Impact Unveiled: A Comprehensive Study on Collision Severity in NYC

Taruna Verma

December 17, 2024

Abstract

This report delves into the Motor Vehicle Collisions dataset from New York City's Open Data portal to uncover patterns and risk factors associated with traffic incidents across various urban scenarios. Analyzing critical factors such as vehicle types, environmental conditions, and times of day, the study employs advanced machine learning techniques, including a Random Forest model, to pinpoint major predictors of severe accidents. Our findings show significant risks with two-wheelers and during poor lighting conditions. These insights are vital for crafting targeted safety measures and regulatory policies to enhance road safety. The research informs policymaking, proposing improvements and laying the groundwork for continuous advancements in traffic management and urban safety.

Introduction

This project is dedicated to examining the underlying factors driving the severity of motor vehicle collisions in New York City. Given the complexity of urban traffic dynamics and the significance of road safety, this analysis aims to derive

actionable insights from NYC's collision data that can directly support data-informed enhancements in traffic safety measures.

The core goal of this project is to pinpoint the most influential factors linked to severe injuries and fatalities in NYC collisions. By delving into the specific attributes and circumstances surrounding high-severity incidents, this project aspires to generate valuable insights that can shape city planning efforts, inform road safety strategies, and raise public awareness on reducing severe outcomes in urban traffic environments.

Motivation

My motivation for selecting this project is grounded in its potential real-world impact on public safety. New York City, a vibrant metropolitan hub with dense vehicular and pedestrian activity, faces a high volume of road accidents. Analyzing the severity of collisions in such a complex urban environment is not only crucial for NYC but can also yield insights applicable to other cities. Through this project, I aim to craft a data narrative that uncovers patterns and trends in collision data, with the broader goal of contributing meaningfully to urban safety and enhancing the well-being of city residents.

Data Description

Data Overview: Motor Vehicle Collisions - Crashes

For this research, I will be using **the New York City Motor Vehicle Collisions – Crashes dataset**, which provides extensive details on individual collision incidents in NYC.

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. The police report (MV104-AN) is required to be filled out for collisions where

someone is injured or killed, or where there is at least $1000 worth of damage.

This dataset includes both structured and unstructured data on various aspects of collisions, such as time, location, contributing factors, and outcomes for individuals involved (e.g., pedestrians, cyclists, motorists). The data is stored in a CSV format and comprises thousands of rows. The dataset is moderately large and allows for a comprehensive analysis of trends, factors, and patterns influencing collision severity in an urban setting.

Rows: **2.13M**, Columns: **29** Each row is a: **Motor Vehicle Collision**

Data Cleasing and Preprocessing

1. **Load and clean**

Initial steps in preparing this dataset involves extensive cleaning.

Originally dataset has Rows: **2.13M**, Columns: **29**

Each row is a: **Motor Vehicle Collision.**

Specific cleaning tasks include:

2. **Handling Missing values:**

Calculated the percentage of missing values in each column so that we can decide what should be done the following columns have the max number of missing values

OFF STREET NAME (82.91%)

CONTRIBUTING FACTOR VEHICLE 3-5 (92.81%, 98.37%, 99.56%)

VEHICLE TYPE CODE 3-5 (93.08%, 98.42%, 99.57%).

3. **We decided to Drop a few Columns:**

- These columns have a very high percentage of missing data, indicating limited usefulness for analysis.

- Dropping them may be the best option unless specific patterns are needed.
- Some columns (such as VEHICLE TYPE CODE 3-5 (93.08%, 98.42%, 99.57%), OFF STREET NAME (82.91%),CONTRIBUTING FACTOR VEHICLE 3-5 (92.81%, 98.37%, 99.56%)) are nearly entirely empty. We'll remove those.
- Also, we will not be using some columns (e.g. collision_id, on_street_name, off_street_name, cross_street_name) so we can drop them completely.

4. **Impute Missing Values:**

- For essential columns (BOROUGH, ZIP CODE, contributing factors, and vehicle types), we impute missing values to retain information and make the dataset more complete.
- Drop rows where LATITUDE, LONGITUDE, or LOCATION are missing.
- Rename Columns for Consistency and Readability.
- After these steps I checked for Number of duplicate rows: 1565
- It's a good idea to remove these duplicates to ensure the integrity of the data.
- **Dataset shape after removing duplicates: (1890937, 19)**
- **Convert Time column to datetime.time format, handling inconsistent formats.**
- Extract hour of the day from the Time column
- Convert Time column to datetime.time format, handling inconsistent formats

Number of missing values in 'time' after conversion: 0

Now the time column has following format of dataset.

Number of missing values in 'Time' after conversion: 0

```
In [90]: # Verify the Time column
         print("Sample of the 'time' column after processing:")
         print(data['Time'].head())

         Sample of the 'time' column after processing:
         3    09:35:00
         4    08:13:00
         6    17:05:00
         7    08:17:00
         8    21:10:00
         Name: Time, dtype: object
```

The Borough values are in uppercase (e.g., BROOKLYN) lets standardize casing for consistency in visualizations and analysis.

Create a separate dataset for rows with 'Unknown' borough

unknown_borough_data = data[data['Borough'] == 'Unknown']

5. **Outlier Detection in Numerical Columns**

Apply log transformation to the 'Persons_Injured' column

Number of outliers in log-transformed Persons_Injured: 326517

Dataset size after adjusted outlier removal: (1104075, 20)

Creating a severity_score Column

The severity_score combines the number of injuries and fatalities to provide a single measure of collision severity. Categorizing severity_score- to analyze collisions by severity categories (e.g., low, medium, high severity), create bins for the severity_score.

6. **Cleaning Steps for Contributing_Factor_1:**

- Remove or Replace "Unspecified" Values
- Replace "Unspecified" with "Unknown" or drop rows where this is present if it constitutes a small portion of the dataset.

