DSCI 5260

BUSINESS PROCESS ANALYTICS

Group-5

Dr. Sameh Shamroukh

PROJECT PROPOSAL

**Traffic Collisions​ in Chicago**

​

Shivani Patlolla (11644461​)

Abhinav Soma (11656519​)

Adithi Chintha (11637571​)

Jahnavi kommineni (11641025​)

Pulkam vasanth kumar (11639822​)

# TABLE OF CONTENTS

[ACKNOWLEDGEMENT **4**](#_ids6jw9d1dva)

[PROJECT GOALS 7](#_6pf0cyf94lh4)

[RESEARCH PROBLEM 8](#_s2cm87exwura)

[RESEARCH OBJECTIVE 9](#_xp7xxg3tktmz)

[POTENTIAL STAKEHOLDERS 10](#_hqleruz69wgg)

[RESEARCH QUESTIONS 11](#_30j81y1g5lfj)

[LITERATURE REVIEW 11](#_g70dcj9yqbzu)

[DEMOGRAPHIC AND TRAFFIC CRASHES  **14**](#_fobkbhwv5xhe)

[RESEARCH DESIGN AND METHODOLOGY 20](#_4eg6zp463jx5)

[DATA 24](#_cwo949yjfdep)

[PARTICIPANT DESCRIPTION: 26](#_j1ojk6wpjqkq)

[PROCEDURES 27](#_z98u2qvqcelg)

[POSSIBLE PROCEDURES 27](#_i187yl3qe8xr)

[RESEARCH MEASURES: DEPENDENT AND INDEPENDENT VARIABLES 29](#_misyhd1b943b)

[LIST OF REFERENCES **83**](#_go8rvmsddu1i)

# 

# 

# ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to Dr. Sameh Shamroukh for his invaluable mentorship and guidance throughout our research journey. His advice and guidance throughout the ‘Business Process Analytics” course has given us the knowledge and abilities we need to finish our final research project.

We also express our sincere thanks to every group member of our team “Group-5” for their dedication, collaboration, and meaningful contributors to our project.

It is our hope that this research achieves its intended goals and makes a meaningful contribution to the field of data science, providing guidance to students navigating their path towards achieving their aspirations in this dynamic domain

# 

# 

# 

# 

# 

# 

# 

**Abstract**

This study takes a closer look at traffic accidents in Chicago, with a special focus on pedestrian and bicycle safety. The goal is to uncover patterns and factors behind these accidents to create more effective safety measures. The majority of crashes, including rear-end collisions, turning accidents, and collisions with parked automobiles, occur in moderate speed zones, particularly those with speed restrictions between 20 and 40 mph. Due to the high volume of vehicles, bikers, and pedestrians in these regions, encounters and accidents occur frequently. Although there are fewer incidents in low-speed zones (such as residential or school neighbourhoods), vulnerable road users, such as cyclists and children, are frequently involved. However, collisions at high-speed zones are more serious but less common. Compared to regions with only stop or yield signs, traffic signals significantly improve road safety by lowering the frequency and intensity of collisions. The study also investigates how weather and time of the week affect collisions. Thursdays have the most collisions, but weekends have fewer collisions but a higher likelihood of serious ones, probably as a result of riskier driving practices like speeding or intoxicated driving. Unexpectedly, weather also affects accidents. The majority of collisions occur at speed limits of 25 to 30 mph, which are typically in cities. This emphasizes the need for improved pedestrian crossing, signs, and traffic calming techniques in these congested regions.

Using data visualizations like box plot and heatmaps the study also identifies critical trends, outliers, and correlations. By analyzing these patterns, the research offers practical insights to help reduce crashes, injuries and fatalities while improving traffic flow in Chicago’s busy streets.

**INTRODUCTION**

Metropolitan areas face difficulties addressing traffic accidents, one of the primary causes of fatalities and injuries globally, due to their large populations and concentration of various types of transportation, including motorcycles. In cities like Chicago, where cars, buses, bicycles, motorcycles, and pedestrians share the same infrastructure, the likelihood of accidents rises sharply. Effectively managing traffic safety in this dynamic environment requires a thorough understanding of the factors contributing to collisions. Carefully collecting and evaluating the raw data is necessary to produce insightful findings. Not only do traffic accidents result in fatalities and injuries, but they also have long-term effects by raising medical expenses and negatively affecting the economy. Additionally, it lowers productivity by increasing medical costs. The Chicago police department uses the Illinois Traffic Crash Reporting form (SR1050) to gather comprehensive data on every traffic accident in the city to solve these issues.

This data includes information on every unit involved in the collision, including motorcycles, vehicles, bikes, and pedestrians. Direction, speed, and the dynamics that led to the accident are among the distinctive characteristics of the data. Each variable in the dataset has distinct characteristics that aid in precisely comprehending how collisions occur and how various vehicle types or non motorized individuals contribute to the incident. This thorough approach makes it much easier for municipal planners and traffic specialists to comprehend crash trends and develop targeted solutions to reduce incidents and improve safety. The goal is to transform Chicago's roadways into safer, more efficient movement networks for everyone who uses them, including drivers, cyclists, pedestrians, and public transportation users. As Chicago's transportation system evolves, the analysis's conclusions can also be applied to urban planning and development plans, guaranteeing that safety and sustainability are given priority.

# PROJECT GOALS

Chicago traffic accidents frequently result in severe injuries and fatalities, especially for vulnerable populations like walkers and cyclists. This causes them a serious threat to public safety. These incidents highlight the significance of understanding the factors that contribute to collisions and the efficacy of the safety measures put in place to guarantee the security of all drivers as we review the crash data for Chicago. Our main goal is to identify the primary causes of traffic accidents and assess how modern safety features, such as bike lanes, crosswalks, and traffic signals, prevent them. We can delve deeper by concentrating on the areas where non motorized users, such as cyclists and pedestrians, frequently get into collisions—an in-depth analysis of these incidents' dynamics. The information and understanding are crucial for improving roadway safety.

Our data will ultimately help to guide decisions on where the city should invest in more significant safety regulations, better infrastructure, and better urban area design to reduce the chance of accidents and save lives.

**Some of the project goals are:**

**Identifying key contributing factors to traffic accidents:**

In-depth analysis of these incident's dynamics. The information and understanding are crucial for improving roadway safety. To reduce the chances of accidents and save lives, our data will ultimately guide decisions on where the city may invest in more significant safety regulations, better infrastructure, and better urban area design.

**Evaluate the effectiveness of current safety control:**

Assess the efficiency of the safety measures currently in place. The primary objective is to evaluate the effectiveness of the current safety protocols, such as bike lanes, speed restrictions, crosswalks, traffic signals, etc., in averting collisions and safeguarding drivers. The goal of the study is to assess whether the safety infrastructure that is currently in place is operating as anticipated. Do safety measures like bike lanes and crosswalks lower the number of accidents?

The project's main objective is to reduce traffic accidents and improve road safety for all users, especially the most vulnerable. It will do this by analyzing crash data to determine the cause of accidents and assess the efficacy of present safety measures.

# RESEARCH PROBLEM

One of the biggest threats to public safety in places like Chicago is traffic accidents. Collisions are far more likely when there is a dense population and shared roads for automobiles, buses, bikes, motorbikes, and pedestrians. Particularly vulnerable are pedestrians and cyclists, frequently as a result of poor infrastructure, inclement weather, speeding, and distracted driving. In addition to causing injuries and fatalities, these incidents have long-term financial repercussions. Deeper imbalances throughout the city are exacerbated by growing medical costs, lost productivity, and the disproportionate impact on lower income. Neighborhoods frequently have higher crash rates. The main cause of traffic accidents in Chicago is the city’s complicated transportation system, which includes a variety of different traffic patterns and significant congestion. Traffic accidents continue to be a serious concern despite efforts to address these issues. By enhancing infrastructure, developing better regulations, and increasing public awareness, initiatives like Vision Zero seek to eradicate fatalities and serious injuries. Initiatives like bike lanes, crosswalks, and traffic signals have resulted from the Chicago police department’s use of tools like the Illinois Traffic Crash Reporting form (SR1050), which has helped reveal accident accident trends. But the issue has not been resolved yet by these actions. To handle the particular difficulties faced by Chicago, safety measures must be continuously reviewed and modified. International initiatives like the United Nations Decade of Action for Road Safety emphasize the value of local solutions because tactics that are successful in one location might not be in another. To further analyze the crash trends, assess existing safety protocols, and close any gaps in Chicago’s traffic safety planning, more study is necessary. It also evaluates how well the current safety measures are working. The ultimate goal is to offer practical insights that can help shape smarter investments in infrastructure, better policies, and new safety initiatives.

# RESEARCH OBJECTIVE

This research aims to measure the efficiency of presently implemented safety measures and determine the influence of the environment on the occurrence of accidents and the rates at which the accidents occur, as well as the time of day, weather conditions, and road surface characteristics. Traffic accidents are the subject of a broad literature review concerning several aspects of road safety; however, the principal focus of the study is on pedestrian and bicycle safety in Chicago. If the causes of specific types of crashes, road structures, danger zones, and specific crash patterns are analyzed, people can understand the reasons behind the accidents. This approach will allow the successful creation of targeted and effective safety interventions, upgrades, designs, and policies relevant to decreasing transportation incidents in Chicago. In achieving these objectives, this research strives to improve strategies for traffic safety, reduce crashes, injuries, and fatalities, and address traffic system congestion in Chicago's transportation modes.

# POTENTIAL STAKEHOLDERS

The key stakeholders in this research problem are diverse groups with significant stakes in traffic safety in Chicago. These are the primary stakeholders in the City of Chicago government: the policymakers, the Chicago Department of Transportation (CDOT), and the Chicago Police Department (CPD). The transport customers, including pedestrians and cyclists, and the mass transport providers, such as the Chicago Transit Authority (CTA), are just as relevant. Urban planners, traffic engineers, local safety advocacy groups, insurance companies, and healthcare providers treating crash victims also have vested interests in the research outcomes.

The purpose of such an approach is to discover the patterns, causes, and factors that characterize traffic crashes in Chicago, specifically those affecting pedestrians and cyclists. This research aims to assess the effectiveness of existing safety practices and identify environmental factors contributing to accident occurrence and frequency.

Based on this research's findings, these implementations would help the stakeholders by providing data-driven insights into designed safety interventions and public measures. They would assist them in identifying risk factors and allocating resources for infrastructure upgrades and safety programs. Thus, participants can explore what safety measures exist to prevent future crashes, injuries, and fatalities more effectively.

The findings would help make informed choices about traffic safety initiatives and investments in transportation systems. This could result in fewer crash costs, better transportation safety and productivity for Chicago citizens and travelers, and enhanced quality of life for everybody in the city. Finally, this research enables stakeholders to make the transportation environment safer and more efficient in Chicago.

# RESEARCH QUESTIONS

1. What is the relationship between speed limits and crash types?

2.a. How effective are traffic signals in reducing crashes compared to other control devices?

b. How does a traffic control device affect the chances of rear-end collisions?

3. How do crash severity and contributory causes vary by time and day of the week?

4. Effects of Weather on Motor Vehicle Accidents: How do weather characteristics affect crash rates and severity?

# LITERATURE REVIEW

The unique opportunities of motor vehicles cause the general traffic growth trend worldwide. Evaluating highway safety is essential because more than 500 million cars and trucks are on the world's highways. WHO has estimated that annually, over 50 million people get injured in road accidents, while 1.3 million die. Traffic accidents are not only a threat to the lives of motorists and their passengers but also produce losses of tangible and intangible property and financial and social costs ([WHO](https://www.who.int/health-topics/road-safety#tab=tab_1)).

Worrying things have worsened, and every country has declared road safety the number one agenda item. Global partners such as the United Nations (UN), World Health Organization (WHO), and so on started the Decade of Action for Road Safety 2021-2030. As a result of this strategic planning, this global plan is part of the plan to ensure that fewer people are dying or experiencing permanent disability due to traffic incidents. As all these efforts are being made, traffic accidents claim thousands of lives and millions of injuries per year.

The effect of traffic loss owing to accidents is significant. This is not to mention the numerous losses of thousands of lives and suffering individuals undergo; accidents also result in hefty losses. Every year, traffic accidents are reported to have caused thirty-seven thousand and nine hundred deaths in addition to one million and fifty thousand injuries in the European Union only. These include direct costs: health, people not working, damaged cars and properties, and police services. The treatment and its hassle fall under tangible costs, while other expenses such as distress, grief, or suffering are considered intangible and contribute to the costs. Analyzing this effect shows how desperately road safety improvement must become one of the highest priorities.

**1. Trends in Traffic Crashes**

Eight thousand six hundred fifty are expected to be killed in Q1 2017, according to the National Highway Traffic Safety Administration ([NHTSA](https://www.nhtsa.gov/press-releases/2024-Q1-traffic-fatality-estimates)), which is 3.2% less than 8,935 deaths in 2023. Traffic deaths have been reducing for eight subsequent quarters since the second quarter of 2022. VMT for January-March was 0.6% higher than the preceding quarter, while the fatality rate slid from 1.18 to 1.13 fatalities per 100 million VMT.

Among the variables, 30 states plus Puerto Rico had fewer deaths, and 19 states plus the District of Columbia had the opposite result. These projections relate to the National Roadway Safety Strategy (NRSS), currently being executed by the United States Department of Transportation. The strategy will offer $1.7 billion through Safe Roads and Roads for All programs and the development of automated emergency brakes and other safety technologies.

The motor vehicle deaths in the United States have been increasing from 4,200 to 46,027 in the year 2022; the statistic has risen by 996 percent. However, due to changes in behavior on the road and enhanced safety in cars, the mortality rate has significantly reduced. The 1937 death rate decreased by 55%, with a rate of 13.8 deaths per 100,000 people in 2022. The fatality rate per 10,000 cars was reduced by 95% from 1913 at 33.38 to 2022 at 1.50 per car. The rises in the past few years, starting from 2019 up to 2022, further stress the need to enhance road safety and constant improvement. However, the previous years' numbers suggest a rising safety improvement ([NHTSA](https://www.nhtsa.gov/press-releases/2024-Q1-traffic-fatality-estimates))

**2. Contributing Factors to Traffic Crashes:**

Multiple factors lead to traffic accidents, but some of the most vital factors include cases of speed, inattention, distraction, and alcohol and drug influence. NHTSA predicts that 94% of highway crashes are due to human error, and 41% of these errors stem from recognition errors: drivers being distracted and not noticing hazards in their environment. Parallel changes were observed in Tennessee, which saw an increase in inattentive driving accidents from 12,934 in 2014 to 15,299 in 2017. ([OEB](https://www.wreckintoacheck.com/faqs/what-are-the-major-contributing-factors-to-traffic-accidents/))

They were drunk, which is a severe cause of the many accidents on the roads that result in fatalities. In 2015, there were 10,265 collision deaths involving alcohol, or one for every 51 minutes. Sixty-four percent of the fatalities involved drivers who had 0.08 percent and above. According to high school-going students are most vulnerable to occasions that involve impaired driving, seeing that 5.5% of them admitted to having driven after consuming alcohol.

Another major cause of accidents is distraction. This means that drivers who use their GPS, text, or eat, among other actions, while driving hazards are very high, will have an accident. In Tennessee, for instance, food-based distractions such as eating and talking caused crashes from 13,302 to 14,835 in 2015 and 2016. Another common cause is speeding, which results in 9,557 deaths in the year 2015 ([OEB](https://www.wreckintoacheck.com/faqs/what-are-the-major-contributing-factors-to-traffic-accidents/)). The same trends were recorded in Tennessee; the speed-related fatalities dropped from 220 in 2014 to 173 in 2016.

Some groups are worse affected—like teenagers, for instance: they are relatively new on the road and more likely to take reckless risks. Young drivers aged 16–17 who drive with other young people produce a twofold increased likelihood of fatal crashes, although this raises fourfold if three or more passengers are included. Male teenagers are also more likely to be involved in risky driving and speeding, and more so when pressure comes from peers.

