**MASTER OF SCIENCE IN**

**COMPUTER SCIENCE AND ENGINEERING**



**Predicting Injury Severity in New York City Car Accidents: An Analysis of Factors Impacting Crash Severity Using Machine Learning.**

Author

**Md Mahamudul Hasan Khan**

CISC 602

Supervisor

**Narasimhaswamy Banavara, Ph.D.**

**Program Director - Graduate Computer Science**

**Department of Mathematics and Computer Sciences**

**School of Liberal Arts**

**A capstone project report submitted to the Department of CSE**

**in fulfillment of the requirements for the degree of M.Sc. in CSE**

**Declaration**

I solemnly declare that this project represents my own original work, derived from the calculations and analyses I have performed. I have duly acknowledged the contributions of other researchers and sources by providing appropriate citations and references. Furthermore, this report, or any portion thereof, has not been submitted elsewhere for the purpose of obtaining a Degree or Diploma.

*Author:*



Md Mahamudul Hasan Khan

Student ID: 20297374

*Supervisor:*



Narasimhaswamy A Banavara Ph.D.

Program Director - Graduate Computer Science

Department of Mathematics and Computer Sciences

School of Liberal Arts

07 May 2023

**Acknowledgement**

I am deeply grateful to the Almighty for providing me with the strength and guidance necessary to complete this project. I would also like to extend my heartfelt appreciation to my supervisor, Dr. Banavara, whose invaluable advice and unwavering support have been instrumental in the project's success. His generous investment of time has truly made a difference.

In addition, I would like to express my gratitude to my parents and wife for their unwavering support and belief in my abilities.

Finally, I wish to acknowledge everyone who has contributed to my personal and professional growth throughout my life. Thank you all.

**Abstract**

In recent years, a growing trend of personal vehicle ownership has emerged, leading to a predicted increase in road accident rates. This presents significant challenges to governments, individuals, and communities, as car accidents are often life-threatening and pose a hazard to society. This paper investigates the main factors contributing to the rise in car accident rates using a dataset collected from traffic accidents in New York City between 2012 and 2023. This paper aims to dig deep to explore the main factors contributing to the increase in car accidents rate. This study employed four models to predict car accident severity, focusing on the primary factors contributing to road accidents. Results indicated that the main factors affecting car accident rates are work rush hour traffic and population density. The research findings can be used to develop solutions to reduce car accidents in New York City, such as promoting remote work, encouraging self-driving vehicle ownership, improving public transportation infrastructure, and distributing rush hours throughout the day.

**Table of Contents**

[**Chapter 1 5**](#_gyugsd8ebt5j)

[1.1 Background Information 5](#_olwj0gsncev3)

[1.2 Statement of the problem 6](#_o1jgohabqdgr)

[1.3 Project goals 6](#_cmips6aupb8q)

[1.4 Methodology 7](#_j43xg257l0k)

[1.5 Study Limitations 8](#_27ykbya8nzao)

[**Chapter 2 - Literature Review 8**](#_1nwkd9gh5uzw)

[**Chapter 3 - Project Description 11**](#_8slotwnmuouj)

[3.1 Data Collection 11](#_3asoot1ox988)

[The dataset contains 29 different attributes; table 1 includes the name of the attributes, descriptions, and types. 11](#_6v9drcpj945t)

[**Chapter 4 Project Analysis 14**](#_3ynbrrxl16r)

[4.1 Exploratory Data Analysis 14](#_ieu3hqlh3yy0)

[4.2 Data Preprocessing 15](#_zck7hit1gceb)

[4.2.1 Data Cleaning 15](#_7se6j3lj8olj)

[4.2.2 Data Selection 15](#_5naoy615oty)

[4.2.3 Data Transformation 16](#_p7olx8zglqtw)

[4.2.4 Feature Engineering 17](#_y1ukdkcdgnj)

[4.4 Model Building 27](#_5g1pfobcp7u1)

[4.4.1 Decision Tree Classifier 28](#_rvy3wios21k)

[4.4.2 Random Forest Classifier 28](#_te8yd937uktu)

[4.4.3 Logistic Regression 28](#_n888kkpnxvgp)

[4.4.4 Multilayer Perceptron (MLP) Classifier 29](#_fpi4jtvibkr0)

[4.5 Comparison of Different Models 29](#_25mdul1orjt3)

[**Chapter 5 - Conclusion 33**](#_b0l7t45ctxn7)

[5.1 Conclusion 33](#_tsz5iyajcp61)

[5.2 Possible Solutions 33](#_wpklsyzmrcb)

[5.3 Future Work 34](#_mqht85al4veq)

[**References 35**](#_epzmj8109opz)

# **Chapter 1**

## **1.1 Background Information**

The adoption of cars in cities has surged in recent years, driven by rapid technological advancements and the fast-paced lifestyle of urban residents. The U.S. Department of Transportation (DOT, 2021) reported an increase in driving licenses issued from 190,625 in 2010 to 228,687 in 2020. This rise in vehicle ownership and drivers on the road has led to a linear relationship with car accident rates, particularly in metropolitan areas like New York City.

Car accidents can occur for various reasons, including controllable factors such as speeding, tailgating, improper lane changes, fatigue, or reckless driving, and uncontrollable factors like vehicle failure, adverse weather conditions, lack of road signs, or poor road conditions. The most common type of car accident is collisions, which can involve single-car accidents, rear-end collisions, sideswipe collisions, multiple-vehicle collisions, or car rollovers. These accidents can result in severe consequences, including injuries, disabilities, fatalities, property loss, and vehicle damage, leading to both personal and social losses.

The World Health Organization (WHO, 2020) estimates that 1.3 million deaths and approximately 50 million injuries occur annually due to car accidents. The French Epidemiological Center (Ce´piDC, 2015) indicated that individuals aged 15-24 have the highest average mortality rate from car accidents. These incidents have significant financial implications for governments, insurance companies, employers, and individuals. The National Highway Traffic Safety Administration (NHTSA, 2010) estimated that losses due to accidents could amount to $1 trillion, factoring in lost productivity and the value of life.

In light of these findings, it is crucial to conduct a thorough analysis of road accidents in New York City to identify the primary factors contributing to the increased accident rates. By understanding these factors, appropriate solutions and preventative measures can be developed to minimize accidents and their associated consequences.

## **1.2 Statement of the problem**

There is a need for a comprehensive analysis of the factors impacting car accident severity in New York City, using a large and diverse dataset to identify patterns and accurately predict injury severity. This analysis should consider aspects such as population density, time-related factors, and city layout, including the presence of bike lanes and pedestrian crosswalks. By addressing these limitations in previous research, this project aims to generate valuable insights that can inform policy decisions and contribute to reducing the frequency and severity of road accidents in New York City.

## **1.3 Project goals**

The primary goal of this paper is to equip government officials with the necessary information to develop effective solutions addressing the rapid increase in car accidents and minimizing their occurrence. This will create a safer environment for everyone by tackling the causes of these life-threatening accidents, subsequently reducing the financial burden on governments, individuals, and businesses. The main objective of the project is to analyze all available factors and gain a comprehensive understanding of the elements contributing to car accidents in New York City.