- Group Similar Factors:Combine similar contributing factors into broader categories for better interpretation (e.g., "Driver Inattention/Distraction" and "Fatigued/Drowsy" can be grouped as "Driver Issues").
- Standardize Case: Ensure all factor values are consistent (e.g., all lowercase or title case).
- Remove or Merge Low-Frequency and Erroneous Categories
- Low-frequency entries can be grouped into broader categories (e.g., "Driver Distraction" for texting, eating, or using headphones).
- Erroneous entries ("80" and "1") can be removed.

Approach and Methods

Feature Engineering

Creating a severity_score Column: The severity_score combines the number of injuries and fatalities to provide a single measure of collision severity.

Formula Used: data['severity_score']=data['Persons_Injured']+(data['Persons_Killed'] * 10

What this tells us:

Higher Weight for Fatalities:

This formula amplifies the effect of fatalities (killed persons) compared to injuries. A single fatality contributes significantly more to the severity score than a single injury. For example:

- If there are 1 injury and 0 fatalities, the severity score will be **1** (1 injury).
- If there are 0 injuries and 1 fatality, the severity score will be **10** (1 fatality $\times$ 10).
- If there are 10 injuries and 1 fatality, the severity score will be **20** (10 injuries + 1 fatality $\times$ 10).

Categorizing severity_score to analyze collisions by severity categories (e.g., low, medium, high severity), create bins for the severity_score.

Exploratory Data Analysis and Visualization

Below is the snapshot of column names in the original dataset:

```
In [8]: list(data.columns)

Out[8]: ['CRASH DATE',
         'CRASH TIME',
         'BOROUGH',
         'ZIP CODE',
         'LATITUDE',
         'LONGITUDE',
         'LOCATION',
         'ON STREET NAME',
         'CROSS STREET NAME',
         'OFF STREET NAME',
         'NUMBER OF PERSONS INJURED',
         'NUMBER OF PERSONS KILLED',
         'NUMBER OF PEDESTRIANS INJURED',
         'NUMBER OF PEDESTRIANS KILLED',
         'NUMBER OF CYCLIST INJURED',
         'NUMBER OF CYCLIST KILLED',
         'NUMBER OF MOTORIST INJURED',
         'NUMBER OF MOTORIST KILLED',
         'CONTRIBUTING FACTOR VEHICLE 1',
         'CONTRIBUTING FACTOR VEHICLE 2',
         'CONTRIBUTING FACTOR VEHICLE 3',
         'CONTRIBUTING FACTOR VEHICLE 4',
         'CONTRIBUTING FACTOR VEHICLE 5',
         'COLLISION_ID',
         'VEHICLE TYPE CODE 1',
         'VEHICLE TYPE CODE 2',
         'VEHICLE TYPE CODE 3',
         'VEHICLE TYPE CODE 4',
         'VEHICLE TYPE CODE 5']
```

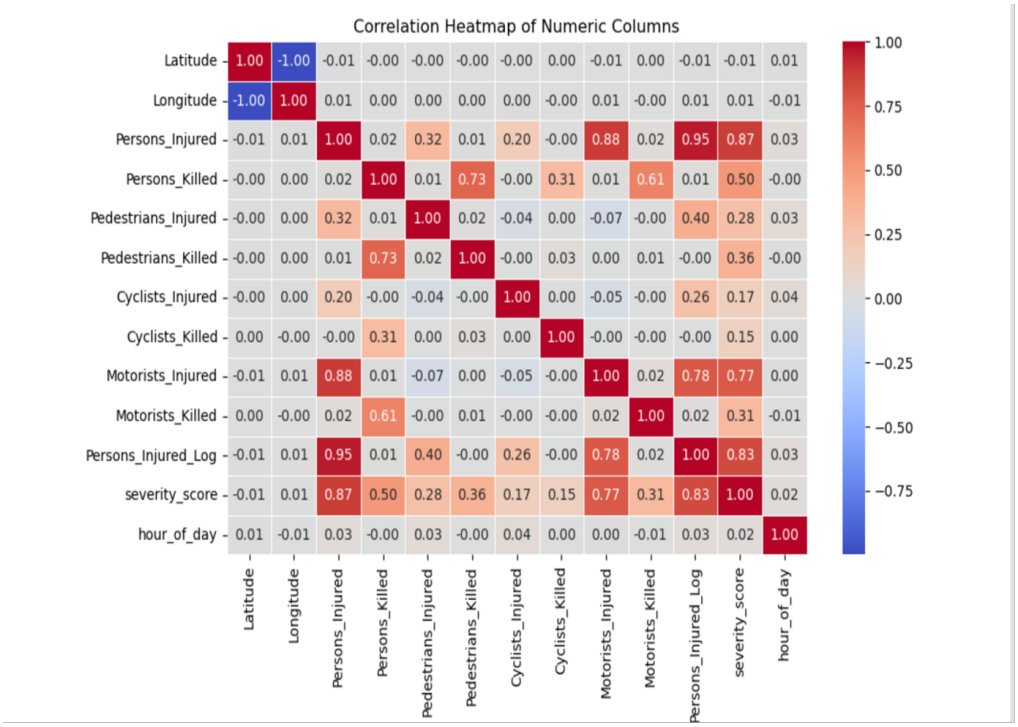
Dataset Review:

```
In [5]: data.head()
```

Out[5]:

	CRASH DATE	CRASH TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	...	CONTRIBUTING FACTOR VEHICLE 2	CONTRIBUTING FACTOR VEHICLE 3	C
0	09/11/2021	2:39	NaN	NaN	NaN	NaN	NaN	WHITESTONE EXPRESSWAY	20 AVENUE	NaN	...	Unspecified	NaN	
1	03/26/2022	11:45	NaN	NaN	NaN	NaN	NaN	QUEENSBORO BRIDGE UPPER	NaN	NaN	...	NaN	NaN	
2	06/29/2022	6:55	NaN	NaN	NaN	NaN	NaN	THROGS NECK BRIDGE	NaN	NaN	...	Unspecified	NaN	
3	09/11/2021	9:35	BROOKLYN	11208.0	40.667202	-73.866500	(40.667202, -73.8665)	NaN	NaN	1211 LORING AVENUE	...	NaN	NaN	
4	12/14/2021	8:13	BROOKLYN	11233.0	40.683304	-73.917274	(40.683304, -73.917274)	SARATOGA AVENUE	DECATUR STREET	NaN	...	NaN	NaN	

Correlation heatmap of Numeric columns



High Correlations: Persons_Injured is highly correlated with Persons_Injured_Log (0.95), which is expected since Persons_Injured_Log is derived from Persons_Injured.

