Northeastern University
College *of* Professional Studies

ALY 6110 : DATA MANAGEMENT & BIG DATA

MOTOR VEHICLE COLLISION CRASH ANALYSIS

Presented by-

RICHA UMESH RAMBHIA

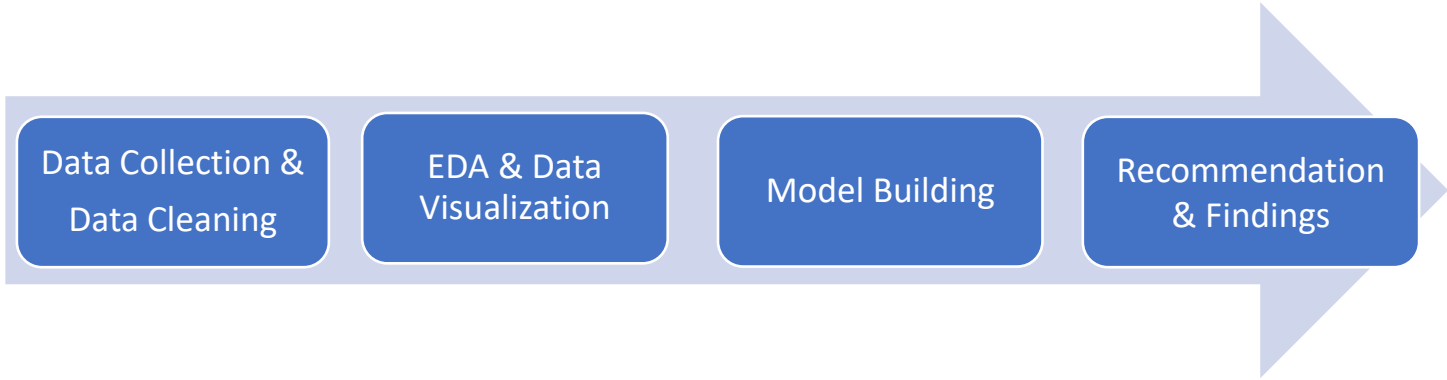
VAIDEHI CHAUHAN

Motor Vehicle Collision Crash

The data for the vehicle collisions and accidents have been collected which helps in analysis such that the incidents can be prevented, and safety measures can be taken accordingly. The goal here is to identify the factors that are causing the vehicle collisions and recommend which factor would lead to a severe incident or a non-severe incident such that appropriate measures can be taken to order to prevent the collisions and accidents and ensure safety of the public.

Problem Statement:

- To analyze the dataset using EDA to gain further insights about the collisions data.
- To determine the total number of deaths and injuries due to the collision crash.
- Analyzing the number of persons injured and killed based on the various contributing factor and vehicle count factor to recommend those factors causing the accident such that it can be prevented.



Tasks Performed:

1. Data Collection & Data Preparation
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Visualization
5. Pre-Modeling Steps
6. Model Building
7. Recommendation & Findings

Motor Vehicle Collision Crash

Data Collection & Preparation

- The dataset collected is from **NYC Open Data** which contains information about the various crashes that took place having various attributes describing each incident.
- This dataset is used in analyzing the crashes and understanding whether someone was injured or killed and what kind of injury did they face.
- It consists data of various crashes having **1931867 rows** of data and about **29 field values.**
- In order to better understand the dataset, **Descriptive & Statistical Analysis** is performed on this collisions data to gain further insights.

```
#displaying the number of rows and columns of the dataset
print("Total number of Rows and Columns:",collision_data.shape)
```

```
Total number of Rows and Columns: (1931867, 29)
```

```
Column Names:
 Index(['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'],
      dtype='object')
```

Motor Vehicle Collision Crash

Descriptive & Statistical Analysis - EDA

The descriptive analysis gives a broad overview of the dataset and the statistical analysis helped in understanding the statistics of the dataset.

- As observed in the statistical analysis, the average count of the various parameters in the dataset are computed which clarifies the total count of the number of persons killed based on the contributing factors.
- If we consider the parameters of the number of persons killed and injured, the average count is as mentioned in the figure above.
- Also, the minimum and maximum values of the variables are computed which is 0.00 and 43.00 respectively for the number of persons killed injured which helps us understand the total number of persons injured on an average.
- Thus, the statistical analysis helps understand the statistics of the parameters of the dataset.

Statistical Analysis of the Dataset

	LATITUDE	LONGITUDE	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	COLLISION_ID
count	1708825.00	1708825.00	1931849.00	1931836.00	1931867.00	1931867.00	1931867.00	1931867.00	1931867.00	1931867.00	1931867.00
mean	40.64	-73.77	0.29	0.00	0.05	0.00	0.03	0.00	0.21	0.00	3049602.52
std	1.88	3.56	0.68	0.04	0.24	0.03	0.16	0.01	0.65	0.03	1502895.11
min	0.00	-201.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.00
25%	40.67	-73.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3096853.50
50%	40.72	-73.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3602131.00
75%	40.77	-73.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4085358.50
max	43.34	0.00	43.00	8.00	27.00	6.00	4.00	2.00	43.00	5.00	4568716.00

Motor Vehicle Collision Crash

Data Cleaning

- It is observed that the various location and street columns have maximum of null values of the entire dataset and thus those columns can be dropped.
- Apart from that, the remaining columns that have maximum null values and that cannot be dropped are explored further in order to replace the null values with some value suitable for the data points.
- Lastly, the number of persons killed and injured have some null values within the field value and thus the rows of data can be dropped as it is a minimal amount of data as compared to the entire dataset.

CRASH DATE	0
CRASH TIME	0
BOROUGH	599304
ZIP CODE	599538
LATITUDE	223042
LONGITUDE	223042
LOCATION	223042
ON STREET NAME	401257
CROSS STREET NAME	710430
OFF STREET NAME	1622980
NUMBER OF PERSONS INJURED	18
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	5814
CONTRIBUTING FACTOR VEHICLE 2	287618
CONTRIBUTING FACTOR VEHICLE 3	1796783
CONTRIBUTING FACTOR VEHICLE 4	1901886
CONTRIBUTING FACTOR VEHICLE 5	1923845
COLLISION_ID	0
VEHICLE TYPE CODE 1	11325
VEHICLE TYPE CODE 2	347740
VEHICLE TYPE CODE 3	1801212
VEHICLE TYPE CODE 4	1902849
VEHICLE TYPE CODE 5	1924070

