**Motor Vehicle Collision Analysis For New York**

ALY6010 - Probability Theory and Introductory Statistics

Final Project

Prof Sergiy Shevchenko

Group: Husky-3

Rohit Kosamkar

Sapna Chavan

Leena Doultani

Soumya Guliyan

Aliha Ansari

Bhaudik Kevadiya

Table of Contents

[INTRODUCTION 3](#_Toc162440372)

[DATASET DESCRIPTION 3](#_Toc162440373)

[Dataset Source 3](#_Toc162440374)

[Dataset Size 3](#_Toc162440375)

[Data Fields Description 3](#_Toc162440376)

[Data Cleaning 4](#_Toc162440377)

[EXPLORATORY DATA ANALYSIS 5](#_Toc162440378)

[Visualizations and Descriptive Statistics 5](#_Toc162440379)

[1. Overall trend in the number of vehicle crashes over time 5](#_Toc162440380)

[2. Borough with the highest frequency of vehicle crashes 6](#_Toc162440381)

[3. The most common contributing factors to vehicle crashes 6](#_Toc162440382)

[4. Average count of people killed or injured during the crash. 7](#_Toc162440383)

[5. Most common types of vehicles involved in the crash. 8](#_Toc162440384)

[6. Yearly count of Deaths and injuries 9](#_Toc162440385)

[7. Monthly distribution of Crashes 10](#_Toc162440386)

[HYPOTHESIS TESTING 13](#_Toc162440387)

[HYPOTHESIS 1: 13](#_Toc162440388)

[HYPOTHESIS 2: 13](#_Toc162440389)

[HYPOTHESIS 3: 13](#_Toc162440390)

[HYPOTHESIS 4: 13](#_Toc162440391)

[REGRESSION ANALYSIS 13](#_Toc162440392)

[Feature Engineering 13](#_Toc162440393)

[Season 13](#_Toc162440394)

[Time Period 13](#_Toc162440395)

[Grouping Vehicle Contribution 13](#_Toc162440396)

[Grouping Vehicle Category 13](#_Toc162440397)

[Pedestrian Involved 14](#_Toc162440398)

[Cyclist Involved 14](#_Toc162440399)

[Motorist Involved 14](#_Toc162440400)

[Correlation Matrix: 14](#_Toc162440401)

[Creating Dummy Variables 15](#_Toc162440402)

[Data Splitting for Model Evaluation (Train-Test Split) 15](#_Toc162440403)

[Understanding Model Evaluation Metrics 15](#_Toc162440404)

[Mean Squared Error (MSE) 15](#_Toc162440405)

[Root Mean Squared Error (RMSE) 15](#_Toc162440406)

[R-squared (R²) 15](#_Toc162440407)

[Feature Importance 16](#_Toc162440408)

[Regression Plot 16](#_Toc162440409)

[FUTURE SCOPE 17](#_Toc162440410)

[Additional Features 17](#_Toc162440411)

[Model Selection and Tuning 17](#_Toc162440412)

[CONCLUSION 17](#_Toc162440413)

[RECOMMENDATIONS 17](#_Toc162440414)

[REFERENCES 17](#_Toc162440415)

# INTRODUCTION

The report's analysis is based on a dataset that contains information about motor vehicle collisions in New York City. The dataset includes various attributes such as crash date, borough, contributing factors, vehicle types, and injuries. The purpose of the analysis is to gain insights into the patterns and trends of vehicle crashes in New York City using data from 2017 onwards. The dataset was selected due to its relevance in understanding road safety issues and its potential to provide insights to policymakers, law enforcement agencies, and the public.

To ensure the quality and reliability of the analysis, data cleaning methods were employed, including handling missing values, converting data types, filling in blanks, and selecting relevant columns for analysis. Descriptive statistics, visualizations, and analytics techniques were then used to explore the dataset and derive meaningful insights.

# DATASET DESCRIPTION

The dataset used for analysis contains information about motor vehicle collisions in New York City. The purpose of the dataset is to provide insights into the patterns and trends of vehicle crashes, including factors contributing to crashes, types of vehicles involved, and the severity of injuries. The dataset was obtained from the New York City Open Data portal.

