Data Analytics Research Project

(Motor Vehicles Collisions- Crashes)

NEHA REDDY YENUGU

**Abstract**

Motor vehicle collisions (MVCs), commonly known as car crashes, are a prevalent and concerning issue that poses a significant threat to road safety worldwide, claiming the lives of approximately 1.25 million people worldwide each year. Despite having just half of the world's vehicles, low- and middle-income countries account for a staggering 90% of global road fatalities. Moreover, half of those killed on the world's roads are considered "vulnerable road users," including pedestrians, cyclists, and motorcyclists. The disproportionate impact on these vulnerable groups underscores the need for targeted interventions to enhance their safety. These incidents, unfortunately, occur frequently, disrupting daily life and causing a range of consequences, from minor injuries to life-altering disabilities and even fatalities. These incidents stem from a range of factors, including driver error, unfavorable weather conditions, or mechanical malfunctions. The consequences of a collision extend beyond the initial impact, encompassing the repair of damaged vehicles, medical attention for any injured individuals, and the legal proceedings to establish fault and determine liability. This alarming figure highlights the urgent need for effective measures to address road safety. Despite ongoing efforts to enhance road safety through various measures, MVCs remain a persistent challenge, demanding a comprehensive understanding of their causes, consequences, and preventive strategies.

Keywords: Motor Vehicle Collisions, Accident severity, Pedestrians, Motorcyclist, Cyclist, New York City.

**Introduction**

Motor vehicle collisions, commonly known as car crashes, involve the forceful impact of one or more vehicles resulting in varying degrees of damage, injuries, or even fatalities. are a significant public health and safety concern. The advent of motor vehicles revolutionized transportation, providing unparalleled convenience and connectivity. However, this technological marvel comes with a dark side—motor vehicle collisions. Every year, millions of lives are disrupted, and countless others are lost due to the repercussions of these collisions. The consequences are far-reaching, affecting individuals, families, and societies at large. In 2019, an estimated 1.35 million people died in road traffic accidents worldwide, and over 50 million people were injured.

**Causes of MVCs:**

There are many factors that can contribute to MVCs, including:

* Driver error: This is the most common cause of MVCs, and it can include factors such as distracted driving, speeding, and driving under the influence of alcohol or drugs.
* Road conditions: Poor Road design, maintenance, and weather conditions can also contribute to MVCs.
* Vehicle factors: Faulty vehicles or defective vehicle parts can also play a role in MVCs.
* Distracted driving: This is a major cause of accidents, with using a mobile phone being one of the most common distractions.
* Speeding: Driving above the speed limit significantly increases the risk of an accident and the severity of injuries if one occurs.
* Driving under the influence of alcohol or drugs: Impairment significantly impairs judgment and reaction times, making it much more likely to cause an accident.
* Drowsy driving: Driving while fatigued is like driving under the influence of alcohol or drugs, as it impairs judgment and reaction times.
* Failure to use seatbelts: Seatbelts are the most effective way to reduce the risk of death and serious injury in a collision.

**Impacts of MVCs:**

Motor vehicle collisions (MVCs) are a prevalent and devastating societal issue, extending far beyond the immediate damage to vehicles. Their repercussions ripple through individuals, families, and communities, leaving an indelible mark on lives and livelihoods. While the physical toll of MVCs is evident in the immediate aftermath, the true impact extends far deeper, encompassing a myriad of personal, financial, and emotional consequences.

They can cause death, injury, property damage, and emotional distress. MVCs can also have a significant economic impact, costing billions of dollars each year in medical expenses, lost wages, and productivity.

**Related Work**

The above reports mainly focus on the effect of motor vehicle collisions on the pedestrian and pedal cyclist as well as the reasons for the crashes along with the social and economic impact of their collisions.

The reports sheds light on numerous critical elements impacting traffic accidents, with a special emphasis on instances involving youths and heavy motor vehicles. An initial item to examine is that, despite the large difference in driving circumstances, many collision studies tend to disregard the contrast between daylight and nighttime occurrences. Environmental conditions, traffic patterns, and driving behaviors all change significantly between day and night, resulting to different risk levels. Notably, nighttime driving relates to difficulties such as slick roads, driver exhaustion, alcohol intake, and speeding, all of which increase the risk of an accident. This dynamic, together with the increasing frequency of work zone collisions, emphasizes the importance of additional study into injury severity.

Furthermore, the data digs into the prevalent tendency in the United States of selecting big motor vehicles, notably light trucks such as SUVs and pickup trucks. While this preference shift is remarkable, it raises worries about the rising number of pedestrian and bike deaths. External cost assessments show that the expenses of driving these bigger cars are significant. Surprisingly, despite these worries, sales of huge motor vehicles have increased while sales of regular passenger cars have decreased.

