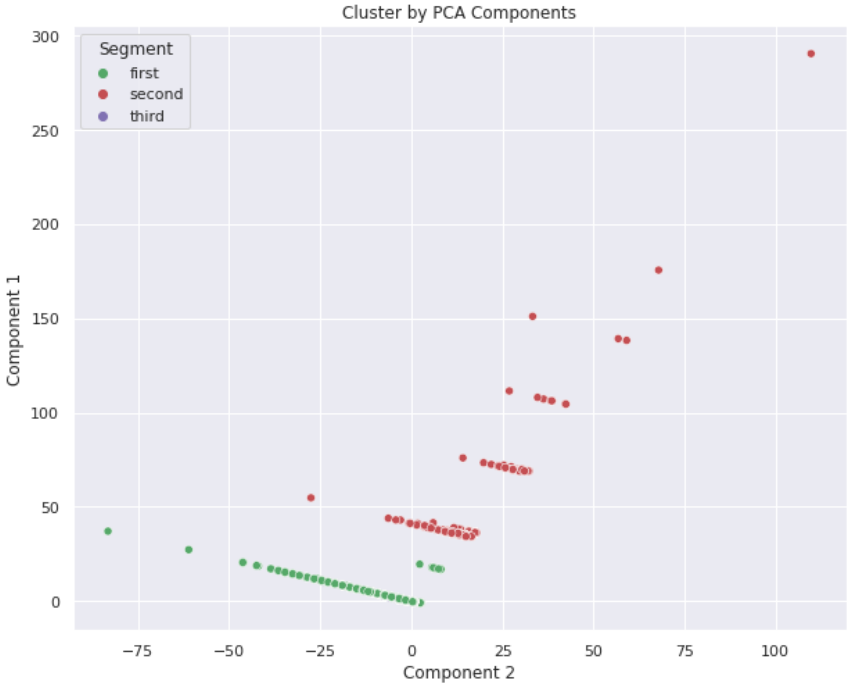
Since the K-means requires us to provide a k-value, the Elbow Method was used to fit the model with the values of k ranging from 1 to 10. The point of inflection of the curve is a reasonable indication that the data fits the model best at that particular value which was k in our case as can be seen in the plot below: .



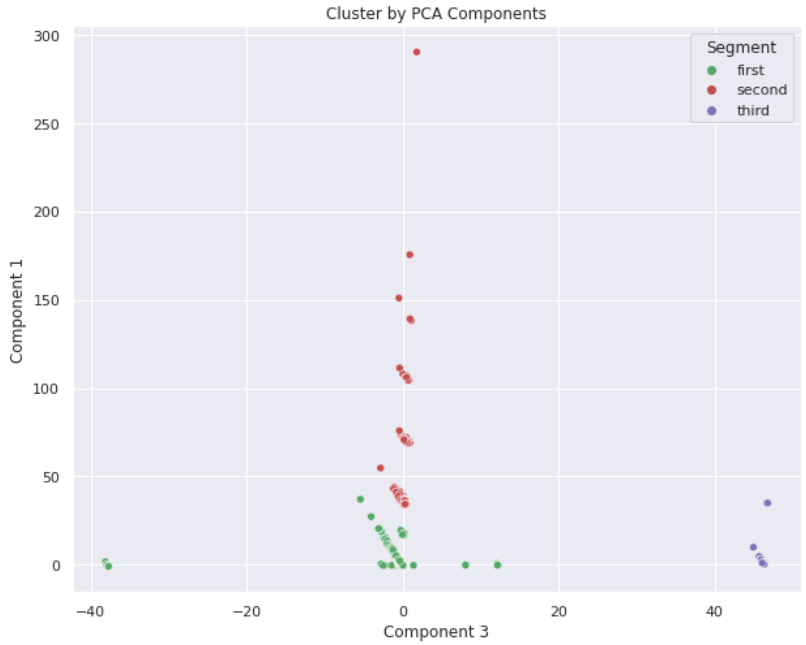
Data Segmentation and Visualization using PCA