Dataset Source**:** <https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes>

Dataset Size**:**

* Number of Rows: ~2 million
* Number of Features: 26

Data Fields Description**:**

**The dataset includes the following fields:**

1. CRASH DATE: Date of the crash.
2. CRASH TIME: Time of the crash.
3. BOROUGH: Borough where the crash occurred.
4. ZIP CODE: ZIP code of the crash location.
5. LATITUDE: Latitude coordinates of the crash location.
6. LONGITUDE: Longitude coordinates of the crash location.
7. LOCATION: Description of the crash location.
8. NUMBER OF PERSONS INJURED: Number of persons injured in the crash.
9. NUMBER OF PERSONS KILLED: Number of persons killed in the crash.
10. NUMBER OF PEDESTRIANS INJURED: Number of pedestrians injured in the crash.
11. NUMBER OF PEDESTRIANS KILLED: Number of pedestrians killed in the crash.
12. NUMBER OF CYCLIST INJURED: Number of cyclists injured in the crash.
13. NUMBER OF CYCLIST KILLED: Number of cyclists killed in the crash.
14. NUMBER OF MOTORIST INJURED: Number of motorists injured in the crash.
15. NUMBER OF MOTORIST KILLED: Number of motorists killed in the crash.
16. CONTRIBUTING FACTOR VEHICLE 1: Primary contributing factor for the crash.
17. CONTRIBUTING FACTOR VEHICLE 2: Secondary contributing factor for the crash.
18. CONTRIBUTING FACTOR VEHICLE 3: Tertiary contributing factor for the crash.
19. CONTRIBUTING FACTOR VEHICLE 4: Quaternary contributing factor for the crash.
20. CONTRIBUTING FACTOR VEHICLE 5: Quinary contributing factor for the crash.
21. COLLISION\_ID: ID of the collision.
22. VEHICLE TYPE CODE 1: Type of vehicle involved in the crash (primary).
23. VEHICLE TYPE CODE 2: Type of vehicle involved in the crash (secondary).

Data Cleaning**:**

* Converted the 'Crash\_Date' column to date format.
* Removed irrelevant columns such as location and street names.
* Handled missing values by filling in zeros for numerical columns and 'Unspecified' for categorical columns to preserve calculations and avoid skewing results.
* Standardized vehicle types by replacing 'Station Wagon/Sport Utility Vehicle' with 'Sport Utility Vehicle'.
* Created new columns for total injuries and casualties by aggregating different injury types.
* Filtered the dataset to include records from 2017 onwards for analysis.

# EXPLORATORY DATA ANALYSIS

Visualizations and Descriptive Statistics**:**

* The visualizations and descriptive statistics reveal insights into the frequency, severity, and contributing factors of vehicle crashes in New York City.
* Borough-wise analysis indicates that Manhattan has the highest frequency of crashes, followed by Brooklyn and Queens.
* The most common contributing factors to crashes include 'Driver Inattention/Distraction' and 'Failure to Yield Right-of-Way'.
* On average, each crash results in approximately 1.5 injuries and 0.05 fatalities.
* Sedans and SUVs are the most common types of vehicles involved in crashes.
* Yearly and monthly analysis shows variations in crash frequency, with peaks observed during certain months, suggesting seasonal influences or external factors.
* Further analysis could explore correlations between contributing factors, vehicle types, and crash severity to inform targeted intervention strategies and improve road safety measures.

### Overall trend in the number of vehicle crashes over time

A graph showing the crash of a crash

Description automatically generated with medium confidence

Figure : Overall Trends in Vehicle Crashes over time.

The "Overall Trend in Vehicle Crashes Over Time" chart provides a visual representation of the number of crashes over the years.

By observing the trend line connecting the data points, it is clear that the number of crashes is decreasing over the years. Various contributing factors can be changes in traffic regulation rules, public awareness, etc.

