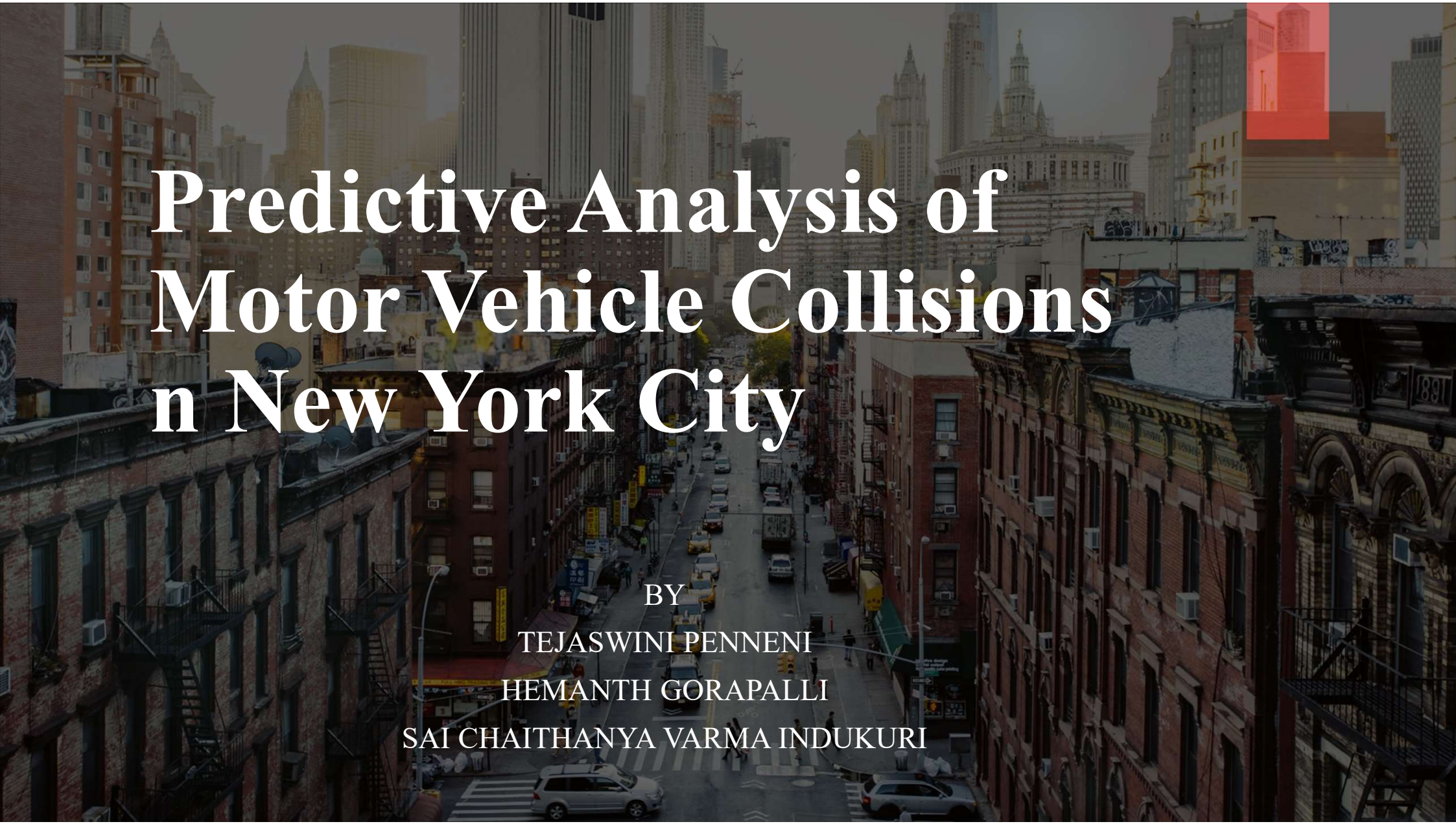


Predictive Analysis of Motor Vehicle Collisions n New York City

BY

TEJASWINI PENNENI

HEMANTH GORAPALLI

SAI CHAITHANYA VARMA INDUKURI

Project Overview

- The project aims to develop a predictive model for accurately forecasting motor vehicle collision (MVC) severity in New York City, driven by the urgent need to enhance public safety. Through in-depth analysis of MVC datasets using big data analytics, the project will identify critical factors such as timing, location, and hazards. Leveraging advanced predictive modeling techniques including logistic regression, decision trees, and random forests, the research aims to uncover patterns and trends in MVC data. The ultimate goal is to provide actionable insights and recommendations for state authorities to implement targeted interventions aimed at reducing the frequency and severity of MVCs, thereby improving road safety in densely populated urban environments.

