Module 5: Capstone Project Draft Report

Professor: Daya Rudhramoorthi

Course: ALY6140: Python & Analytics Systems Technology

Submitted by

Group 6

Shyam Kumar Chittaluru

Sameeksha Santhosh

10-12-2023

Introduction

Title: Motor Vehicle Collisions – Crashes Public Safety

In this comprehensive analysis of motor vehicle collisions in New York City (NYC), the project aims to shed light on critical aspects of road safety. Leveraging the extensive Motor Vehicle Collisions dataset, we strive to uncover patterns, contributing factors, and trends that are essential for evidence-based strategies to enhance public safety. This analysis aligns with the overarching goal of providing valuable insights for targeted interventions and informed policy recommendations, contributing to a safer traffic environment in NYC.

Dataset Source: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Columns in this Dataset

1. Column Name: Description

2. CRASH DATE: Occurrence date of collision

3. CRASH TIME: Occurrence time of collision

4. BOROUGH: Borough where the collision occurred

5. ZIP CODE: Postal code of incident occurrence

6. LATITUDE: Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees

7. LONGITUDE: Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees

8. ON STREET NAME: The street on which the collision occurred

9. CROSS STREET NAME: Nearest cross street to the collision

10. OFF STREET NAME: Street address if known

11. NUMBER OF PERSONS INJURED: Number of persons injured

12. NUMBER OF PERSONS KILLED: Number of persons killed

13. NUMBER OF PEDESTRIANS INJURED: Number of pedestrians injured

14. NUMBER OF PEDESTRIANS KILLED: Number of pedestrians killed

15. NUMBER OF CYCLIST INJURED: Number of cyclists injured

16. NUMBER OF CYCLIST KILLED: Number of cyclists killed

17. NUMBER OF MOTORIST INJURED: Number of vehicle occupants injured

18. NUMBER OF MOTORIST KILLED: Number of vehicle occupants killed

19. CONTRIBUTING FACTOR VEHICLE 1: Factors contributing to the collision for the designated vehicle

20. CONTRIBUTING FACTOR VEHICLE 2: Factors contributing to the collision for the designated vehicle

21. CONTRIBUTING FACTOR VEHICLE 3: Factors contributing to the collision for the designated vehicle

22. CONTRIBUTING FACTOR VEHICLE 4: Factors contributing to the collision for designated vehicle

24. CONTRIBUTING FACTOR VEHICLE 5: Factors contributing to the collision for designated vehicle

25. COLLISION_ID: Unique record code generated by system. Primary Key for Crash table.

26. VEHICLE TYPE CODE 1: Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

27. VEHICLE TYPE CODE 2: Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

28. VEHICLE TYPE CODE 3: Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

29. VEHICLE TYPE CODE 4: Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

30. VEHICLE TYPE CODE 5: Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

Goals of the Project:

1. Borough-wise Analysis: Understand the distribution of motor vehicle collisions across different boroughs in NYC and identify areas with higher incident rates.

2. Temporal Trends: Explore temporal patterns in motor vehicle collisions over time, including monthly and yearly variations, to identify trends and inform seasonal safety measures.

3. Spatial Analysis: Pinpoint high-incidence locations of collisions in NYC, contributing to the identification of areas requiring focused interventions and infrastructure improvements.

4. Contributing Factors: Investigate the primary contributing factors to motor vehicle collisions, with a focus on the top factors influencing road safety.

5. Injury Severity by Borough: Analyze the severity of injuries resulting from collisions across different boroughs to prioritize safety measures.

Questions to Investigate:

1. What are the primary contributing factors to motor vehicle collisions in different boroughs of NYC?

2. How has the implementation of Traffic Stat and Vision Zero initiatives affected the trends in traffic safety over the years?

3. Can we identify specific intersections or streets with a high incidence of collisions, and what factors contribute to these locations?

4. Are there seasonal or temporal patterns in collisions, and how do they correlate with external factors such as weather conditions?

These questions aim to uncover insights into the causes, trends, and patterns of motor vehicle collisions in New York City, providing valuable information for targeted interventions and policy recommendations.

Predictive Models:

1**. Random Forest Model**:

Purpose: Predict and classify the likelihood of collisions based on various features.

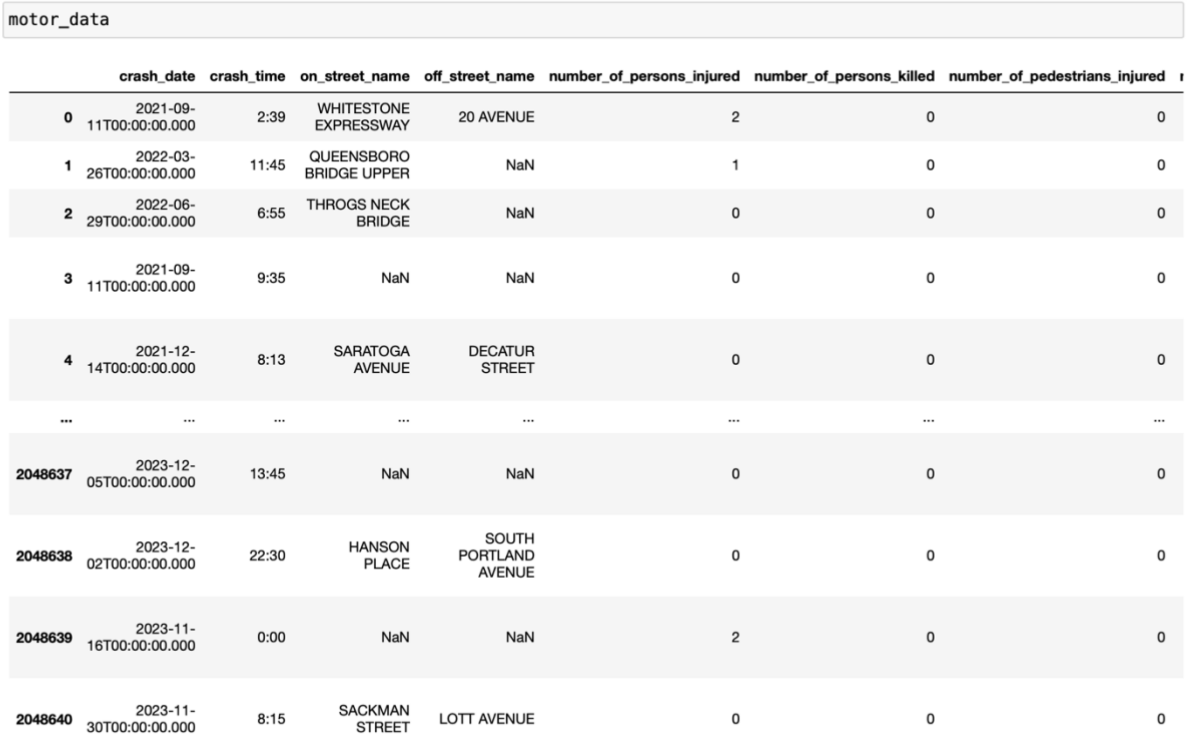
Method: Ensemble learning algorithm combining multiple decision trees for accurate predictions.