Four are seat belts; teenage drivers use seatbelts the least out of all age groups; about 53% of teenage drivers killed in auto-related incidents in 2016 were not wearing seatbelts, let alone those who died on the weekend or evening ([OEB](https://www.wreckintoacheck.com/faqs/what-are-the-major-contributing-factors-to-traffic-accidents/)). As research found that 50% of fatal crash incidents that involve young people happen on country roads, it can be deduced that night or rural driving increases the accident risk for these drivers.

This is why it is critical to ensure strict compliance with road safety laws. Distracted driving, speeding, inattentive and impaired driving, and risky behavior, including those among teenagers, become prolific factors in a dangerous driving terrain.

# 3. Demographics and Traffic Crashes

The facts and fiction regarding men's or women's propensity for auto accidents provide some relevant statistics. More men get into an accident than women do; however, due to the aggressive way they drive, the accident they do have is fatal. Even though women drive less often than men (40% of the time compared to 60%), they are involved in 68.1% of all accidents, according to a survey of 6.5 million cases. There are, however, such statistical differences as the number of miles per year men drive—16550 instead of 10142 for women.Also, they are more likely to drink and drive, drive at high speeds, or not use seat belts; these factors contribute to the seriousness of their accidents. At some point in 2017, male drivers were involved in 37.477 fatal crashes, while female drivers were involved in 13.502 fatal crashes. Men are also more at risk in terms of accidents because they are proposed to have more traffic tickets and DUIs ([IEEE](https://ieeexplore.ieee.org/document/8954865/citations#citations)) Although women drive more safely than men, they are sometimes more at risk. Frontal and left-shoulder crashes are often the result of unequal participation and greater male predilection for driving large vehicles.

It is worth pointing out that the root causes of these gender differences are that men tend to drive longer distances and take more risks. At the same time, women's accidents are less severe and usually referred to as 'minor scrapes' or 'fender benders.' Nevertheless, driving exposes both sexes to severe perils; male Lord's risk-taking behavior was the leading cause of the collision, as was that of female drivers.

There are also differences by age in the crash rate. Teens are disproportionately involved in disproportionate types of accidents, particularly those aged 16 and 19 years. Though they constituted only 3.6% of the licensed drivers, they account for 6.1% of the fatal accidents and 9.1% of all the accidents. On the other hand, senior drivers (65-74), who constitute 13.4% of the licensed drivers, were involved in 7.1% of all collisions and 8.8% of fatal collisions ([IEEE](https://ieeexplore.ieee.org/document/8954865/citations#citations)). Information derived from such a collective rate indicates that although annual miles driven by elderly persons are comparable to those of young persons, older drivers are wealthy from experience and are safer; consequently, the aggregate crash rate per 100,000 licensed drivers gradually decreases with the driver's age.

Similarly to the overall rise in fatal crashes, there is a concrete upward trend in deadly crashes involving drivers 75+ years old. Younger drivers are those aged sixteen to twenty-four; their number of licensed drivers increases while the crash rate decreases. Young drivers between 16 and 19 record a fatal crash rate, after which the rate begins to decline with age 20 and above.

To sum up, young drivers have the most significant crash risks, although drivers aged 65-74 have the lowest crash risk. The risk of being killed in a crash is higher among drivers over 75. The rate of accidents continuously varies according to people's age and gender.

Bias in road safety research will likely significantly influence study results and conclusions. These include publication bias, which sees only positive results published; selection bias, which compromises the representativeness of a study; cultural bias, which impacts both the study's design and its analysis; and reporting bias, in which only partial results are reported. Such biases affect preschool intervention effectiveness, ignore crucial crash factors, confine extensibility, and misguide investigation strategies. To reduce such problems, researchers must extensively search the literature, use critical appraisal tools, declare limitations, and adopt graphical displays such as funnel plots to assess publication bias. This paper aims to identify and discuss some of these biases as follows: their critical importance in the improvement of validity and reliability, and therefore improve the methods used in road safety research; hence, improving the type of interventions and policies put in place to support the improvement of traffic safety across these various scenarios.

**4. Effectiveness of road safety interventions:**

The article Campbell Systematic Reviews published by John Wiley and Sons Ltd in 2024 on behalf of Campbell Collaboration provides information about road traffic injuries, which are one of the most common causes of death every year; 1.35 million people die because of road accidents, with most deaths occurring in low- and middle-income countries. Although most of the global literature focuses on LMICs, the effort of cities like Chicago, through implementing various interventions, improved infrastructure, traffic law enforcement, and public education campaigns, is critical to reducing road injuries. In Chicago, the effectiveness of these strategies is assessed to show how such interventions borrowed from global research can be adapted in various urban settings ([Malmam)](https://www.malmanlaw.com/malman-law-injury-blog/who-causes-more-car-accidents-men-or-women/). However, studies have shown that interventions applied in high-income countries may not work in low- and middle-income countries; therefore, reviewing how healthy measures work locally is essential. Further, while international research points out several gaps in road safety for LMICs, there may be similar gaps in research in Chicago. This project would thus serve as a chance to look at local data, assess interventions, and identify areas where changes could be made. The review relates the global findings to local research, connecting international research to Chicago's traffic safety effort. It also recognizes that solutions should be adapted according to context. ([Malman](https://www.malmanlaw.com/malman-law-injury-blog/who-causes-more-car-accidents-men-or-women/))

Chicago’s Vision Zero action plan aims to reduce road accidents by 2026, particularly severe crashes that result in severe injury or death. Continuous measures should be taken to achieve this. The first step towards road safety is to create an intervention policy even before the problem is recognized and to provide proper infrastructure and strict rules and regulations. The goal is to get zero road accidents by 2050, which will take a lot of funding and planning to get to that point. It is a challenging task, so dividing the intervention into phases makes it easier to establish.

**Fundamentals of Vision Zero**

In Chicago, everyone has the right to walk on roads, bike, use public transit, and drive on streets, no matter who they are or where they reside. Vision Zero provides value to the resources to enhance traffic safety in high-crash areas. High-crash areas around the city communities experience higher rates of accidents than Chicago's average. HCA is being developed for specific site plans, including this plan for the CBD ([Vision Zero](https://doi.org/10.1016/j.ssci.2017.11.005))

**4.1 Policy Interventions**

By using modern technology in the automotive industry, it is mandatory to use the vehicle in the city and should only be able to register. With the minimum basic required features like automatic emergency braking, cars do not move without seat belt fasteners, and line assist with adaptive lane change. This would solve unexpected road traffic accidents. For this to happen, the city has to paint the lanes. Using NFC technology before the speed limit reduces the vehicle's speed and allows pedestrians to pop notifications on their devices to maintain caution upon crossing. To avoid crashes in the accident-prone area. By increasing the ticket price, everyone will start to follow the rules. Even the funds from the tickets issued can be used for modern infrastructure for road safety, such as using an automated barrier gate during pedestrian crossings at the traffic lights and using flyovers to avoid traffic congestion. They are improving driving training and equitable traffic safety enforcement policies. The main problem we are facing today in Chicago is road traffic accidents. The intervention policies we can bring include building infrastructure like bikes, pedestrians, public transport, and private vehicle lanes with specific speed limits. If the rules are not followed, it will result in a considerable ticket to pay.

Effective post-crash care to avoid fatalities by having quick emergency access for the paramedic team and fire service and immediately making spontaneous decisions to divert traffic near the crash site Grade separation at an intersection by providing under or overpasses with on and off ramps to reduce traffic at the intersection. Roadside barrier systems for pedestrians and bike line safety are needed to avoid collisions when vehicles leave the roadway and cross into opposite lanes that are out of control due to winter weather conditions. Intervention for the management of roads according to weather conditions, especially during winter, is being prepared for snow conditions by sprinkling salt before snow for easy and fast clearance of the streets, which is safely available for people to use during snow conditions.

**4.2 Public awareness campaigns**

We are arranging awareness campaigns in schools and universities and especially for teenagers about road safety, explaining the effects by showing videos and teaching them things to avoid while on the road while walking, not to use headphones, and not to disturb the driver while driving and using billboards, and teaching road signs in class from schooling benefits of good road sense to avoid unwanted experiences. We should also do social media campaigns to use lines on the road and follow road traffic rules irrespective of the method of transport (pedestrian, bike, public transport, or private transport).

Rewarding them with souvenirs will attract and encourage everyone attending to remember things they have learned from the campaign. Arrange a campaign in collaboration with the DMV before issuing driver's licenses, making it mandatory to complete certain hours of safe driving with the instructor to be eligible to get the driver's license in the learning period.

# RESEARCH DESIGN AND METHODOLOGY

This research uses a quantitative method with an emphasis on descriptive research design to address the research questions concerning Chicago traffic crash patterns and the extent to which existing measures have averted the causes of the crashes. Data analysis considers the probable routes for crashes and assesses various facilities needed for safety with owners of hazardous properties for pedestrians and cyclists. By drawing descriptive statistics and inferential measures, the study explores factors such as speed limit, day or night, favorable or adverse weather conditions, and geographical regions.

The methodological approach includes several steps toward the goal. First, data cleaning and preprocessing will handle missing values, outliers, and data inconsistencies for data credibility. After that, visualizations, including box and whisker plots, histograms, and heat plots, are useful for analyzing the features and distribution patterns. Many results of correlation analysis point to the multicollinearity problem, which concerns the coincidence between factors that impact crash severity. For example, machine learning techniques including logistic regression or decision tree analysis can be used to predict the relationship between various factors like speed limits, weather conditions, safety measures, and the severity/probability of accidents.

The implications are expected to highlight levels that require more safety intervention efforts. Such findings can help direct city planning to allocate resources such as bikeways, better crossings, and better traffic signs to minimize accidents in the public domain. The analysis plan for this study is realistic and elaborate in design. A comprehensive set of statistical and machine-learning methods helps to find essential patterns in the extensive data set. As with most data analyses, each presented analysis has been chosen to address specific aspects of the research questions and to optimize information content from the data.

**1. Correlation Analysis:**

Approach: Pearson correlation coefficients for continuous variables; Chi-square tests of independence for categorical variables.

Justification: Correlation analysis reveals the strength and direction of relationships between variables. For instance, if one can investigate files searching for trends linking bad weather with the number of casualties, one can determine how lousy weather aggravates crash intensity. This step is essential in predicting for more complex analysis, plus it is a way of discovering other existing relationships that deserve a deeper look.

**2. Regression Analysis:**

Approach: Multiple linear regression for continuous outcomes; logistic regression for binary outcomes (e.g., injury vs. no injury).

Justification: Regression analysis measures independent variables affecting a dependent variable net of other influence factors. This method helps to determine which predictor is significant with others. For example, selecting the relationship between variables such as posted speed limits, total extent of injuries, and roadway conditions while accounting for other factors gives out meaningful info concerning the aspects of the highest significance in injury prevention.

**3. Predictive Modeling:**

Various predictive models will be implemented to capture complex relationships and provide insights into crash dynamics:

**a. Logistic Regression:**

Approach: A binary classification model predicting the probability of injury occurrence.

Justification: Suitable for modeling binary outcomes (injury/no injury), providing interpretable odds ratios for each predictor.

**b. Logistic Regression with SMOTE:**

Approach: Addresses class imbalance in the dataset.

Justification: Improves model performance when dealing with rare events (e.g., severe injuries), ensuring better prediction accuracy for less common but critical outcomes.

**e. Decision Tree:**

Approach: a tree-based model for classification and regression.

Justification: Justification captures non-linear relationships and interactions between variables, providing an intuitive representation of decision pathways leading to injuries.

**f. Decision Tree with SMOTE:**

Approach: Combines decision tree with SMOTE for balanced class representation.

Justification: Improves the model's ability to accurately classify common and rare injury outcomes.

**g. Random Forest:**

Approach: Ensemble of decision trees.

Justification: This technique aggregates multiple decision trees to improve prediction accuracy and reduce overfitting, providing robust predictions across crash scenarios.

**h. Random Forest with SMOTE:**

Approach: Integrates SMOTE with the Random Forest algorithm.

Justification: This enhancement enhances the model's performance on imbalanced datasets, ensuring accurate predictions across the spectrum of injury severities.

**I. Random Forest with Grid Search:**

Approach: Hyperparameter optimization for Random Forest.

Justification: Fine-tunes the model to achieve optimal performance, ensuring the most accurate

and reliable predictions possible.

These analysis approaches help to examine the data from different perspectives in multiple ways. By incorporating the ordinary statistical approach with the state-of-the-art machine learning technique, one can identify not just overarching but unique trends and patterns in crashes and their injuries. These multiple instruments increase the reliability of our conclusions and create a base for developing specific measures for traffic safety improvement.

**Heading**

**Traffic Control Devices:** Hypothesis: Poor traffic control device injury totals will be lower when properly controlled traffic flow.

Rationale: Traffic control devices work effectively through their design, and proper functioning can help avoid collisions as they provide guides on how traffic should be managed.

Hypothesis: Posted speed limits will be even higher in high-speed limit zones, increasing the severity of accidents.

Rationale: Higher velocities imply that the combined kinetic energy in case of an impact is higher, and hence, the resultant effects on the users are worse.

# DATA

**DATA SOURCING:**

The Chicago Police Department's electronic crash reporting system (E-Crash) is the primary data source for this research. This system was selected for several compelling reasons:

**Comprehensiveness:** E-Crash observes all reported traffic crashes within the City of Chicago starting from September 2017, and the high case volume makes the dataset rich and extensive.

**Reliability:** Although E-Crash is a police database for reporting crash statistics, data submitted are strictly checked for validity, thus increasing the credibility of the data collected.

**Richness of Information:** The system captures many significant parameters, including weather conditions, traffic signals and signs, and the degree of impact a crash occasioned. Thus, it is an essential source of information on different aspects affecting traffic safety.

**Consistency:** The standardized reporting format uniformly measures the quantity and quality of data from all recorded incidents and allows for more effective analysis.

**Accessibility:** The data is provided for research activities, making it easy to analyze with professionalism without violating the rights of the participants or the law.

**Temporal Relevance:** The dataset covers data starting from September 2017 that gives current-day traffic trends and safety measures.

The choice of this data source does not offend the study's objectives. Crash data is helpful because it is comprehensive, accurate, and detailed, allowing for an extensive description of factors leading to crashes and their consequences.

**DATA COLLECTION METHOD:**

The data collection process for E-Crash involves two primary methods:

**Self-Reports by Drivers:** Crash participants can report to police stations. This method captured some events that may not warrant the actual police attendance at the scene but which, when investigated, revealed some infringement of the law.

**On-Scene Reports by Police Officers:** In some crashes, police are called to the scene, and the officers write detailed reports based on what they witnessed and learned from the involved parties.

This dual approach to data collection was chosen for several reasons:

**Comprehensive Coverage:** It guarantees that, even in cases of minor accidents, data will be collected, giving a broader view of the crash situation.

**Reduction of Reporting Bias:** Using self-reports combined with officer reports minimizes any biases in reporting.

**Immediate and Retrospective Data:** On-scene reports record current information, and self-reports permit the incorporation of information that may occur after the incident.

**Enhanced Accuracy:** Where self-reports involve the emotion of the incident, officer reports give a more objective account of the crash scenes.

One of the method's main drawbacks is that other minor accidents that people did not report may occur. However, for this study, which aims to identify factors influencing injury severity, the general applicability of the model poses a slight bias that should not seriously affect this research.

# PARTICIPANT DESCRIPTION:

The database includes 881316 records, each referring to a singular crash incident and almost 48 variables concerning all aspects of a crash. The large sample size drawn from this study is essential to increase the statistical power necessary for the planned analysis.