Key research questions for this project include:

1. Which contributing factors have the highest frequency in NYC car accidents?
2. What is the relationship between the number of injured persons and the time of day?
3. Which day of the week experiences the highest number of accidents?
4. What are the most dangerous intersections in NYC, and what factors contribute to their riskiness?
5. What vehicle types are most often involved in severe accidents?
6. How have the frequency and severity of car accidents in NYC changed over time?
7. How has implementing Vision Zero policies in NYC affected the frequency and severity of car accidents, and what more can be done to improve road safety in the city?
8. Which contributing factors are most common across all boroughs in NYC?
9. In which season do most accidents occur?

By addressing these research questions, this project aims to provide valuable insights that can inform policy decisions and contribute to reducing the frequency and severity of road accidents in New York City.

## **1.4 Methodology**

Numerous factors contribute to road accidents, and to identify the primary causes, a thorough study must be conducted to build a comprehensive understanding of how each factor affects the rate at which road accidents increase. CRISP-DM provides an excellent methodology for carrying out a data analytics project.

The first step of this methodology is Data Understanding: the dataset is a collection of accidents in New York City from 2012 to 2023 and consists of almost 2 million observations and various attributes. We performed feature engineering on the dataset obtained from NYC OpenData to clean and generate new data for our analysis.

The second step is Data Preparation, in which the data is modified by removing unwanted attributes, adding new attributes that will be helpful during the evaluation phase, and addressing missing values. Once the preparation is complete, different visualizations are performed to understand the relationships between variables and accidents.

The third step is Modeling. After performing visualizations and understanding the relevance between accidents and the chosen attributes, four different algorithms will be applied and compared to predict accident severity: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and MLP Classifier.

The final step consists of evaluating the models and visualization results. Once the conclusions are drawn, concerned parties can use these findings to develop effective solutions to address the primary problem of the rapid increase in road accidents in New York City.

## **1.5 Study Limitations**

The limitations of this research paper arise from the incomplete information in the dataset. The dataset does not include driver profile information such as age, gender, or race, which could aid in identifying road accident factors and detecting patterns. Additionally, the dataset does not contain weather information, which could be a significant contributing factor.

# **Chapter 2 - Literature Review**

Every year, billions of dollars are spent on addressing car crashes. A study by Miller et al. (2011) revealed that governments allocate around $35 billion annually to cover medical expenses, social welfare, and forgone taxes related to injured individuals and fatalities. Road traffic crashes pose significant economic and health challenges, particularly for developing countries (Chen, 2010). Various factors contribute to accidents; for instance, Reyner and Horne (1998) found that many road accidents occur due to drivers falling asleep, with many of these incidents being work-related. Knipling and Wang (1994) noted that each lorry in the United States is likely involved in at least one sleep-related crash.

Reyner and Horne (1998) discovered that male drivers under 30 are more prone to sleep-related accidents in the early morning due to poor sleeping patterns. However, such accidents shift to drivers aged 50 and above later in the afternoon. They concluded that self-awareness is crucial in preventing these incidents, and sleeping detectors would not help; instead, the best solution is to pull over and stop driving.

Celik and Oktay (2014) argued that less educated drivers are more susceptible to fatal accidents, and their study indicated that the driver's age and driving time are strong predictors of such incidents. Tadege (2020) stated that driver age is a significant contributing factor behind fatal car accidents. Furthermore, driver inexperience raises the likelihood of human error leading to severe or fatal accidents. Tadege (2020) also reported that male drivers caused around 99.6% of fatal accidents. Abu Jadayil et al. (2020) explained that young male drivers are more likely to cause accidents than older adults due to their tendencies to be reckless and exceed speed limits, which in turn increases the chances of crashing.

Car accidents are mainly associated with two factors: traffic and human error (Gicquel et al., 2017). Human errors encompass various forms, including poor decision-making, sleeping, tailgating, alcohol and drug consumption, mobile phone usage, and more. Abu Jadayil et al. (2020) noted that human error is the primary cause of accidents, accounting for 96.8% of 97,981 incidents. Fan (2015) asserts that over-speeding is a common cause of accidents, as drivers often struggle to avoid collisions when time is limited and decisions are difficult to make.

Waylen and McKenna (2008) found that driving under the influence of drugs or alcohol triples the chances of car accidents, impairing drivers' concentration, reflexes, and awareness. The widespread use of mobile phones has become a significant distraction, with texting and driving increasingly prevalent, particularly among young people, leading to numerous severe accidents (Saifuzzaman et al., 2015). Drivers using their phones often neglect to pay attention to vehicles in front of them or road signs like traffic signals and stop signs.

Accidents may also result from extreme weather conditions, poor road conditions, or vehicle malfunctions. Fan (2015) concluded that brake failure, steering system failure, light failure, and tire bursts are unavoidable factors in car accidents, although they are relatively rare occurrences. Chen et al. (2019) argued that weather and road conditions substantially impact drivers' behavior, creating potentially hazardous situations. Low visibility due to snow and dense fog can affect drivers' behavior. Mao et al. (2019) emphasized the importance of various factors in car accidents, such as driver age, gender, weather, traffic density, vehicle speed, lane changes, vehicle type, time of day, and day of the week.

Road lighting and visibility are crucial factors influencing driver behavior (Farooq and Juhasz, 2019). Mobile phones and blind spots are primary contributors to reduced driver visibility and reaction time, resulting in accidents. Drivers often cannot brake or avoid collisions when faced with visibility issues (Boyce, 2003). Jägerbrand and Sjöbergh (2016) found that vehicle speed in clear weather conditions and daylight is higher than during hours of darkness. Rain significantly impacts fatalities and serious injuries, although rush hours and extreme nighttime conditions were excluded from their study.

Car accidents are undeniably a pressing issue that not only affects people's lives but also results in governments and institutions spending significant amounts on addressing the damages caused by these accidents. Factors contributing to accidents include drowsy driving, mobile phone usage, inadequate road, and lighting conditions, and adverse weather conditions. Previous research has demonstrated that age, gender, and human errors are the most influential factors contributing to road accidents. However, these studies are not without limitations, such as small sample sizes and incomplete information regarding traffic hours, vehicle speed, location, extreme weather conditions, and driver profiles.

# **Chapter 3 - Project Description**

## **3.1 Data Collection**

The quality, source, and reliability of data are crucial factors in data analysis. The data has been collected through the Motor Vehicle Collisions crash table, which contains information from all police-reported motor vehicle collisions in New York City. Police officers are required to complete an MV-104AN form for collisions where someone is injured or killed, or where there is at least $1000 worth of damage. Initially, the Traffic Accident Management System (TAMS) was implemented in July 1999 to collect basic intersection traffic crash statistics in a uniform method across the city. However, in March 2016, TAMS was replaced with the Finest Online Records Management System (FORMS) to address the need for more comprehensive traffic data. FORMS allows police officers to electronically enter all MV-104AN data fields using a department cellphone or computer, and stores this information in the department's crime data warehouse for detailed traffic safety analyses. It has 2 million observations and 29 attributes.