2. **Gradient Boosting Model:**

Purpose: Utilize boosting technique to improve prediction accuracy.

Method: Gradient boosting algorithm that builds decision trees sequentially to enhance model performance.

Dataset Glimpse:



```
motor_data
```

	crash_date	crash_time	on_street_name	off_street_name	number_of_persons_injured	number_of_persons_killed	number_of_pedestrians_injured
0	2021-09-11T00:00:00.000	2:39	WHITESTONE EXPRESSWAY	20 AVENUE	2	0	0
1	2022-03-26T00:00:00.000	11:45	QUEENSBORO BRIDGE UPPER	NaN	1	0	0
2	2022-06-29T00:00:00.000	6:55	THROGS NECK BRIDGE	NaN	0	0	0
3	2021-09-11T00:00:00.000	9:35	NaN	NaN	0	0	0
4	2021-12-14T00:00:00.000	8:13	SARATOGA AVENUE	DECATUR STREET	0	0	0
...	...	...	...	...	...	...	...
2048637	2023-12-05T00:00:00.000	13:45	NaN	NaN	0	0	0
2048638	2023-12-02T00:00:00.000	22:30	HANSON PLACE	SOUTH PORTLAND AVENUE	0	0	0
2048639	2023-11-16T00:00:00.000	0:00	NaN	NaN	2	0	0
2048640	2023-11-30T00:00:00.000	8:15	SACKMAN STREET	LOTT AVENUE	0	0	0

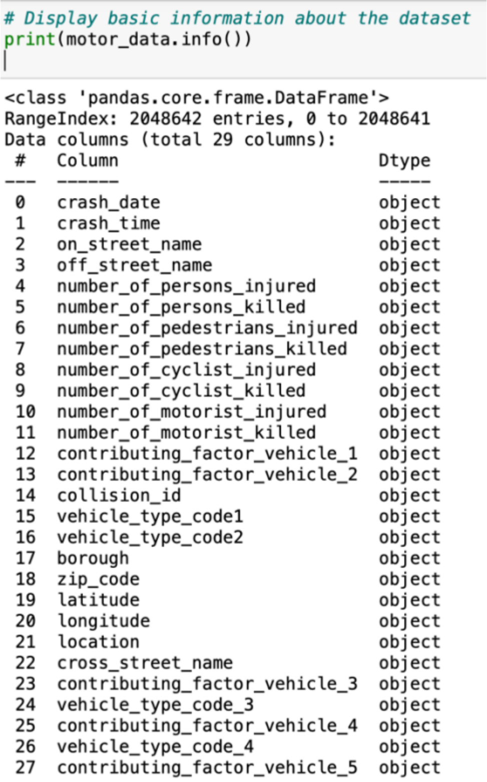
Exploratory Data Analysis

Data Loading Overview: In this analysis, we harnessed the power of the New York City Open Data API through the Socrata API client to retrieve valuable information on motor vehicle collisions. Leveraging this API allowed us to access a comprehensive dataset with over 2 million records, providing a detailed account of traffic incidents in New York City.

The process involved establishing a connection to the API, and we efficiently retrieved the dataset within seconds. The Socrata API client streamlined the data acquisition, offering a seamless and authenticated interface to explore and analyse real-world traffic collision data.

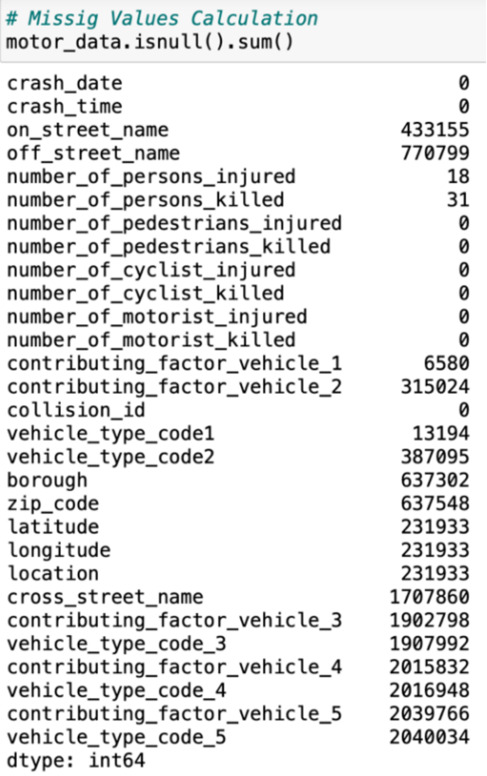
This dataset serves as the foundation for our in-depth examination of traffic patterns, safety trends, and other critical insights, contributing to a more informed understanding of urban mobility and road safety in New York City.

We have utilized the info() method to provide a concise summary of the dataset. It includes information such as the total number of entries, the data types of each column, and the count of non-null values. This output helps in understanding the structure of the dataset and identifying potential data types or missing values.



```
# Display basic information about the dataset
print(motor_data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2048642 entries, 0 to 2048641
Data columns (total 29 columns):
 #   Column                         Dtype
---  ------                         -----
 0   crash_date                     object
 1   crash_time                     object
 2   on_street_name                 object
 3   off_street_name                object
 4   number_of_persons_injured      object
 5   number_of_persons_killed       object
 6   number_of_pedestrians_injured  object
 7   number_of_pedestrians_killed   object
 8   number_of_cyclist_injured      object
 9   number_of_cyclist_killed       object
 10  number_of_motorist_injured     object
 11  number_of_motorist_killed      object
 12  contributing_factor_vehicle_1  object
 13  contributing_factor_vehicle_2  object
 14  collision_id                   object
 15  vehicle_type_code1             object
 16  vehicle_type_code2             object
 17  borough                        object
 18  zip_code                       object
 19  latitude                       object
 20  longitude                      object
 21  location                       object
 22  cross_street_name              object
 23  contributing_factor_vehicle_3  object
 24  vehicle_type_code_3            object
 25  contributing_factor_vehicle_4  object
 26  vehicle_type_code_4            object
 27  contributing_factor_vehicle_5  object
```



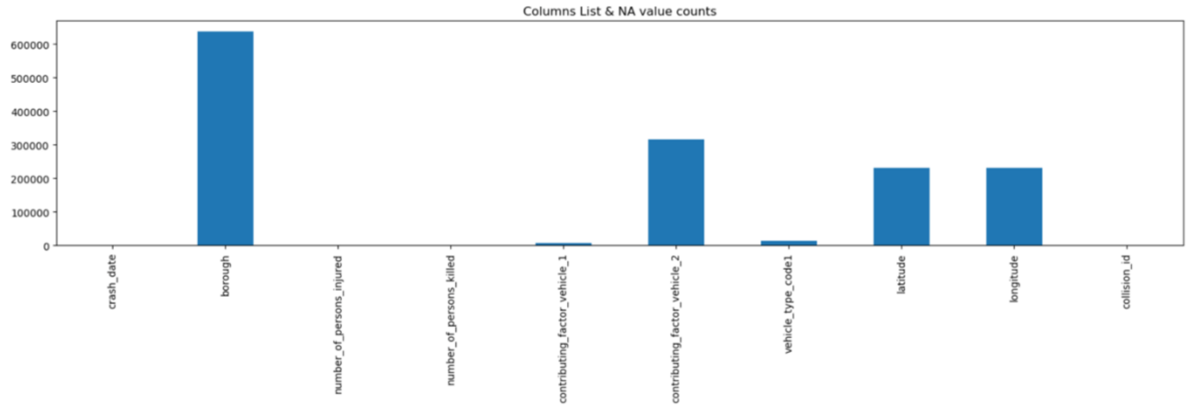
```
# Missig Values Calculation
motor_data.isnull().sum()

crash_date                             0
crash_time                             0
on_street_name                    433155
off_street_name                   770799
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       6580
contributing_factor_vehicle_2     315024
collision_id                           0
vehicle_type_code1                 13194
vehicle_type_code2                387095
borough                           637302
zip_code                          637548
latitude                          231933
longitude                         231933
location                          231933
cross_street_name                1707860
contributing_factor_vehicle_3    1902798
vehicle_type_code_3              1907992
contributing_factor_vehicle_4    2015832
vehicle_type_code_4              2016948
contributing_factor_vehicle_5    2039766
vehicle_type_code_5              2040034
dtype: int64
```