Total number of null values in each column of the dataset

Motor Vehicle Collision Crash

Data Cleaning

```
#changing datatype of the variables

collision_data['CRASH DATE'] = collision_data['CRASH DATE'].astype('datetime64[ns]')
collision_data['CRASH TIME'] = collision_data['CRASH TIME'].astype('datetime64[ns]')
collision_data['BOROUGH'] = collision_data['BOROUGH'].astype('str')
#collision_data['ZIP CODE'] = collision_data['ZIP CODE'].astype('int')
collision_data['ON STREET NAME'] = collision_data['ON STREET NAME'].astype('str')
collision_data['CROSS STREET NAME'] = collision_data['CROSS STREET NAME'].astype('str')
collision_data['OFF STREET NAME'] = collision_data['OFF STREET NAME'].astype('str')
collision_data['CONTRIBUTING FACTOR VEHICLE 1'] = collision_data['CONTRIBUTING FACTOR VEHICLE 1'].astype('str')
collision_data['CONTRIBUTING FACTOR VEHICLE 2'] = collision_data['CONTRIBUTING FACTOR VEHICLE 2'].astype('str')
collision_data['CONTRIBUTING FACTOR VEHICLE 3'] = collision_data['CONTRIBUTING FACTOR VEHICLE 3'].astype('str')
collision_data['CONTRIBUTING FACTOR VEHICLE 4'] = collision_data['CONTRIBUTING FACTOR VEHICLE 4'].astype('str')
collision_data['CONTRIBUTING FACTOR VEHICLE 5'] = collision_data['CONTRIBUTING FACTOR VEHICLE 5'].astype('str')
collision_data['VEHICLE TYPE CODE 1'] = collision_data['VEHICLE TYPE CODE 1'].astype('str')
collision_data['VEHICLE TYPE CODE 2'] = collision_data['VEHICLE TYPE CODE 2'].astype('str')
collision_data['VEHICLE TYPE CODE 3'] = collision_data['VEHICLE TYPE CODE 3'].astype('str')
collision_data['VEHICLE TYPE CODE 4'] = collision_data['VEHICLE TYPE CODE 4'].astype('str')
collision_data['VEHICLE TYPE CODE 5'] = collision_data['VEHICLE TYPE CODE 5'].astype('str')

#filling null values with 0 and then converting the datatype
collision_data['NUMBER OF PERSONS KILLED'] = collision_data['NUMBER OF PERSONS KILLED'].fillna(0)
collision_data['NUMBER OF PERSONS INJURED'] = collision_data['NUMBER OF PERSONS INJURED'].fillna(0)

collision_data['NUMBER OF PERSONS KILLED'] = collision_data['NUMBER OF PERSONS KILLED'].astype('int')
collision_data['NUMBER OF PERSONS INJURED'] = collision_data['NUMBER OF PERSONS INJURED'].astype('int')

print("Datatype conversion completed!")
```

✓ 4.7s

```
Datatype conversion completed!
```

Datatype conversion

Replacing and dropping values of the dataset

```
#replacing null values/blanks of Borough column with 'No Value'
collision_data['BOROUGH'] = collision_data['BOROUGH'].replace('NaN', 'No Value')
print("Replace Successful")
```

[7] ✓ 0.1s Python

```
Replace Successful
```

```
#replacing null values/blanks of Contributing Factor 1 column with 'No Value'
collision_data['CONTRIBUTING FACTOR VEHICLE 1'] = collision_data['CONTRIBUTING FACTOR VEHICLE 1'].replace('NaN', 'No Value')
print("Replace Successful")
```

[8] ✓ 0.9s Python

```
Replace Successful
```

```
#replacing null values/blanks of Vehicle Type Code 1 column with 'No Value'
collision_data['VEHICLE TYPE CODE 1'] = collision_data['VEHICLE TYPE CODE 1'].replace('NaN','No Value')
print("Replace Successful")
```

[9] ✓ 0.1s Python

```
Replace Successful
```

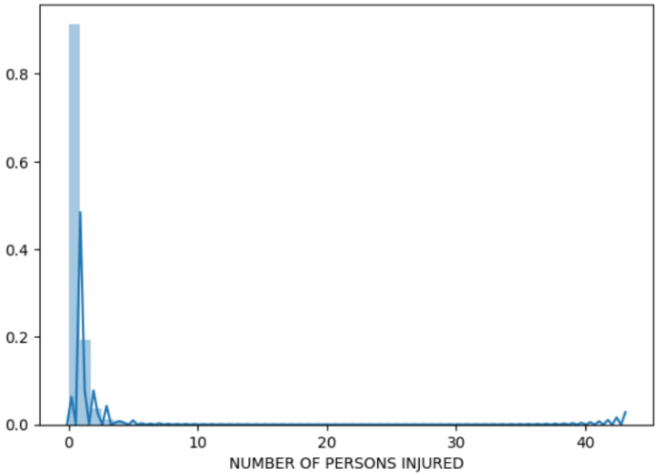
```
#dropping columns which have maximum null values and are not required for analysis
collision_data = collision_data.drop(['ZIP CODE', 'LATITUDE', 'LONGITUDE', 'LOCATION', 'ON STREET NAME', 'CROSS STREET NAME', 'OFF STREET NAME', 'CONTRIBUTING FACTOR VEHICLE 2',
'CONTRIBUTING FACTOR VEHICLE 3','CONTRIBUTING FACTOR VEHICLE 4','CONTRIBUTING FACTOR VEHICLE 5', 'VEHICLE TYPE CODE 2', 'VEHICLE TYPE CODE 3', 'VEHICLE TYPE CODE 4',
'VEHICLE TYPE CODE 5'], axis=1)
print("Variables dropped!")
```

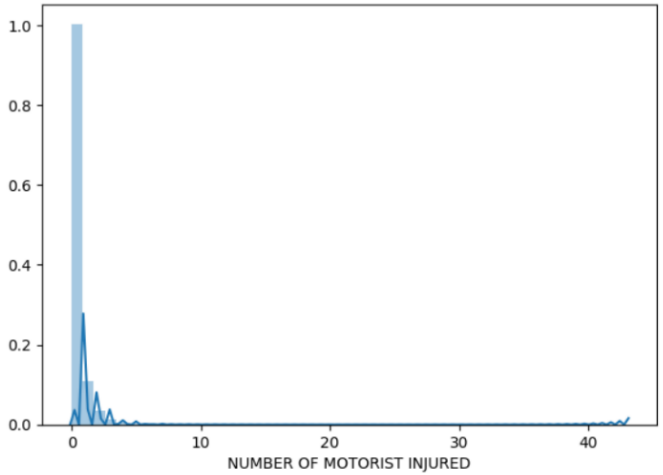
[10] ✓ 1.1s Python

```
Variables dropped!
```