**3.2 Data Exploration**

## The dataset contains 29 different attributes; table 1 includes the name of the attributes, descriptions, and types.

| # | Attribute | Description | Type |
| --- | --- | --- | --- |
| 1 | CRASH DATE | Occurrence date of collision | Continuous |
| 2 | CRASH TIME | Occurrence time of collision | Continuous |
| 3 | BOROUGH | Borough where the collision occurred | Nominal |
| 4 | ZIP CODE | Postal code of incident occurrence | Continuous |
| 5 | LATITUDE | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | Discrete |
| 6 | LONGITUDE | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | Discrete |
| 7 | LOCATION | Latitude, Longitude pair | Discrete |
| 8 | ON STREET NAME | Street on which the collision occurred | Nominal |
| 9 | CROSS STREET NAME | Nearest cross street to the collision | Nominal |
| 10 | OFF STREET NAME | Street address if known | Nominal |
| 11 | NUMBER OF PERSONS INJURED | Number of persons injured | Continuous |
| 12 | NUMBER OF PERSONS KILLED | Number of persons killed | Continuous |
| 13 | NUMBER OF PEDESTRIANS INJURED | Number of pedestrians injured | Continuous |
| 14 | NUMBER OF PEDESTRIANS KILLED | Number of pedestrians killed | Continuous |
| 15 | NUMBER OF CYCLIST INJURED | Number of cyclists injured | Continuous |
| 16 | NUMBER OF CYCLIST KILLED | Number of cyclists killed | Continuous |
| 17 | NUMBER OF MOTORIST INJURED | Number of vehicle occupants injured | Continuous |
| 18 | NUMBER OF MOTORIST KILLED | Number of vehicle occupants killed | Continuous |
| 19 | CONTRIBUTING FACTOR VEHICLE 1 | Factors contributing to the collision for designated vehicle | Nominal |
| 20 | CONTRIBUTING FACTOR VEHICLE 2 | Factors contributing to the collision for designated vehicle | Nominal |
| 21 | CONTRIBUTING FACTOR VEHICLE 3 | Factors contributing to the collision for designated vehicle | Nominal |
| 22 | CONTRIBUTING FACTOR VEHICLE 4 | Factors contributing to the collision for designated vehicle | Nominal |
| 23 | CONTRIBUTING FACTOR VEHICLE 5 | Factors contributing to the collision for designated vehicle | Nominal |
| 24 | COLLISION\_ID | Unique record code generated by system. Primary Key for Crash table. | Continuous |
| 25 | VEHICLE TYPE CODE 1 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | Nominal |
| 26 | VEHICLE TYPE CODE 2 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | Nominal |
| 27 | VEHICLE TYPE CODE 3 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | Nominal |
| 28 | VEHICLE TYPE CODE 4 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | Nominal |
| 29 | VEHICLE TYPE CODE 5 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | Nominal |

Table 1: List of attributes and their description

# **Chapter 4 - Project Analysis**

## **4.1 Exploratory Data Analysis**

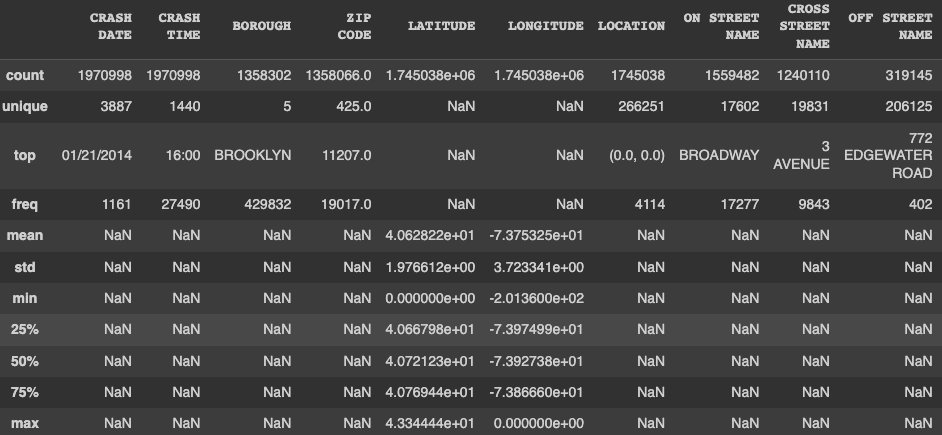
The dataset contains 1,970,998 records of traffic collisions, with attributes such as the crash date, crash time, borough, zip code, latitude, longitude, location, street names, and various statistics on persons injured or killed. The data is available for five different boroughs and 425 unique zip codes. The mean latitude and longitude are approximately 40.63 and -73.75, respectively. The average number of persons injured in a collision is around 0.3, with an average of 0.0014 persons killed. The dataset also includes contributing factors for up to five vehicles involved in the collision and the vehicle types. The most common contributing factor is unspecified, with 681,653 occurrences for the first vehicle and 1,409,041 for the second vehicle. The summary statistics provided also include the mean, standard deviation, minimum, and quartiles for the numeric attributes. 

Figure 1: Data Summary

## **4.2 Data Preprocessing**

In this research paper, we performed several critical preprocessing steps on the dataset, which included data cleaning, data selection, data transformation, and feature engineering. These steps were essential to prepare the data for further analysis and to build a robust and accurate predictive model.

### **4.2.1 Data Cleaning**

The first step in the preprocessing was to handle the missing values in the dataset. We began by replacing "nan" strings with numpy's NaN values, which are easier to work with. We then removed rows with missing values in the 'CONTRIBUTING FACTOR VEHICLE 1', 'LONGITUDE', 'BOROUGH', 'ON STREET NAME', 'CROSS STREET NAME', 'CONTRIBUTING FACTOR VEHICLE 2', 'VEHICLE TYPE CODE 1', and 'VEHICLE TYPE CODE 2' columns, as these are crucial features for our analysis. Additionally, we filled the missing values in 'NUMBER OF PERSONS INJURED' and 'NUMBER OF PERSONS KILLED' columns with zeros, assuming that missing values indicate no injuries or fatalities.

### **4.2.2 Data Selection**

To focus on the most relevant features for our analysis, we selected a subset of the columns available in the dataset. The final set of features we chose included ['BOROUGH', 'DAY\_OF\_WEEK', 'SECONDS\_SINCE\_MIDNIGHT', 'SEASON', 'ACCIDENT\_COUNT', 'CONTRIBUTING FACTOR VEHICLE 1', 'CONTRIBUTING FACTOR VEHICLE 2','VEHICLE TYPE CODE 1','VEHICLE TYPE CODE 2','INJURY\_CATEGORY']. These features were considered significant for our analysis based on their potential impact on the occurrence and severity of accidents.