Also we employ the isnull() method to create a binary matrix where True indicates a missing value, and False indicates a non-missing value. The sum() function is then applied to each column, counting the number of missing values in each. The resulting output is a summary of the count of missing values for each attribute in the dataset. This information is crucial for assessing the completeness of the dataset and deciding on strategies for handling missing data, such as imputation or removal.

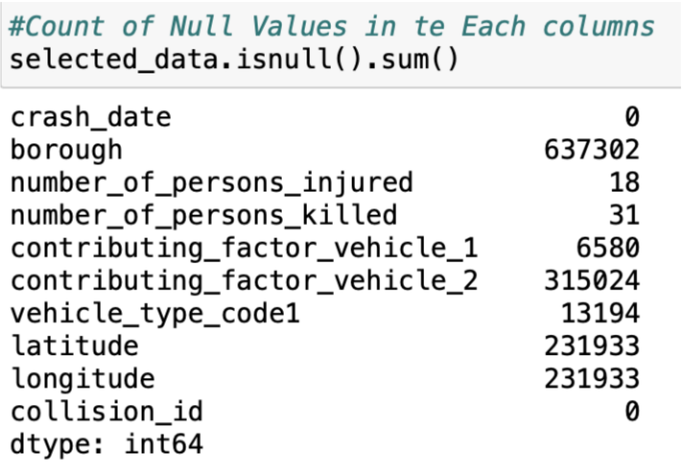
By combining these techniques, we gain insights into the dataset's structure and identify any potential data quality issues, paving the way for effective pre-processing and analysis.

In this part, a list of variables (selected_vars) relevant to the problem statement is defined. The dataset (motor_data) is then subsetted to include only the selected variables using the Pandas DataFrame indexing. Then we calculate the count of null values for each selected column and visualize the result using a bar chart. The x-axis represents the column names, and the y-axis represents the count of null values. The purpose is to identify columns with a high number of missing values, which may impact the analysis.



The above bar chart helps identify columns with missing values. In this case, the 'borough' column has the highest number of null values, indicating that a substantial portion of data for this variable is missing. The 'contributing_factor_vehicle_2' column has the second-highest count of null values. This information is valuable for deciding how to handle missing data in subsequent analysis, such as imputation or removal of rows/columns with missing values.

The below snapshot indicates the count of null values for each column in the ‘selected_data’ DataFrame:



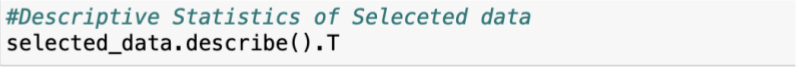
```
#Count of Null Values in te Each columns
selected_data.isnull().sum()
```

```
crash_date                           0
borough                         637302
number_of_persons_injured           18
number_of_persons_killed            31
contributing_factor_vehicle_1     6580
contributing_factor_vehicle_2   315024
vehicle_type_code1               13194
latitude                        231933
longitude                       231933
collision_id                         0
dtype: int64
```

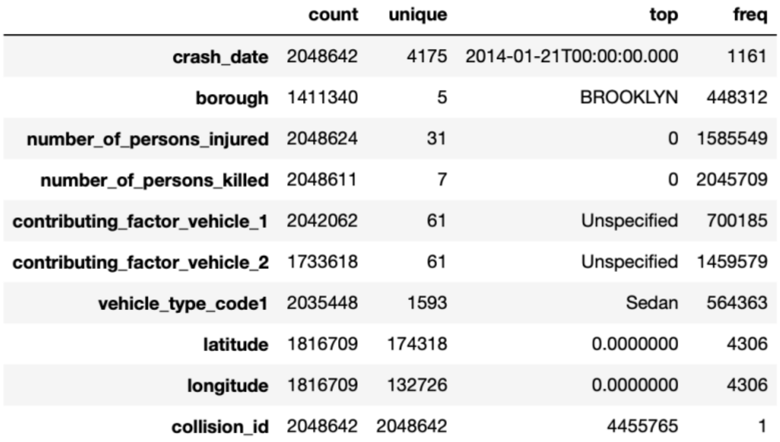
- The 'borough' column has a significant number of null values (637,302), indicating that a large portion of data for this variable is missing.

- 'contributing_factor_vehicle_2' and 'latitude'/'longitude' also have a considerable number of null values (315,024 and 231,933, respectively).
- Other columns have a relatively small number of null values.

Understanding the distribution of missing values is crucial for making informed decisions about data imputation, removal, or other pre-processing steps in subsequent analysis.



```
#Descriptive Statistics of Seleceted data
selected_data.describe().T
```



	count	unique	top	freq
crash_date	2048642	4175	2014-01-21T00:00:00.000	1161
borough	1411340	5	BROOKLYN	448312
number_of_persons_injured	2048624	31	0	1585549
number_of_persons_killed	2048611	7	0	2045709
contributing_factor_vehicle_1	2042062	61	Unspecified	700185
contributing_factor_vehicle_2	1733618	61	Unspecified	1459579
vehicle_type_code1	2035448	1593	Sedan	564363
latitude	1816709	174318	0.0000000	4306
longitude	1816709	132726	0.0000000	4306
collision_id	2048642	2048642	4455765	1

The describe() function provides descriptive statistics of the numeric columns in the ‘selected_data’ DataFrame. By examining these statistics, we can gain insights into the central tendency, spread, and distribution of the numerical variables in your dataset. This information is valuable for understanding the overall characteristics of the selected data.

Data Cleaning: The provided data cleaning script addresses several aspects of preparing the dataset for analysis:

Data Type Conversion:

- The 'crash_date' column is converted to a datetime format for proper handling of date-related operations.
- 'number_of_persons_injured' and 'number_of_persons_killed' are converted to numeric types and subsequently to integers, ensuring they are treated as whole numbers.