Persons_Injured and Motorists_Injured (0.88) show a strong positive correlation, suggesting that a significant proportion of injuries in accidents involve motorists.

Severity_Score shows a strong correlation with Persons_Injured (0.87) and Persons_Injured_Log (0.83), indicating that these features contribute significantly to severity.

Moderate Correlations:

Pedestrians_Injured and Persons_Injured (0.32): Pedestrian injuries have a noticeable, albeit weaker, contribution to total injuries.

Pedestrians_Killed and Persons_Killed (0.73): Fatalities among pedestrians are strongly associated with total fatalities.

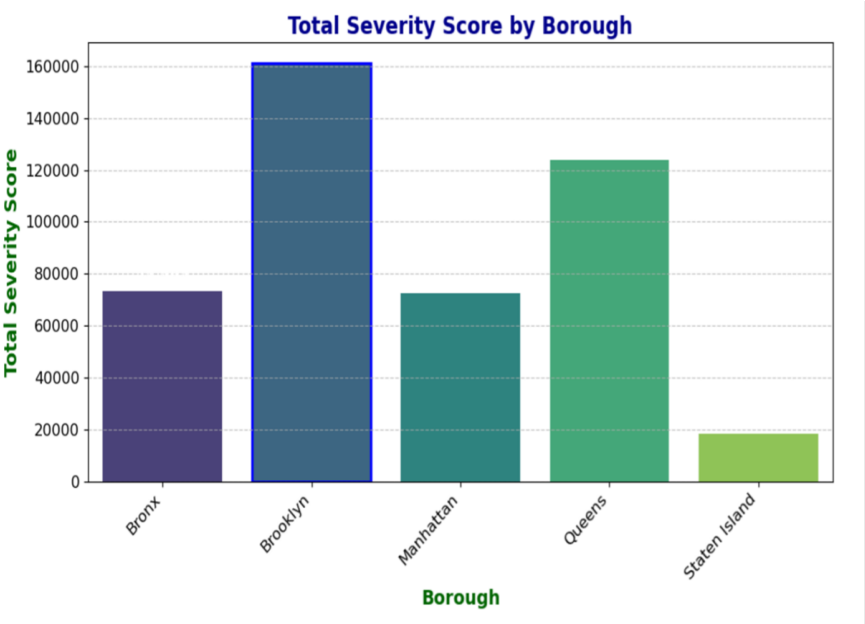
Motorists_Injured and Severity_Score (0.77): Motorist injuries are closely tied to accident severity.

Low or Negligible Correlations:

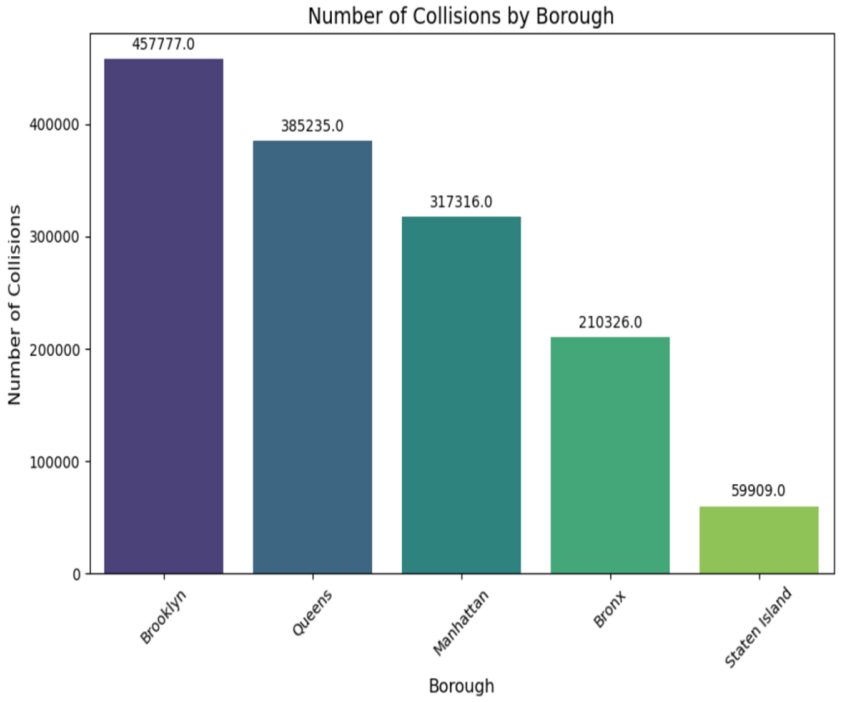
Geographic attributes like Latitude and Longitude have very weak correlations with other features, implying they don't directly influence injuries or severity in a straightforward linear relationship.

1. **Which boroughs have the highest number of accidents and potential hotspots using geographic data.**

Brooklyn stands out with the highest severity score, significantly surpassing the other boroughs. Queens follows as the second highest, while Manhattan and the Bronx have notably lower scores. Staten Island shows the lowest severity score among all the boroughs.