The 48 variables capture a wide range of information: add all 48 columns names

Temporal data (date, time of crash)

Spatial data (location, type of roadway)

Environmental conditions (weather, lighting)

Vehicle characteristics (type, make, model)

Driver attributes (age, gender, condition)

Crash dynamics (kind of collision, point of impact)

Outcomes (injuries, fatalities, property damage)

These diverse measures provide rich sets of variables through which to investigate patterns of crash factors and consequences, making the study possible to answer a range of research questions related to traffic safety.

Although the study's primary analysis will be quantitative, the study design also permits using qualitative data due to the large sample size. If such data were obtained through interviews or questionnaires, responding individuals would be involved in the taped crash occurrences. This could include the driver, passengers, pedestrians, especially those who witnessed the incident, or any other form of person. The qualitative part shall help to refine the understanding of people's behaviors and actions contributing to crashes compared to the E-Crash data analysis based on statistical methods.

The sampling method for recruiting the participants in any of the qualitative parts would be stratified random sampling to increase the generality of the study across the various crash characteristics, crash severity, and driver demographics. This approach would help gather a broad spectrum of data and provide a more objectively comprehensive and detailed review of the situation.

# PROCEDURES

1. Data preprocessing: Null values, Missing values, outliers, value capping, duplicates

2. Feature Engineering: -Using data visualizations

3. Model building: regression, smote, decision tree, random forest, and other advanced models.

# POSSIBLE PROCEDURES

While the primary data source for this study exists in E-Crash and is used for analysis, the following steps would be followed if a survey or interview were to be conducted. These procedures ensure ethical compliance, data integrity, and participant protection:

**1. Ethical Approval:** This information collected should be presented to the Institutional Review Board (IRB) before the direct collection of the data. This will help ensure ethical treatments are correct when conducting human subject research.

**2. Participant Recruitment:** Potential participants will likely be enrolled from the E-Crash database and contacted by mail or phone, respecting their privacy. A sample would be employed that would cover many aspects, such as age, gender, and kind of crash.

**3. Informed Consent:** Every respondent would be expected to go through an informed consent process. This would involve:

A detailed explanation of the study's purpose, procedures, and potential risks and benefits.

Clarifying: Providing protection. The voluntary nature of participation and the right to withdraw at any time.

Explaining data confidentiality measures and how the information will be used and stored.

Obtaining written consent before proceeding with data collection.

**4. Data Collection Instruments:** Some questions from more standardized questionnaires would be ideal, or semi-structured interview guides would be designed to avoid unclear questions in their respective contexts.

**5. Interview/Survey Administration:**

For surveys, participants would be given the choice to complete the survey online using a secure website or by mailing paper-based questionnaires.

Depending on the participants and organizational context, interviews might be face-to-face or videoconference.

**6. Data Security:** All data collected would be stripped of identifiers and stored on password-protected encrypted servers and systems. Paperwork would only be stored in locked cabinets in secure areas.

**7. Follow-up Procedures:** The participants would be given the contact details of the research team if they wanted to contact the specific team after their participation.

**8. Debriefing:** The participants will receive a debriefing statement concerning the study's aim and objective and where to get information on traffic safety measures. These procedures ensure that any additional data collection is conducted ethically, systematically, and in a manner that protects participant rights and data integrity.

# RESEARCH MEASURES: DEPENDENT AND INDEPENDENT VARIABLES-

The study applies several variables to fully capture several potential causes of crash-related injuries. These variables have been carefully selected based on their theoretical relevance and practical significance in traffic safety research.

**Predictive Relationships:**

This study assumes several meaningful predictive relationships of the independent variables with the dependent variable (Injuries Total). These hypotheses are grounded in existing traffic safety literature and logical deductions:

Weather Conditions:

Hypothesis: Unfavorable climate conditions, including rain, snow, or fog, will likely be followed by high accident statistics.

Rationale: Rain reduces visibility and thus forces or limits control of the vehicles involved, resulting in more severe accidents.

**Dependent Variable:**

**Injuries Total (INJURIES\_TOTAL):** This continuous variable measures the number of injuries adopted regarding the crash incident. It is the most helpful measure chosen, as it quantifies the extent of the crash. This investigation selected this variable as the dependent measure because it can provide insights into public health or safety and policy decisions to decrease crash-related injuries.

**Independent Variables:**

**1. Weather Condition (WEATHER\_CONDITION):**

Type: Categorical

Categories: Clear, Rain, Snow, Fog, etc.

Justification: Road and weather conditions refer to driving conditions with poor visibility due to bad weather. Bad weather is believed to positively affect injuries because of loss of control and poor visibility. This variable allows us to determine how much varying meteorological conditions affect the crash consequences.

**2. Traffic Control Device (TRAFFIC\_CONTROL\_DEVICE)**

Type: Categorical

Categories: Traffic Signal, Stop Sign, Yield Sign, No Controls, etc.

Justification: Both the availability and kind of traffic control are critical since they determine the constructive patterns of traffic circulation and avoidance of crashes. This variable allows for evaluating the relationship between traffic management measures and the incidence of injuries and could be helpful in urban design and the design of traffic systems.

**3. Posted Speed Limit (POSTED\_SPEED\_LIMIT):**

Type: Continuous

Unit: Miles per hour

Justification: Specifically, the speed at which a car travels is essential in the level of crash occurrence. High speeds, therefore, are expected to go hand in hand with higher degrees of injury because of the relatively higher energy impact from the accident. This variable becomes valuable by quantifying the connection between allowed speed levels and the experience of injury outcomes.

**4. First Crash Type (FIRST\_CRASH\_TYPE):**

Type: Categorical

Categories: rear-end, sideswipe, head-on.

Justification: The kind of contact between objects involved in a particular car crash can determine the kinds of injuries that occur and the extent of the same. Various crash types imply different force distributions, locations of the impacting surfaces, and expected dissimilarities in injury patterns. It permits examination of the connection between selected aspects of crash characteristics and the nature of the outcomes sustained by the occupants.

**5. Roadway Surface Condition (ROADWAY\_SURFACE\_COND):**

Type: Categorical

Categories: Dry, Wet, Ice, Snow, etc.

Justification: The road surface imprinted on or sensed by the tires plays a vital role in determining the friction between the tires in contact with the road surface and the vehicles. Sources of risk are likely to worsen and enhance the possibility of losing control of the car, causing a worse crash. This enabled the researchers to establish the extent to which various degrees of road conditions trigger the likelihood of an accident.

These variables were chosen to give an adequate picture of various aspects that shape crash consequences. They are weather-related, road characteristics, traffic signals, speed strips, and other factors relating to the accident. This provides a broad approach to understanding multiple factors that interact with crash injuries.

These specific variables have been selected for the following reasons: first, all these variables are relevant based on the traffic safety literature; second, these variables can be used to design practical interventions. For example, knowledge of the degree and nature of the association between speed limits and injury severity can be applied directly in developing speed management policies. Understanding the impact of different traffic control devices can be utilized to decide on the appropriate infrastructure investment.

**DATA CLEANING AND DATA PREPROCESSING :**

The initial step of data cleaning involved analyzing and addressing the missing values. During the analysis a significant number of missing values were identified. The total count of missing values across all columns is 9,048,938. And the highest missing values are found in the column WORKERS\_PRESENT\_I which is 891688.

The missing values in each column :

CRASH\_RECORD\_ID 0

CRASH\_DATE\_EST\_I 827002

CRASH\_DATE 0

POSTED\_SPEED\_LIMIT 0

TRAFFIC\_CONTROL\_DEVICE 0

DEVICE\_CONDITION 0

WEATHER\_CONDITION 0

LIGHTING\_CONDITION 0

FIRST\_CRASH\_TYPE 0

TRAFFICWAY\_TYPE 0

LANE\_CNT 693957

ALIGNMENT 0

ROADWAY\_SURFACE\_COND 0

ROAD\_DEFECT 0

REPORT\_TYPE 27660

CRASH\_TYPE 0

INTERSECTION\_RELATED\_I 687989

NOT\_RIGHT\_OF\_WAY\_I 852282

HIT\_AND\_RUN\_I 612878

DAMAGE 0

DATE\_POLICE\_NOTIFIED 0

PRIM\_CONTRIBUTORY\_CAUSE 0

SEC\_CONTRIBUTORY\_CAUSE 0

STREET\_NO 0

STREET\_DIRECTION 4

STREET\_NAME 1

BEAT\_OF\_OCCURRENCE 5

PHOTOS\_TAKEN\_I 880793

STATEMENTS\_TAKEN\_I 872438

DOORING\_I 890144

WORK\_ZONE\_I 887973

WORK\_ZONE\_TYPE 889110

WORKERS\_PRESENT\_I 891688

NUM\_UNITS 0

MOST\_SEVERE\_INJURY 1980

INJURIES\_TOTAL 1966

INJURIES\_FATAL 1966

INJURIES\_INCAPACITATING 1966

INJURIES\_NON\_INCAPACITATING 1966

INJURIES\_REPORTED\_NOT\_EVIDENT 1966

INJURIES\_NO\_INDICATION 1966

INJURIES\_UNKNOWN 1966

CRASH\_HOUR 0

CRASH\_DAY\_OF\_WEEK 0

CRASH\_MONTH 0

LATITUDE 6424

LONGITUDE 6424

LOCATION 6424

And the total null values found in the data are 9048938.

We have renamed the CRASH\_TYPE column to SEVERE to make it clearer that this column indicates the severity of a crash. Then, we mapped the values in the SEVERE column to a binary format: 1 for severe crashes (INJURY AND / OR TOW DUE TO CRASH) and 0 for all other crash types. This mapping simplifies the data for analysis and modeling.

We converted the CRASH\_DATE column to a datetime format using pd.to\_datetime(). This ensures that the date values are properly recognized and can be used for time-based analysis, such as filtering or calculating time intervals.

And then we have extracted time-based features from the CRASH\_DATE column. We created three new columns:

* **HOUR**: Extracted the hour of the crash using .dt.hour.
* **DAY\_OF\_WEEK**: Extracted the day of the week using .dt.dayofweek (where Monday is 0 and Sunday is 6).
* **MONTH**: Extracted the month of the crash using .dt.month.

These features provide valuable insights into crash patterns based on time, which can be useful for analysis and modeling.

Continuing with data transformation, we converted categorical values in specific columns to binary format using lambda functions:

* **INTERSECTION\_RELATED\_I**: Mapped 'Y' to 1 (indicating the crash is intersection-related) and all other values to 0.
* **NOT\_RIGHT\_OF\_WAY\_I**: Mapped 'Y' to 1 (indicating a right-of-way violation) and all other values to 0.
* **HIT\_AND\_RUN\_I**: Mapped 'Y' to 1 (indicating a hit-and-run) and all other values to 0.

These changes simplify the data, making these columns more interpretable and suitable for analysis or modeling tasks.

Next, we handled missing values and converted the WORK\_ZONE\_I column to a binary format.

* First, we replaced all null values in the column with 0 to ensure consistency and avoid errors in processing.
* Then, we mapped the values using a lambda function: 'Y' was converted to 1 (indicating the crash occurred in a work zone), and all other values were converted to 0.

This transformation standardizes the data, making the WORK\_ZONE\_I column easier to interpret and use in analysis or modeling.

Continuing the data cleaning process, we addressed null values in key columns related to injuries and lane counts:

* The columns **INJURIES\_TOTAL**, **INJURIES\_FATAL**, **INJURIES\_INCAPACITATING**, **INJURIES\_NON\_INCAPACITATING**, **INJURIES\_REPORTED\_NOT\_EVIDENT**, **INJURIES\_NO\_INDICATION**, and **LANE\_CNT** had missing values that were replaced with 0.

Replacing null values with 0 indicates no reported injuries or unrecorded lane counts, depending on the context of each column.

Continuing with the handling of missing data, we addressed null values in the **MOST\_SEVERE\_INJURY** column:

* Missing values in this column were replaced with the label 'Unknown'.

The label 'Unknown' indicates cases where the severity of the injury was not recorded, allowing for clearer interpretation and analysis without losing the integrity of the dataset.

To refine the dataset and focus on relevant information, we created a list of columns to drop. These columns were either redundant, had too many missing values, or were not critical for analysis.the columns we dropped are given below :

'CRASH\_DATE\_EST\_I','RD\_NO','REPORT\_TYPE','DATE\_POLICE\_NOTIFIED', 'STREET\_DIRECTION','BEAT\_OF\_OCCURRENCE','PHOTOS\_TAKEN\_I','STATEMENTS\_TAKEN\_I','DOORING\_I','WORK\_ZONE\_TYPE','WORKERS\_PRESENT\_I','LATITUDE', 'LONGITUDE','CRASH\_RECORD\_ID','INJURIES\_UNKNOWN','STREET\_NO', 'MOST\_SEVERE\_INJURY','SEC\_CONTRIBUTORY\_CAUSE','LOCATION','STREET\_NAME'

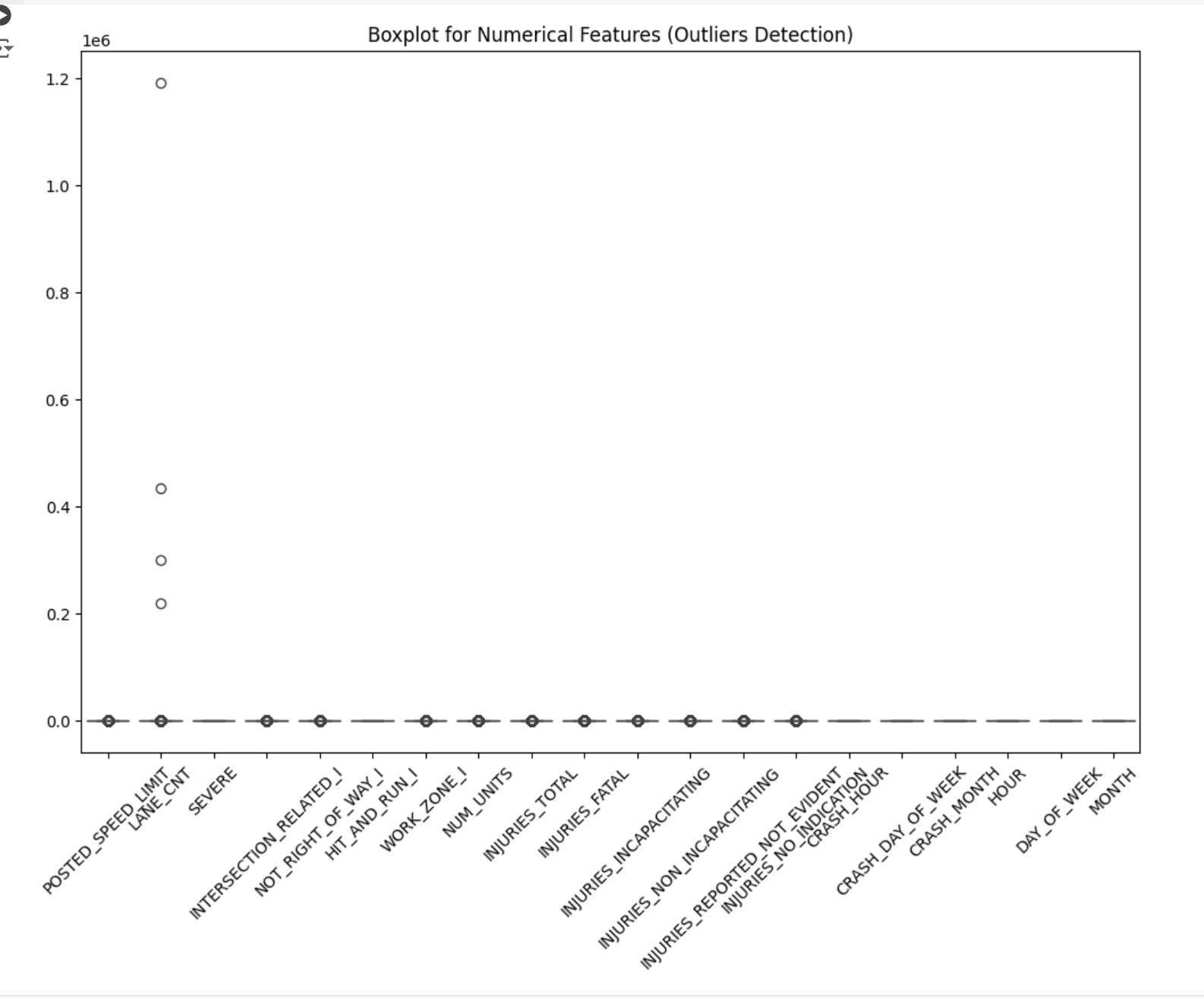
To finalize the column removal process, we ensured that only columns present in the dataset were dropped:

* we filtered the predefined drop\_list to include only columns that exist in the DataFrame using a list comprehension.
* Then, removed these columns from the DataFrame with the drop() method, ensuring no errors occurred if certain columns were already missing from the dataset.