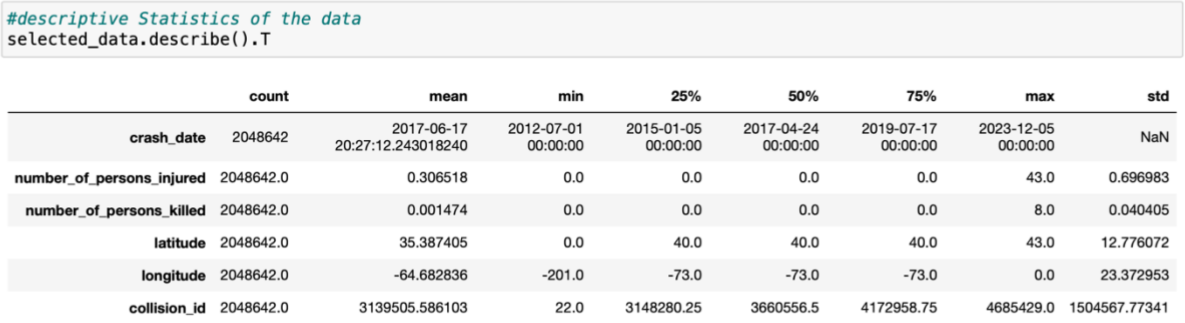
Handling Categorical Variables:

- For categorical variables like 'contributing_factor_vehicle_1', 'contributing_factor_vehicle_2', 'borough', and 'vehicle_type_code1', missing values are filled with the mode (most frequently occurring value) of each respective column.

Numeric Conversion and Handling Errors:

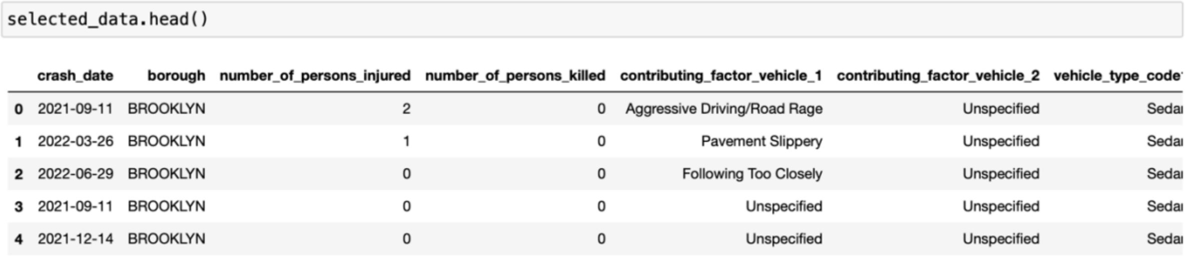
- 'latitude' and 'longitude' are converted to numeric types, with errors being coerced to NaN. NaN values are then filled with 0 for the time being.
- 'latitude', 'longitude', and 'collision_id' are further converted to integers.

These data-cleaning steps ensure that the selected variables have appropriate data types, handle missing values, and are ready for subsequent analysis. It's important to note that the choice of strategies for handling missing values (e.g., filling with the mode or 0) and the conversion of data types may depend on the specific requirements of the analysis and the characteristics of the dataset.



```
#descriptive Statistics of the data
selected_data.describe().T
```

	count	mean	min	25%	50%	75%	max	std
crash_date	2048642	2017-06-17 20:27:12.243018240	2012-07-01 00:00:00	2015-01-05 00:00:00	2017-04-24 00:00:00	2019-07-17 00:00:00	2023-12-05 00:00:00	NaN
number_of_persons_injured	2048642.0	0.306518	0.0	0.0	0.0	0.0	43.0	0.696983
number_of_persons_killed	2048642.0	0.001474	0.0	0.0	0.0	0.0	8.0	0.040405
latitude	2048642.0	35.387405	0.0	40.0	40.0	40.0	43.0	12.776072
longitude	2048642.0	-64.682836	-201.0	-73.0	-73.0	-73.0	0.0	23.372953
collision_id	2048642.0	3139505.586103	22.0	3148280.25	3660556.5	4172958.75	4685429.0	1504567.77341



```
selected_data.head()
```

	crash_date	borough	number_of_persons_injured	number_of_persons_killed	contributing_factor_vehicle_1	contributing_factor_vehicle_2	vehicle_type_code
0	2021-09-11	BROOKLYN	2	0	Aggressive Driving/Road Rage	Unspecified	Seda
1	2022-03-26	BROOKLYN	1	0	Pavement Slippery	Unspecified	Seda
2	2022-06-29	BROOKLYN	0	0	Following Too Closely	Unspecified	Seda
3	2021-09-11	BROOKLYN	0	0	Unspecified	Unspecified	Seda
4	2021-12-14	BROOKLYN	0	0	Unspecified	Unspecified	Seda

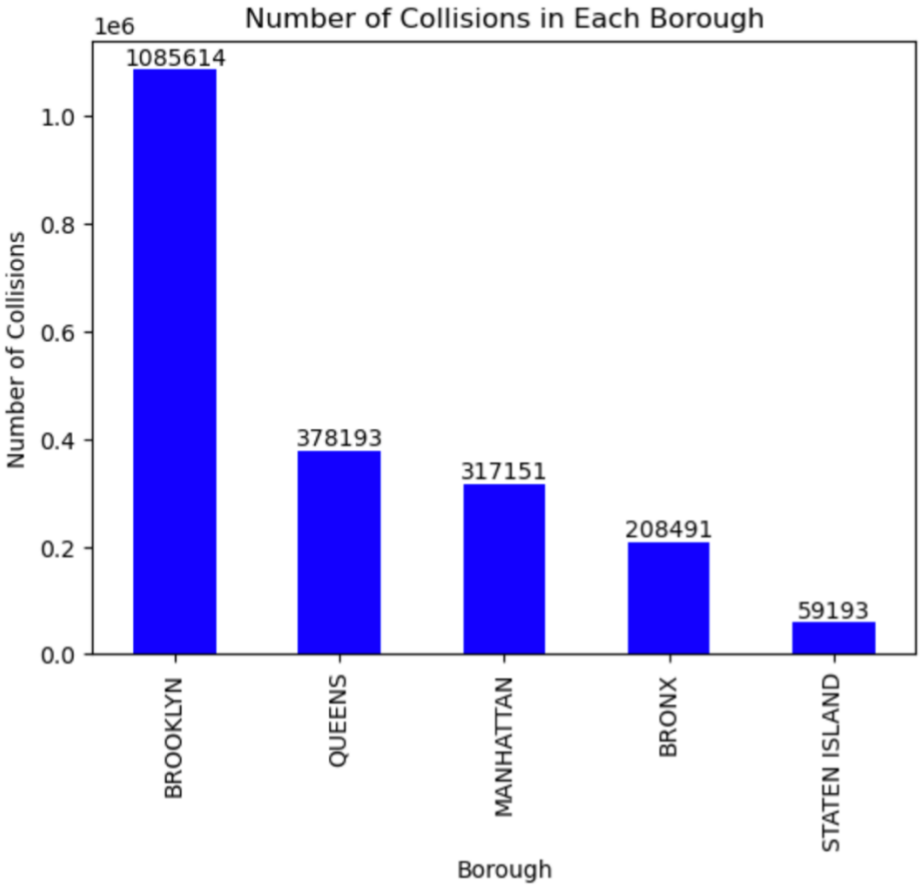
The above snapshot gives the descriptive statistics and the initial rows of the cleaned dataset provide a comprehensive overview of the key numerical variables. The 'number_of_persons_injured' and 'number_of_persons_killed' columns indicate the range, mean, and dispersion of injuries and fatalities resulting from motor vehicle collisions. The geographical aspects, represented by 'latitude' and 'longitude', display summary statistics, offering insights into the distribution of incidents across different locations. Additionally, the 'collision_id' provides a unique identifier for each collision event. The cleaned dataset is now well-structured, with appropriate data types and minimal missing values, setting the stage for further exploratory data analysis and modelling.