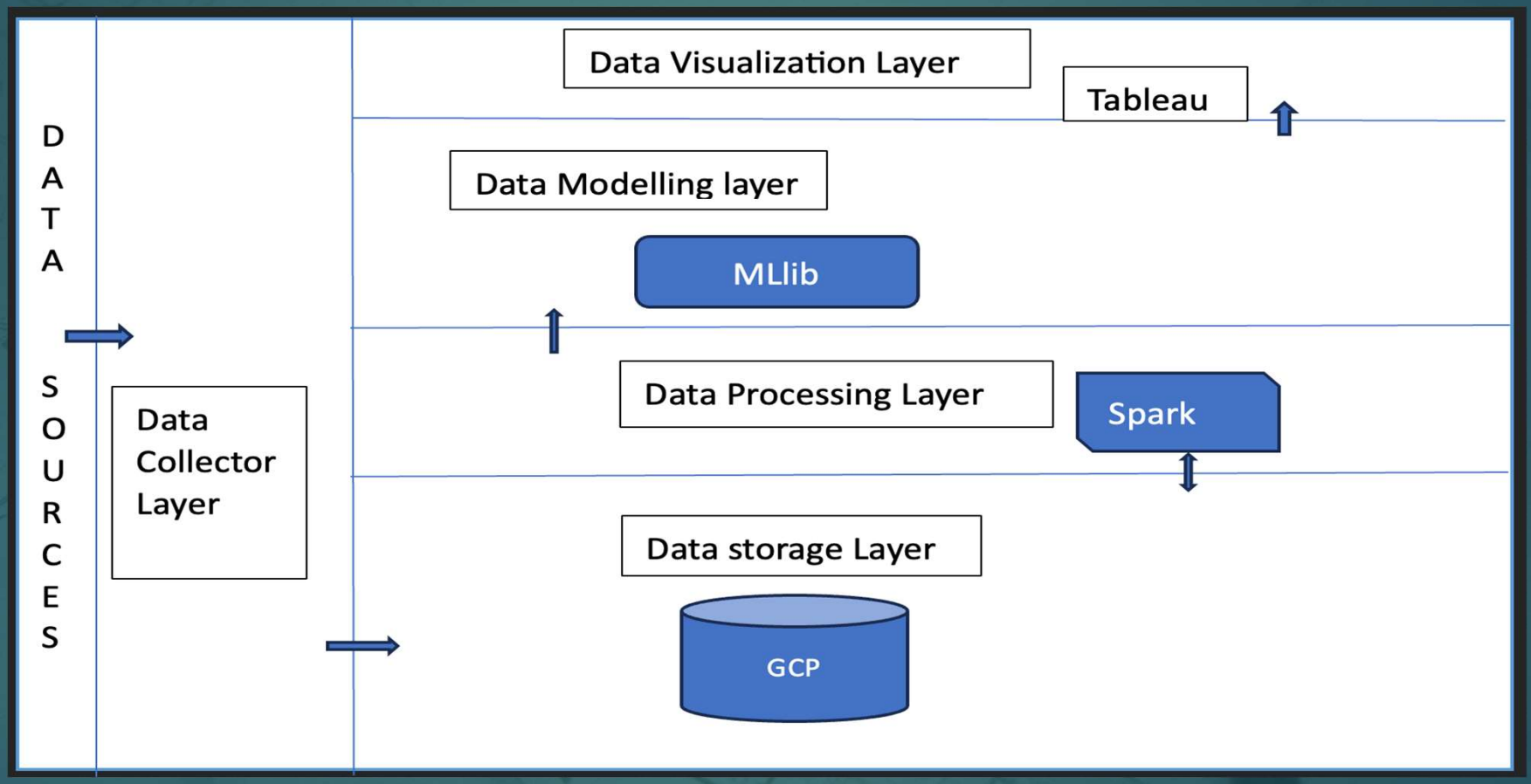
Problem Statement

- Motor vehicle crashes are part of the cause of injury-related mortality among New York State citizens. This investigation aims to investigate and comprehend the trends behind these collisions in New York City. By addressing specific business concerns, the objective is to deliver insights that might help individuals and organizations reduce the risk of accidents.