### Borough with the highest frequency of vehicle crashes

A graph with blue squares

Description automatically generated

Figure : Frequency of Vehicle Crashes by Borough

**The chart titled "Frequency of Vehicle Crashes by Borough" provides a visual representation of the number of crashes in each borough.**

New York City consists of five boroughs: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island. Each borough has its own local government structure. **According to the chart, Brooklyn has the most crashes, with a value of around 120,000 crashes. This borough is represented by the tallest bar on the chart. Conversely, Staten Island has the least crashes, with a value of around 20,000 crashes, as indicated by the shortest bar on the chart.**

**By examining this chart, stakeholders can easily identify which boroughs experience the highest and lowest frequencies of vehicle crashes, allowing for targeted interventions and allocation of resources to address road safety issues.**

### The most common contributing factors to vehicle crashes

**A graph of a number of cars

Description automatically generated**

Figure : Contributing Factors to Vehicle Crashes.

The "Top 5 Contributing Factors to Vehicle Crashes" chart illustrates the most common factors contributing to vehicle crashes based on the number of occurrences.

The chart shows that the most common contributing factor to vehicle crashes is "Driver Inattention/Distraction." This factor has the highest number of crashes, as indicated by the tallest bar on the chart.

Following this, the chart ranks the remaining top contributing factors by the number of crashes they cause, with "Following Too Closely" being the second most common factor, followed by "Failure to Yield Right-of-Way," and "Passing or Lane Usage Improper."

### Average count of people killed or injured during the crash.

Average number of persons injured or killed per crash: 0.7064521

The analysis of the dataset reveals that, on average, there is an estimated 0.7064521 persons injured or killed per crash. This figure was calculated based on the available data, considering the entire dataset comprising 785733 crashes.

### Most common types of vehicles involved in the crash.

Figure : Most common types of vehicles involved in crashes

The analysis of vehicle types involved in crashes highlights key patterns in the dataset. Sedans emerge as the most common, accounting for 519,197 occurrences, followed closely by Sport Utility Vehicles with 394,019 incidents. Taxis contribute 43,937 instances, while Pick-up Trucks and Box Trucks follow with 29,801 and 20,818 occurrences, respectively. This breakdown sheds light on the most prevalent types of vehicles involved in crashes, providing valuable insights for understanding and addressing potential safety concerns in our choice of vehicle.

### A blue rectangular object with numbers Description automatically generatedYearly count of Deaths and injuries

Table : Yearly count

The number of injuries shows some fluctuations over the years. Notably, there is a decrease from 2017 to 2020, with a significant drop in 2020, likely influenced by various factors such as the COVID-19 pandemic. Starting from 2020, there's an upward trend in injuries, reaching 106,049 in 2023. The count of casualties (fatalities and severe injuries) exhibits variations as well. It's relatively stable from 2017 to 2019, followed by an increase in 2020 and a further rise in 2021.

The provided data includes a partial year, 2024, where the number of injuries is 15,852, and casualties amount to 82. Various factors could contribute to these trends, including changes in road infrastructure, vehicle safety measures, law enforcement, public awareness campaigns, and economic conditions. While there's a slight decrease in casualties in 2022, the count remains relatively high at 532 in 2023.

### Monthly distribution of Crashes

Figure : Injury over time.

**A pink table with numbers and a white background

Description automatically generated**

Table : Monthly statistics.