Motor Vehicle Collision Crash

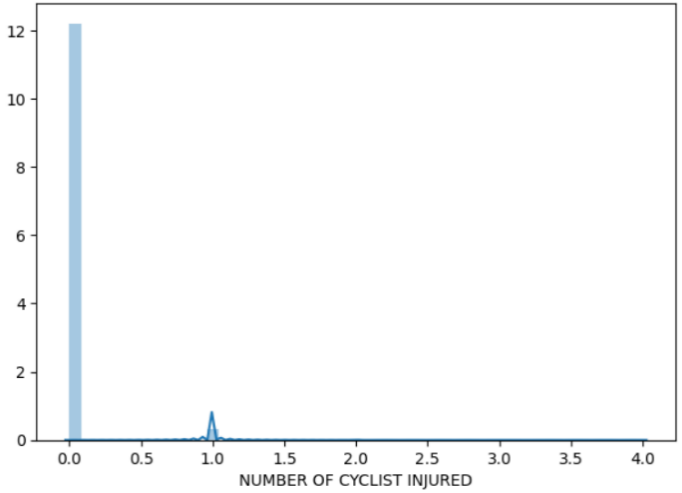
Data Visualization

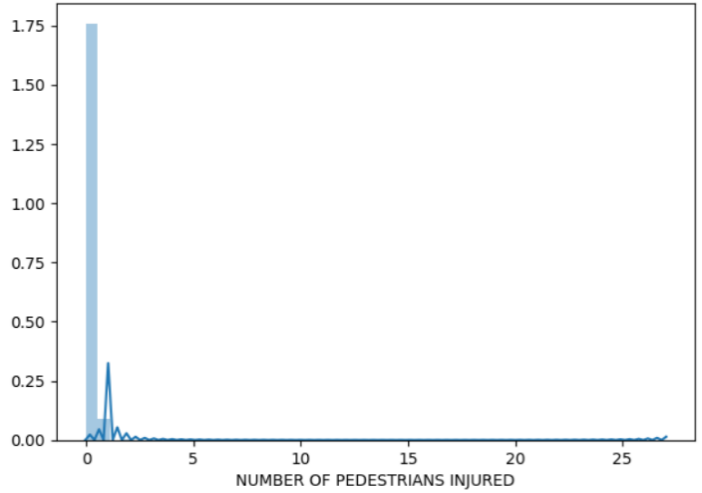




Distribution Plot for number of persons injured and number of motorist injured

The distribution plot for the various field values of the dataset show that the data for the data values is skewed.



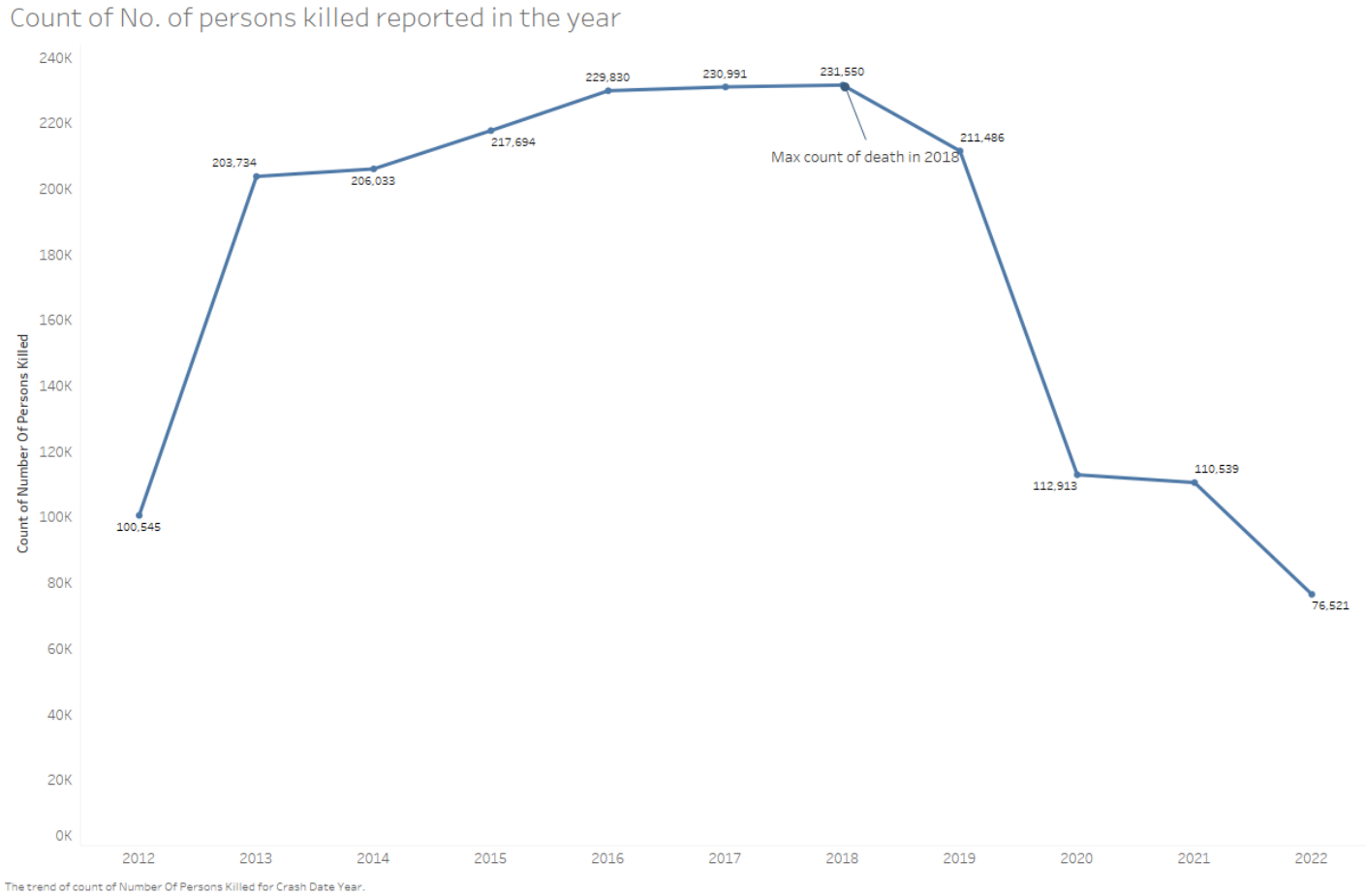


Distribution Plot for number of cyclist injured and number of pedestrians injured

Motor Vehicle Collision Crash

Data Visualization

- The visual represents the count of number of persons killed which were reported in the year from **2012 to 2022**.
- As it can be observed, the count of persons killed **increased rapidly after 2012**, and kept increasing for the rest of the years.
- In the year **2018**, the count of persons killed is the highest as compared to the other years which is **231,550,** after which the count decreased from the year **2019 to 2022.**
- The year **2022** encountered minimum of death rates, i.e., **76,521**, as compared to others and thus it is observed that appropriate measures were taken.