To identify potential outliers in the dataset, we created a boxplot for the numerical features:

* First, we selected all numerical columns in the dataset using select\_dtypes(include=[np.number]).
* Using the sns.boxplot() function, we plotted a combined boxplot for these features to visualize their distributions and detect outliers.
* The plot was styled for clarity with a title, axis adjustments, and rotated x-axis labels for better readability.

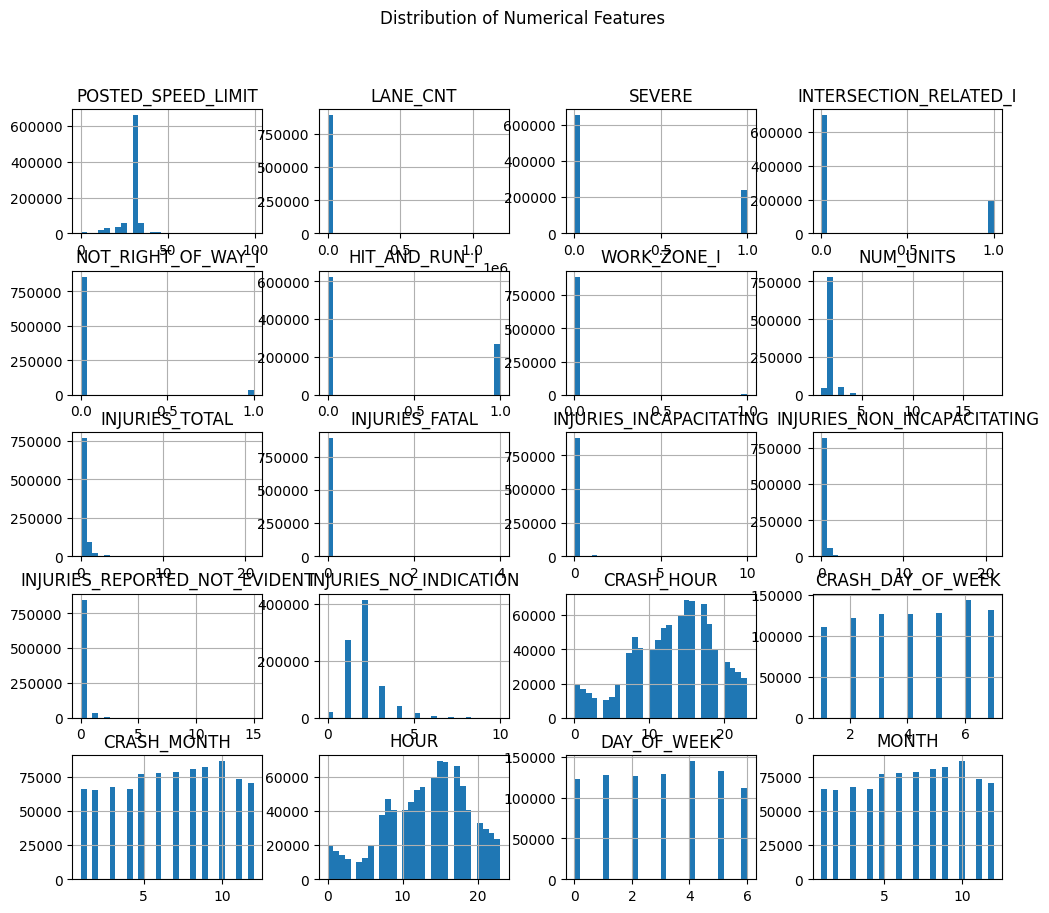
Significant outliers are evident in features like "POSTED\_SPEED\_LIMIT" and "LANE\_CNT," indicating unusually high values. Most other features show minimal variation with flat distributions near the x-axis, suggesting low variability or distant outliers. The presence of these extreme values implies potential data entry errors or rare events, which could impact the analysis if not addressed.



Continuing with exploratory data analysis, we performed univariate analysis to examine the distribution of numerical features:

* We plotted histograms for all numerical columns using hist(). This provides a clear visualization of the frequency distribution of each variable.
* The figure size was set to (12, 10) for better readability, and we used 30 bins to capture the detailed spread of values.
* A title, *'Distribution of Numerical Features'*, was added to summarize the purpose of the visualization.

Most variables, such as "SEVERE" and "LANE\_CNT," are heavily skewed, with the majority of data clustered at the lower end, suggesting low frequencies of extreme values. Features like "HOUR" and "MONTH" have more evenly distributed data, reflecting expected periodic trends in crashes across hours of the day and months of the year. Features such as "POSTED\_SPEED\_LIMIT" show a pronounced peak, indicating frequent occurrences within a narrow range. This univariate analysis reveals important data characteristics for further exploration.

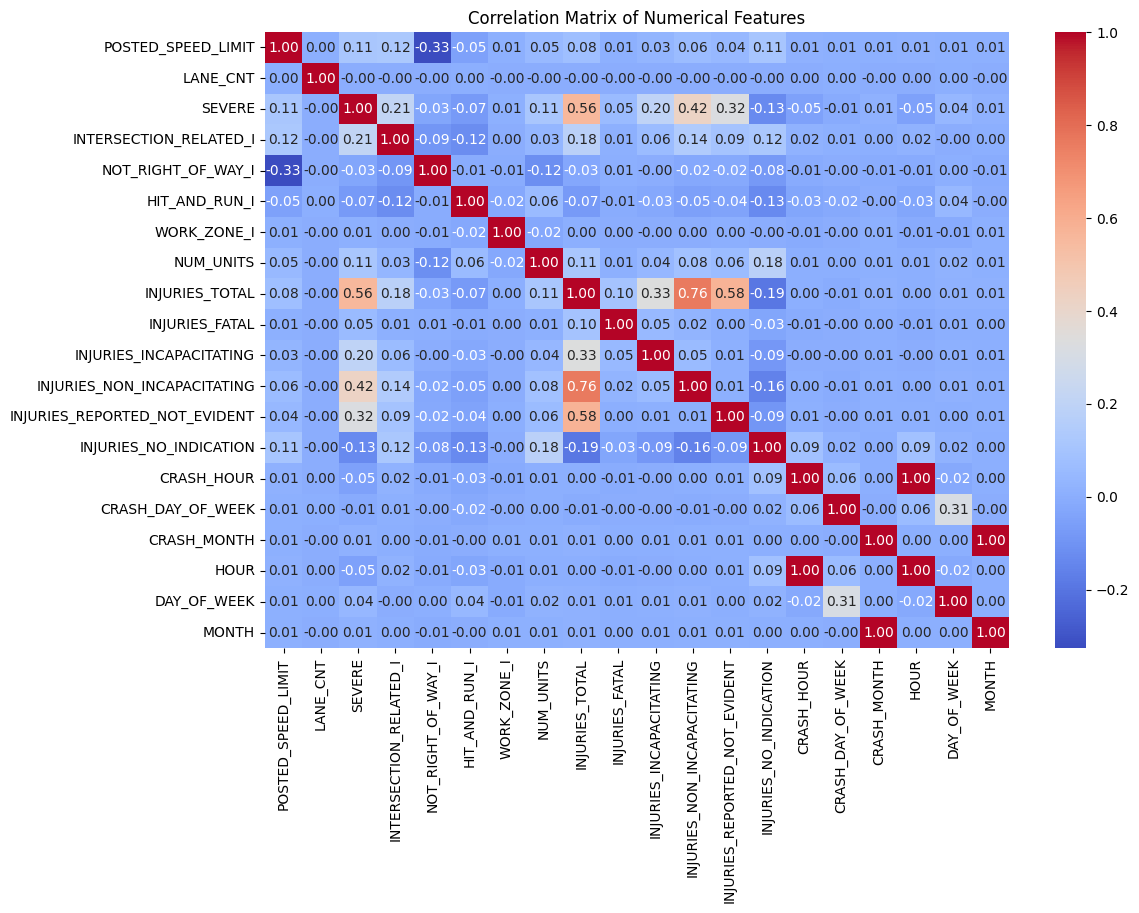


To further explore relationships between numerical features, we conducted bivariate analysis by creating a correlation matrix:

* We calculated the correlation coefficients for all numerical columns in the dataset to measure the strength and direction of their relationships.
* A heatmap was generated using sns.heatmap() to visually represent these correlations. The color scheme, *'coolwarm'*, highlights positive and negative correlations, while annotations provide the exact correlation values.
* The plot was titled *'Correlation Matrix of Numerical Features'* for clarity, and the figure size was adjusted to ensure readability.

Strong positive correlations are seen between "INJURIES\_TOTAL" and "INJURIES\_NON\_INCAPACITATING" (0.76), indicating that non-incapacitating injuries are a major component of total injuries.

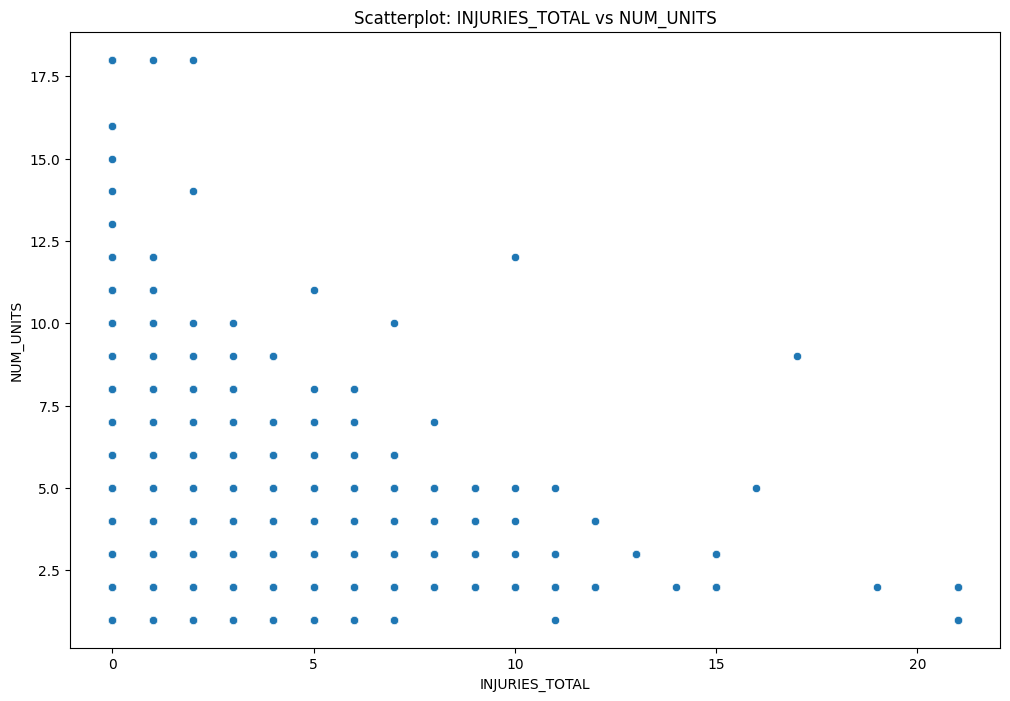
Similarly, "NUM\_UNITS" has moderate positive correlations with "INJURIES\_TOTAL" (0.33), suggesting crashes involving more vehicles tend to result in more injuries. Notably, most correlations are weak, with values close to zero, suggesting limited linear relationships among many features. These findings highlight the complex and mostly non-linear interdependencies between different crash characteristics, informing further modeling considerations.



To further explore relationships between specific numerical features, we created a scatter plot:

* This plot shows the relationship between **INJURIES\_TOTAL** (total injuries) and **NUM\_UNITS** (number of units involved in a crash).
* Using sns.scatterplot(), we plotted INJURIES\_TOTAL on the x-axis and NUM\_UNITS on the y-axis.
* The chart includes labels for both axes and a title, *'Scatterplot: INJURIES\_TOTAL vs NUM\_UNITS'*, to clearly communicate the relationship being visualized.

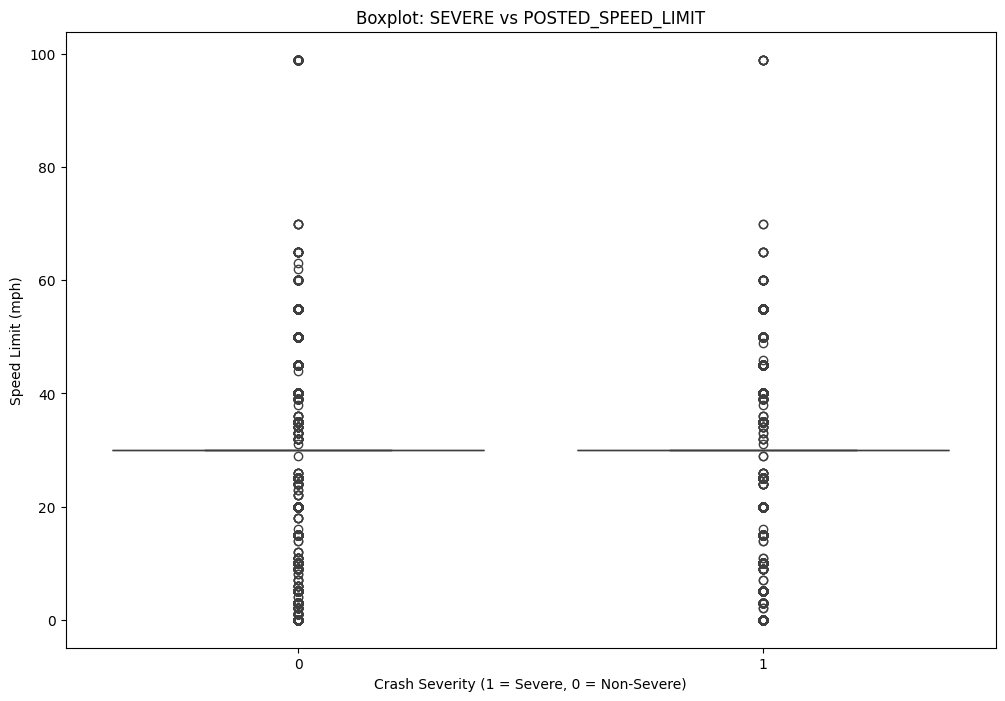
The plot shows a positive trend where an increase in the number of units involved (i.e., vehicles) generally corresponds to higher total injuries, although the relationship is not strictly linear. The data points are scattered with a higher density of incidents involving fewer units and injuries. Outliers can be observed where the number of units is high but the total injuries remain relatively low, suggesting that not all multi-vehicle crashes result in significant injuries.



To analyze the relationship between crash severity and speed limits, we created a boxplot:

* The boxplot shows how **POSTED\_SPEED\_LIMIT** varies across the two categories of **SEVERE**: 1 (severe crashes) and 0 (non-severe crashes)
* Using sns.boxplot(), we plotted SEVERE on the x-axis and POSTED\_SPEED\_LIMIT on the y-axis.
* The chart includes a title, *'Boxplot: SEVERE vs POSTED\_SPEED\_LIMIT'*, and labeled axes for clarity.