Now that we have our prepared data and optimal k-value, we segmented the data into three PCA components after normalizing it with sklearn's StandardScaler. Our final cluster plots comparing the three components pairwise were as followed:

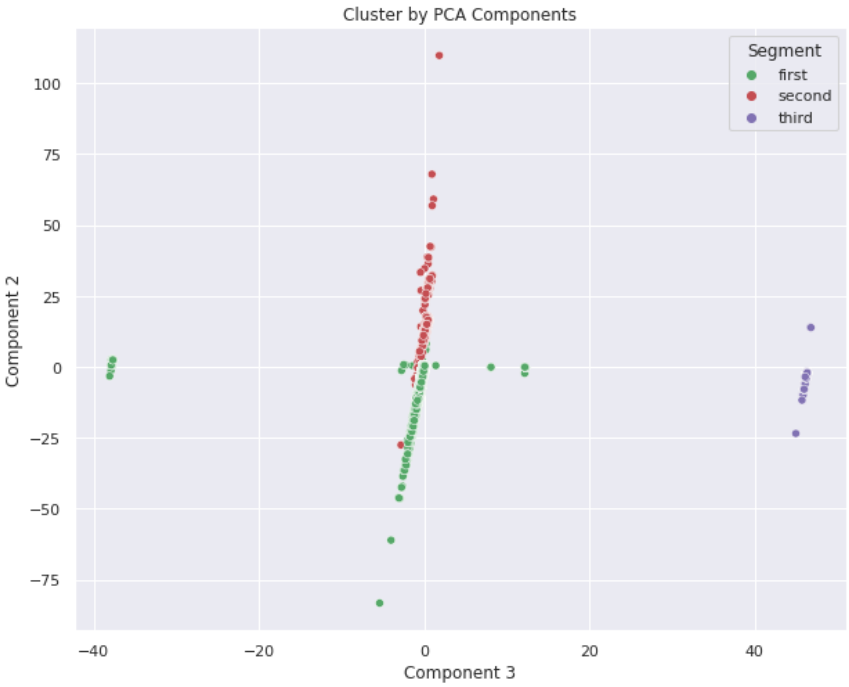
Component 1 vs. Component 2



Component 1 vs. Component 3



Component 2 vs. Component 3



2.2 Frequent Pattern Mining with the FPGrowth Algorithm

Procedure

Now that we are done with clustering, we will use the FPGrowth algorithm to find frequent patterns in our data. We choose this over Apriori, due to its performance superiority to find frequent patterns in our large dataset. To do so, we will drop the time related columns as those are more or less unique values as well as the location column as it is already covered in the Longitude and Latitude columns. We will also drop the people killed and injured columns as they are difficult to distinguish when working with patterns so will not offer a lot of useful information. We will then convert our dataframe to a list of lists and run the FPGrowth algorithm with a minimum support of our choice. Since this is a large dataset, we chose a min support of 0.1.

Results

We got the following frequent itemsets of length 2 with support of 0.1 or more:

```
Driver Inattention/Distraction station wagon/sport utility vehicle
support: 0.11921540694147924

Driver Inattention/Distraction sedan
support: 0.14782025407934027

QUEENS passenger vehicle
support: 0.12160271218634994

QUEENS station wagon/sport utility vehicle
support: 0.12234615017836674

QUEENS sedan
support: 0.13634886314318576

BROOKLYN station wagon/sport utility vehicle
support: 0.11739573541651666

BROOKLYN passenger vehicle
support: 0.13118696427540422

BROOKLYN sedan
support: 0.15281997188429094

sedan station wagon/sport utility vehicle
support: 0.15862800010639078

passenger vehicle passenger vehicle
support: 0.1316605329788704

sedan sedan
support: 0.12517004684437819
```

Findings

Results from the FPGrowth algorithm revealed some interesting patterns in our dataset:

1. There is a pattern amongst Sedan and Station wagon/sport utility vehicle drivers to be involved in an accident due to inattention and distraction.
2. There is also a pattern for Sedan, Passenger vehicle and Station wagon/sport utility vehicle drivers to be involved in accidents in the Queens and Bronx boroughs.
3. There is a pattern for Sedans to be involved in accidents with other Sedans or Station wagon/sport utility vehicles.
4. There is a pattern of Passenger vehicles being involved in accidents with other Passenger vehicles.