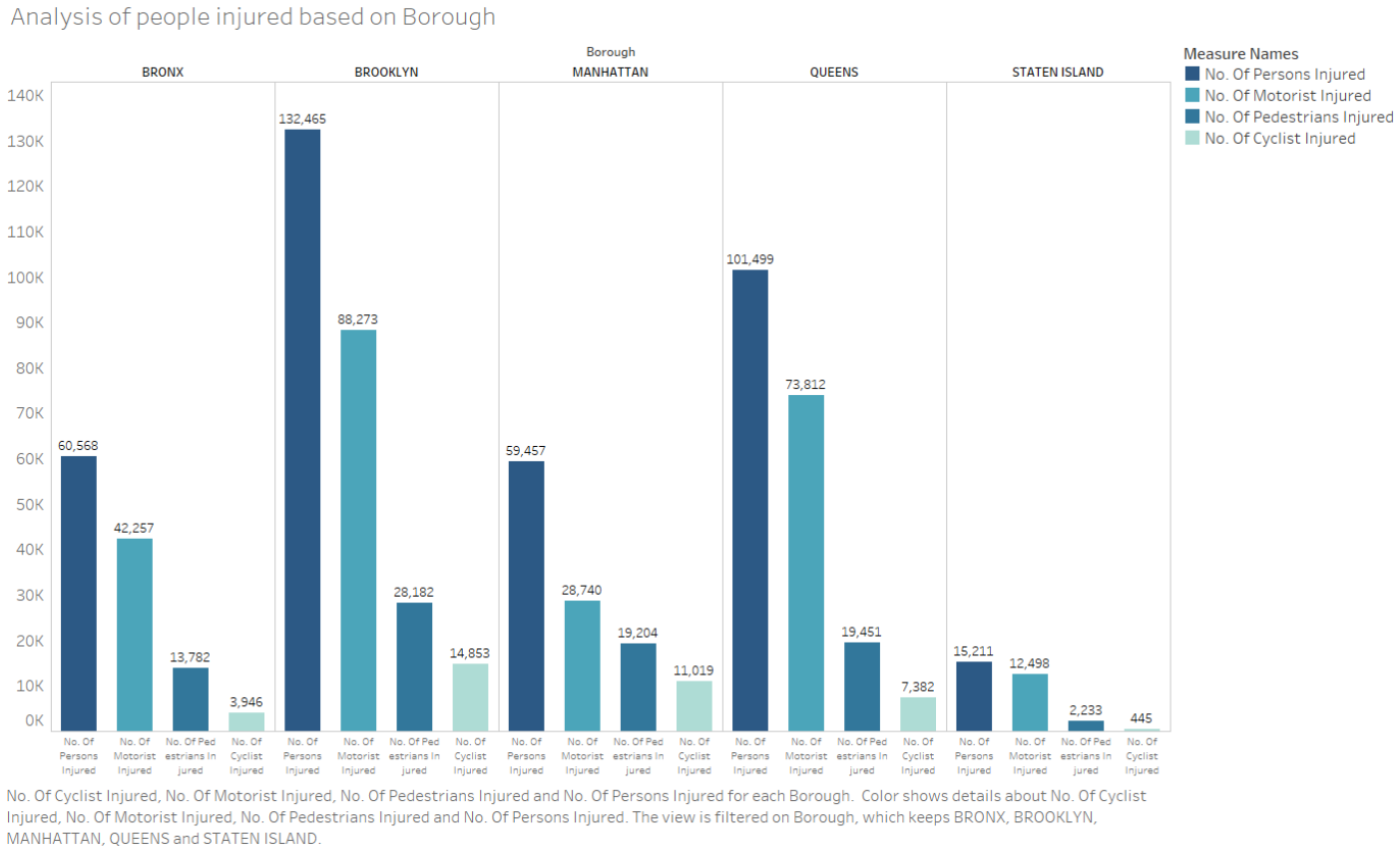


Count of number of persons killed reported in the year

Motor Vehicle Collision Crash

Data Visualization

- The analysis of people injured based on Borough gives an idea about the persons injured, the motorist injured, the cyclist injured, and the pedestrians injured which is filtered based on the Borough.
- Thus, it is observed that **Brooklyn** has the highest number of persons injured which is **132,465** whereas **Staten Island** has the least number of persons injured, i.e., **15,211**.

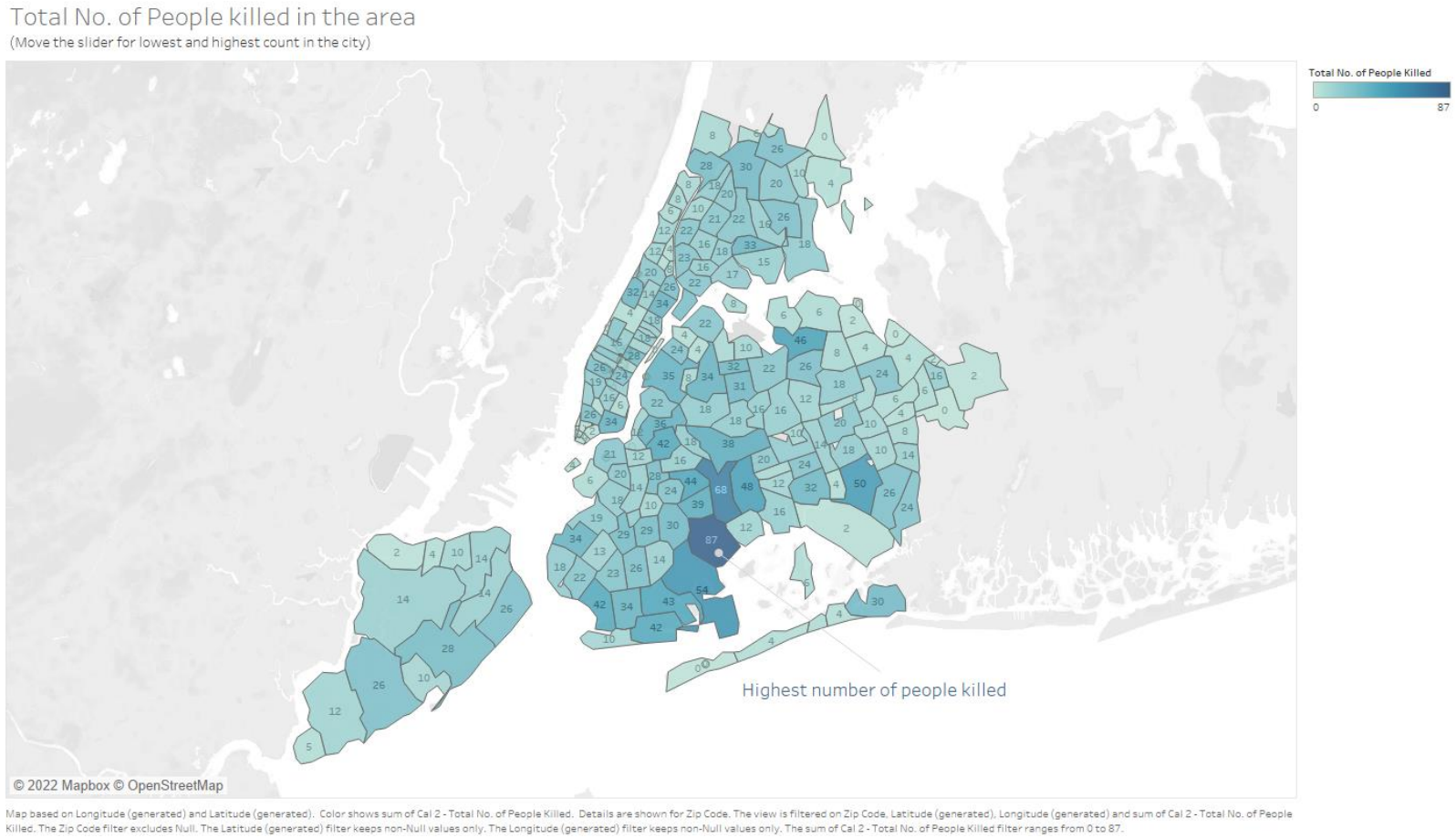


Analysis of people injured based on Borough

Motor Vehicle Collision Crash

Data Visualization

- Here, the sum of number of persons killed, number of motorist killed, number of cyclist killed, and number of pedestrians killed are taken into consideration for an overall analysis to understand the average of number of people killed in the New York state, by area.
- The highest number of people killed is in the area with zip code 11236, which is 87. For this graph, a slider is enabled which would help analyze the data in an effective way for a particular area.