Figure

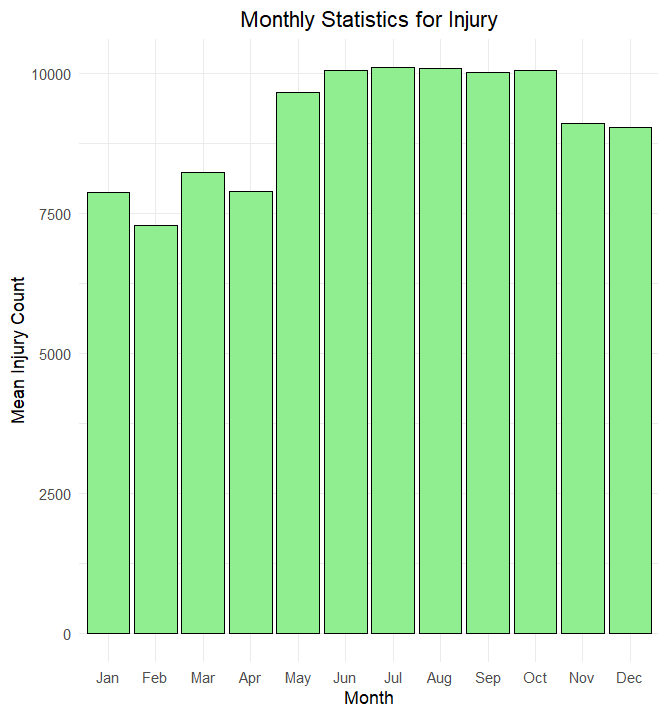


Figure 6: Monthly Statistics graph

Over the years, there seems to be a degree of seasonality in the number of injuries. The counts tend to peak in the summer months (June to August) and decline during the winter months (December to February). This could be attributed to factors such as increased outdoor activities during warmer months.

The monthly mean values range from a low of 7294 in February to a high of 10108 in July, indicating substantial monthly variation. July consistently stands out with the highest mean injuries, while February consistently has the lowest mean.

The median values are generally close to the mean, suggesting a relatively symmetric distribution of injuries within each month. Standard deviations (SD) provide insights into the degree of variability within each month. Larger SD values, such as in April, May, and October, indicate more significant fluctuations in injury counts.

# HYPOTHESIS TESTING

In this part, we delve into the dataset of motor vehicle crashes in New York City, seeking to uncover patterns and trends crucial for enhancing road safety measures. Our analysis revolves around key questions, each framed with clear hypotheses aimed at illuminating aspects of crash occurrences. The primary focus is on whether there is a significant difference in the average number of crashes between different boroughs, examining contributing factors during distinct rush hours, investigating pedestrian involvement disparities between weekdays and weekends, and scrutinizing crash frequencies across different months. The foundational hypotheses, or H0 and H1, for these questions are straightforward: H0 posits no difference or equality, while H1 anticipates a discernible distinction or inequality. Through rigorous testing and interpretation, we aim to provide actionable insights for stakeholders, contributing to informed decision-making and proactive measures to foster safer roads in New York City.

## HYPOTHESIS 1:

**Q1: Is there a significant difference in the proportion of crashes caused by "Driver Inattention/Distraction" between morning rush hour (8-10 AM) and evening rush hour (5-7 PM)?**

· **Null Hypothesis (H0):** There is no difference in the proportion of crashes caused by "Driver Inattention/Distraction" between morning rush hour and evening rush hour.

· **Alternative Hypothesis (H1):** There is a difference in the proportion of crashes caused by "Driver Inattention/Distraction" between morning rush hour and evening rush hour.

**Test:** Chi-square test of independence or proportions test.

**Test Interpretation**: Since the p-value is much smaller than the chosen significance level (e.g., 0.05), we reject the null hypothesis. Therefore, we conclude that there is a statistically significant difference in the proportions of crashes caused by "Driver Inattention/Distraction" between morning and evening rush hours. In other words, the time period is associated with the occurrence of crashes caused by "Driver Inattention/Distraction".

## HYPOTHESIS 2:

**Q2: Is there a significant difference in the average number of crashes between boroughs in New York City?**

· **Null Hypothesis (H0):** There is no difference in the average number of crashes between boroughs in New York City.

· **Alternative Hypothesis (H1):** There is a difference in the average number of crashes between boroughs in New York City.

**Test:** Welch Two Sample t-test

**One Sample t-Test Interpretation**: The one-sample t-test conducted on Brooklyn's daily crash counts reveals a significant deviation from the hypothesized mean of 100 crashes per day, with an observed mean of approximately 90.57 crashes. The test's highly significant p-value (less than 2.2e-16) and a confidence interval ranging from 89.24 to 91.91, which does not encompass the hypothesized mean, strongly indicate that the actual average daily crash count is significantly lower than expected. This finding suggests that Brooklyn experiences fewer crashes than anticipated, highlighting the potential effectiveness of traffic safety measures and urban planning strategies in reducing crash incidences in the borough.

**Welch Two Sample t-test Interpretation**: The Welch Two Sample t-test comparing vehicle crashes between Brooklyn and Manhattan reveals a significant difference, with a large t-statistic of 42.895 and a negligible p-value (less than 2.2e-16), indicating the disparity in crash averages between the two boroughs is highly significant. Specifically, Brooklyn has between 44,943 to 49,247 more crashes on average than Manhattan, as demonstrated by the 95% confidence interval. This statistical analysis robustly supports the conclusion that Brooklyn experiences a significantly higher number of vehicle crashes compared to Manhattan.