- **Business Question**

1. What are the key factors influencing MVC severity, and how can they be effectively integrated into predictive modeling techniques?
2. Are there specific months , specific days and specific hours of the day that are more accident-prone than others?
3. How do weather conditions influence the frequency and severity of MVCs throughout the year, and are certain months more prone to adverse weather-related accidents?

Proposed Architecture



Technologies Used

- **Google Cloud Platform:** "Provides a secure, scalable environment for storing and accessing large volumes of traffic data, facilitating easy integration with analytical tools."
- **Apache Spark:** "Enables rapid data transformations and aggregations, crucial for preprocessing tasks in real-time analytics.“
- **Pandas:** Used for data manipulation, cleaning, and preprocessing.
- **MLlib:** "Offers a range of machine learning algorithms that enhance our capability to build sophisticated models predicting MVC severity."
- **Tableau:** "Allows us to create interactive and dynamic visualizations, making complex data more accessible and understandable for stakeholders.“
- **GitHub:** Used for Code documentation

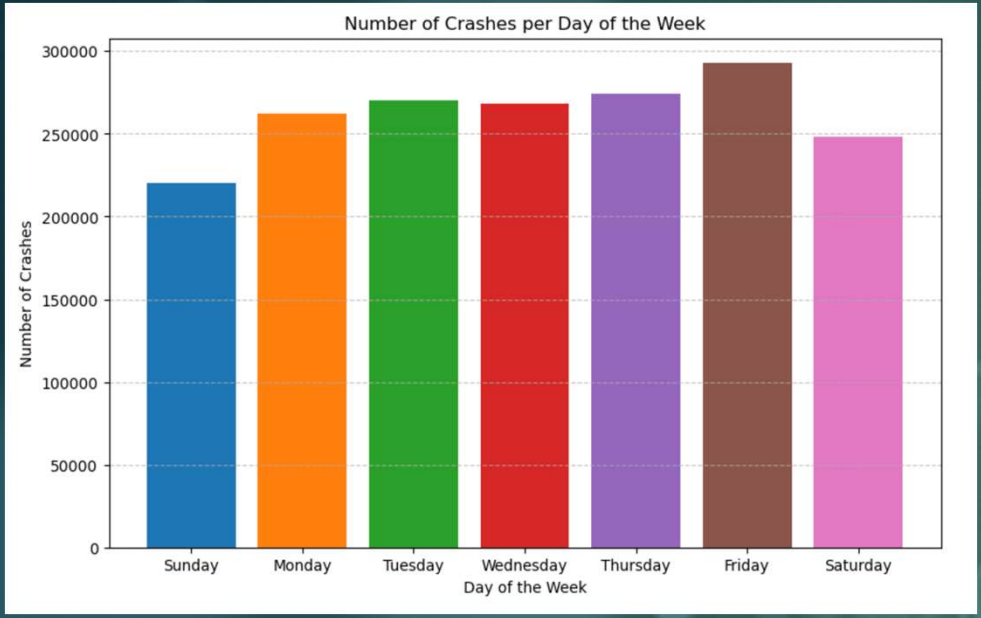
Integration and Workflow

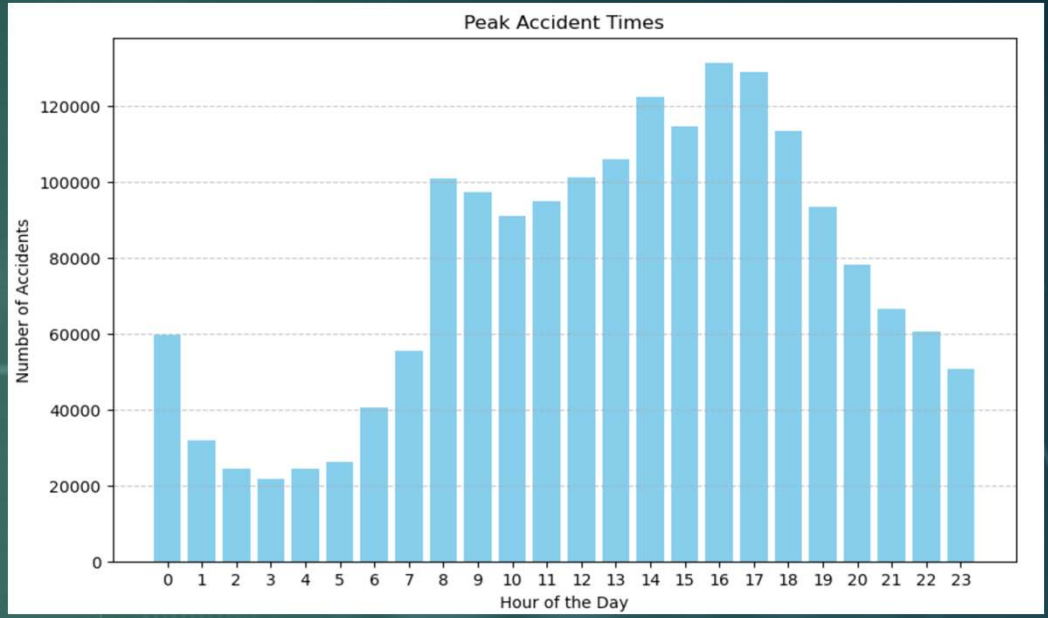
- Data Workflow: Data stored in GCP is processed through Apache Spark for cleaning and analysis. The processed data is then modeled using MLlib, and the insights are visualized using Tableau."
- **Efficiency and Streamlining:** "This integrated workflow ensures that data moves seamlessly from storage to visualization, maximizing efficiency and reducing latency in data handling."

Data Management and Processing

- Managing and preprocessing large datasets involves addressing issues of data quality such as missing values, inconsistencies, and noise.
- We employ techniques such as data imputation, feature engineering and data balancing to prepare the dataset for accurate modeling.