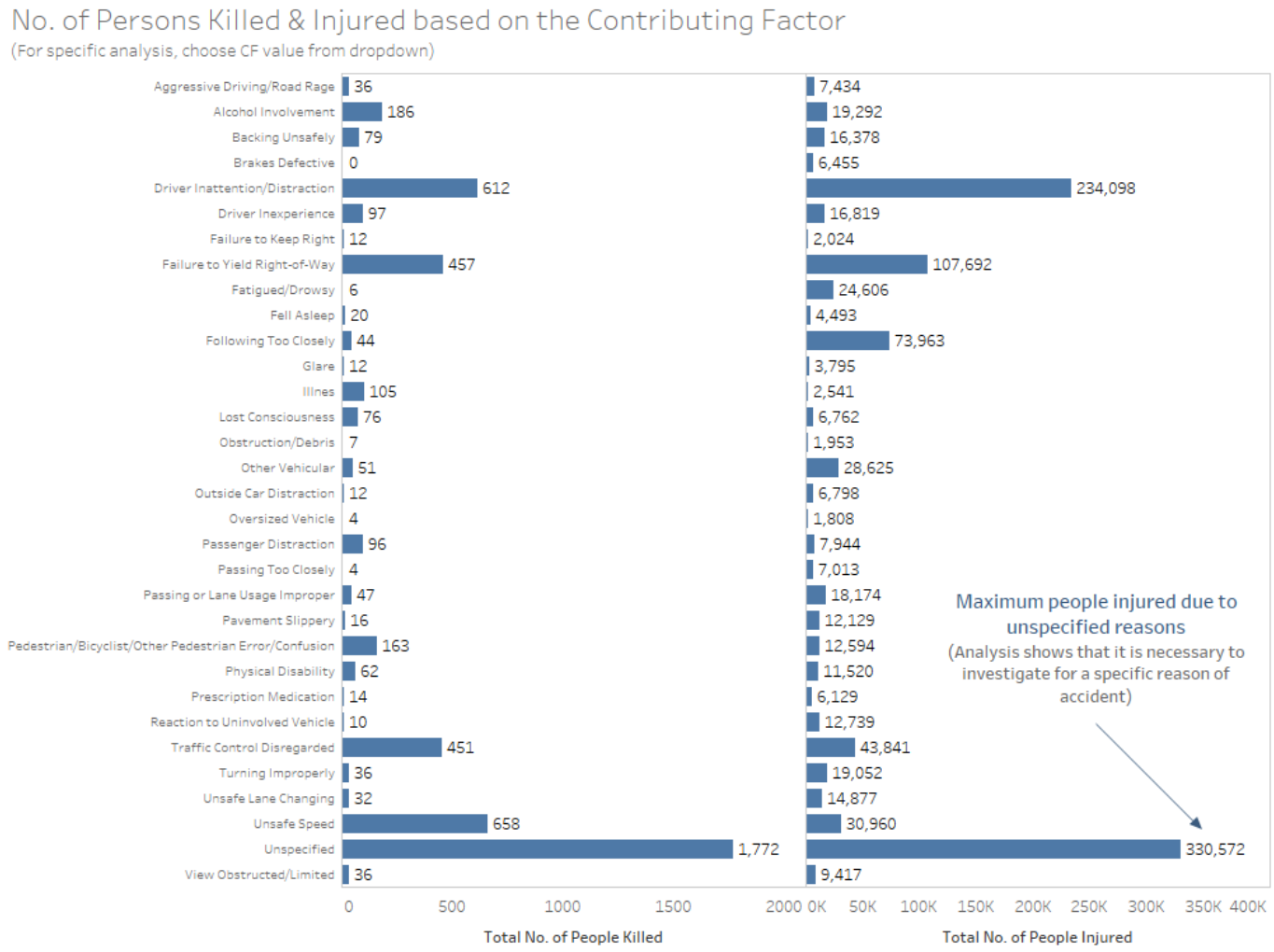


Total number of people killed in the area

Motor Vehicle Collision Crash

Data Visualization

- Contributing factors are the features that help understand why the accident or the crash collision happened in the first place.
- The dropdown function here would help in analyzing the number of persons killed and injured based on the specific contributing factor.
- But the observation here is that the highest number of persons killed and injured are due to **unspecified reasons**, and thus it is recommended to investigate these incidents in order to have a specific reason of accident.