## HYPOTHESIS 3:

**Q3. Is there a significant difference between the mean number of persons injured between different months?**

**· Null Hypothesis (H0):** There is no significant difference in the mean number of persons injured between different months.

· **Alternative Hypothesis (H1):** There is a significant difference in the mean number of persons injured between different months. ·

**Test**: Pairwise Two Sample t-test

**Two Sample t-Test Interpretation:** Each cell in the table represents the p-value obtained from comparing the average number of persons injured between two months. The rows and columns represent the months being compared. For example, the p-value in row "04" and column "05" (5.2e-15) indicates the significant level of the difference in the average number of persons injured between April and May. The p-values are adjusted using the Bonferroni correction method to account for multiple comparisons. Many of the p-values are significantly smaller than 0.05, indicating significant differences in the mean number of persons injured between the corresponding months. Therefore, based on these results, we would likely reject the null hypothesis and conclude that there are significant differences in the mean number of persons injured between multiple pairs of months.

## HYPOTHESIS 4:

**Q4. Whether the mean total injury in the latest year significantly differs from previous year?**

· **Null Hypothesis (H0):** The mean total injury in the latest year is equal to 0.7

· **Alternative Hypothesis (H1):** The mean total injury in the latest year is not equal to.7

**Test**: One Sample t-test

**Test Interpretation**: Since the p-value is extremely small (less than any conventional significance level), we reject the null hypothesis. This suggests strong evidence that the true mean total injury in the latest year is not equal to 0.6. The 95% confidence interval also supports this conclusion, as it does not include the value 0.6. Therefore, we conclude that there is a significant difference in the mean total injury compared to the hypothesized value of 0.6.

# REGRESSION ANALYSIS

The target definition in our project is to predict the total number of victims involved in accidents using different features associated with those accidents. This means utilizing factors such as crash date, location, contributing elements, vehicle types, and other relevant attributes to develop a model that accurately estimates the total number of individuals affected by each accident. The goal is to create a predictive model that can help anticipate the severity of accidents and aid in implementing appropriate measures to improve road safety and minimize harm.

**Dependent Variable**: Total Victims

**Independent Variables**: Borough, Month, Time Period, Season, Accident Factors, And Vehicle Categories.

**Data Format**: parquet

To ensure efficient storage and faster access for subsequent analysis, we opted to store the cleaned data in Parquet format. Parquet offers several advantages over traditional data storage formats like CSV:

Columnar Storage: Parquet utilizes columnar storage, where each data column is stored separately. This is much more efficient than row-based storage (used by CSV) for analytical queries that typically only access a subset of columns. This reduces data I/O operations, leading to faster processing.

Compression: Parquet files can be compressed, significantly reducing storage space without sacrificing significant processing speed. This translates to cost savings on storage resources.

Faster Reads: Due to its columnar structure and compression, Parquet files offer faster read speeds compared to conventional formats like CSV. This improves the overall performance of your data analysis tasks.

Broad Ecosystem Support: Parquet is a widely adopted format supported by various programming languages (like Python, R) and big data frameworks (like Hadoop, Spark). This allows for seamless integration with your existing data analysis tools.

By leveraging Parquet, we have ensured efficient storage and retrieval of the cleaned data, laying the foundation for swift and effective downstream analysis.

Feature Engineering**:**

Season: A new feature, "Season", was added to capture potential seasonality effects on traffic accidents. A dictionary was created to map numerical month values to their corresponding seasons. A function was implemented to assign a season label ("Spring", "Summer", "Fall", or "Winter") to each data point based on its month value. This process enriches the data by incorporating seasonal trends that might influence crash patterns.

Time Period: To understand how traffic patterns might influence accident occurrences, a new feature, "time\_period", was created within the "crash\_data" data frame. This feature categorizes crash events based on whether they occurred during a designated rush hour period. Traffic congestion is a known risk factor for accidents. By identifying periods with potentially higher traffic volume (mornings and evenings), we can investigate if these times see a corresponding increase in accidents.