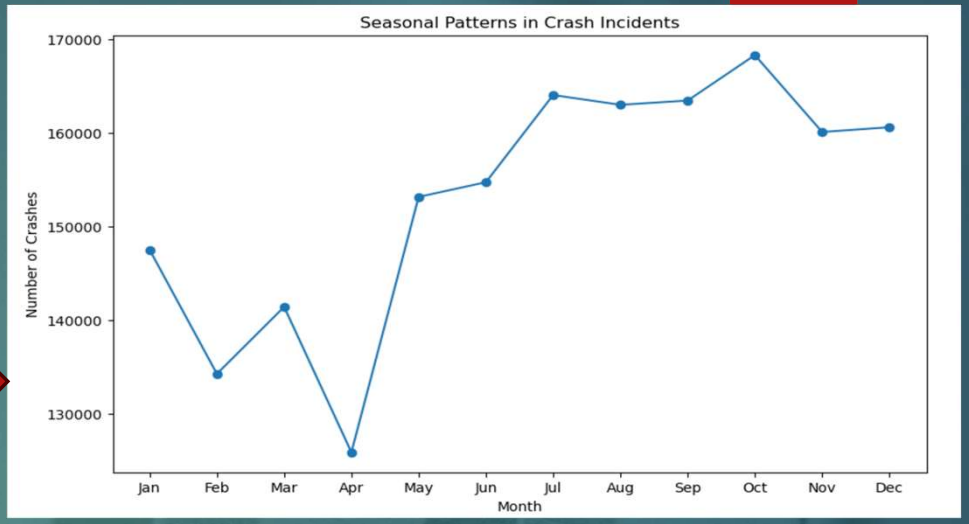
Data Analysis



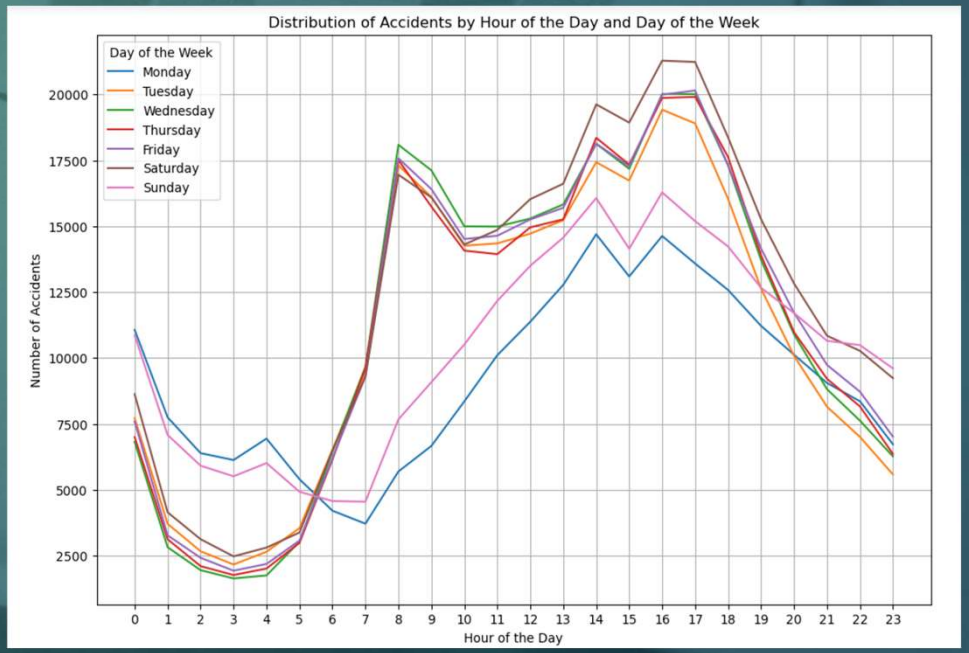


Data Analysis

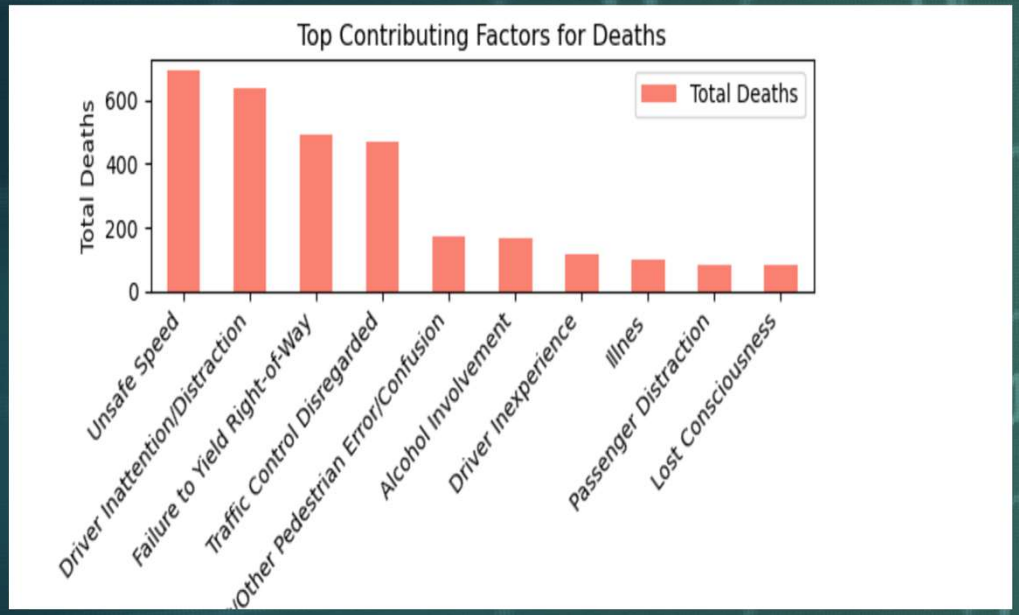
The line chart provides a clear visual depiction of the fluctuation in motor vehicle collisions over time, from months with the highest frequency of accidents to those with the lowest occurrence. Notably, there is a significant decline in the number of crashes observed in February, followed by a notable increase in March and a peak in June. Subsequently, there is a surge in collision numbers during the summer months, before a decline in September, followed by a subsequent increase in November.