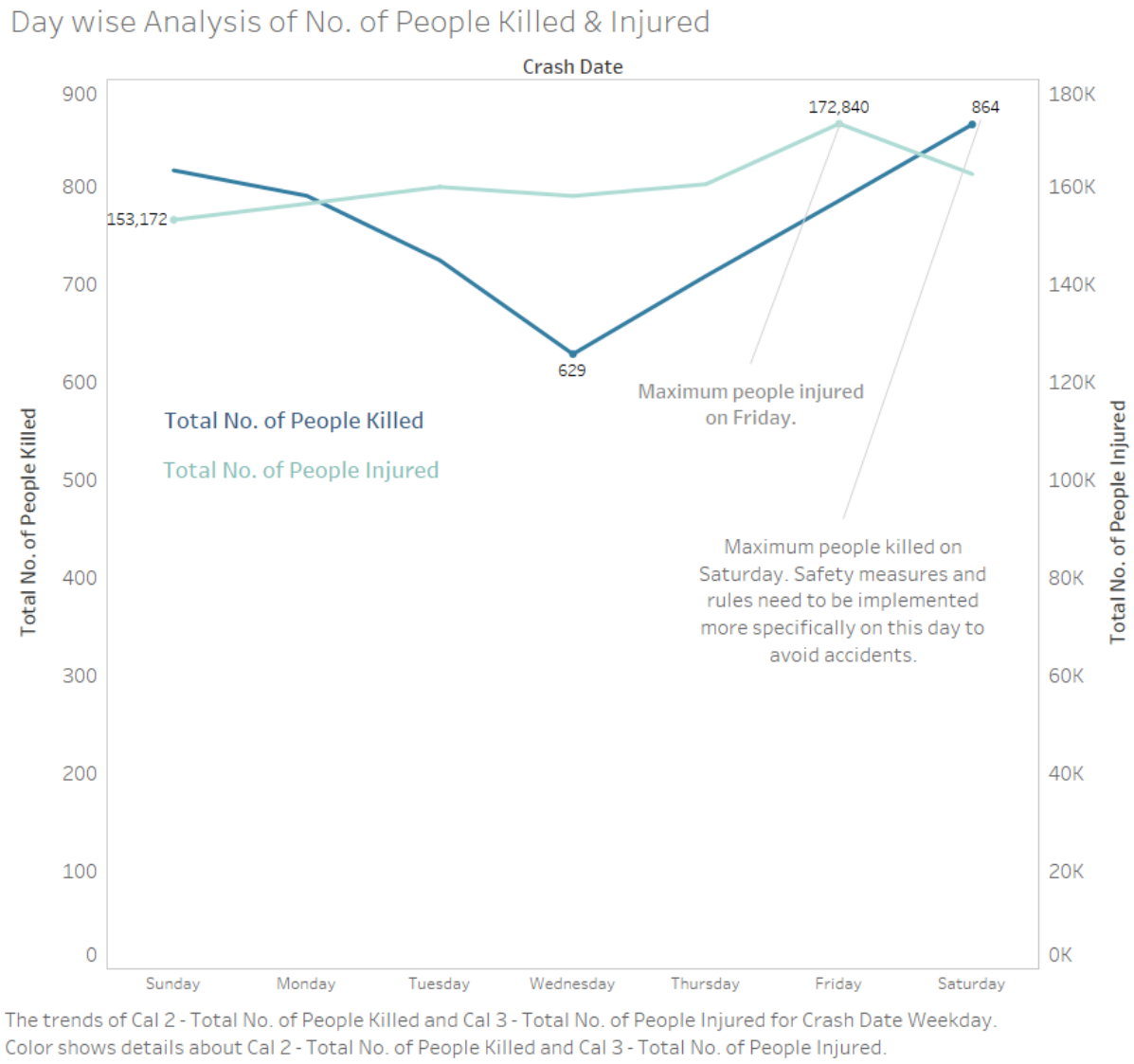


Number of persons killed & injured based on contributing factor

Motor Vehicle Collision Crash

Data Visualization

- The day wise analysis of number of people killed and injured help in analyzing the crash collisions on a regular basis, specifically analyzing in a day wise manner.
- The graph thus represents both the total number of persons killed and injured during each day and we can see that *maximum people meet with an accident* on **Saturday** and are *highest number of people are injured* on **Friday**.
- This recommends implementing strict actions and rules specifically on these days which will help avoid the crash collisions and reduce the deaths and injuries.

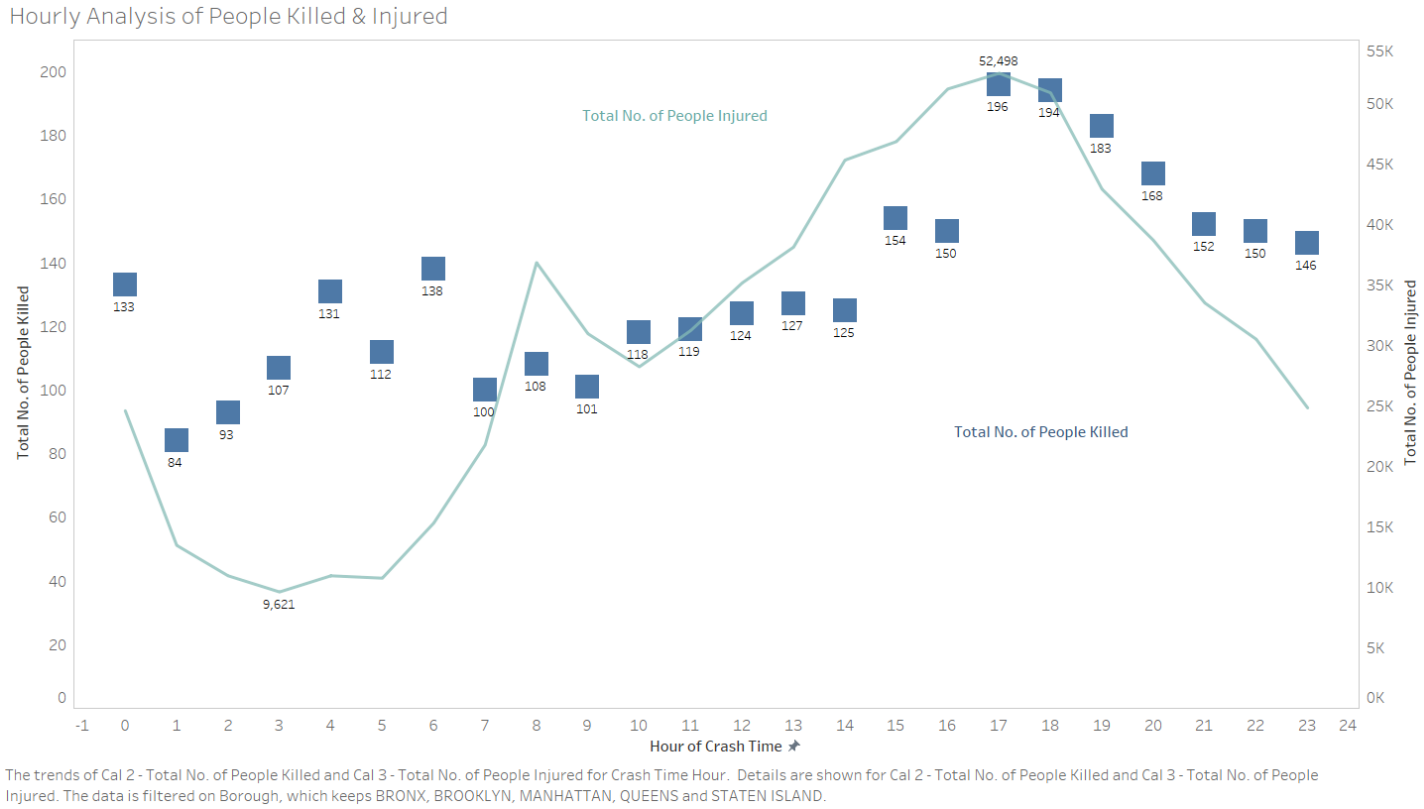


Day wise Analysis of number of people killed & injured

Motor Vehicle Collision Crash

Data Visualization

- Like the day wise analysis of the number of people killed and injured, an hourly analysis of the same is implemented in order to understand the graph with respect to the hourly manner and understand the trend of the highest and lowest accidents in each hour of the day.
- The square shapes indicate the total number of people killed and the line graph indicates the total number of people injured, thus the visual is a combination graph.
- The highest & lowest number of people killed & injured are indicated in the graph.



Hourly Analysis of people killed & injured

Motor Vehicle Collision Crash

Pre-Modeling Steps

1. Label Encoding
2. Correlation Plot
3. Feature Selection & Extraction

The features such as '**Borough**', '**Contributing Factor Vehicle**', and '**Vehicle Type Code**' are label encoded to numerical form which can be considered as features to training of the model.