### **4.2.3 Data Transformation**

We converted the 'CRASH DATE' column to a datetime object, which allowed us to extract additional temporal features easily. We also created a new column 'DAY OF WEEK' to store the day of the week when the accident occurred. To capture the time of the day when the accidents took place, we converted the 'CRASH TIME' column to a datetime object and calculated the total number of seconds since midnight for each record. This transformation facilitated easier analysis and modeling of the time-related patterns in accidents. In the data transformation process, we also incorporated a new feature called "SEASON" by categorizing the date of the accident into four distinct seasons: Spring, Summer, Fall, and Winter. This categorization was based on the month of the accident, with the assumption that the months' of March, April, and May belong to Spring; June, July, and August to Summer; September, October, and November to Fall; and December, January, and February to Winter. This new feature enriches the dataset by providing an additional dimension for analysis, as it allows us to explore if there are any seasonal patterns or trends in the occurrence of accidents. We also standardize the street names in the dataset to improve data consistency and facilitate analysis. This process involved converting all street names to uppercase and removing any leading or trailing whitespaces. Additionally, we replaced common street name abbreviations with their full forms to ensure uniformity. For example, we replaced 'ST.' with 'STREET', 'AVE.' with 'AVENUE', 'BLVD.' with 'BOULEVARD', 'RD.' with 'ROAD', 'PL.' with 'PLACE', 'PKWY.' with 'PARKWAY', 'DR.' with 'DRIVE', 'LN.' with 'LANE', and 'CT.' with 'COURT'. This standardization was applied to both 'ON STREET NAME' and 'CROSS STREET NAME' columns.

By implementing this transformation, we ensured that the dataset would be more easily interpretable and could be used more effectively for analysis or comparison with other datasets.

### **4.2.4 Feature Engineering**

To analyze the severity of accidents, we created a new column 'INJURY\_SEVERITY' by combining 'NUMBER OF PERSONS INJURED' and 'NUMBER OF PERSONS KILLED' columns, where the latter was multiplied by 10 to give more weight to fatalities. We then categorized the injury severity levels into 'No Injury', 'Mild Injury', 'Moderate Injury', 'Severe Injury', and 'Fatal' using the 'INJURY\_SEVERITY' column values. This categorization enabled us to analyze the distribution of accident severities and build a model that could predict the severity of an accident.

The preprocessing steps of data cleaning, data selection, data transformation, and feature engineering were crucial for preparing the dataset for further analysis and modeling. By addressing missing values, selecting relevant features, transforming time-related data, and engineering new features to capture accident severity, we ensured that the dataset was robust and suitable for our research objectives.

**4.3 Visualizations**  
In our comprehensive analysis of car accidents in New York City, we discovered several interesting trends and patterns.

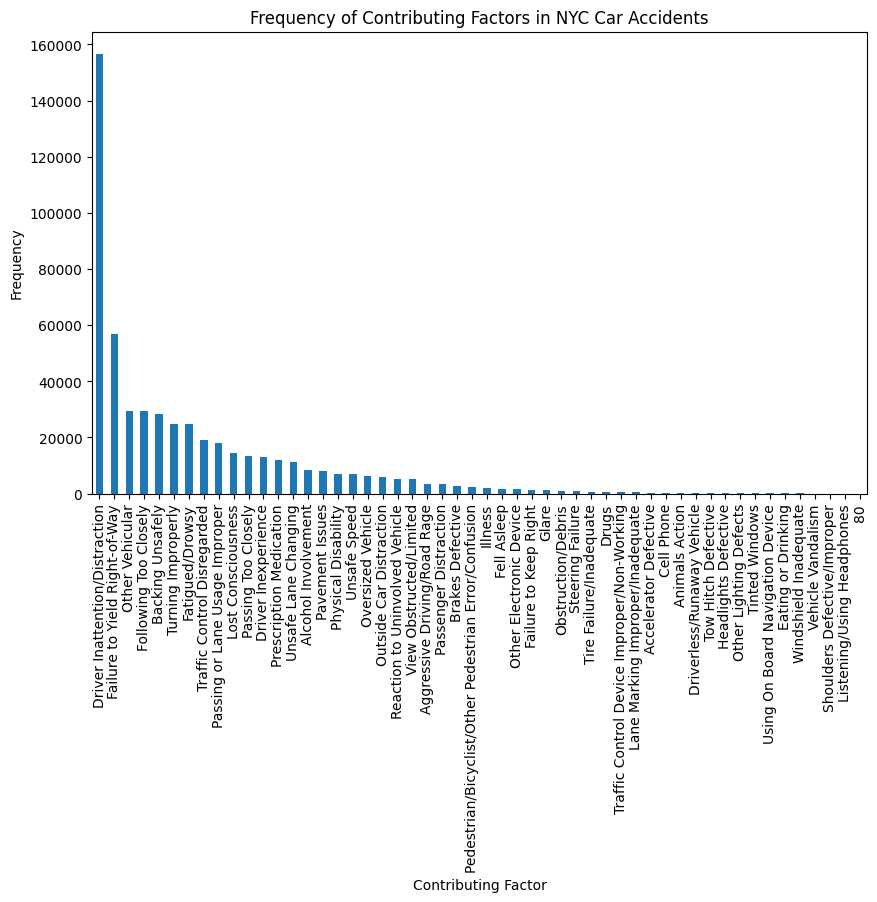


Figure 2: Frequency of Contributing Factors in NYC Car Accidents

Figure 2 depicts the frequency of various contributing factors in NYC car accidents. We found that "Driver Inattention/Distraction" was the leading cause of accidents, followed by "Failure to Yield Right-of-Way" and "Following Too Closely." This information highlights the importance of driver attentiveness and adherence to traffic rules to prevent accidents.

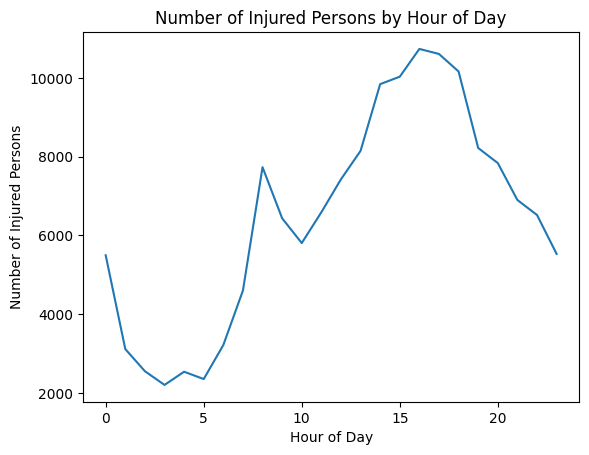


Figure 3: Number of Injured Persons by Hour of Day

Figure 3 shows the relationship between the number of injured persons and the time of day. We observed that the number of injuries peaked during rush hours, particularly in the late afternoon and early evening. This pattern suggests that congestion and increased traffic during these times may contribute to a higher risk of accidents and injuries.