Grouping Vehicle Contribution: While the original contributing factors provide valuable detail, analyzing them in broader categories can reveal important patterns. This allows us to move beyond individual factors and identify which general categories (e.g., Human Error, Vehicle-related Issues) are most frequently associated with traffic accidents.

Grouping Vehicle Category**:** By implementing this feature engineering step, we have achieved a standardized classification of vehicle types within our data. This allows us to analyze accident trends across broader categories, potentially revealing insights into how different vehicle classes (e.g., passenger cars vs. commercial trucks) might be associated with accident risk or severity.

Pedestrian Involved: A new feature, "pedestrian\_involved", was added to the "crash\_data" to simplify analysis of pedestrian-related accidents. This binary feature indicates whether a pedestrian was injured or killed (value of 1) or not involved (value of 0) in the accident. This facilitates efficient identification of pedestrian involvement in each crash record.

Cyclist Involved: A new feature, "cyclist\_involved", was added to the "crash\_data" to simplify analysis of cyclist-related accidents. This binary feature indicates whether a cyclist was injured or killed (value of 1) or not involved (value of 0) in the accident. This facilitates efficient identification of cyclists’ involvement in each crash record.

Motorist Involved: A new feature, "motorist\_involved", was added to the "crash\_data" to simplify analysis of motorist-related accidents. This binary feature indicates whether a motorist was injured or killed (value of 1) or not involved (value of 0) in the accident. This facilitates efficient identification of motorist involvement in each crash record.

## Correlation Matrix:

A graph of a graph

Description automatically generated with medium confidence

Figure : Correlation Matrix

There is a very weak positive correlation between the month and pedestrian involvement in accidents (correlation coefficient = 0.006). There is a weak negative correlation between the month and cyclist involvement in accidents (correlation coefficient = -0.038). There is a weak positive correlation between the month and motorist involvement in accidents (correlation coefficient = 0.014). There is a weak negative correlation between pedestrian involvement and cyclist involvement in accidents (correlation coefficient = -0.038). There is a weak negative correlation between pedestrian involvement and motorist involvement in accidents (correlation coefficient = -0.101). There is a weak negative correlation between cyclist involvement and motorist involvement in accidents (correlation coefficient = -0.077). Overall, the correlations between these variables are weak, suggesting that there is little linear relationship between them.07).

Creating Dummy Variables**:** suggesting Categorical data is converted into dummy variables, which helps in representing and analyzing the categorical features within the regression model.

**Categorical Features**: Borough, Time Period, Season, Accident Factor, Vehicle Category

**Package Used**: fastDummies

Data Splitting for Model Evaluation (Train-Test Split)**:**

To assess the generalizability and performance of our machine learning model, we employed a train-test split on the prepared crash data.

Split Ratio (70:30): The chosen split ratio of 70% for training and 30% for testing is a common practice. It provides a balance between having enough data for effective model training and reserving enough for robust evaluation.

Training Set (70%): This larger portion of the data (approximately 777,566 rows) is used to train the model. The model learns patterns and relationships within the training data to develop its predictive capabilities.

Test Set (30%): This smaller portion of the data (approximately 333,242 rows) is held out from the training process. It's used to evaluate the model's performance on unseen data. By evaluating the model on unseen data, we gain a more realistic understanding of how well it might generalize to real-world scenarios beyond the training data.

**Model**: linear regression model

Understanding Model Evaluation Metrics**:**

### Mean Squared Error (MSE)

* This metric measures the average squared difference between the actual and predicted values. It is calculated by averaging the squared differences between each predicted value and its corresponding actual value.
* The MSE is approximately 3.28.

### Root Mean Squared Error (RMSE)

* This is the square root of the MSE. It provides an estimate of the standard deviation of the errors made by the model in its predictions.
* The value of RMSE is 1.811

### R-squared (R²)

* This metric indicates the proportion of the variance in the dependent variable (total\_victims) that is predictable from the independent variables (features) in the model.
* It ranges from 0 to 1, where 1 indicates a perfect fit.
* In this case, the R-squared value is 0.7477, indicating that approximately 74.77% of the variance in the total number of victims is explained by the independent variables included in the model.