The plot compares speed limits across severe and non-severe crashes. Both categories exhibit similar median speed limits, with most crashes occurring in lower speed zones (around 25–35 mph). However, numerous outliers are present, especially at higher speed limits (above 40 mph), which are more common in both severe and non-severe crashes. This suggests that while higher speeds are associated with severe crashes, non-severe crashes also occur at high speed limits, indicating multiple contributing factors beyond speed alone.



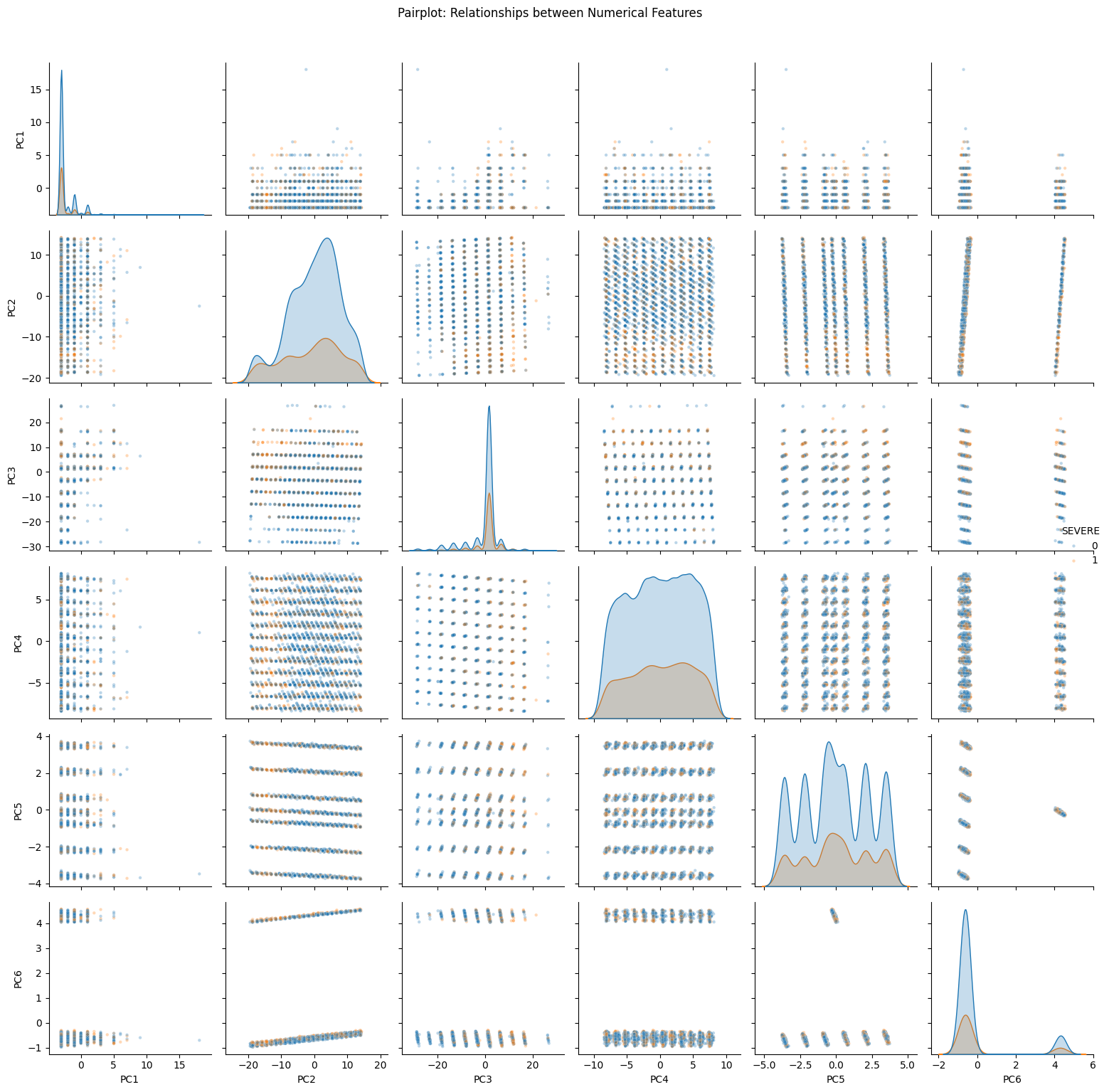
To explore relationships among multiple numerical features and crash severity, we performed multivariate analysis using pairplots and dimensionality reduction:

When the number of numerical features exceeded six, we applied **Principal Component Analysis (PCA)** to reduce dimensionality to six principal components. This step simplifies high-dimensional data while retaining most of its variability.The resulting principal components were stored in a new DataFrame, and the **SEVERE** column was added to link crash severity.

If the dataset had more than 10,000 rows, we randomly sampled 10,000 entries to optimize the performance of pairplot visualization while maintaining representativeness.

we generated a **pairplot** using sns.pairplot() to explore relationships between numerical features and crash severity (SEVERE), where different hues distinguished crash severity levels.The pairplot also included optimized settings, such as reduced marker size and transparency (alpha), to handle potential overlap in dense data.

The scatterplots reveal various patterns among the principal components, highlighting differences between severe and non-severe crashes. While most pairwise relationships do not show clear separation, some distinct trends can be seen, with clusters appearing based on crash severity in certain components (e.g., PC1 and PC4). The diagonal histograms further indicate the density distributions, showing different distributions for severe and non-severe crashes in several components.



Please place this viz in correct place

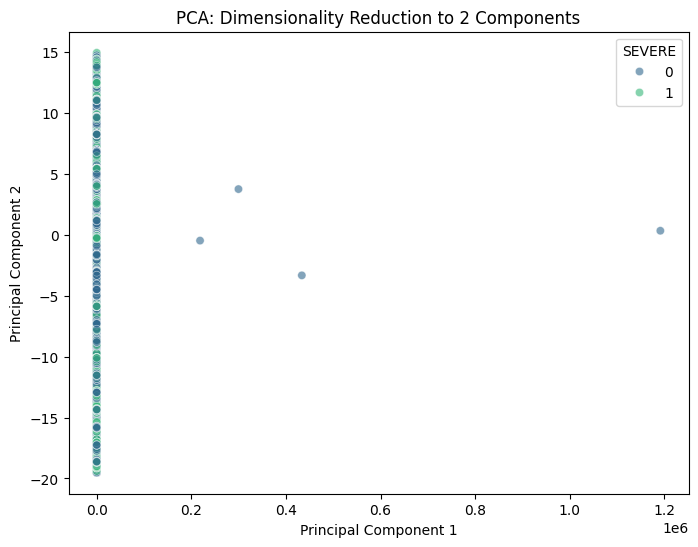
To reduce the complexity of the dataset and visualize high-dimensional relationships, we applied Principal Component Analysis (PCA):

we reduced the dataset's numerical features to two principal components using **PCA**, capturing the most significant variance in the data

A new DataFrame was created to store the two principal components, named **PC1** and **PC2**.The **SEVERE** column (indicating crash severity) was added to this DataFrame for visualization purposes.

Using a scatter plot, we plotted **PC1** on the x-axis and **PC2** on the y-axis, with crash severity (SEVERE) represented as different hues.The plot uses a **'viridis'** color palette for better distinction, and transparency (alpha=0.6) was applied to handle overlapping points.

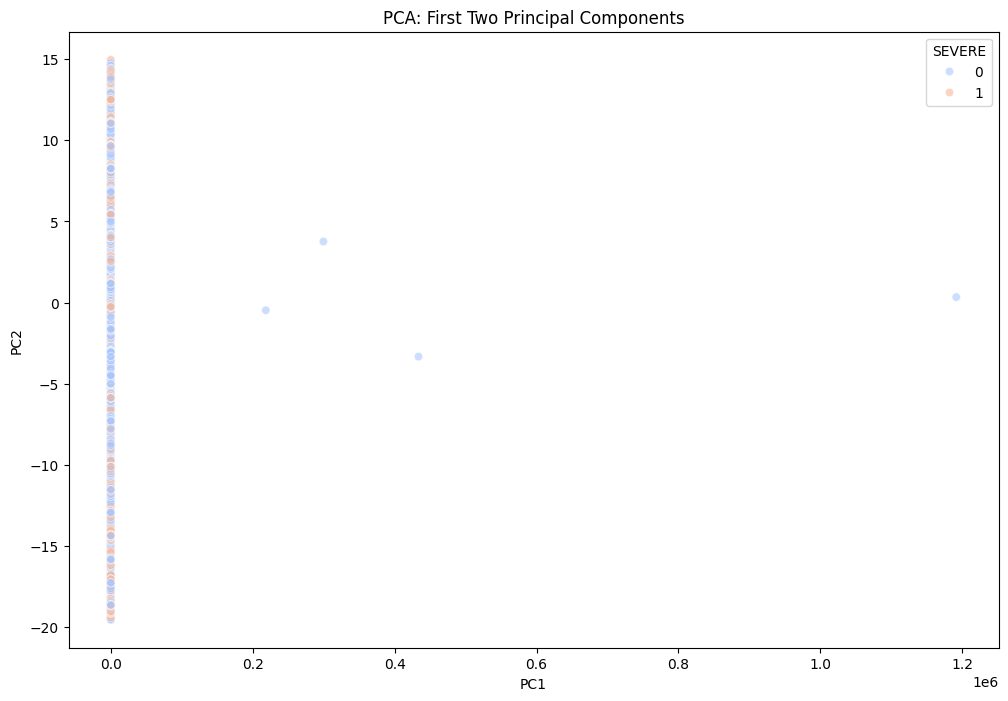
The plot differentiates between severe (1) and non-severe (0) crashes. Most data points are clustered along the vertical axis (PC2), indicating that the variance captured by PC1 is minimal for the majority of observations. There are a few outliers where PC1 captures more variance. This suggests that the first principal component may not add substantial discriminatory information, while PC2 captures more spread among the observations.



To visualize the results of the PCA and understand the relationship between the two principal components and crash severity, we created a scatter plot:

The **first principal component (PC1)** was plotted on the x-axis, and the **second principal component (PC2)** was plotted on the y-axis.  
Crash severity (**SEVERE**) was represented as a categorical hue in the scatterplot, with distinct colors from the **'coolwarm'** palette distinguishing between severe and non-severe crashes.Transparency (alpha=0.6) was applied to reduce clutter and handle overlapping points.  
The plot includes clear axis labels (PC1 and PC2) and a title, *'PCA: First Two Principal Components'*.

Most of the data points are densely clustered along the PC2 axis, suggesting minimal variance captured by PC1 for the majority of the dataset. This clustering indicates that the separation between severe and non-severe crashes is not distinct in the first two principal components, as severe (1) and non-severe (0) crashes appear intermixed.

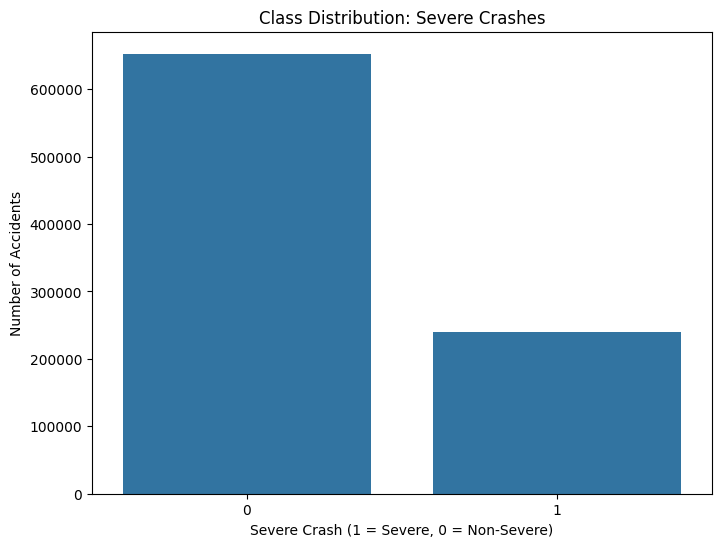


To evaluate the class distribution in the target variable, **SEVERE** (indicating whether a crash is severe), we checked for any class imbalance:

We used **value\_counts()** to check the number of occurrences of each class (severe vs. non-severe crashes) in the target variable. This helps identify whether the dataset is imbalanced.The results were printed to assess the distribution of the classes.

We created a **countplot** using sns.countplot() to visualize the distribution of severe and non-severe crashes. The x-axis represents the **SEVERE** variable (1 for severe, 0 for non-severe), and the y-axis shows the number of accidents in each category. And we calculated and printed **summary statistics** for all numerical columns using describe(), providing key insights into their central tendencies (mean, median), spread (standard deviation, min, max), and distribution (quartiles).

The bar plot highlights a notable class imbalance, where non-severe crashes (class 0) significantly outnumber severe crashes (class 1). The count of non-severe crashes is approximately three times that of severe crashes. Such class imbalance can introduce bias in machine learning models, leading to poor performance in predicting severe crash instances. Addressing this imbalance through techniques like oversampling, undersampling, or using balanced algorithms may be necessary to ensure an effective classification model.

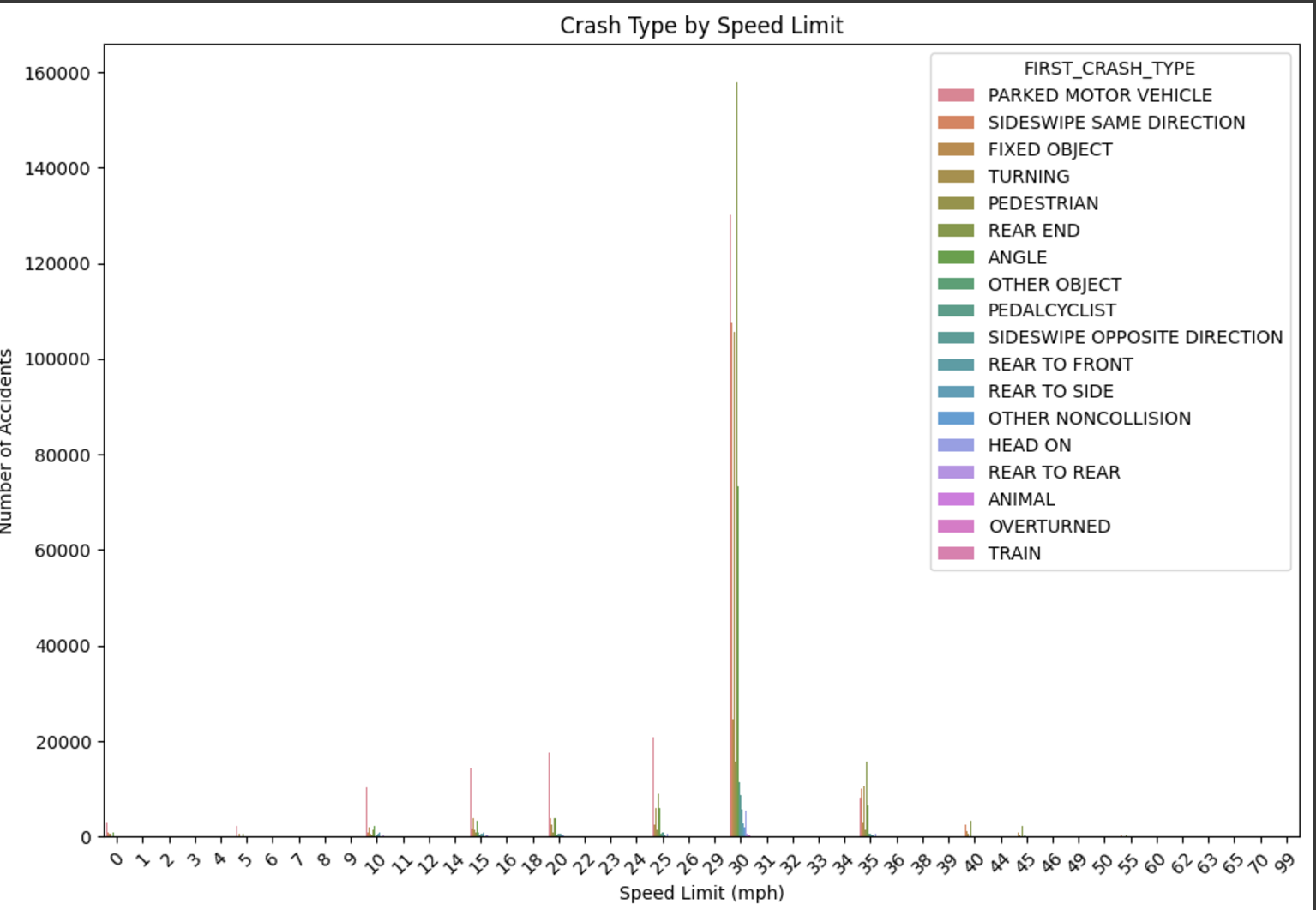


**DATA ANALYSIS**

This data set is analyzed based on examining the factors contributing to their occurrence and evaluating the effectiveness of current safety measures implemented by local authorities. Among the project objectives, the following can be listed: the determination of the basic cause of traffic accidents; evaluation of the availability of existing safety equipment; and the identification of

problem areas with increased rates of traffic accidents.