Brooklyn leads with the highest number of collisions, closely followed by Queens. Manhattan, while densely populated, shows a significantly lower number of collisions compared to Brooklyn and Queens. The Bronx and Staten Island exhibit the fewest collisions, with Staten Island having notably fewer incidents than any other borough. This distribution highlights areas with higher traffic safety concerns and suggests potential targets for focused traffic management and accident prevention initiatives.



2. **What is the relationship between time-related factors and collision severity?**

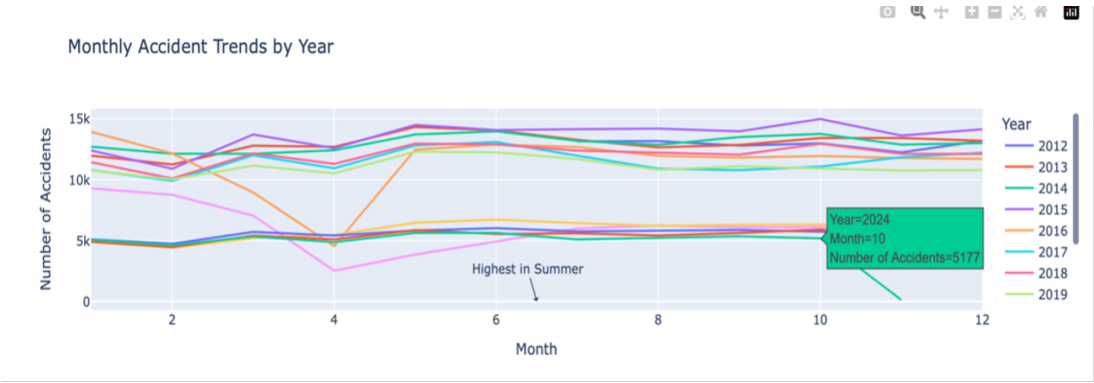
Rising and Falling Trends: There was a sharp increase in accidents from 2012 to 2014, followed by a period of stability until a gradual decline began after 2018.There was a significant decrease in accidents post-2020, likely attributable to pandemic-related changes in traffic patterns.

Monthly accident trends by year seasonal variations:

Accidents peak mid-year, correlating with summer travel and activities.

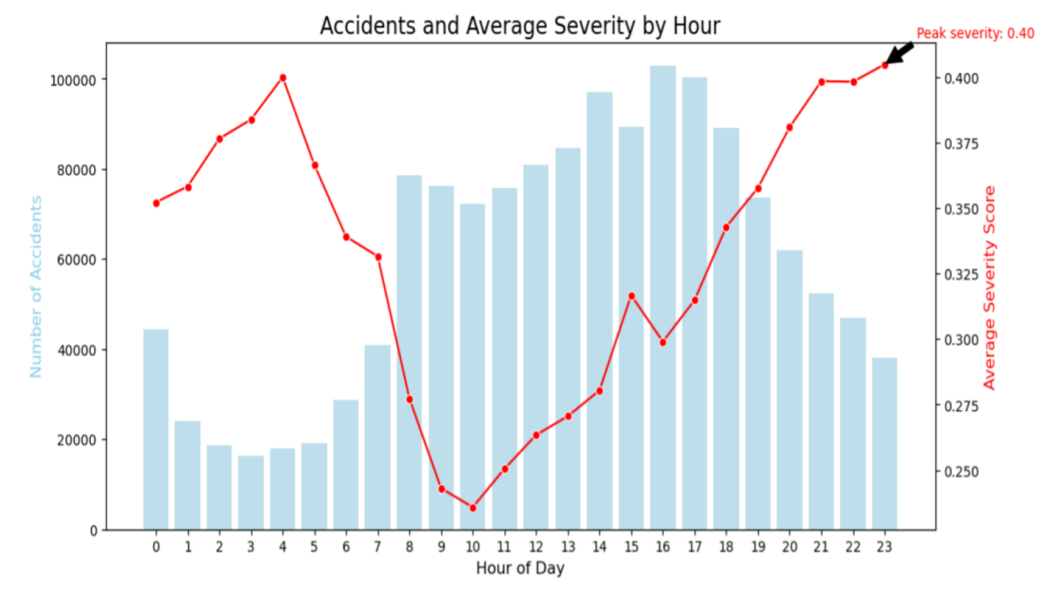
End-of-year decline:

Noticeable reduction in accidents toward year-end, more pronounced in recent years.



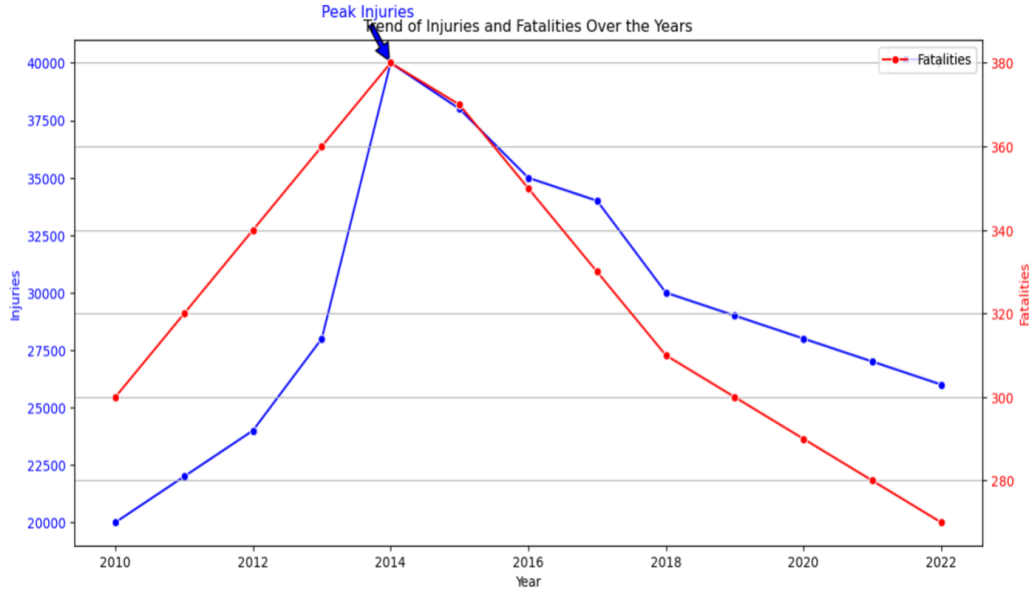
Analysis of Hourly Patterns

The number of accidents peaks in the morning hours, sharply decreases, and then gradually increases again during the evening rush hour, peaking at 21:00. Interestingly, while the frequency of accidents decreases after the morning peak, the average severity of accidents consistently increases from the early afternoon until it reaches its peak at 23:00. This suggests that accidents later in the day, though less frequent, tend to be more severe.