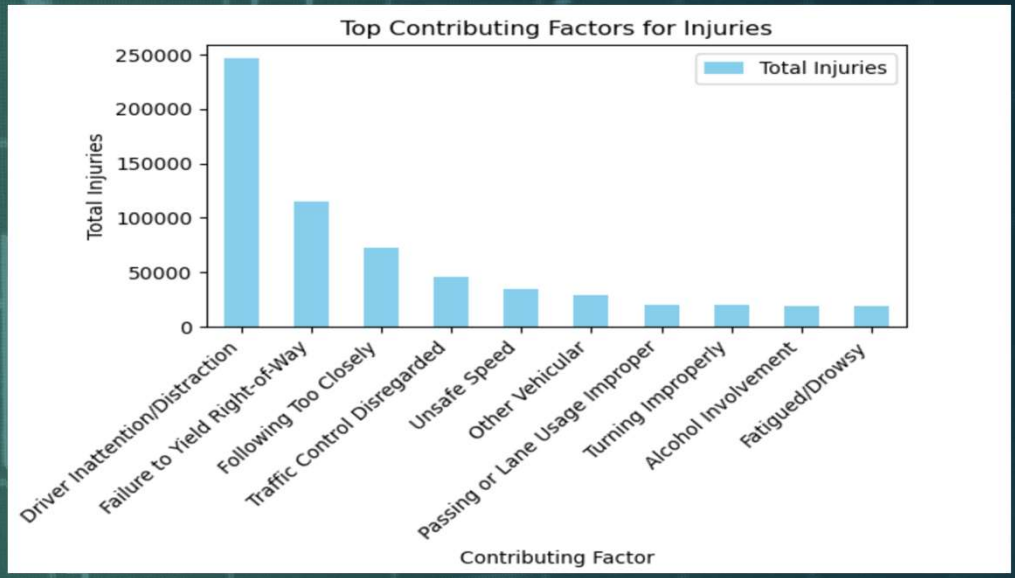


This comprehensive line chart illustrates the number of traffic accidents that occur at different times throughout the day, segmented by days of the week. Notably, all days share a common pattern where accident frequencies start to increase from the early morning hours, peak during the late afternoon (between 14:00 and 18:00), and gradually decline throughout the evening.