**1. What is the relationship between speed limits and crash types?**



The visualization above illustrates the relationship between speed limits (in mph) and the frequency of various crash types. The speed limit is represented on the X-axis and the number of accidents on the Y-axis for various crash types.

Relationship between speed limit and crash types:

Urban speed limit: (20-40 mph)

These speed limits see the highest number of crashes, specifically involving parked vehicles, rear end collisions, and turning accidents.

Highway speed limit (50+mph):

Crashes are less frequent but may involve more severe types, such as overturned vehicles or head-on collisions.

Low speed limits(0-10 mph):

Crashes are rare.

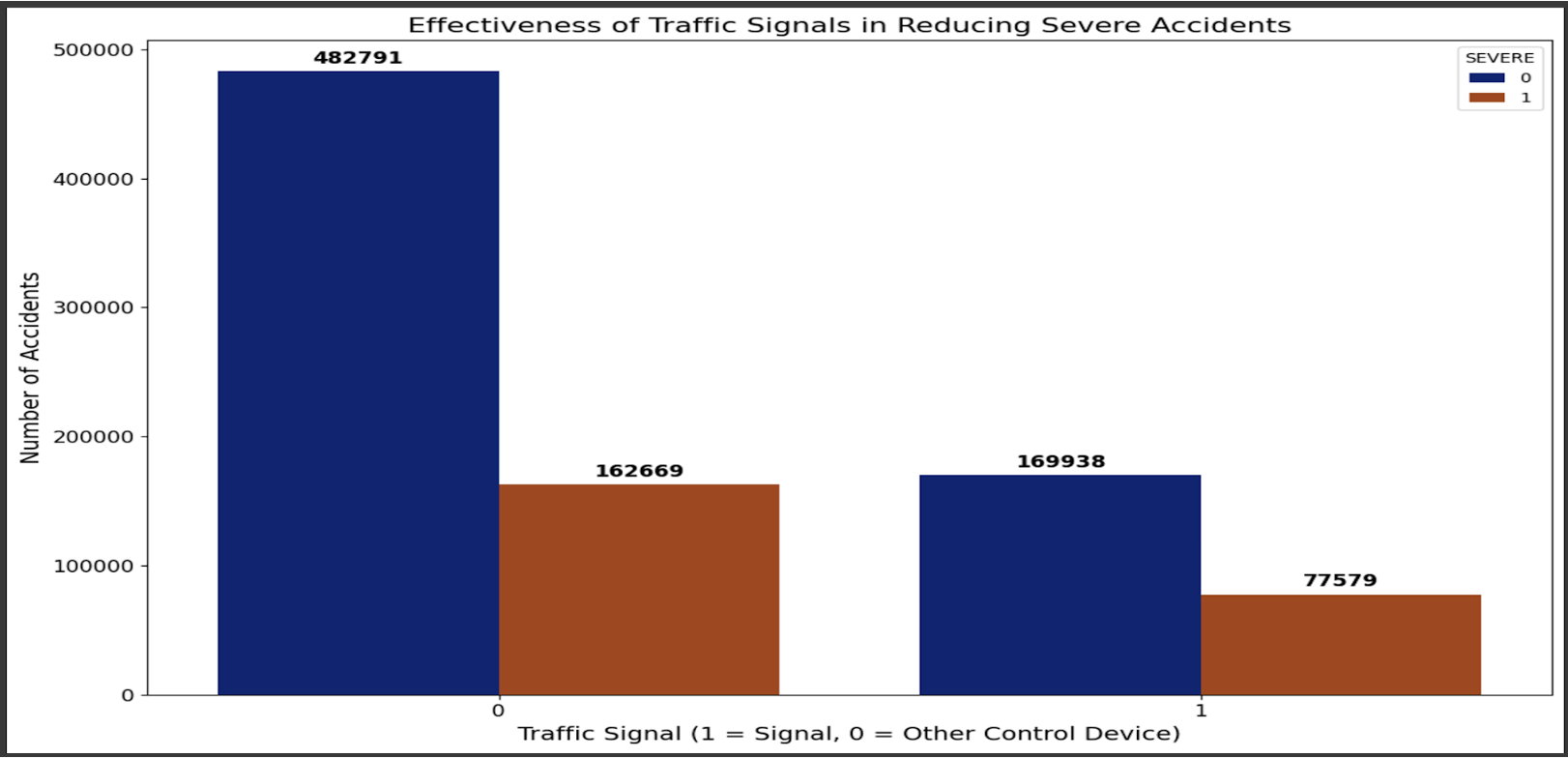
The visualisation depicts that the crashes are more frequent in moderate speed zones, likely due to higher traffic density. The most traffic accidents happen at 30 mph, which is a typical speed limit in cities like Chicago. This implies that the crashes occur more frequently in moderate speed zones, most likely as a result of increased traffic density, frequent pauses, vehicle-to-vehicle contacts, and the presence of cyclists and pedestrians. At this speed, the most common crash types are “Rear End” parked motor vehicles, and turning events, which represent common driving difficulties in metropolitan areas such abrupt braking, close following distances and lane changes. Although there are fewer collisions at lower speed limit (0-20mph), more vulnerable road users, such cyclists or pedestrians, may be involved in residential or school zones .

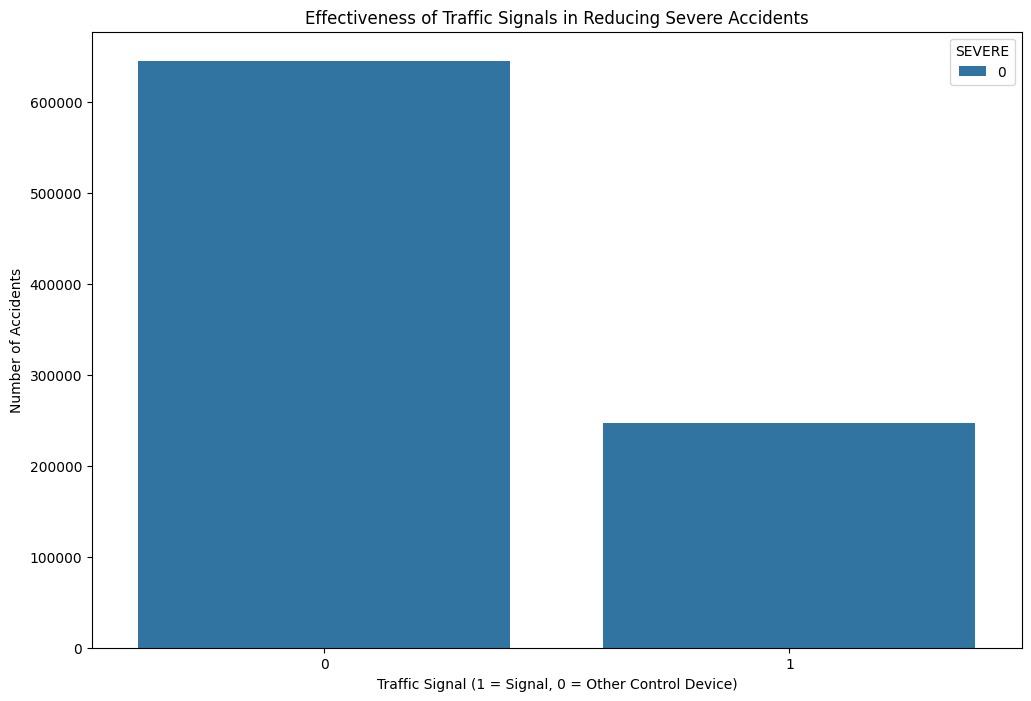
Conversely crash rates decreased dramatically at higher speed limits(35 mph and beyond), yet collisions at these speeds can be more serious. Rare occurrences like train or “overturned” collisions are infrequent and not closely related to speed limitations, although less common crash types including “sideswipe same direction” and “fixed object” are spread throughout different speed limits. The need for focused safety measures in these areas, like better signage, upgraded pedestrian infrastructure, and traffic- calming measures, is highlighted by the step increase in collisions at 30 mph. In order to reduce hazards and enhance general traffic safety, these findings highlight the significance of concentrating safety efforts on regions with high accident frequency.

**2.a. How effective are traffic signals in reducing crashes compared to other control devices?**

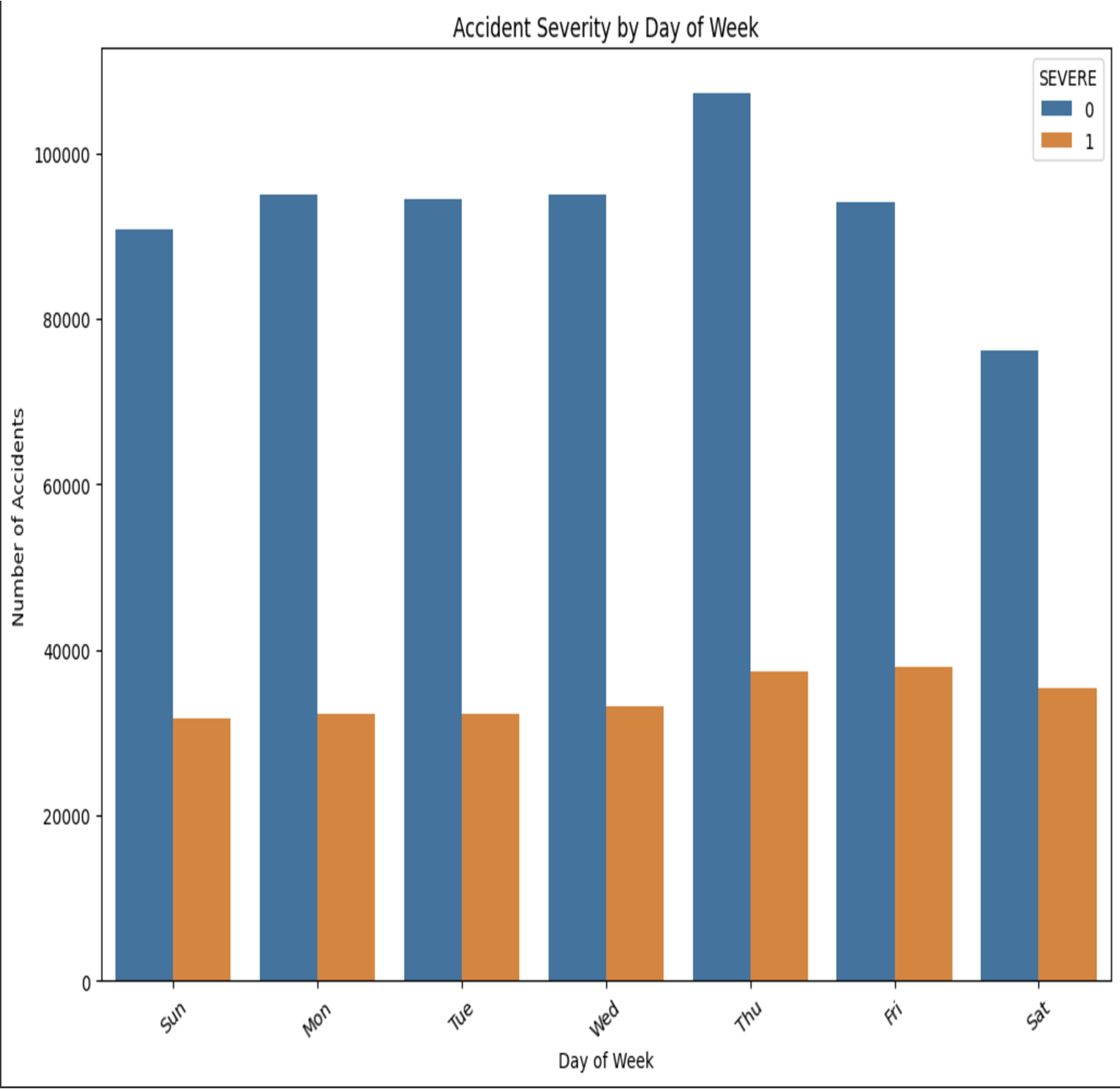
Traffic signals (coded as 1 and other control devices (coded as 0) are compared in the visualization for their ability to reduce crashes, especially serious ones. According to the graphic, there are significantly more accidents overall in regions without traffic signals (control devices other than traffic lights), with 482,791 non-severe wrecks and 162,669 serious crashes. With 169,938 non-severe crashes and 77,579 serious crashes, intersections or zones with traffic lights, on the other hand, have a significantly lower accident frequency. This suggest that both the total number of crashes and the percentage of serious crashes can be decreased with the help of the traffic signals. In particular, the presence of a signal significantly lowers the number of serious collisions because fewer than half of all catastrophic crashes occur when traffic lights are in place managed by other control mechanisms. The visualizations emphasize the critical role traffic signals play in managing traffic flow and enhancing safety, particularly in high-risk or busy locations. It implies that increasing the usage of traffic signals in regions where serious collisions occur could enhance safety results even more.