Annual Trends in Traffic Injuries and Fatalities

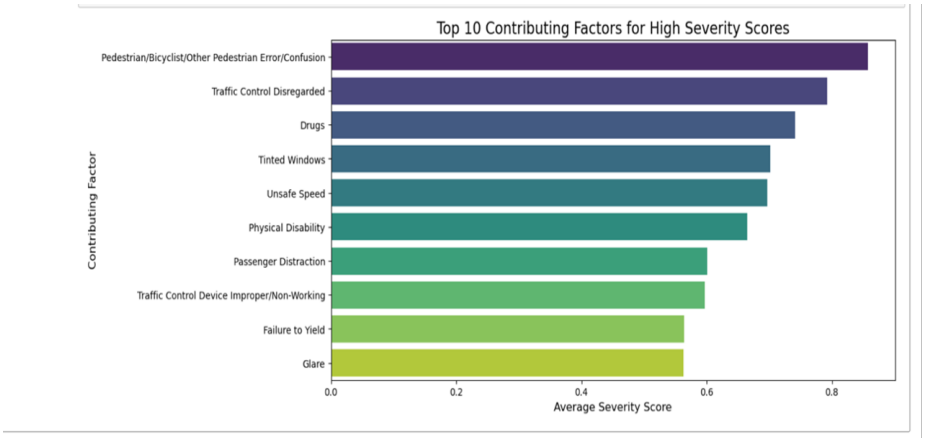
Injuries saw a sharp increase, peaking in 2014, before declining consistently thereafter. Fatalities, represented by the red line, followed a similar upward trend until 2014, but the decline in fatalities was less steep compared to injuries. This suggests that while the number of injuries significantly dropped over the years, the decrease in fatalities, though present, was less pronounced.



3. **What factors contribute most significantly to severe injuries and fatalities in NYC motor vehicle collisions?**

Top Contributing Factors for Severe Accidents

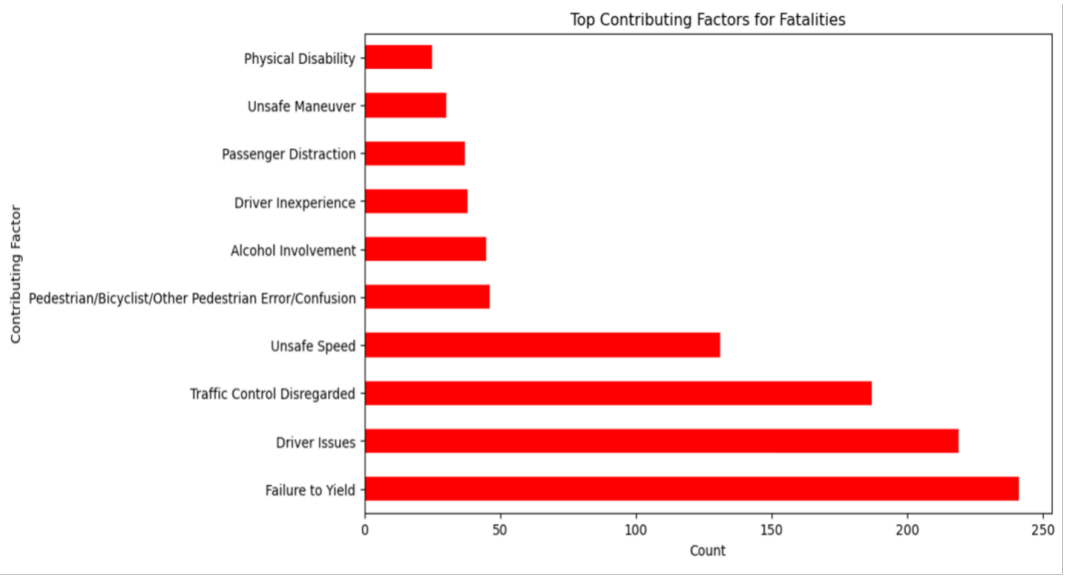
Pedestrian or bicyclist error/confusion is identified as the leading contributor to severe accidents, followed by disregarding traffic controls, and drug-related influences. Factors like tinted windows and unsafe speeds also significantly contribute to accident severity. The lower end of the spectrum includes factors such as failure to yield and glare, which, while still impactful, are less frequently associated with high severity scores.