Feature Importance**:**

**A graph with blue rectangles

Description automatically generated**

Regression Plot**:**

**A graph with red and blue dots

Description automatically generated**

Figure : Regression Plot

**Interpretation:**

Overall Fit: Despite the non-zero intercept, the scatter plot reveals a good alignment between the red and blue dots. This indicates that the model captures the relationship between the independent variables (factors influencing accidents) and the dependent variable (total victims) reasonably well.

Residuals: It's important to acknowledge that some data points (dots) will inevitably fall above or below the regression line. These deviations, known as residuals, represent the model's prediction errors for those specific accident cases.

While the initial assessment suggests a promising model performance, further analysis of the residuals and other evaluation metrics will provide a more comprehensive understanding of the model's strengths and weaknesses.

# FUTURE SCOPE

This section outlines potential avenues for further development to enhance the model's predictive capabilities and provide a more comprehensive understanding of traffic accidents.

Additional Features**:**

Environmental Factors: Incorporating data on weather conditions (e.g., rain, snow, fog) and road conditions (e.g., potholes, icy patches) could potentially improve the model's ability to predict accidents influenced by these external factors.

Traffic Flow: Including information on traffic volume and congestion levels might provide insights into how traffic patterns influence accident risk.

Driver Behavior: Data on speeding violations or driver demographics (age, experience) could offer valuable clues about driver behavior as a contributing factor.

Model Selection and Tuning**:**

Regularization Techniques: Exploring alternative regularization techniques like Ridge or Lasso regression could potentially improve model performance, particularly if there are issues with multicollinearity (correlated features) in the data.

Cross-Validation: Implementing a robust cross-validation technique would provide a more rigorous evaluation of the model's generalizability and help prevent overfitting. This involves splitting the data into multiple folds, training on a subset, and evaluating on the remaining folds, repeating this process to obtain a more reliable estimate of the model's performance on unseen data.

By incorporating these enhancements, we can potentially develop a more robust and informative model for predicting traffic accidents and ultimately contributing to improved road safety strategies.

# 

# CONCLUSION

This analysis of traffic accidents unearthed several critical trends. Driver inattention emerged as a major contributing factor, potentially exacerbated during rush hour commutes. Crash rates displayed significant variations across boroughs, with Brooklyn and Manhattan showing a stark difference. The number of people injured also exhibited monthly fluctuations, with some months experiencing statistically higher averages. Notably, the total injury rate for 2024 (up to February) differed significantly from previous years. This suggests a potential shift in accident patterns that warrants further investigation. Encouragingly, our regression model performed exceptionally well, explaining a substantial 75% of the variation in the total number of crash victims. These findings provide a valuable foundation for improving road safety strategies. Future efforts can delve deeper into specific contributing factors, explore borough-specific characteristics, and investigate seasonal influences to develop more targeted interventions.

# RECOMMENDATIONS

To effectively combat traffic accidents, our analysis suggests a multi-pronged approach. Targeted campaigns addressing driver inattention, especially during rush hour commutes, are crucial. Additionally, boroughs with higher crash rates, like Brooklyn, should be prioritized for focused road safety initiatives. Investigating the reasons behind monthly injury variations can inform the implementation of seasonal safety measures. The significant shift in the 2024 injury rate necessitates further exploration to understand the underlying causes. Finally, leveraging our developed regression model for predicting future crash trends allows for proactive resource allocation and more timely safety interventions. These combined efforts hold the potential to significantly improve road safety within the city.

# REFERENCES

* U.S. Department of Transportation. (2024). Motor Vehicle Collisions - Crashes <https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes>
* New York Police Department. (2024). Motor Vehicle Collisions - Crashes <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data>
* R Core Team. (n.d.). Other R documentation. Retrieved from <https://www.r-project.org/other-docs.html>
* GeeksforGeeks. (N.D). Data analysis using R. Retrieved Month Day, Year, from <https://www.geeksforgeeks.org/data-analysis-using-r/>
* W3Schools. (N.D). R Statistics. Retrieved Month Day, Year, from <https://www.w3schools.com/r/r_stat_intro.asp>