**b. How does a traffic control device affect the chances of rear-end collisions?**



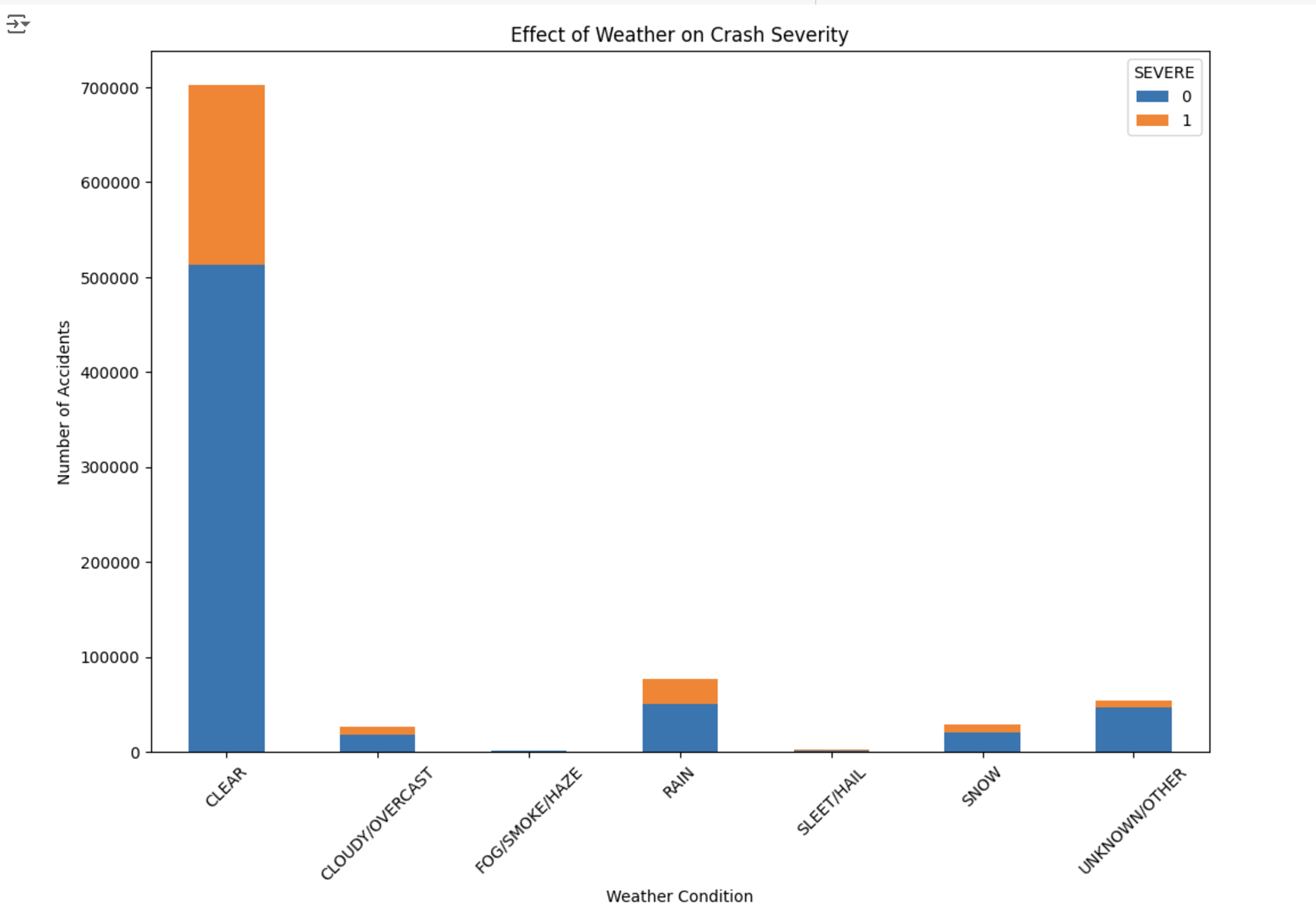


**3. How do crash severity and contributory causes vary by time and day of the week?**



The aforementioned graphic, which separates non-severe (blue bars) from severe (orange bars) incidents, looks at the correlation between collision severity and the day of the week. According to the figure, the number of crashes, regardless of their severity, is pretty constant throughout the week, with Thursday having the most accidents overall. Orange bars indicate severe crashes which are marginally less common than non-severe crashes but equally consistent. Notably the overall number of accidents is slightly lower on weekends-sundays and saturday in particular than on any other weekdays; yet the percentage of severe accidents seems to be slightly greater on these days. This may indicate more dangerous driving practices or altered weekend traffic patterns, including faster or intoxicated driving, which could lead to more severe outcomes.

Regular commuting and higher traffic density are probably the reasons for the steady number of crashes that occur over the week, although contributing factors like road conditions, driver behaviour, and the time of day may have an impact on the severity of the crashes. All things considered, the graph emphasises the necessity of focused safety measures, especially on Thursdays and weekends when traffic patterns and behaviour appear to have a distinct impact on the frequency and severity of the crashes.



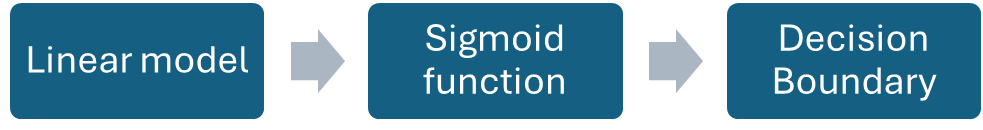
**4)Effects of Weather on Motor Vehicle Accidents: How do weather characteristics affect crash rates and severity?**

The above visualisation is divided into two groups “severe=0” and “severe=1” (severe accidents), shows the correlation between different meteorological conditions and the severity of motors vehicle accidents different weather condition such as "CLEAR," "SNOW," "UNKNOWN/OTHER," "RAIN," "CLOUDY/OVERCAST," "FOG/SMOKE/HAZE," and "SLEET/HAIL," are represented by the X-axis while the number of accidents is displayed on the Y-axis. With over 500,000 non severe incidents and over 200,000 severe accidents the data shows that most accidents happen during “CLEAR” weather. Compared to incidents in other weather conditions this is far higher. Although they are significantly less common than in clear weather “RAIN” and “CLOUDY/OVERCAST” circumstances also play a significant number of incidents. Under these circumstances, the number of severe accidents is proportionately larger than that of non-severe ones. There are comparatively fewer accidents when weather conditions like “SNOW”, “FOG/SNOW/HAZE,” and “SLEET/HAIL” are present.

Model Validation:

Logistic regression:

Logistic regression is the most common and easy to use algorithm in machine learning by maintaining its performance and mainly used to predict binary classification output i.e one or two possible outcomes(yes/no).

Firstly, the dataset is divided into training and testing dataset and logistic regression model is applied to predict the severity of the traffic collision. Although the dataset before SMOTE is imbalanced with more non severe cases, the model performed well and efficiently.

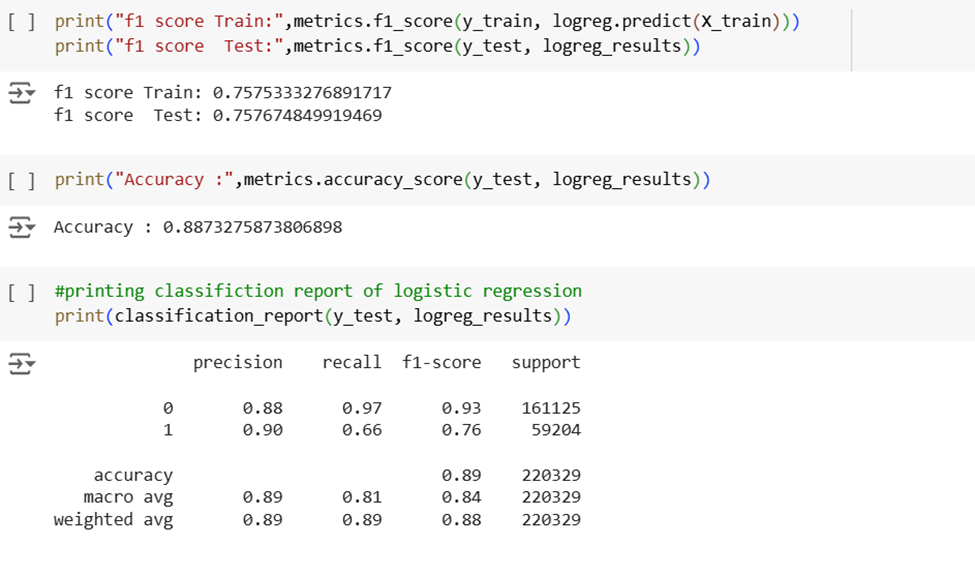


Figure: 1.7 Metrics and classification report of logistic Regression

From the above figure 1.7, we can observe f1 score for train and test are consistent of around ~0.75 and good accuracy of 88.7%

From the above figure we can also observe that the model has high precision of 0.9 which signifies that the model is accurately predicting the severity of cases, but the constraint is low recall of 0.65.

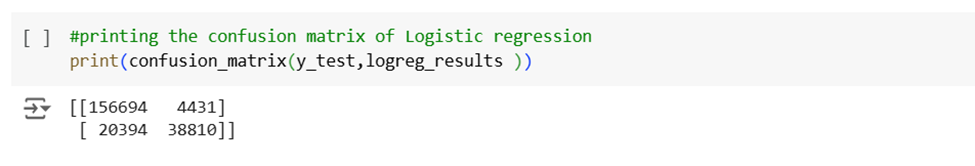


Figure: 1.8: Confusion Matrix of Logistic Regression

From the above figure 1.8, Confusion matrix shows that there are many missing severe cases, almost 20394 severe accidents were incorrectly classified.

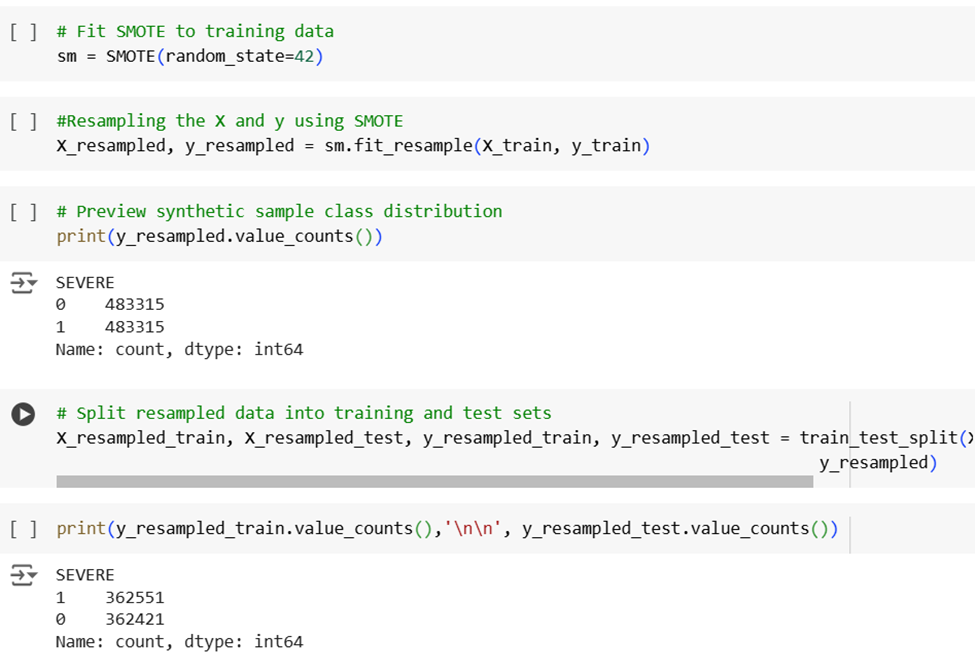


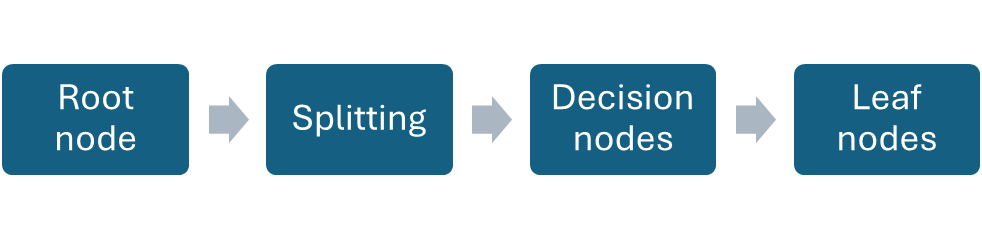
Figure: 1.9 : Training and Testing sets count

SMOTE has been implemented with training data to deal with imbalance issues. From above figure 1.9, we can observe that after smote has been implemented, severe cases of class 0 and class 1 are equal to count 483315 before splitting. After splitting into train and test data, the counts of class 1, class 0 are equal to 362551 and 36421 which is equal to 483315 which balances the imbalance data.

Overall, the logistic regression model performed well and is a good start stating our trained model predicts the expected output of the severity, moving ahead the main focus is to address the imbalance in the dataset by which we can get much more efficient accuracy.

**Decision Tree:**

Decision trees work by splitting data into smaller subsets in a tree-like structure to make the predictions for both classification and regression.



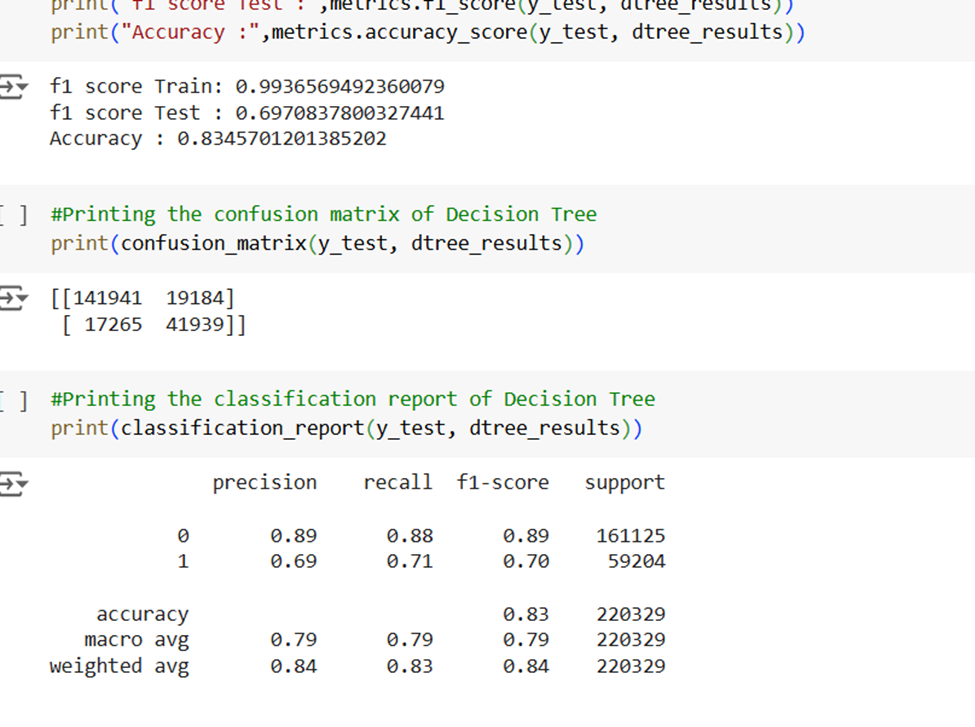


Figure 2.0: Confusion and Classification report of Decision Tree

From the above figure 2.0, we can observe that the F1 score of train and test is not consistent of 0.99 and 0.69 respectively with accuracy of 0.83. The decision tree outperformed logistic regression in identifying these crucial cases, achieving a higher recall (0.74) for severe situations, but if precision is important, logistic regression can be considered as the best model.

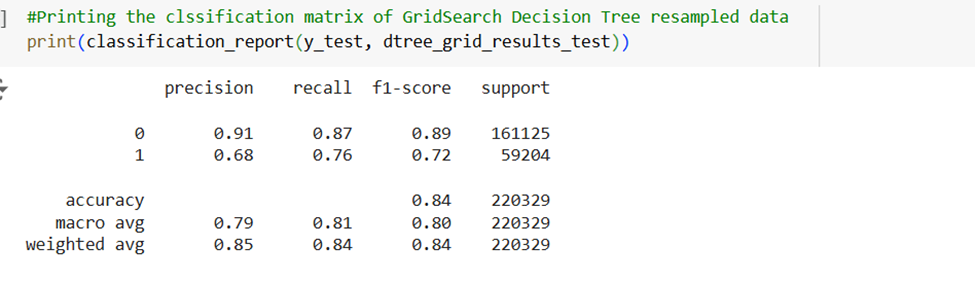
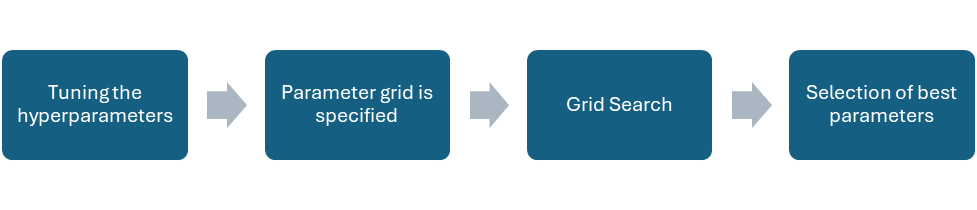


Figure: 2.1 : GridSearch Decision Tree Classification Matrix

Decision tree mainly functions on hyperparameters, so to have better accuracy and to search the best combination of hyperparameters GridSearch is the advantageous solution, from the above figure 2.1, we can observe the performance metrics of gridsearch Decision trees where precision, recall and f1-score of class 0 is significantly low than class 1 but for severe cases it maintains the improved recall,balanced metrics, reduced false negatives and false positives .In Gridsearch we can tune the model by setting certain parameters to trees which helps in optimization of model performance, systematic exploration and reproducibility.