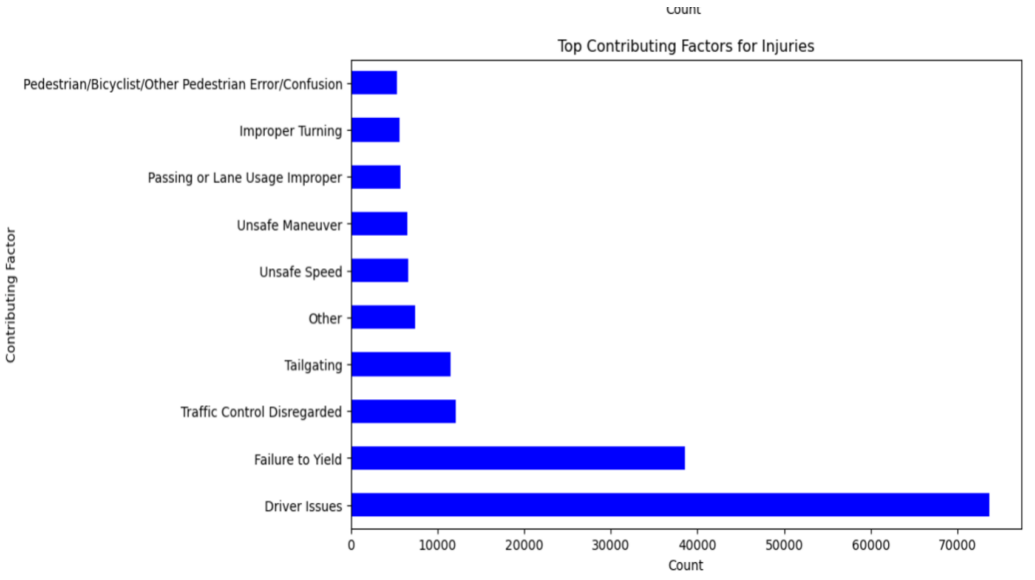
In the "Top Contributing Factors for Fatalities" chart, 'Failure to Yield' leads as the most significant contributor, followed by 'Driver Issues' and 'Traffic Control Disregarded'. These factors suggest major issues in driving behavior and adherence to traffic laws.

The "Top Contributing Factors for Injuries" chart shows 'Driver Issues' as the predominant cause, significantly outpacing other factors like 'Failure to Yield' and 'Traffic Control Disregarded'. The high frequency of 'Driver Issues' in both fatalities and injuries underscores the need for enhanced driver education and stricter enforcement of traffic laws to mitigate these risks.

Top Contributing Factors for Fatalities:

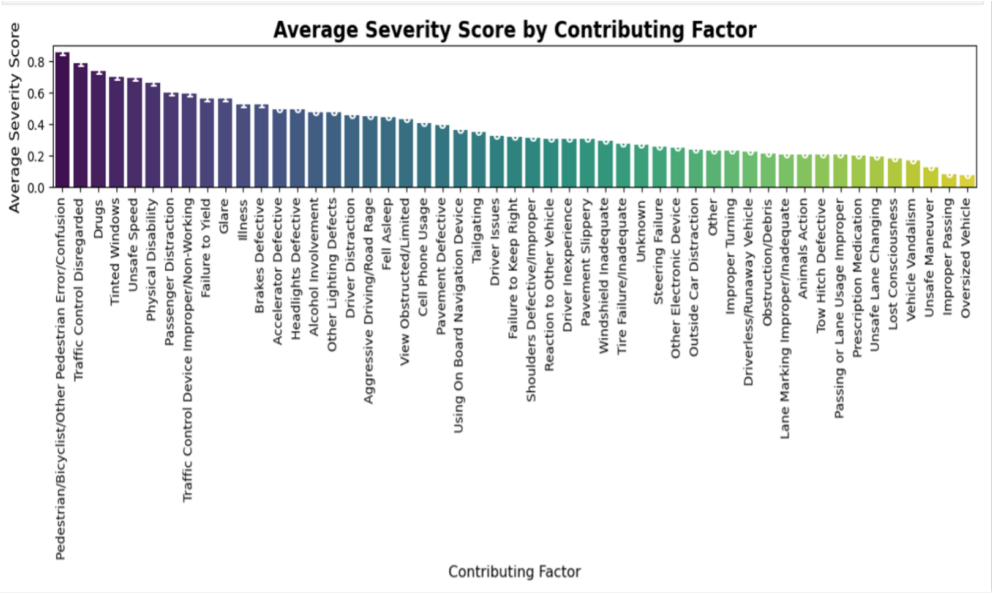


Top Contributing Factors for Injuries:



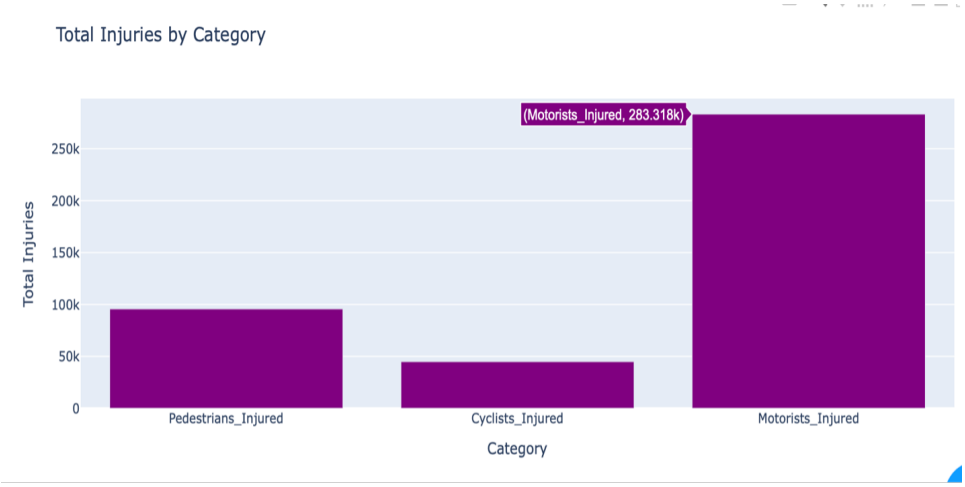
The chart displays a gradient of average severity scores for numerous factors, from pedestrian/bicyclist error/confusion, which has the highest severity score, gradually decreasing through factors like traffic control device issues, physical