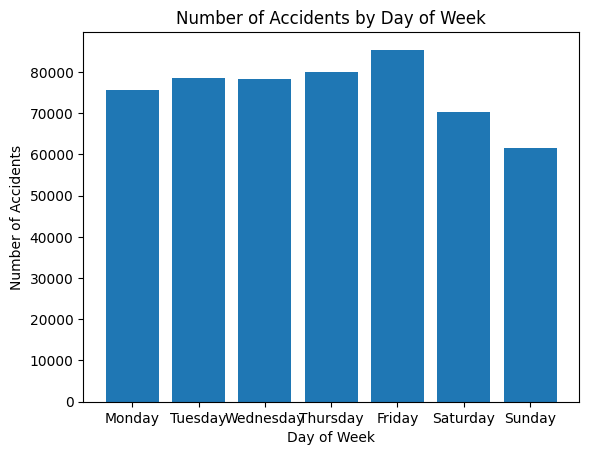


Figure 4: Number of Accidents by Day of Week

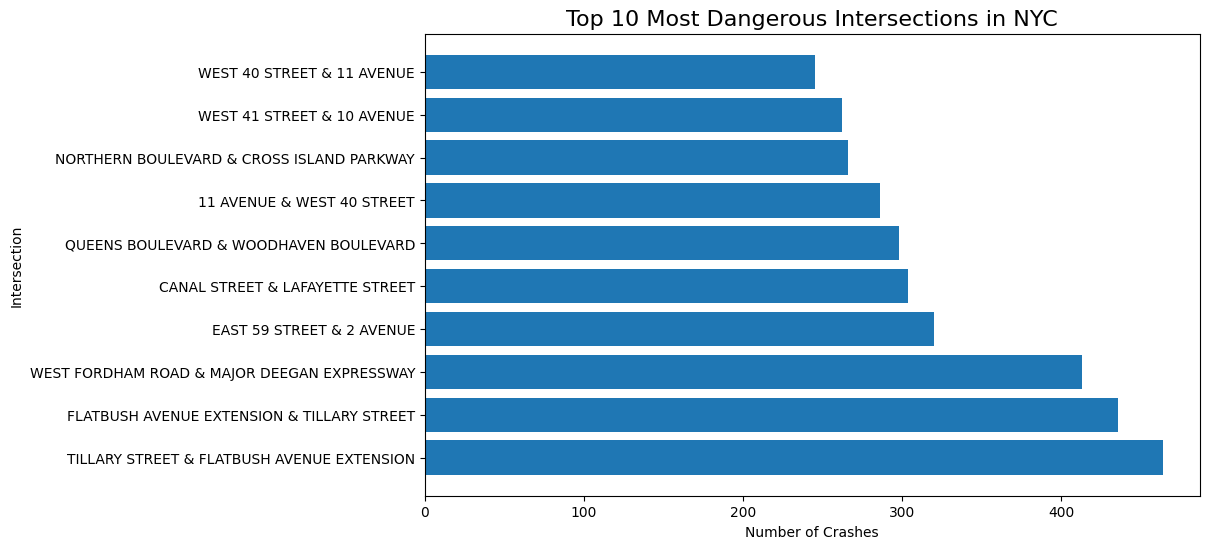
Figure 4 focuses on the days of the week with the highest number of accidents. Surprisingly, Fridays had the most accidents, while weekends witnessed a lower number of accidents. This could be attributed to increased traffic and rush during the end of the workweek, while fewer people may be on the roads during weekends.

Figure 5: Top 10 Most Dangerous Intersections in NYC

Figure 5 investigates the most dangerous intersections in NYC and the factors contributing to their riskiness. We identified the top 10 most dangerous intersections based on the number of accidents and analyzed the main contributing factors at each location. By understanding these factors, city planners and authorities can implement targeted interventions to improve road safety at these high-risk spots.

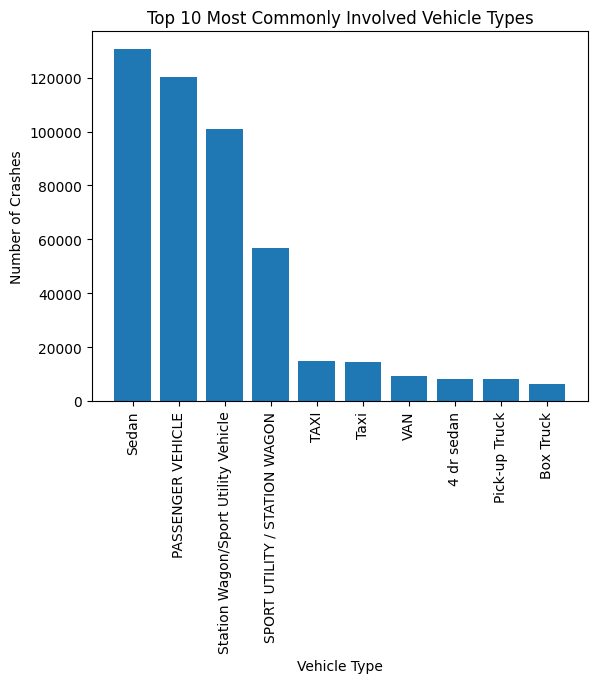


Figure 6: Top 10 Most Commonly Involved Vehicle Types

Figure 6 looks into the most dangerous vehicle types involved in car accidents in NYC. We found that "Sedan" was the most common vehicle type involved in accidents, followed by "Station Wagon/Sport Utility Vehicle" and "Taxi." This information can be used to raise awareness among drivers of these vehicle types and encourage them to take extra precautions while driving.

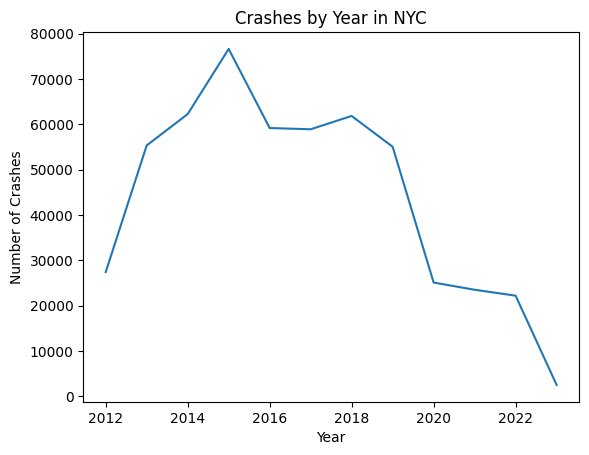


Figure 7: Crashes by Year in NYC

Figure 7 presents a bar chart that illustrates a general decrease in the number of car crashes in New York City from 2013 to 2021. The chart shows that the number of crashes has been on a downward trend during this period, indicating that driving conditions or enforcement policies may have improved over time, leading to a safer environment for both drivers and pedestrians.