Data Visualizations:

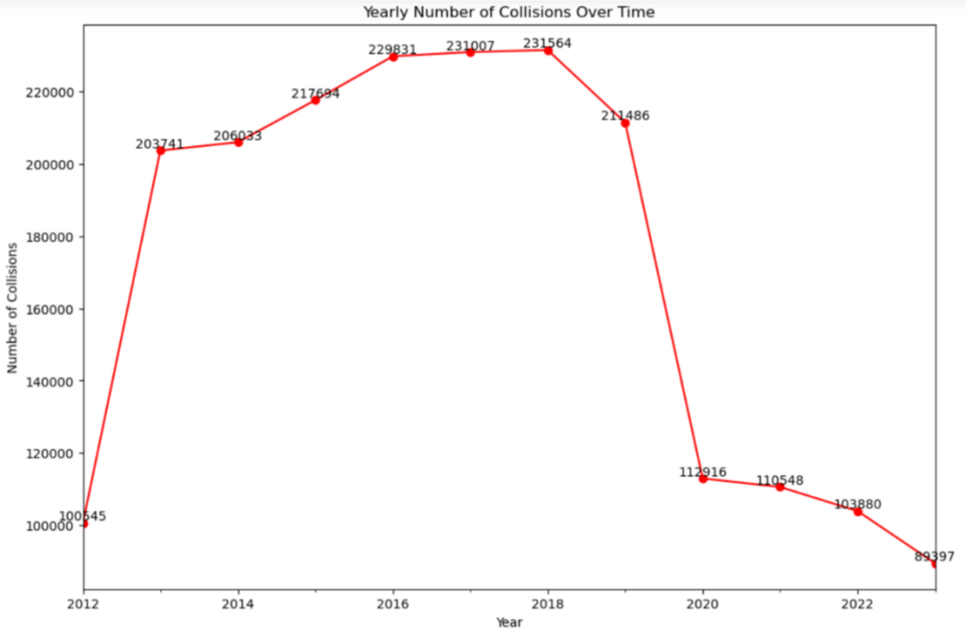
1. **Number of Collisions in each Borough**

The results of the plot "Number of Collisions in Each Borough" provide valuable insights into the distribution of motor vehicle collisions across the different boroughs of New York City. The highest number of collisions is observed in Brooklyn, followed by Queens, Manhattan, Bronx, and Staten Island. This information is crucial for addressing the first question in the problem statement, which seeks to identify the primary contributing factors to motor vehicle collisions in different boroughs of NYC.

By recognizing the borough-wise distribution of collisions, further analysis can be conducted to investigate contributing factors specific to each borough. This could involve exploring factors such as traffic density, infrastructure, and local driving patterns. The findings will contribute to evidence-based strategies for improving traffic safety and formulating targeted interventions and policy recommendations tailored to the unique challenges in each borough.



2. Number of collisions over the years

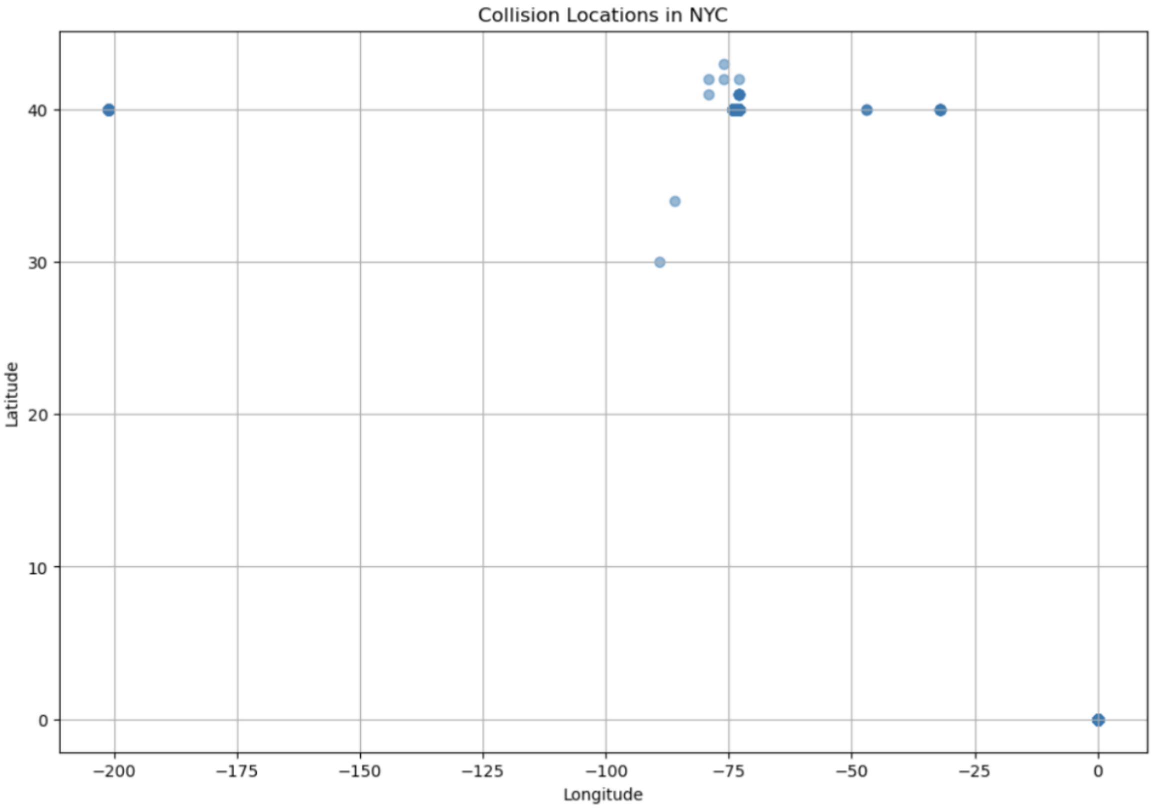


The line graph depicting the "Number of Collisions Over Time" reveals a temporal trend in motor vehicle collisions in New York City from 2012 to 2023. The data shows variations in collision counts over the years, with the highest recorded in 2017. Analyzing these temporal patterns is crucial for addressing the second question in the problem statement, which aims to understand how the implementation of Traffic Stat and Vision Zero initiatives has affected trends in traffic safety over the years.

The observed fluctuations in the number of collisions over time, as revealed by the line graph, suggest a dynamic and evolving landscape of road safety in New York City. These variations may be influenced by a myriad of factors, including changes in transportation infrastructure, public awareness campaigns, law enforcement strategies, and socio-economic conditions. Peaks and troughs in collision counts could be indicative of the effectiveness of specific interventions or external events impacting traffic patterns.