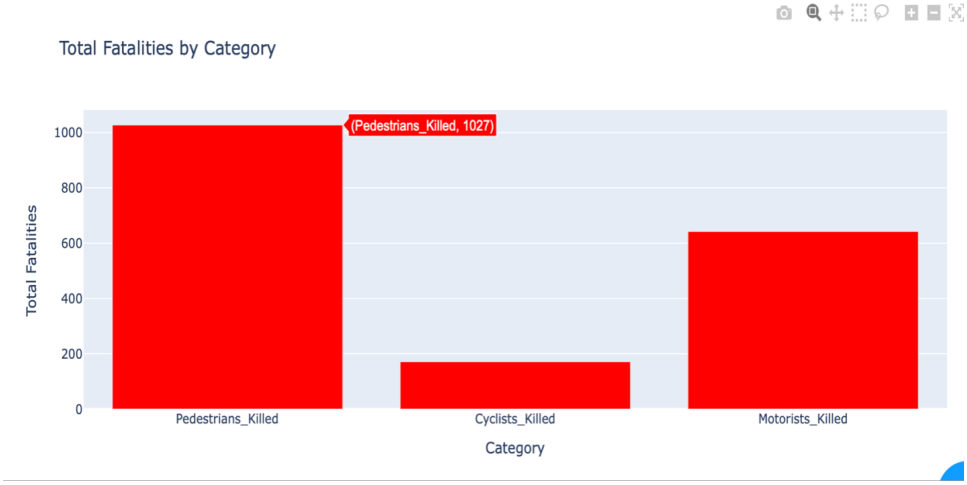
disabilities, and passenger distraction. The spectrum extends to less severe factors such as passing improperly and oversized vehicles. This comprehensive analysis helps in identifying key areas where interventions can be prioritized to reduce the severity of traffic accidents.



4. **Are certain types of road users (pedestrians, cyclists, motorists) more prone to injuries or fatalities?**

The first graph, "Total Injuries by Category," shows that motorists sustain the majority of injuries, significantly outnumbering injuries to pedestrians and cyclists. The second graph, "Total Fatalities by Category," highlights a concerning trend: pedestrians face a substantially higher fatality rate compared to cyclists and motorists. This discrepancy suggests that pedestrians are exceptionally vulnerable in traffic incidents, indicating a critical area for targeted safety interventions and policies.

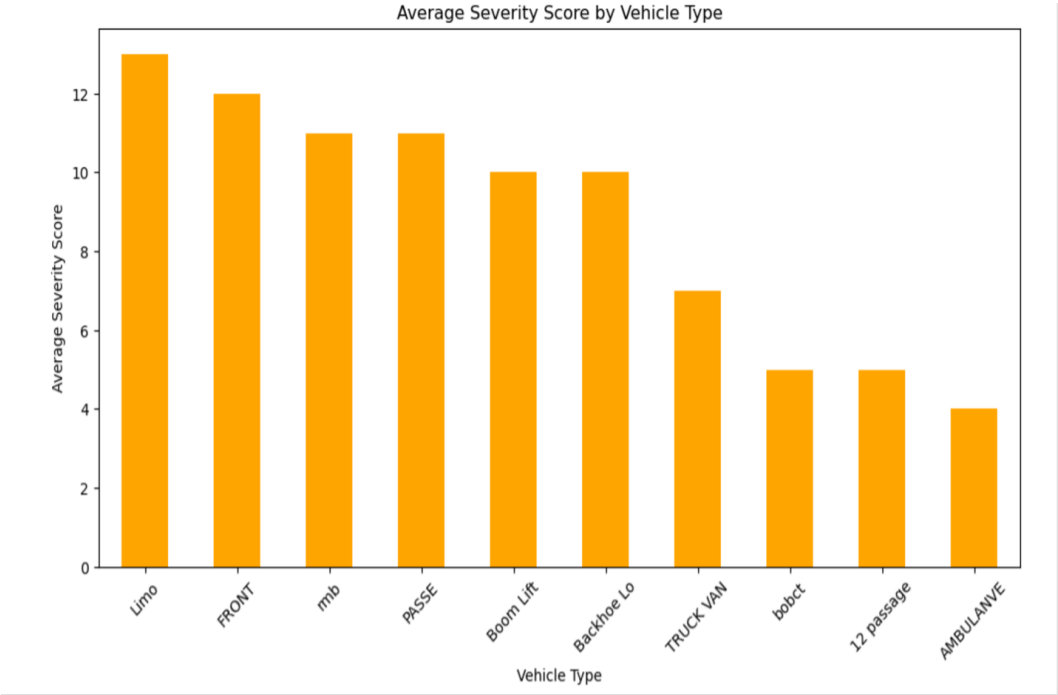




5. **How do different vehicle types and their interactions influence the likelihood of severe outcomes?**