Feedback Incorporation

Based on the feedback we received from our assigned TA, we made the following improvements to our report.

Deliverable 1:

1. We added more background of our dataset, what we can achieve from it and how it will be helpful.
2. Removed redundant information from the report.
3. In the map added boundaries of boroughs and improved presentation by separately labelling boroughs.
4. Removed latitude longitudes outliers from the map plot.
5. Instead of dropping rows for missing values of injured and killed persons, we filled them by aggregating pedestrian, motorist and cyclist data.
6. Changed the axis in borough wise accidents plot to make them more relevant.
7. Added plots for the number of persons injured and killed as well as boroughs.
8. Added borough wise plots for accidents to population ratio.

Deliverable 2:

1. Updated the interpretation for clustering with vehicle type and contributing factor.
2. Added missing captions and titles for graphs.
3. Added reasons for why we chose to drop certain attributes before clustering.
4. Updated formatting and structure of headings.
5. Instead of having all clusters on a single plot, we spaced them onto different plots making it easier to study any trends that show up.

3. Recommendations and Findings for Policy Makers

Based on our in-depth analysis, we got some insightful findings that policy makers can use to better limit accidents in New York City:

1. There are certain vehicles such as Station wagon/sport utility vehicles which are strongly correlated with occurrence of accidents. Policy makers should look into why this is the case. One likely reason would be that these vehicles transport goods, especially sports goods, around in a hurry since New York City is known for it's love for sports, especially baseball.
2. Although the accident count per borough is directly proportional to the borough's population, when we normalize this count by population, we see Staten Island has the highest ratio of accidents to population despite having lower population. This number is over 8 while the bigger borough's all have ratios of below 2. This is a startling result and so policy makers should allocate more attention and resources to improving the transportation network in Staten Island.
3. Failure to yield right of way is the second most common cause of accidents in NYC after driver inattention. Drivers in NYC should be given training about the correct method to give way to other motorists, pedestrians and cyclists. Policy makers can incorporate this in the driving tests required to get a licence in NYC.
4. Our K-Means clustering revealed that in the In the borough of Manhattan, the most severe accidents are likely to occur in the morning hours and their most common cause is driver inattention. This could be because people are commuting to work in a hurry.

   To overcome this, policy-makers can enforce stricter regulation in the morning by deploying more traffic constables. They can also introduce stricter penalties such as higher fines during peak hours in the morning so that people are more cautious at that time.

5. This is a somewhat intuitive result but accidents fell to half their average count in 2020 and a similar trend can be expected in 2021 given the Covid 19 pandemic. This gives policy makers the leeway to focus on other pressing domains like healthcare and worry less about transportation given the pandemic has resulted in lower casualties.
6. We learned from our analysis that accidents are most likely to happen between 2pm and 7pm. This makes sense as this is a busy time in the city with most commercial activity peaking. Similarly, we can see that accidents are least likely to happen between 1am and 7am which is when most of the city is resting. Policy makers could thus allocate more resources such as traffic police during rush hour compared to having 24/7 duties for police even though accidents are low at night.

7. We noticed that the borough entries in our data set sometimes did not match the location given via longitude and latitude. We dealt with these points as outliers, however policy makers should ensure that data entry is done more carefully so that they do not lose out on valuable data to gain useful inferences from.
8. Since human errors like distraction and driving too close were the most common reasons for accidents in New York City, this is an indicator that policy makers should implement stricter laws with higher penalties to encourage drivers to drive more mindfully.