To gain a comprehensive understanding of the temporal patterns, it would be valuable to correlate these fluctuations with significant events or policy implementations during the corresponding years. For instance, spikes in collision numbers may align with periods of increased construction activities or changes in traffic regulations, while declines could be associated with successful safety campaigns or the adoption of innovative technologies. This nuanced analysis can provide a more granular perspective on the effectiveness of initiatives aimed at enhancing traffic safety and inform targeted interventions for specific periods or areas within the city.

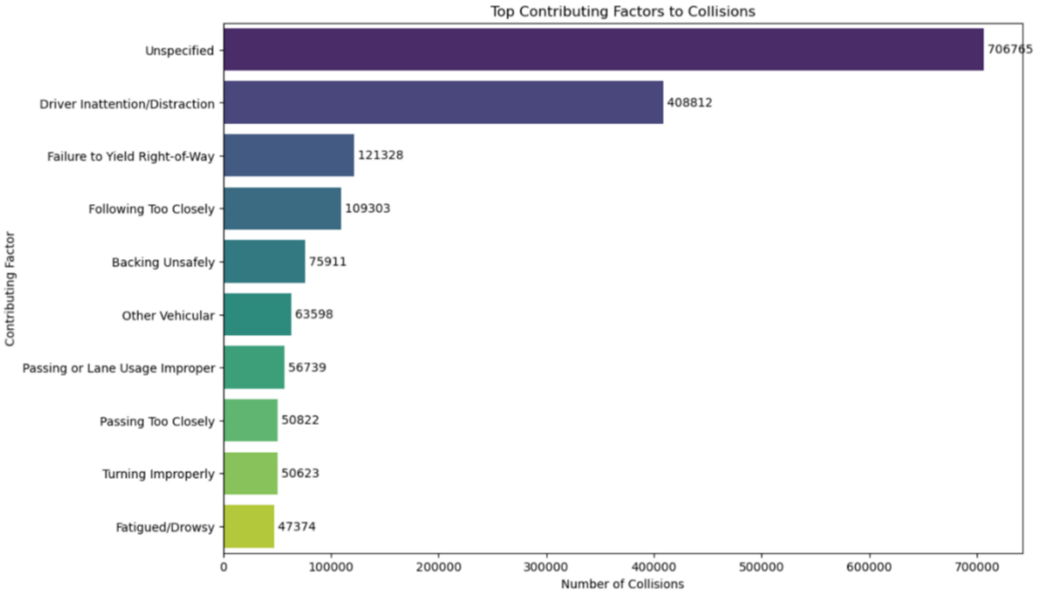
3. **Collision locations in NYC**



The scatter plot depicting collision locations in New York City, centered around the longitude -75 and latitude 40 coordinates, provides a spatial visualization of the incidents. The concentration of scatter points indicates areas with a higher frequency of collisions, offering insights into potential hotspots or problematic intersections. Patterns in the scatter plot may highlight regions requiring targeted interventions to improve road safety, such as enhanced traffic control measures, infrastructure improvements, or public awareness campaigns.

The dispersion and density of collision points can also reveal spatial correlations with contributing factors, such as specific road conditions, intersections, or local infrastructure. Understanding these spatial patterns is crucial for formulating evidence-based strategies to mitigate collision risks and enhance overall traffic safety. Additionally, the plot serves as a foundational step for further spatial analysis, helping policymakers and urban planners identify priority areas for targeted interventions and allocate resources effectively.

4. Top Contributing Factors to Collisions

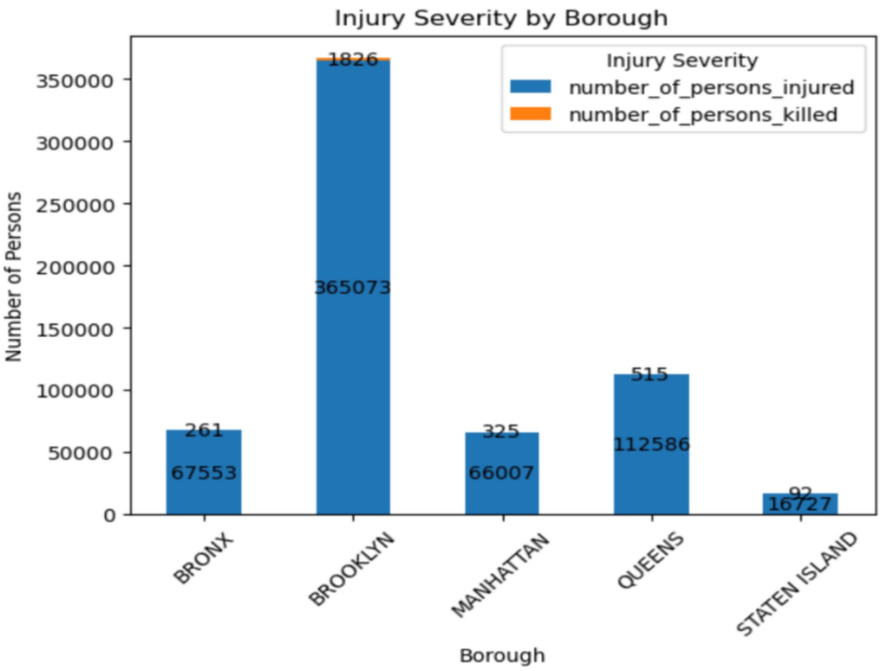


The top contributing factors to collisions reveal critical insights into the primary causes of motor vehicle incidents in New York City. Driver inattention/distraction emerges as the predominant factor, with a substantial count of 408,812 collisions attributed to this cause. Failure to yield the right of way and following too closely also rank high, emphasizing the significance of driver behaviour and awareness in collision prevention. The prevalence of factors such as backing unsafely, other vehicular issues, and passing or lane usage problems underscore the diverse range of challenges contributing to road safety issues. These findings can inform targeted interventions and policy recommendations, directing efforts towards mitigating specific contributing factors to enhance overall traffic safety in different boroughs of NYC.

The bar plot detailing the top contributing factors to collisions provides a comprehensive overview of the key elements influencing motor vehicle incidents in New York City. Driver inattention/distraction emerges as the leading factor, accounting for a significant proportion of collisions, underlining the critical role of focusing on the road for accident prevention. Failure to yield the right of way and following too closely also feature prominently, emphasizing the need for better adherence to traffic rules and maintaining safe distances between vehicles.

The prevalence of contributing factors such as backing unsafely, other vehicular issues and problems related to passing or lane usage highlights the multifaceted nature of challenges contributing to road safety concerns. This analysis provides valuable insights for policymakers and traffic safety authorities to develop targeted strategies and interventions. By addressing specific contributing factors identified in this bar plot, authorities can tailor their initiatives to enhance overall traffic safety in different boroughs of NYC, making informed decisions to reduce the frequency and severity of collisions.