The paper emphasizes the significant economic consequences of road accidents, citing a startling cost of $340 billion in 2019. This cost includes a variety of components such as medical costs, productivity losses, legal fees, and property damage. Importantly, a significant portion of these costs are incurred by those who are not directly engaged in the accidents, as seen by higher insurance premiums, increased tax burdens, and congestion-related charges.

Furthermore, the data emphasizes the major risk variables that contribute to teenage car accidents. These risks include the use of alcohol and drugs, inattentive driving, the presence of other minors in the car, unsafe driving habits, and a troubling lack of seat belt usage among youths. Teenagers have a heightened proclivity for dangerous conduct, which typically escalates when peers are around. Furthermore, the increased collision risk linked with evening and weekend driving is stressful, particularly among the younger generation. The research also emphasizes the increased hazard presented by rural roads to young drivers, which requires special attention.

In summary, the data emphasizes the need to tackle these multiple issues in order to improve road safety and reduce the significant economic burden caused by traffic accidents in the United States. The findings presented here form the basis for future study and policy development targeted at minimizing the many hazards and problems involved with road safety.

**Objective**

This project seeks to comprehensively understand and address critical aspects of motor vehicle collisions in New York City by exploring three key questions. Firstly, it aims to examine the influence of the time of day on collision frequency and ascertain whether specific periods correlate with increased collision severity. Secondly, in this project, we will investigate the most common contributing factors to collisions, analyzing variations across different boroughs to inform localized road safety strategies. Lastly, it will explore the relationship between vehicle types involved in collisions and the severity of injuries or fatalities sustained by pedestrians, cyclists, and motorists. And can we predict the severity of motor vehicle collisions based on factors such as time of day, contributing factors, and borough. By providing insights into these questions, the project aims to contribute data-driven recommendations for targeted interventions and policies, ultimately enhancing road safety and reducing the impact of motor vehicle collisions in New York City.

**Proposed Selected Dataset**

The dataset employed in this study provides a comprehensive overview of motor vehicle collisions in New York City, encompassing a wealth of information related to each recorded crash event. The dataset comprises 32 columns, meticulously capturing key details such as the occurrence date and time, location particulars including borough, zip code, and geographical coordinates, as well as precise street information. These granular details lay the foundation for a nuanced analysis of when and where collisions occur.

Crucial indicators of the severity of each collision are meticulously documented through dedicated columns, including the number of individuals injured or killed, further categorized by pedestrians, cyclists, and motorists. This data facilitates the examination of human impact, allowing for a thorough understanding of the differential impact of collisions on vulnerable road users. Additionally, contributing factors to each collision are systematically recorded, providing insights into the circumstances surrounding these incidents.

Furthermore, the dataset includes comprehensive information about the types of vehicles involved, captured through vehicle type codes. This classification, encompassing a wide range of vehicles from ATVs and bicycles to cars, trucks, and motorcycles, enables a detailed examination of the role of different vehicle types in collisions. Moreover, the dataset incorporates a time-of-day column, offering a temporal dimension to the analysis. Lastly, a unique identifier for each collision is provided, facilitating data linkage, and referencing.

These rich and diverse data points collectively pave the way for a multifaceted exploration of motor vehicle collisions in New York City. The project will leverage this comprehensive dataset to address key questions related to the time of day's impact, prevalent contributing factors, and the interplay between vehicle types and collision severity. The inclusion of geographical coordinates also opens avenues for spatial analysis, enabling a holistic understanding of the factors influencing road safety in different areas of the city.

The given dataset has 32 attributes with 280k rows where the dataset describes vehicle collision during the period 2012 to 2023.

It describes when and at what time did collision occur and how many people were injured, or killed on that date and time and a specific vehicle column is provided and how much severity is there, and what are the most contributing factors (i.e. the reason for an accident). SEVERITY is the predictor column where it predicts the degree or level of seriousness, intensity, or impact associated with an event or observation. In my dataset severity column is likely to represent the severity of traffic collisions. It's a crucial variable that provides information about how severe or impactful each recorded collision was. The severity level can help in understanding the potential consequences of the collision and may influence decision-making in areas such as traffic management and safety improvement.

**Proposed Development Platforms**

There are several development platforms which are going to be considered for this project.

**Python**

We will start by using python which is an easy-to-use programming language which is compatible with most other platforms we are going to use.

**R Studio**

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics.

**SQL**

Structured query language (SQL) is a programming language for storing and processing information in a relational database.

**AWS**

Amazon Web Services, Inc. is a subsidiary of Amazon that provides on-demand cloud computing platforms and APIs to individuals, companies, and governments, on a metered, pay-as-you-go basis.