****

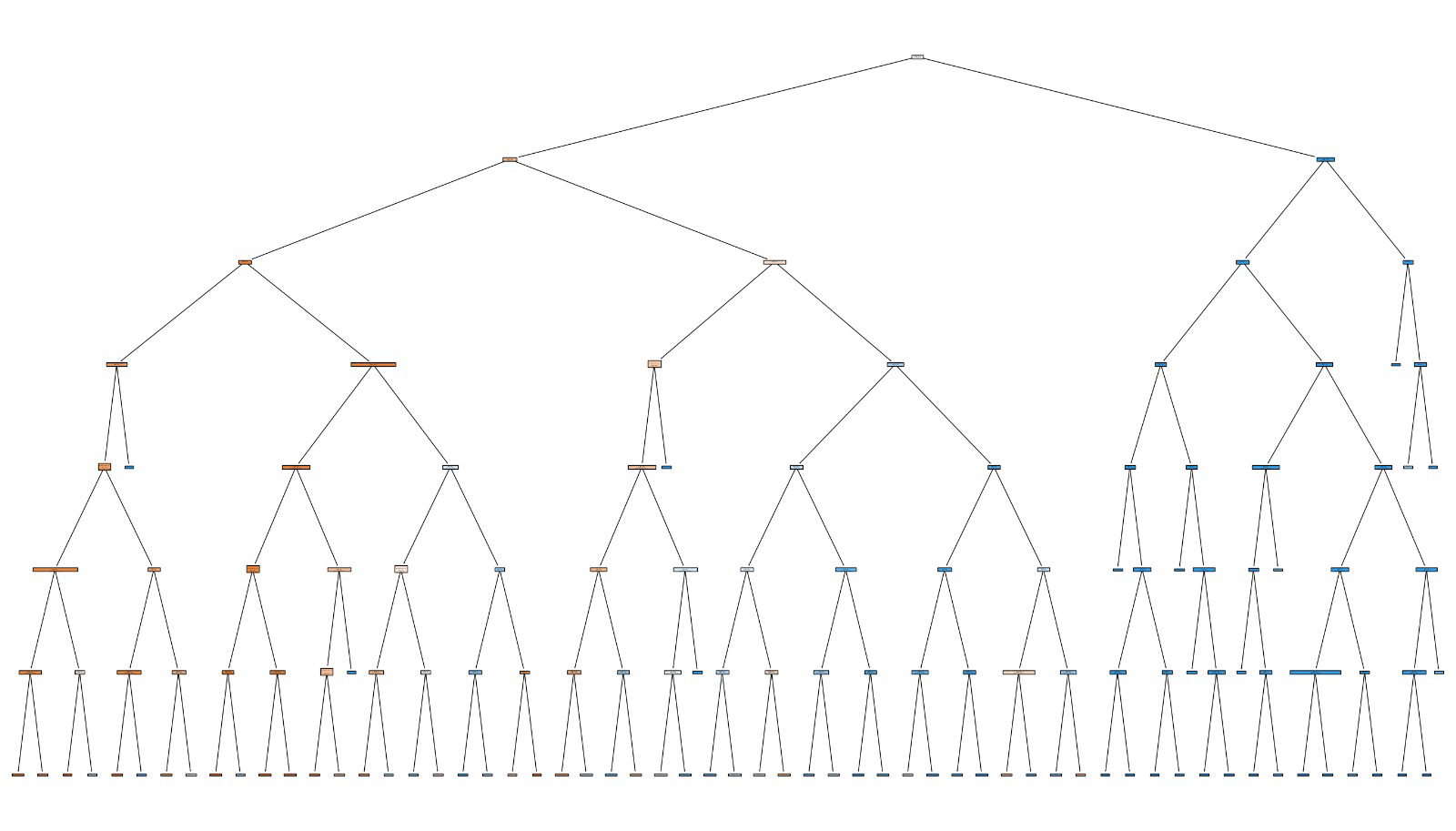
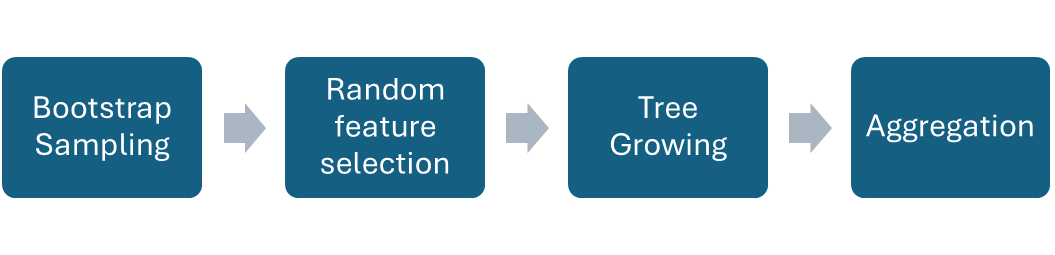


Figure: 2.2 : Visualization of Grid Search Decision Tree

Random Forest Algorithm :

Upon testing the logistic regression model and decision tree performance, moved ahead to the Random Forest algorithm, which is one of the powerful and efficient machine learning algorithms. To explain in crisp, random forest algorithm follows the assembling of decision trees approach, which is nothing but collection of decision trees designed during training and aggregates the respective outputs mainly for classification and regression.



Why random forest algorithm? The main agenda in going with random forest algorithm is to reduce variance in predictions, reduce the high correlation in trees, and mainly to average the predictions, this can be done by bootstrapping, random feature selection, tree growing and aggregation respectively as shown in above figure.

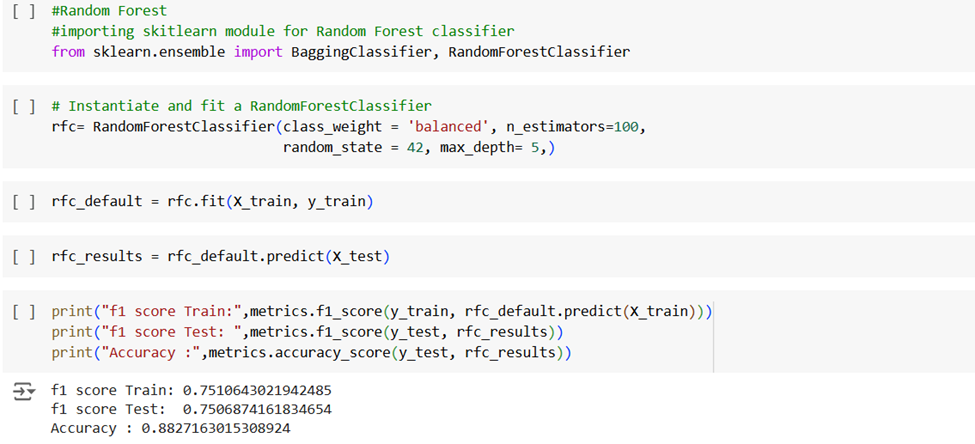


Figure 2.3 : Performance metrics of Random Forest Algorithm

From the above figure 2.3, we can observe a few performance metrics: f1 score and accuracy. We can clearly observe the F1 score of the train and test is approximately equal and near to 1 which is ~0.75. And the accuracy is 0.88 which is acceptable and can assume that this algorithm best suits for the problem statement of predicting the traffic collisions.

Random forest algorithm with SMOTE:

SMOTE is a simple technique which balances the dataset before training. We now observe the metrics for imbalanced and balanced dataset used for training using SMOTE.

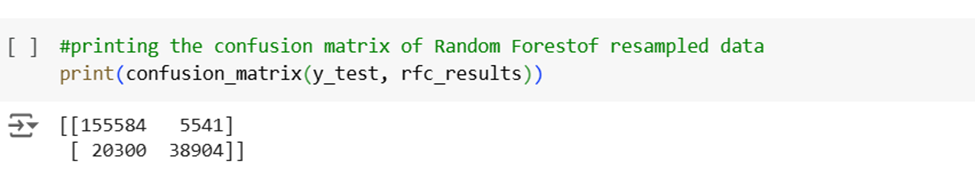


Figure 2.4 Confusion Matrix of Random Forest Algorithm

Above is the results from the confusion matrix,

True Negatives (TN) = 155,584

False Positives (FP) = 5,541

False Negatives (FN) = 20,300

True Positives (TP) = 38,904, by this we can calculate Accuracy

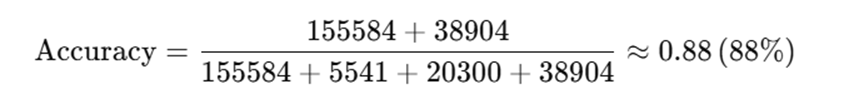


Figure 2.5​​ Accuracy of Random Forest Algorithm

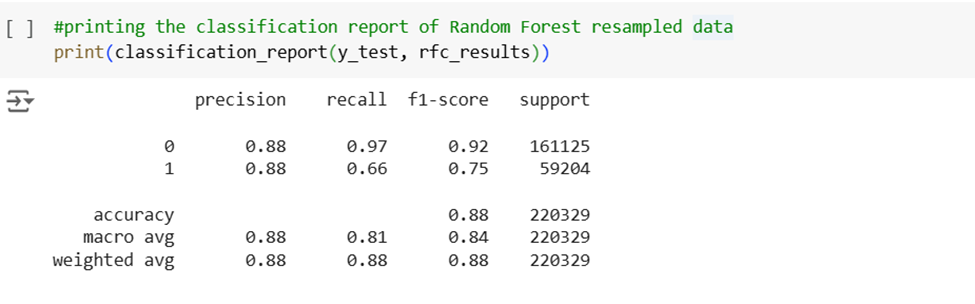


Figure 2.6​​ Performance metrics of Random Forest Algorithm vs Random forest algorithm with SMOTE

From the above figure 2.6, we can observe the classification report of random forest resampled data and can demonstrate that there is a slight difference in class 0 and class 1 performance metrics mainly in terms of recall, f1-score and support. Precision is consistent with respect to both classes meaning high precision with few positive errors but recall for class 0 is 0.97 and recall for class 1 is 0.66 which means there is a significantly lower recall for class 1 meaning misses a substantial number of actual collisions (false negatives). And same with f1-score class 1 0.75 which is significantly lower than class 0: 0.92 meaning there is a jump trade-off between precision and recall. As the support numbers of class 1 and class 0 difference is huge, it states that the data is highly imbalanced before SMOTE. Using SMOTE technique, recall and f1-score are being enhanced. Therefore, we can interpret high accuracy and precision for both classes which make the algorithm efficient and powerful.

Comparison of models:

| Metric Parameter | Logistic Regression (Without SMOTE) | Logistic Regression(With SMOTE) | Grid Search Decision Tree (With SMOTE) | Decision Tree (Without SMOTE) | Decision Tree (With SMOTE) | Random Forest (Without SMOTE) | Random Forest (With SMOTE) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Accuracy | 88.72% | 88.72% | 84% | 84.69% | 84.69% | ~88% | ~88% |
| F1-Score (Class 0) | High | High | 0.89 | 0.92 | ~0.92 | 0.92 | ~0.92 |
| F1-Score (Class 1) | 0.65 | 0.75 | 0.72 | 0.74 | ~0.77 | 0.75 | ~0.77 |
| Precision (Class 0) | High | High | 0.91 | High | High | High | High |
| Precision (Class 1) | 0.9 | 0.9 | 0.68 | 0.71 | 0.71 | High | High |
| Recall (Class 0) | ~0.98 | ~0.98 | 0.87 | ~0.97 | ~0.97 | 0.97 | ~0.97 |
| Recall (Class 1) | 0.65 | ~0.70 | 0.76 | 0.74 | ~0.74 | 0.66 | ~0.70 |
| True Negatives (TN) | ~155,000 | ~155,000 | ~161,125 | ~155,000 | ~155,000 | 155,584 | 155,584 |
| False Positives (FP) | ~5,000 | ~5,000 | Reduced | ~5,000 | ~5,000 | 5,541 | Reduced slightly |
| False Negatives (FN) | 20,498 | Reduced (~19,000) | ~15,000 | 15,478 | Reduced (~14,000) | 20,300 | Reduced (~19,000) |
| True Positives (TP) | ~36,000 | ~38,000 | ~40,000 | ~38,000 | ~40,000 | 38,904 | ~40,000 |

Table 1.0: Models Comparison

From the above table 1.0 after validation, we can conclude that each model has its own advantages and disadvantages. But according to scenario and situation we can choose the best fits in. Our main problem statement and goal is to find the severe accidents that need attention. In this scenario if precision and recall should be balanced, logistic regression with SMOTE is the best choice. If interpretability and recall should be balanced, a Decision tree can be the best choice. But the best model that suits this scenario which handles the imbalance data, good accuracy, maintaining good precision and recall for both Class 0 and class 1 is Random forest Algorithm. Therefore, we conclude that the Random forest algorithm is the best model to predict severe accidents which is giving good performance metrics as well.

Need to explain details with numbers please rewrite ( who wrote this)?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

# LIST OF REFERENCES

Admin. & Admin. (2024, September 19). Who causes more car accidents, men or women? | Malman Law. Malman Law.

<https://www.malmanlaw.com/malman-law-injury-blog/who-causes-more-car-accidents-men-or-women/>

Accident Analysis & Prevention | Vol 43, Issue 1, Pages 1-494 (January 2011) | ScienceDirect.com by Elsevier. (n.d.).

<https://www.sciencedirect.com/journal/accident-analysis-and-prevention/vol/43/issue/1>

Kristianssen, A., Andersson, R., Belin, M., & Nilsen, P. (2017). Swedish Vision Zero policies for safety: A comparative policy content analysis. Safety Science, 103, 260–269.

<https://doi.org/10.1016/j.ssci.2017.11.005>

Khayesi, M. (2006), February, *The Handbook of Road Safety Measures*.<https://pmc.ncbi.nlm.nih.gov/articles/PMC2563494/>

Ma, Y., Xu, J., Gao, C., Mu, M., E, G., & Gu, C. (2022). Review of research on road traffic operation risk prevention and control. International Journal of Environmental Research and Public Health, 19(19), 12115.

<https://doi.org/10.3390/ijerph191912115>

Maslamani, Y. A. (n.d.). Effect of the Structured Traffic Education Program (STEP) on knowledge and commitment toward traffic rules in the first year of legal driving in Saudi Arabia. Informit.

<https://search.informit.org/doi/abs/10.3316/informit.T2024060200001990074894717>

NHTSA estimates traffic fatalities continued to decline in the first quarter of 2024. (2024, June 24). NHTSA.

<https://www.nhtsa.gov/press-releases/2024-Q1-traffic-fatality-estimates>

Ryan. (2018, July 31). What are the major contributing factors to traffic accidents? The law offices of OEB. OEB Law, PLLC.

<https://www.wreckintoacheck.com/faqs/what-are-the-major-contributing-factors-to-traffic-accidents/>

World Health Organization: WHO. (2019, October 2). Road safety.

<https://www.who.int/health-topics/road-safety#tab=tab_1>

World Health Organization: WHO. (2023, December 13). Road traffic injuries. <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>

Will crash experience affect the driver’s behavior? An observation and analysis on time headway variation before and after a traffic crash. (n.d.). TUP Journals & Magazine IEEEXplore.

<https://ieeexplore.ieee.org/document/8954865/citations#citations>