BOROUGH	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	COLLISION_ID	VEHICLE TYPE CODE 1	Borough Labels	Contributing Factor Vehicle 1 Labels	Vehicle Type Code 1 Labels
nan	2	0	0	0	0	0	2	0	Aggressive Driving/Road Rage	4455765	Sedan	5	3	893
nan	1	0	0	0	0	0	1	0	Pavement Slippery	4513547	Sedan	5	39	893
nan	0	0	0	0	0	0	0	0	Following Too Closely	4541903	Sedan	5	21	893
BROOKLYN	0	0	0	0	0	0	0	0	Unspecified	4456314	Sedan	1	56	893
BROOKLYN	0	0	0	0	0	0	0	0	nan	4486609	nan	1	61	1287

Label Encoding

These features would help in the prediction of **total number of persons killed** based on various features selected for training of the model and to understand the various contributing factors that are affecting the death rates and injuries of the person.

```
print(features)
```

✓ 0.4s Python

```
['Borough Labels', 'NUMBER OF PEDESTRIANS INJURED', 'NUMBER OF PEDESTRIANS KILLED', 'NUMBER OF CYCLIST INJURED', 'NUMBER OF CYCLIST KILLED', 'NUMBER OF MOTORIST KILLED', 'Contributing Factor Vehicle 1 Labels', 'Vehicle Type Code 1 Labels', 'NUMBER OF PERSONS INJURED']
```

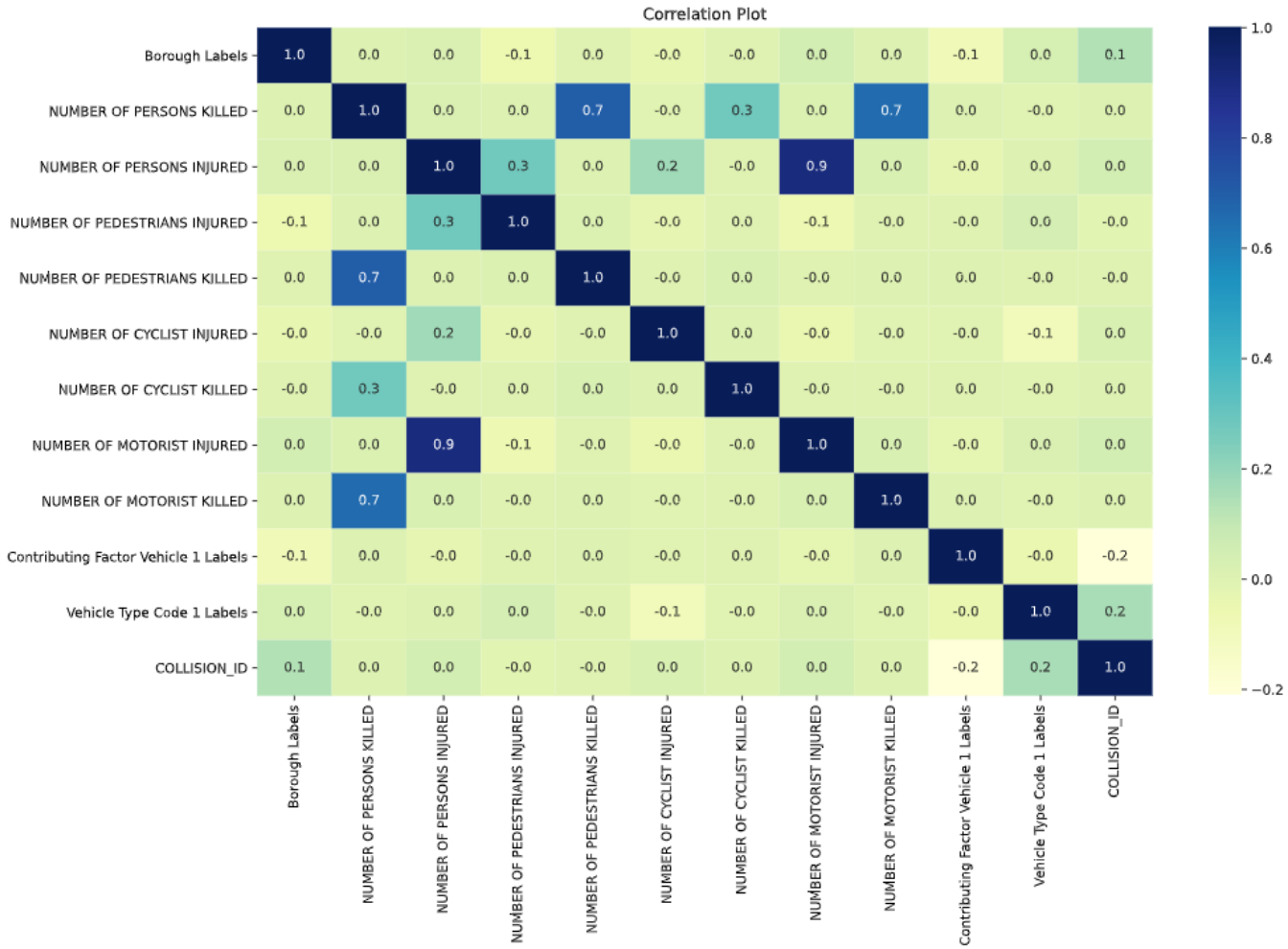
Feature Selection for Modeling

Motor Vehicle Collision Crash

Pre-Modeling Steps

1. Label Encoding
2. Correlation Plot
3. Feature Selection & Extraction

- From the correlation plot it is observed that since there are different variables and parameters which have a **high collinearity** that is existing with a correlation value of **0.9** and **0.7**.
- This would result in multicollinearity and the model would outperform resulting in inaccurate model and results.
- Thus, some of the features having high collinearity are dropped before considering them for training of the model based on which features are important for the model.



Correlation Plot

Motor Vehicle Collision Crash

Model Building

Machine Learning Models Implemented.

- Linear Regressor Model
- Decision Tree Regressor Model
- Random Forest Regressor Model

Linear Regressor Model

- The accuracy obtained for both the training and testing set of the Linear Regression model is **98% and 98.26%** respectively.
- Since this is a regressor type of model, the model evaluation is based on the **MAE, MSE, RMSE and R-Squared** values in order to determine the prediction error of the model implemented.