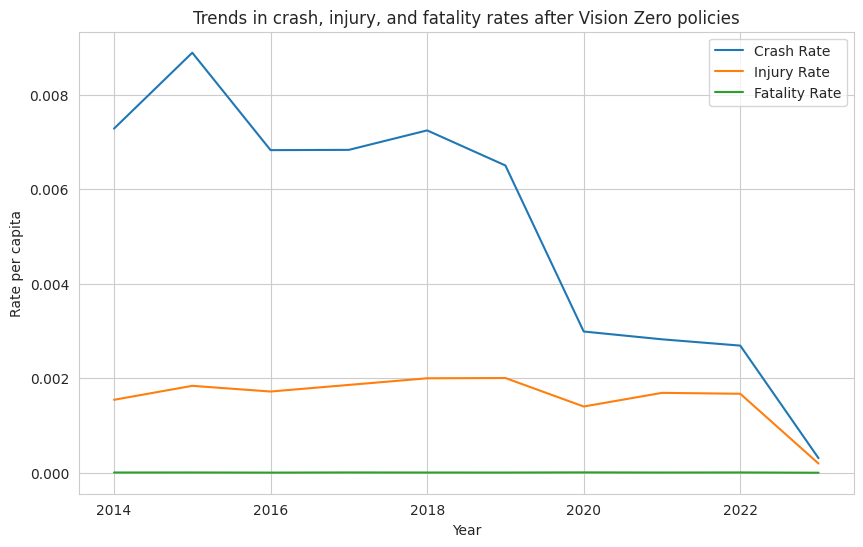


Figure 8: Trends in crash, injury, and fatality rates after Vision Zero policies

Figure 8 displays a line graph that highlights the impact of the Vision Zero initiative on crash, injury, and fatality rates in New York City. The graph demonstrates that after the implementation of Vision Zero policies, there has been a steady decline in crash rates, as well as in the number of injuries and fatalities resulting from those crashes. This suggests that the policies have been effective in promoting road safety and reducing the severity of accidents.

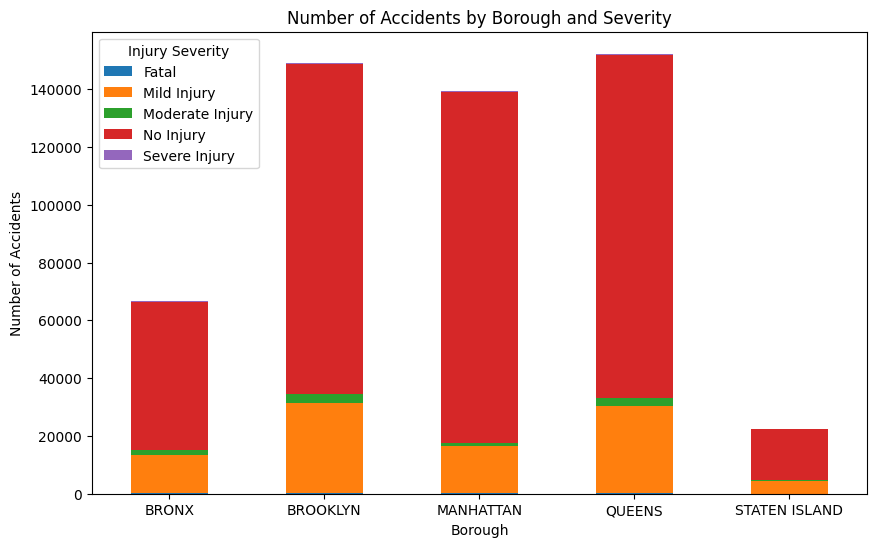


Figure 9: Number of Accidents by Borough and Severity

Figure 9 showcases a bar chart that breaks down the number of accidents in each borough of New York City according to the severity of the injuries sustained in those accidents. The chart provides a clear visual representation of the differences in accident frequencies and severities across the boroughs, allowing for a better understanding of which areas may require more attention or resources to improve road safety.

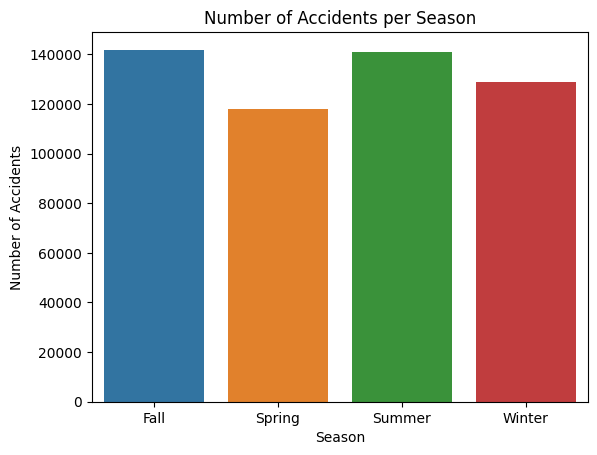


Figure 10 Number of Accidents per Season

Figure 10 is a bar chart that compares the number of accidents in each season of the year. This chart helps to identify any seasonal patterns in car accidents, such as increased accident rates during winter months due to icy roads or decreased rates during summer months when road conditions are generally more favorable.

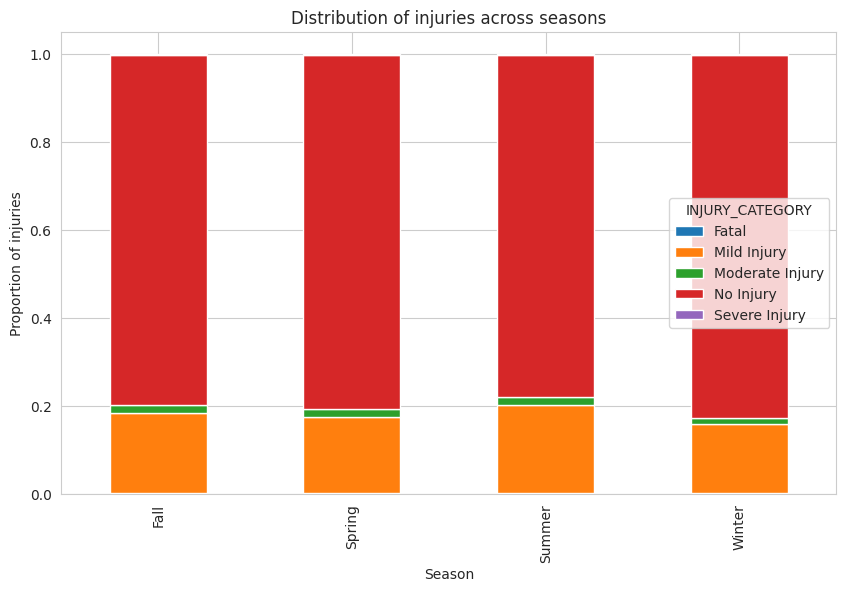


Figure 11 Distribution of injuries across seasons

Figure 11 is a bar chart that displays the proportion of injuries sustained in car accidents during each season. This chart allows for a better understanding of the severity of accidents in each season, and whether certain seasons have a higher prevalence of severe injuries. By examining the distribution of injuries across seasons, it becomes possible to identify potential seasonal risk factors that contribute to more severe accidents, allowing for targeted interventions to enhance road safety.