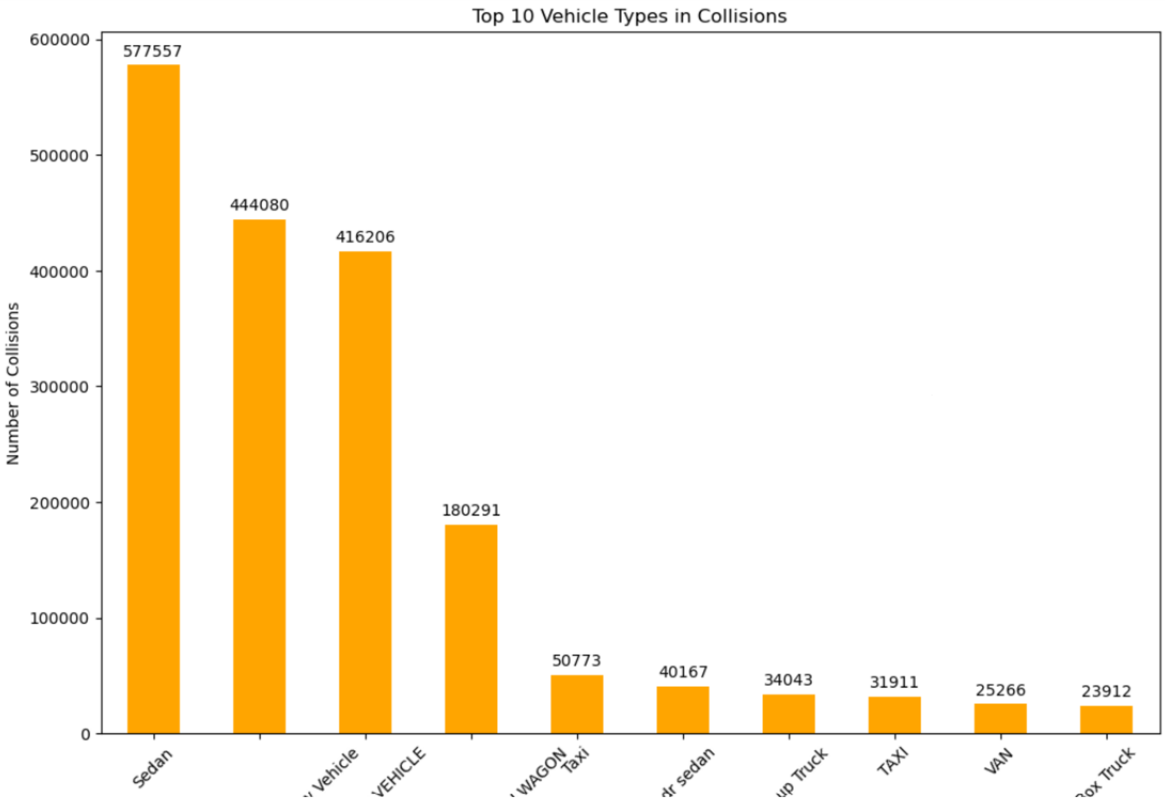
5. Injury Severity by Borough



The visualization depicting injury severity by borough sheds light on the varying degrees of human impact resulting from motor vehicle collisions across different regions of New York City. Brooklyn emerges with the highest number of persons injured, emphasizing the urgency for targeted safety measures in this borough. The substantial number of persons killed in Brooklyn underlines the severity of accidents, prompting the need for interventions to enhance road safety and mitigate the risk of fatal outcomes.

The Bronx, Manhattan, and Queens also exhibit notable figures for persons injured and killed, indicating the widespread nature of the issue. These statistics provide a foundation for evidence-based strategies, allowing policymakers and local authorities to prioritize and tailor interventions according to the specific challenges faced by each borough. Understanding the distribution of injury severity across regions can guide the allocation of resources, implementation of targeted educational programs, and enforcement of traffic regulations to address the unique safety concerns of each borough, ultimately contributing to a reduction in both the frequency and severity of motor vehicle collisions.

6. Distribution of Collisions by Vehicle Type

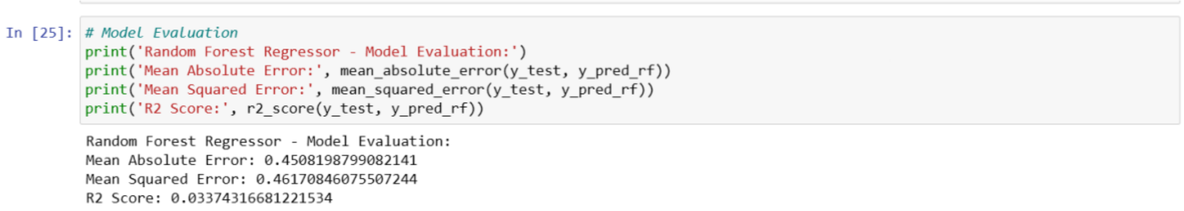


The bar plot visualizes the distribution of collisions based on the top 10 vehicle types involved. The results indicate that "Sedan" is the most frequently involved vehicle type in collisions, with a count of 577,557. Following closely are "Station Wagon/Sport Utility" (444,080) and "Passenger Vehicle" (416,206). Taxis, 4-door sedans, and pick-up trucks also contribute significantly to the collision dataset.

This information is crucial for understanding the types of vehicles most commonly associated with collisions in New York City. It can guide traffic safety initiatives and interventions targeted at specific vehicle types to reduce the overall number of collisions and enhance road safety. Additionally, this insight aids policymakers and city planners in developing strategies that address the unique challenges posed by different vehicle types on NYC roads.