**Data Analysis Methods**

**Data Ingestion**

The first step in our analysis is the Data Ingestion phase. This phase involves loading the dataset into Jupyter and R studio. This dataset provides us with a vast amount of motor vehicles collision date and time along with the location of collisions and number of persons killed or injured during accident and the cause of collisions.

A screenshot of a computer

Description automatically generated

*Fig*. *1*

**Data Exploration**

During the Data Exploration phase, we thoroughly analyze the dataset to gain a comprehensive understanding of its characteristics. This involves examining the data's different features, distributions, and potential patterns.

A screenshot of a computer

Description automatically generated  
*Fig. 2*

**Data Preprocessing**

Preparing the dataset for modeling is a crucial step that involves several operations collectively known as data preprocessing. This includes handling missing values, standardizing numerical features, and encoding categorical variables.

A screenshot of a computer

Description automatically generated

*Fig. 3*

**Methods Used**

During this phase, all research questions are solved and used a Decision Tree to predict the model for predicting accuracy.

**Interpretation**

**How does the time of day affect the number of motor vehicle collisions, and is there a correlation between certain times of the day and increased collision severity?**

Using ggplot, when we map between the number of collisions against the hours of the day, we can observe from the bar plot that the count of collisions is higher during peak commuting hours (8 AM and 5 PM), suggesting increased traffic and potentially higher collision risk during these times. The lowest counts are during late-night and early morning hours.

A graph of a number of collisions

Description automatically generated

*Fig. 4*

A screenshot of a graph

Description automatically generated

*Fig. 5*

.

A graph with numbers and lines

Description automatically generated

*Fig. 6*

Correlation coefficients measure the strength and direction of a linear relationship between two variables. The value ranges from -1 to 1, where:

1: A perfect positive correlation

0: No correlation (variables are independent)

-1: A perfect negative correlation

When we find the correlation between Time of day and Severity, the correlation is very close to zero, indicating a very weak or negligible linear relationship between the time of day and severity. There is little to no linear relationship between the time of day and severity. Indicating that at what time the accident occurred has no relationship with the severity.

Using python, represented a line graph, it represents that at what time the collisions occurred and what is the average severity, we can see that early morning accidents have high severity due to empty roads and less traffic, when collides the severity is high and may cause high damage to persons or the property.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

*Fig. 7*

From the above tables, it shows the number of collisions at time of the day which shows that evening time has highest number of collisions count. And the severity of collisions has highest harm at early mornings hours. Another heat map representation of severity and hour of the day displays the same result.

A graph of different colors

Description automatically generated

*Fig. 8*

Impact of accident and Number of persons injured on severity.

**A screenshot of a computer code

Description automatically generated**

*Fig. 9*

The above regression model tells us that those variables are very significant.

**Interpretation of above model summary:**

**1**. Intercept: When all predictor variables are zero, the estimated SEVERITY is -0.0129040.

2. accident\_severity: For a one-unit increase in accident\_severity, the SEVERITY is estimated to increase by 0.0189463 units.

3. NUMBER.OF. PERSONS.INJURED: For a one-unit increase in the number of persons injured, the SEVERITY is estimated to increase by 0.9940959 units.

4. Overall Fit: The model is statistically significant (p-value < 2.2e-16), and it explains a substantial amount of the variance in SEVERITY (Multiple R-squared = 0.9905).

All the factors that are contributing to motor vehicle collisions are due to various reasons, here are the top 10 factors that are contributing to the highest vehicle collisions. Most of the reasons are not specified or unspecified. Apart from that a significant portion of accidents stem from riders' non-compliance with traffic rules and their lack of situational awareness. These lapses in attention and adherence to regulations can lead to unpredictable behaviors that increase the risk of collisions. Additionally, the factor of mistiming or improper signaling contributes to a substantial number of accidents.

**What are the most common contributing factors to motor vehicle collisions, and how do these factors vary by borough or location within New York City?**

A screenshot of a computer

Description automatically generated

*Fig. 10*

When we plot a heat map between the location and the contributing factors, we can observe that Brooklyn has the highest collisions are not specified and has highest number of collisions at that location. And another highest reason for the next highest collisions at Brooklyn is due to Driver Inexperience and Cell phone use due to driving. At the same location the next succeeding factors at the same location are lack of knowledge in traffic signals and unsafe line changing. Apart from the location Brooklyn, the next highest collisions occurred in New York is Queens, due to the same contributing factors.

A screenshot of a graph

Description automatically generated

*Fig. 11*

**Is there a relationship between the type of vehicles involved in collisions and the severity of injuries or fatalities sustained by** **pedestrians, cyclists, and motorists?**

A graph of text and numbers

Description automatically generated with medium confidence

*Fig. 12*

For the research question, plotted a interactive graph using plotly, where we can observe the severity of the persons against the vehicle type and what is the count of each vehicle type along with the number of collisions occurred. Connected RDS to MySQL and executed all the commands.