## **4.4 Model Building**

This section presents the methodology employed for constructing predictive models using the New York City traffic accident data. The objective is to develop a model capable of effectively predicting accident severity based on the available features within the dataset. To accomplish this, several machine learning algorithms are evaluated and compared in terms of performance to identify the most appropriate model for the analysis.

Four distinct algorithms are selected for evaluation:

### **4.4.1 Decision Tree Classifier**

A widely-used classification algorithm that recursively partitions the data into subsets based on input feature values, learning a series of if-then-else decision rules to predict the target variable (Breiman et al., 1984).

### **4.4.2 Random Forest Classifier**

An ensemble method consisting of multiple decision trees trained on random subsets of the dataset. The algorithm introduces randomness into the decision-making process by selecting a random set of features at each split, resulting in a more robust and accurate model (Breiman, 2001).

### **4.4.3 Logistic Regression**

A linear regression model adapted for classification tasks, estimates the probability of an instance belonging to a specific class by fitting a logistic function to the data (Hosmer Jr et al., 2013).

### **4.4.4 Multilayer Perceptron (MLP) Classifier**

An artificial neural network-based classifier employing a feedforward architecture with multiple layers of nodes, each applying an activation function. The MLP model can learn complex, non-linear relationships between input features and target variables (Rumelhart et al., 1986).

## **4.5 Comparison of Different Models**

In this study, four machine learning algorithms were employed to predict accident severity based on New York City traffic accident data. These models were evaluated and compared to identify the best-performing model.

| Mean Squared Error | Metric | Precision | Recall | F1-score | Support | Accuracy |
| --- | --- | --- | --- | --- | --- | --- |
| 1.0168 |  |  |  |  |  | 69.06% ± 4.06% |
|  | 0 | 0.01 | 0.02 | 0.02 | 152 |  |
|  | 1 | 0.32 | 0.35 | 0.34 | 18921 |  |
|  | 2 | 0.04 | 0.05 | 0.04 | 1752 |  |
|  | 3 | 0.84 | 0.82 | 0.83 | 84835 |  |
|  | 4 | 0.00 | 0.00 | 0.00 | 285 |  |
|  | Macro Avg | 0.24 | 0.25 | 0.25 | 105945 |  |
|  | Weighted Avg | 0.73 | 0.72 | 0.72 | 105945 |  |

Table 2: Decision Tree Classifier classification report

Decision Tree Classifier: The Decision Tree Classifier model achieved an accuracy of 69.06% with a standard deviation of 4.06%. The mean squared error for this model was 1.0168, indicating that the predictions deviated significantly from the actual values. Precision, recall, and f1-score values for each class varied considerably, with the lowest scores for class 4 and the highest for class 3.

| Mean Squared Error | Matric | Precision | Recall | F1-score | Support | Accuracy |
| --- | --- | --- | --- | --- | --- | --- |
| 0.6692 |  |  |  |  |  | 79.83% ± 3.54% |
|  | 0 | 1.00 | 0.00 | 0.00 | 152 |  |
|  | 1 | 0.77 | 0.15 | 0.25 | 18921 |  |
|  | 2 | 1.00 | 0.00 | 0.00 | 1752 |  |
|  | 3 | 0.82 | 0.99 | 0.90 | 84835 |  |
|  | 4 | 1.00 | 0.00 | 0.00 | 285 |  |
|  | Macro Avg | 0.92 | 0.23 | 0.23 | 105945 |  |
|  | Weighted Avg | 0.82 | 0.82 | 0.76 | 105945 |  |

Table 3: Random Forest Classifier classification report

Random Forest Classifier: The Random Forest Classifier showed an accuracy of 79.83% with a standard deviation of 3.54%. The mean squared error for this model was 0.6692, suggesting better performance compared to the Decision Tree Classifier. The precision and recall values were generally higher, particularly for class 3, which had a high f1-score of 0.90.

| Mean Squared Error | Matric | Precision | Recall | F1-score | Support | Accuracy |
| --- | --- | --- | --- | --- | --- | --- |
| 0.7465 |  |  |  |  |  | 79.34% ± 1.23% |
|  | 0 | 1.00 | 0.00 | 0.00 | 152 |  |
|  | 1 | 1.00 | 0.00 | 0.00 | 18921 |  |
|  | 2 | 1.00 | 0.00 | 0.00 | 1752 |  |
|  | 3 | 0.80 | 1.00 | 0.89 | 84835 |  |
|  | 4 | 1.00 | 0.00 | 0.00 | 285 |  |
|  | Macro Avg | 0.96 | 0.20 | 0.18 | 105945 |  |
|  | Weighted Avg | 0.84 | 0.80 | 0.71 | 105945 |  |

Table 4: Logistic Regression Classifier classification report

Logistic Regression: The Logistic Regression model had an accuracy of 79.34% with a standard deviation of 1.23%. The mean squared error was 0.7465, which is higher than that of the Random Forest Classifier. The model demonstrated high precision for all classes, but the recall values were considerably lower, leading to low f1-scores for classes 1, 2, and 4.

| Mean Squared Error | Matric | Precision | Recall | F1-score | Support | Accuracy |
| --- | --- | --- | --- | --- | --- | --- |
| 0.6698 |  |  |  |  |  | 79.94% ± 0.02% |
|  | 0 | 0.00 | 0.00 | 0.00 | 152 |  |
|  | 1 | 0.68 | 0.19 | 0.30 | 18921 |  |
|  | 2 | 1.00 | 0.00 | 0.00 | 1752 |  |
|  | 3 | 0.83 | 0.98 | 0.90 | 84835 |  |
|  | 4 | 1.00 | 0.00 | 0.00 | 285 |  |
|  | Macro Avg | 0.70 | 0.23 | 0.24 | 105945 |  |
|  | Weighted Avg | 0.80 | 0.82 | 0.77 | 105945 |  |

Table 5: Multilayer Perceptron (MLP) Classifier classification report

Multilayer Perceptron (MLP) Classifier: The MLP Classifier achieved an accuracy of 79.94% with a standard deviation of 0.02%, making it the most accurate model among the four. The mean squared error for this model was 0.6698, which is lower than that of the Logistic Regression. Similar to the Random Forest Classifier, the MLP Classifier exhibited high precision and recall values for class 3, resulting in a high f1-score of 0.90.

In conclusion, the Multilayer Perceptron Classifier demonstrated the best performance in terms of accuracy and mean squared error. This suggests that the MLP Classifier is the most suitable model for predicting accident severity using the given dataset. However, it is important to note that the performance of each model may vary depending on the specific features used, the hyperparameters chosen, and the data preprocessing techniques applied. Further experimentation and optimization may lead to improved performance for each of the evaluated models.

# **Chapter 5 - Conclusion**

## **5.1 Conclusion**

This study of New York City's car accidents aimed to identify key contributing factors, patterns, and trends through a rigorous process of data preprocessing, cleansing, transformation, and feature engineering. By employing advanced visualization techniques and predictive modeling, we have uncovered critical insights into the underlying causes of these accidents.