Accuracy of the Linear Regressor Model

```
Accuracy of Linear Regressor model on training set: 0.98
Accuracy of Linear Regressor model on test set:     0.98

0.9826
```

```
Model Evaluation of Linear Regression.
Mean Absolute Error: 0.0
Mean Squared Error: 0.0
Root Mean Squared Error: 0.0
R-Squared value: 0.9826460733537029
```

Model Evaluation

Motor Vehicle Collision Crash

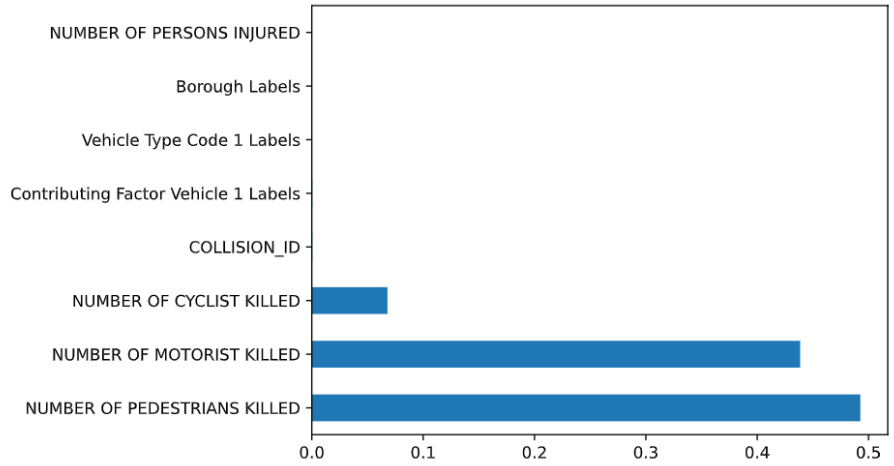
Decision Tree Regressor Model

- The independent variables considered for the implementation of the Decision Tree Regressor are **borough, number of persons injured, number of pedestrians injured, number of pedestrians killed, number of cyclist killed, number of cyclist injured, contributing factors and vehicle type code**.
- The decision tree model is implemented with a train and test data split of **80-20 with a maximum branch depth of 5.**

```
Model Evaluation of Decision Tree Regressor.
Mean Absolute Error: 0.0
Mean Squared Error: 0.0
Root Mean Squared Error: 0.0
R-Squared value: 0.9808530602314666
```

Model Evaluation

- The accuracy of the Decision Tree model obtained for the prediction of number of persons killed is **99%** and **98.1%** for both the **training and testing set** respectively



Feature Importance of the Decision Tree Regressor Model

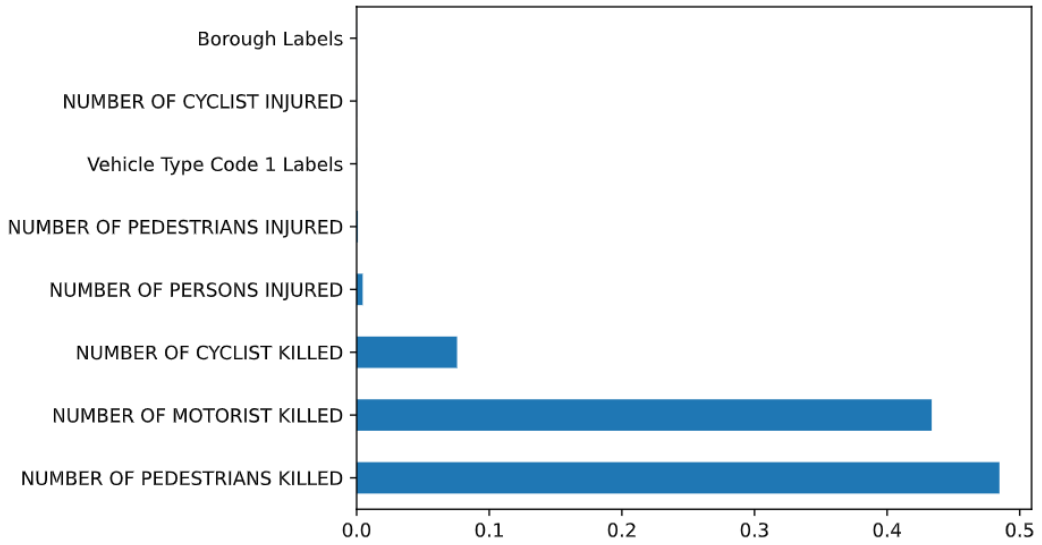
Motor Vehicle Collision Crash

Random Forest Regressor Model

- The data is split into **80-20 ratio** for training and testing of the model where the Random Forest Regressor is implemented with a minimum of **500 tress** and **maximum depth branch of 3.**
- The accuracy of the model obtained is **97% for training data and 97.7% for test dataset.**

```
Model Evaluation of Random Forest Regressor.
Mean Absolute Error: 0.0
Mean Squared Error: 0.0
Root Mean Squared Error: 0.0
R-Squared value: 0.9768055430330816
```

Model Evaluation



Feature Importance of the Random Forest Regressor Model

Motor Vehicle Collision Crash

Model Evaluation & Comparison

- Based on the model evaluation and comparison, it is observed that the accuracy for **Linear Regressor model** is more as compared to other two models.
- Thus, based on the evaluation metrices, it is observed that the **root mean squared error** is less in all the three model which implies that the **prediction rate error is less**. Hence, based on **the accuracy of the model and the r-squared value**, **Linear Regressor Model is recommended** to use for the **prediction of the number of persons killed based on the contributing factors**.

	Machine Learning Regressor Models		
Evaluation Metric	Linear Regressor	Decision Tree Regressor	Random Forest Regressor
Accuracy	98.26%	98.1%	97.7%
Mean Absolute Error	0.0	0.0	0.0
Mean Squared Error	0.0	0.0	0.0
Root Mean Squared Error	0.0	0.0	0.0
R-Squared value	0.9826	0.9808	0.9768

Motor Vehicle Collision Crash

Recommendation

- One of the major causes of death in the US is car crashes. The number of pedestrian and bicycle fatalities is increasing, which is causing more individuals to die unexpectedly. The data and visualization shows that the statistics separated by years, with 2013 having the greatest number of fatalities, followed by 2021.
- After digging a little further to see which **Borough** had the most fatalities, we discovered that while cyclists had the fewest there, drivers **on Staten Island had the most fatalities**. **Brooklyn is the most populated** Borough in New York and has the most fatal pedestrian and bike accidents.
- The evaluation of the regressor models are based on the MAE, MSE, RMSE, and R-squared values.
- The accuracy obtained for the models are **98.3% for Linear Regressor model, 98.1% for Decision Tree** and **97.7% for Random Forest Regressor model**.
- Hence, **Linear Regressor model** can be considered as the **best fit model for the prediction of the number of persons killed.**

Motor Vehicle Collision Crash

Conclusion

- The project deals with analyzing the data to understand the factors due to which the accident took place and the number of **people killed and injured due to the crash**.
- The data and visualization shows that the statistics separated by years, with **2013 having the greatest number of fatalities**, followed by 2021.
- After digging a little further to see which **Borough** had the most fatalities, we discovered that while cyclists had the fewest there, **drivers on Staten Island had the most fatalities**. **Brooklyn** is the most populated Borough in New York and has the most fatal pedestrian and bike accidents.
- Based on the **day wise or hour wise analysis** of the crashes that have taken place, it is observed that **majorly on weekends** the **crash rate increases,** which could possibly be due to the week off, and thus people spend time on outings which could lead to traffic and has a high possibility for crashes to happen.
- **Linear Regressor model** can be considered as the **best fit model** for the prediction of the number of persons killed based on the accuracy and model evaluation.

Motor Vehicle Collision Crash

THANK YOU