A screenshot of a computer

Description automatically generated

*Fig.13*

The above figure shows all the vehicle types which were involved in the collisions. These were the vehicles that were involved in the accident due to various contributing factors.

A screenshot of a computer

Description automatically generated

*Fig. 14*

The above figure shows the accident severity of persons by the number of persons which were categorized according to the collision’s harm level. It was clear that the highest collision severity has the highest effect on the persons as well as the property damage.

A screenshot of a computer

Description automatically generated

*Fig. 15*

Overall number of persons injured involving pedestrians, cyclists, and motorists due to motor vehicle collision between the year 2021- 2023 only in New York City were rated high when compared to the other geographical locations.

A screenshot of a computer program

Description automatically generated

*Fig. 16*

Number of persons injured were more than the number of persons killed which involved pedestrians, cyclists, and motorists in New York City due to various contributing factors.

**Can we predict the severity of motor vehicle collisions based on factors such as time of day, contributing factors, and borough?**

A screenshot of a computer

Description automatically generated

*Fig. 17*

Accuracy is the ratio of correctly predicted instances to the total instances. The overall accuracy of 78% suggests that the model is relatively successful in predicting the severity of motor vehicle collisions with the attributes like Contributing factors, time of the day and borough.

Precision is the ratio of correctly predicted positive observations to the total predicted positives. The high precision (0.78) indicates that when the model predicts low severity. Recall is the ratio of correctly predicted positive observations to all observations in actual class. The high recall (1.00) indicates that the model captures most of the actual instances of low severity.

**Limitation**

Data Quality and Completeness:

The validity of the findings is contingent upon the accuracy and comprehensiveness of the underlying dataset. Inaccuracies or missing data elements could introduce biases and compromise the robustness of the conclusions drawn.

Geographical Focus:

The study is restricted to New York City, limiting the generalizability of the findings to other urban or rural environments with differing traffic patterns and infrastructure.

Categorical Nature of Contributing Factors:

The analysis of contributing factors relies on categorical data, potentially oversimplifying the intricate nature of multifactorial incidents. A more nuanced approach could better capture the complexities inherent in contributing factors.

Dynamic Nature of Road Safety Measures:

The effectiveness of interventions suggested in the study may change over time due to technological advancements, policy modifications, or shifts in societal behaviors, necessitating ongoing evaluation.

**Future Work**

The study gives important insights into the causes and severity of collisions, which may be utilized to build targeted actions to decrease harm and increase public safety. As the project progresses, there are several other solutions that may be implemented to alleviate the issue of route impacts. Incorporating qualitative research methodologies, such as interviews or surveys, alongside quantitative data could provide deeper insights into behavioral aspects contributing to motor vehicle collisions. Additionally, advanced behavioral analytics can be employed to gain a better understanding of driver and pedestrian behavior, such as distracted driving, adherence to traffic rules, and situational awareness, thus paving the way for targeted interventions addressing specific behavioral factors. According to the research, injuries are far more prevalent than deaths, emphasizing the importance of receiving rapid medical assistance in these cases. Therefore, fostering public-private partnerships with entities like insurance companies or technology providers can unlock access to additional data sources and insights, enriching the depth and breadth of data available for analysis.

**References**

1)Motor Vehicle Collisions- Crashes – September29,2023- The City of New York- Data.gov.

[*https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes*](https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes)

2)Thomain- A Ride for a Cause- 24th October 2016-Daily Mirror.

<https://www.dailymirror.lk/print/opinion/Thomian-Ride-A-Ride-for-a-Cause-/172-118029>

3) Zhang Kairan, Hassan Mohamed, August 13,2019, Crashes severity analysis of nighttime and daytime highway work zone crashes.

<https://doi.org/10.1371%2Fjournal.pone.0221128>

4) Edwards Mickey, Leonard Daniel, October 12, 2021, Effects of large vehicles on pedestrian and pedal cyclist injury severity, Visiting Research Specialist, University of Illinois at Springfield, United States,.

<https://doi.org/10.1016/j.jsr.2022.06.005>

5) Nehiba Cody, Tyndall Justin, March 10,2023, Highways and pedestrian deaths in US neighborhoods, National Center for Environmental Economics at the US Environmental Protection Agency, Washington, DC, United States.

<https://doi.org/10.1016/j.regsciurbeco.2023.103938>

6) Factors that Increase the Risk of Crashes, Youth.gov

<https://youth.gov/youth-topics/factors-increase-risk-crashes>

7) Blincoe, L., Miller, T., Wang, J.-S., Swedler, D., Coughlin, T., Lawrence, B., Guo, F., Klauer, S.,

& Dingus, T. (2023, February). The economic and societal impact of motor vehicle crashes, 2019

<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813403>