Our findings revealed that the highest frequency of accidents occurs during rush hour, as commuters travel to and from work or school, with a higher likelihood of accidents in densely populated areas. The analysis also demonstrated that certain days of the week and specific intersections have higher accident rates, which can aid city planners and law enforcement in developing targeted interventions.

By incorporating these insights and leveraging the attributes such as season, time, borough, and street, authorities can make more informed decisions and implement proactive measures to mitigate traffic severity, enhance road safety, and ultimately reduce the frequency and severity of car accidents in New York City.

## **5.2 Possible Solutions**

Recognizing that traffic congestion is a primary factor contributing to road accidents, implementing strategies to reduce traffic will ultimately decrease accident rates in New York City. Several potential solutions can be explored to address this issue. One such approach is promoting remote work policies, as the COVID-19 pandemic demonstrated that a significant number of employees can effectively complete their tasks from home, thereby reducing the volume of vehicles on the roads during peak hours.

Another potential solution involves staggering the start times of schools and businesses, which would help distribute the number of vehicles on the streets across a wider time frame, alleviating traffic congestion during rush hour periods.

Lastly, encouraging and supporting the adoption of autonomous vehicles could significantly reduce accident rates, as the majority of accidents are often attributed to human error resulting from inattention, impaired concentration, or poor decision-making. By employing self-driving vehicles, human errors can be minimized, leading to a safer driving environment and decreased car accidents in New York City.

## **5.3 Future Work**

Subsequent research endeavors could expand the scope of this study by examining road accident data from various countries to determine if similar trends emerge. Additionally, incorporating more detailed information about drivers and their vehicles, such as age, gender, profession, vehicle type, and ownership status, could provide valuable insights. By investigating these factors, researchers can delve deeper into the psychological aspects of drivers and explore how these characteristics may influence individual driving behaviors in the context of New York City and beyond.

# References

National Center for Statistics and Analysis. (2020, December). *Overview of motor vehicle crash- es in 2019*. (Traffic Safety Facts Research Note. Report No. DOT HS 813 060). National High- way Traffic Safety Administration. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ ViewPublication/813060

National Center for Statistics and Analysis. (2020, March). *Pedestrians: 2018 data* (Traf- fic Safety Facts. Report No. DOT HS 812 850). National Highway Traffic Safety Administration. Retrieved from; https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812850

National Center for Statistics and Analysis. (2019, December). *Seat belt use in 2019 – Overall Results* (Traffic Safety Facts Research Note. Report No. DOT HS 812 875). National Highway Traffic Safety Administration. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812875

National Center for Statistics and Analysis. (2020, October). *Preview of motor vehicle traffic fatalities in 2019* (Research Note. Report No. DOT HS 813 021). National Highway Traf- fic Safety Administration. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813021

Horne, J., & Reyner, L. (1999). Vehicle accidents related to sleep: A review. *Occupational and Environmental Medicine*. BMJ Publishing Group.  
https://doi.org/10.1136/oe m.56.5.289

Farooq, D., Juhasz, J. (2020). *Simulation Analysis of Contributing Factors to Rider Visibility Issues for Car-Motorcycle Accidents*, Periodica Polytechnica Transportation Engineering, 48(3), pp. 203– 209. https://doi.org/10.3311/PPtr.13521

Fan, F. (2018). *Study on the Cause of Car Accidents at Intersections*. Open Access Library Journal, 5: e4578. https://doi.org/10.4236/oalib.1104578

Chen, C., Zhao, X., Liu, H., Ren, G., & Liu, X. (2019). *Influence of adverse weather on drivers’ perceived risk during car following based on driving simulations*. Journal of Modern Transportation, 27(4), 282–292. https://doi.org/10.1007/s40534-019-00197-4

Mao, X., Yuan, C., Gan, J., & Zhang, S. (2019). *Risk factors affecting traffic accidents at urban risk factors affecting traffic accidents at urban weaving sections: Evidence from China*. International Journal of Environmental Research and Public Health,16(9).  
https://doi.org/10.3390/ijerph16091542

Jägerbrand, A. K., & Sjöbergh, J. (2016). *Effects of weather conditions, light conditions, and road lighting on vehicle speed*. SpringerPlus, 5(1).  
https://doi.org/10.1186/s40064-016-2124-6

Jadayil, W. A., Khraisat, W., & Shakoor, M. (2020). *Statistical analysis for the main fac- tors causing car accidents*. ARPN Journal of Engineering and Applied Sciences, 15(5), 696–715. Retrieved from: http://www.arpnjournals.org/jeas/research\_papers/rp\_2020/jeas\_0320\_8150.pdf

Miller, T. R., Bhattacharya, S., Zaloshnja, E., Taylor, D., Bahar, G., & David, I. (2011). *Costs of crashes to Government, United States, 2008*. In Annals of Advances in Automotive Med- icine (Vol. 55, pp. 347–355). Retrieved from:  
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3256813/

Tadege, M. (2020). *Determinants of fatal car accident risk in Finote Selam town, Northwest Ethiopia*. BMC Public Health 20, 624. https://doi.org/10.1186/s12889-020-08760-z

Celik AK, Oktay E. (2014). *A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars provinces of Turkey*. Accident Analysis prev. 72:66–77. Retrieved from: https://doi.org/10.1016/j.aap.2014.06.010

Gicquel L, Ordonneau P, Blot E, Toillon C, Ingrand P and Romo L (2017). *Description of Various Factors Contributing to Traffic Accidents in Youth and Measures Proposed to Alleviate Recurrence*. Front. Psychiatry 8:94. https://doi.org/10.3389/fpsyt.2017.00094

Waylen AE, McKenna FP. (2008). *Risky attitudes towards road use in pre-drivers*. Accident Analysis Prev 40(3):905–11. https://doi.org/10.1016/j.aap.2007.10.005

Saifuzzaman M, Haque M, Zheng Z, Washington S. (2015). *Impact of mobile phone use on car- following behavior of young drivers*. Accident Analysis Prev 82:10–9. https://doi.org/10.1016/j.aap.2015.05.001

Knipling RR, Wang J-S. (1994). *Crashes and fatalities related to driver drowsiness/fatigue*. Washington, DC, OYce of Crash Avoid- ance Research, US Department of Transportation, Research note. Retrieved from: https://rosap.ntl.bts.gov/view/dot/2936/dot\_2936\_DS1.pdf

Boyce PR (2003) Human factors in lighting, 2nd edn. Taylor & Francis, London Brodsky H, Hakkert AS (1988) *Risk of a road accident in rainy weather*. Accident Analysis Prev 20(3):161–176. https://doi.org/10.1016/0001-4575(88)90001-2

Michel Bédard, Gordon H. Guyatt, Michael J. Stones, John P. Hirdes. (2002). *The independent contribution of driver, crash, and vehicle characteristics to driver fatalities*, Accident Analysis & Prevention, Volume 34, Issue 6, Pages 717-727, https://doi.org/10.1016/S0001-4575(01)00072-0