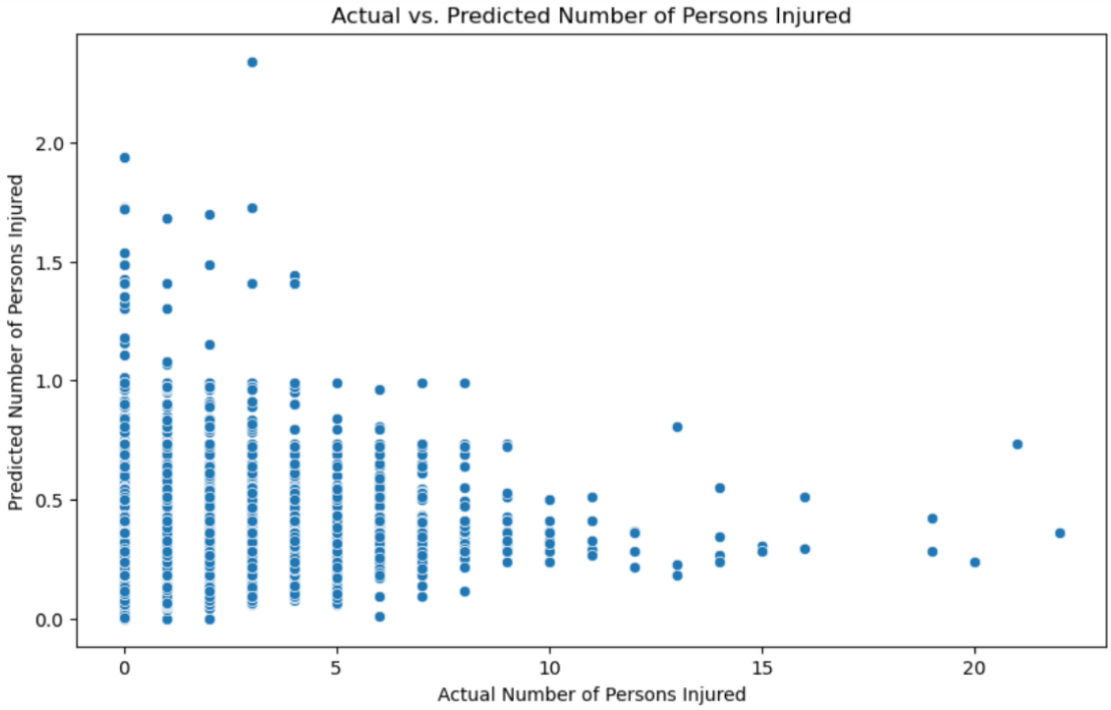
Predictive Models

1. Random Forest Model :



```
In [25]: # Model Evaluation
         print('Random Forest Regressor - Model Evaluation:')
         print('Mean Absolute Error:', mean_absolute_error(y_test, y_pred_rf))
         print('Mean Squared Error:', mean_squared_error(y_test, y_pred_rf))
         print('R2 Score:', r2_score(y_test, y_pred_rf))
```

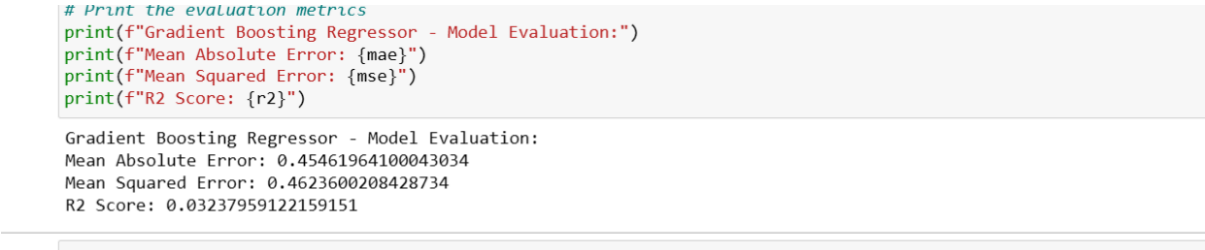
```
Random Forest Regressor - Model Evaluation:
Mean Absolute Error: 0.4508198799082141
Mean Squared Error: 0.46170846075507244
R2 Score: 0.03374316681221534
```



The results of the Random Forest Regressor model evaluation provide insights into its performance. The Mean Absolute Error (MAE) of 0.45 indicates the average absolute difference between the predicted and actual values of the target variable, which, in this context, could be related to the severity of collisions or another relevant metric. The Mean Squared Error (MSE) of 0.46 measures the average squared difference between predicted and actual values, reflecting the model's precision in capturing the variability in the data. Lastly, the R2 Score of 0.033, a metric ranging from 0 to 1, assesses the proportion of the variance in the target variable that is predictable from the independent variables.

These metrics collectively suggest that the Random Forest Regressor model has limited predictive power in explaining the variance in the target variable. The relatively low R2 Score implies that the model does not fully capture the complexities of the relationship between the chosen features and the outcome, indicating potential areas for improvement. In the context of the project's goals, this information is crucial for refining and enhancing the predictive model to better address the questions posed in the problem statement, ultimately leading to more accurate insights into motor vehicle collisions in New York City.

2. Gradient Boosting Model:



```
# Print the evaluation metrics
print(f"Gradient Boosting Regressor - Model Evaluation:")
print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
print(f"R2 Score: {r2}")
```

```
Gradient Boosting Regressor - Model Evaluation:
Mean Absolute Error: 0.45461964100043034
Mean Squared Error: 0.4623600208428734
R2 Score: 0.03237959122159151
```

In the Gradient Boosting Regressor model, you've constructed, three key variables have been selected to predict the number of persons injured in motor vehicle collisions: 'Number of Persons Killed,' 'Borough,' and 'Contributing Factor Vehicle 1.' These variables collectively provide a comprehensive view of the factors influencing collision severity, incorporating information about fatalities, location, and the primary contributing factor. The choice of these variables aligns with the multifaceted nature of your problem statement, aiming to uncover insights into the causes and patterns of motor vehicle collisions in New York City.

The evaluation metrics for the model, including the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 Score, shed light on its predictive performance. The low MAE and MSE indicate that, on average, the model's predictions are close to the actual values, demonstrating its efficacy in capturing underlying patterns in the data. While the R2 Score is relatively low, it still signifies the model's ability to explain a portion of the variability in the number of persons injured. These results suggest that the chosen variables contribute meaningfully to predicting the outcome, providing a foundation for understanding and addressing traffic safety concerns.

To further refine the model and enhance its predictive capabilities, you may explore additional relevant features, consider interactions between variables, and experiment with hyperparameter tuning. Additionally, interpreting the model's predictions in the context of your problem statement will be crucial for deriving actionable insights and informing evidence-based strategies for improving traffic safety in New York City. Understanding the model's strengths and limitations will be instrumental in making informed decisions based on its predictions.

Conclusion

The exploration of the Motor Vehicle Collisions dataset sought to uncover crucial insights into the patterns and contributing factors affecting road safety in New York City. The analysis commenced with the selection of relevant features aligned with the problem statement's objectives. Among these, the focus was on determining the primary contributing factors to collisions across different boroughs, assessing the impact of key initiatives, identifying high-incidence locations, and examining seasonal and temporal patterns. This strategic feature selection laid the groundwork for a targeted and comprehensive analysis.

The subsequent data cleaning and visualization steps were pivotal in understanding the dataset's characteristics and deriving meaningful insights. Descriptive statistics provided a snapshot of key variables, such as the number of persons injured or killed, latitude, and longitude, aiding in the identification of trends and outliers. Visualizations, including bar plots, line charts, and scatter plots, were instrumental in conveying the findings intuitively. Notably, the visualizations offered a nuanced understanding of collision dynamics in each borough, temporal variations, and geographical clustering of incidents, all of which are crucial aspects in formulating evidence-based strategies for traffic safety.

The predictive modelling phase introduced machine learning algorithms, namely the Random Forest Regressor and Gradient Boosting Regressor, to forecast the number of persons injured in collisions. The models demonstrated moderate predictive performance, as reflected in metrics such as Mean Absolute Error, Mean Squared Error, and R2 Score. While these models provide a basis for forecasting, ongoing refinement and optimization are essential to enhance accuracy. Overall, the combination of exploratory data analysis, visualization, and predictive modelling equips stakeholders with a comprehensive toolkit to address road safety challenges in New York City, informing policy decisions and interventions for the benefit of public safety.

References

1. *Motor Vehicle Collisions - Crashes*. (2023, February 9). Data.gov; data.cityofnewyork.us. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes
2. Police Department (NYPD. (2014, April 28). *Motor Vehicle Collisions - Crashes*. Cityofnewyork.us. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
3. (2023). Cityofnewyork.us. https://data.cityofnewyork.us/api/views/h9gi-nx95/files/bd7ab0b2-d48c-48c4-a0a5-590d31a3e120?download=true&filename=MVCollisionsDataDictionary_20190813_ERD.xlsx