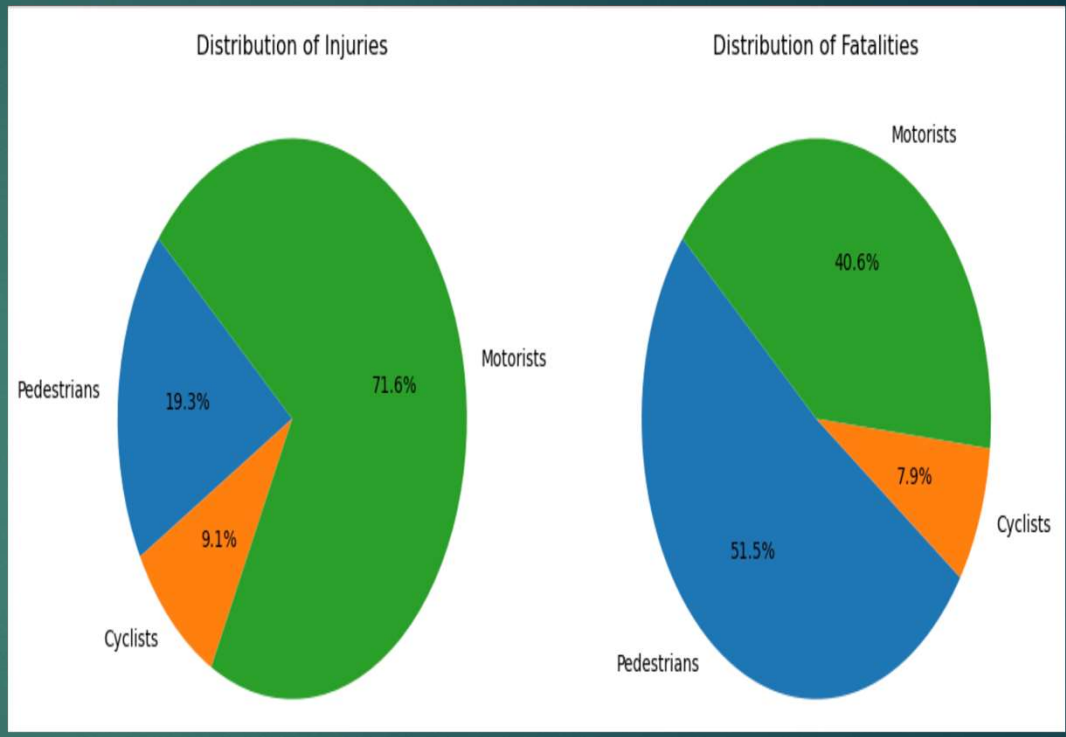
Data Analysis





Data Analysis

From the pie chart analysis, it's evident that the majority of accidents affected motorists, comprising 71.6% of the total incidents, followed by pedestrians at 19.3%, and cyclists at 9.1%. However, when considering the distribution of fatalities, pedestrians bore the brunt with 51.55%, followed by motorists at 41.95%, and cyclists at 7.9%.



Data Modelling - Performance

- Data modeling performance refers to the effectiveness of predictive models in analyzing data and making accurate predictions. In this project, three models—logistic regression, decision tree classifier, and random forest classifier—were evaluated based on metrics like accuracy, precision, recall, and F1-score
- The performance of the models varied, with the decision tree classifier exhibiting the highest overall performance, achieving an accuracy of 84%, precision of 82%, and recall of 84%. The analysis of feature importance revealed that factors such as time of day, weather conditions, and type of intersection significantly influenced collision severity. For example, collisions occurring during rush hours and at poorly lit intersections were more likely to be severe.

Results with Insights and Recommendations

- The analysis of motor vehicle collision data in New York City revealed intriguing patterns and trends. It was observed that the frequency of accidents tends to increase throughout the week, with the lowest occurrences on Sundays and the highest on Fridays.
- Moreover, a distinct rise in collisions was noted during morning rush hours, starting from 7 AM, peaking between 2 PM and 6 PM during the afternoon rush hour.
- Seasonal variations were also evident, with a significant decline in accidents in February followed by an increase in March and a peak in June, coinciding with the summer months
- The predictive modeling phase, particularly the implementation of the random forest model, yielded noteworthy results. The superior performance of the random forest model suggests its effectiveness in capturing the complexity of factors influencing motor vehicle collision (MVC) severity.
- Crashes with vehicle defects includes accelerator defective as primary reason followed by brake defective.
- Furthermore, the analysis identified key risk factors such as time of day and lighting conditions at intersections, which significantly impact MVC severity. These insights offer actionable recommendations for targeted interventions, such as enhancing lighting and adjusting traffic signals during peak hours.

Conclusion

- Our project focused on improving public safety and traffic management in New York City by developing a predictive model for motor vehicle collision severity. Using advanced analytics and machine learning techniques, we identified key factors like time of day and weather conditions that impact collision severity. These insights enable targeted interventions, such as optimizing traffic signals and enhancing lighting in high-risk areas. Also, Public awareness campaigns should focus on safe driving during peak traffic hours and city should consider putting in place measures to address the specific patterns identified, such as increased safety measures during peak travel timings. Looking ahead, integrating real-time data and additional variables can further enhance our model's predictive power, contributing to safer urban environments beyond New York City.

What we learn

- Our project highlighted the significance of thorough data cleaning and exploratory analysis in preparing extensive datasets for predictive modeling. We delved into various machine learning algorithms, including logistic regression, decision trees, and random forests which helped us to understand and develop accurate predictive models for forecasting motor vehicle collision severity. Our analysis unveiled key factors such as time of day, contribution factors, and specific intersection types, which significantly influence collision severity. Understanding these factors is pivotal for implementing targeted interventions aimed at enhancing road safety. Moving forward, we aim to enhance our model by integrating real-time data and including additional socio-economic and behavioral variables.

THANK YOU