The Relationship Between Severity and Other Factors

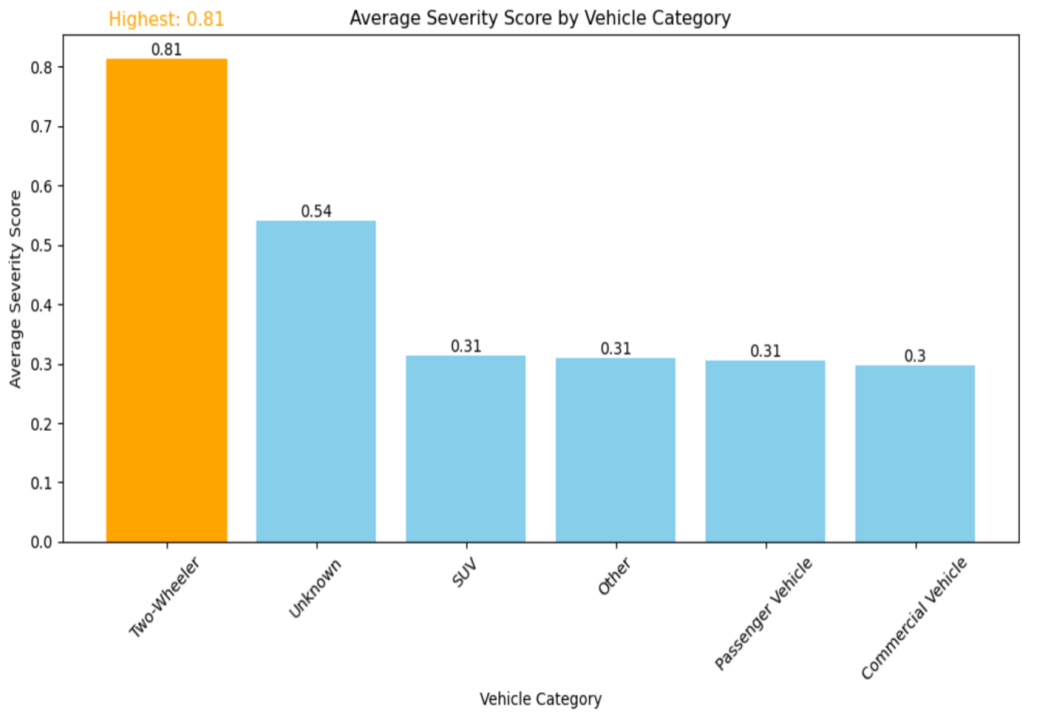
Limousines have the highest average severity score, closely followed by front and midsized vehicles. Passenger vehicles and boom lifts also report significant severity scores. Conversely, trucks, vans, boats, 12-passenger vehicles, and ambulances exhibit relatively lower severity scores. This distribution indicates that larger or specialized vehicles like limousines and boom lifts may be involved in more severe accidents, potentially due to their size and operational contexts.



Categorize vehicle types into broader categories

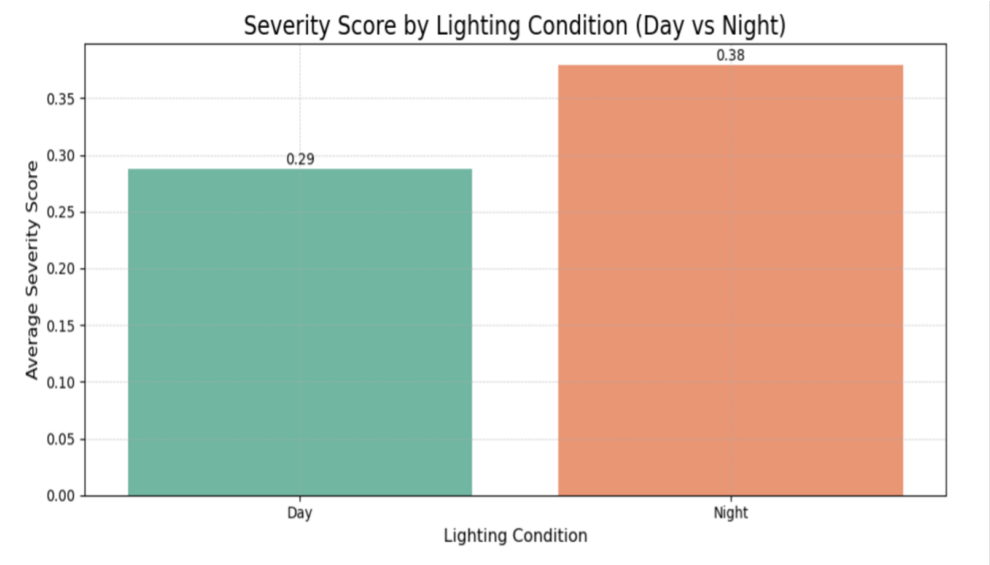
	Vehicle_Category_1	Persons_Injured	Persons_Killed	severity_score
0	Two-Wheeler	3336.0	95.0	0.813283
1	Unknown	13733.0	71.0	0.540572
2	SUV	127825.0	510.0	0.312690
3	Other	47257.0	422.0	0.310278
4	Passenger Vehicle	204553.0	609.0	0.305885
5	Commercial Vehicle	33591.0	166.0	0.296359

Two-wheelers show the highest average severity score, indicating that accidents involving motorcycles or similar vehicles tend to be more severe. This is followed by the 'Unknown' vehicle category, suggesting high variability or underreported data. SUVs also show a notable severity score. In contrast, other categories like passenger vehicles and commercial vehicles have relatively lower severity scores, indicating less severe outcomes in accidents involving these vehicles.


Highest: 0.81
Average Severity Score by Vehicle Category
Average Severity Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.81
0.54
0.31
0.31
0.31
0.3
Two-Wheeler
Unknown
SUV
Other
Passenger Vehicle
Commercial Vehicle
Vehicle Category

6. **How do external conditions, such as weather and lighting, impact the severity of collisions?**

   The score for nighttime is notably higher at 0.38 compared to 0.29 during the day. This suggests that lower visibility or other factors associated with nighttime driving contribute to more severe accidents. These findings highlight the need for enhanced safety measures and possibly greater awareness or improved lighting during nighttime driving conditions.



Data Modeling

Out of Logistic Regression, SVM and Random forest all performed very similar.

The Random Forest model produced excellent results, with nearly perfect accuracy and very high F1-scores across all categories.

The results from the Random Forest model show exceptionally high-performance metrics. The accuracy of the model is nearly perfect at 0.9966. Precision, which measures the accuracy of positive predictions, stands impressively at 0.9972 across categories, while the ROC AUC, indicating the model's ability to distinguish between classes, is at 0.9980. The F1 Score, a balance between precision and recall, is also nearly perfect at 0.9904, demonstrating the model's robustness.

```
Random Forest Model Metrics:
Accuracy: 0.9966167213653346
Precision: 0.9972381570939867
ROC AUC: 0.9980225552508997
F1 Score: 0.9904298736397305
Confusion Matrix:
[[   105      0      3      0]
 [     0  64063      0    854]
 [     0      7    540      6]
 [     0     98      0 220437]]
Classification Report:
              precision    recall  f1-score   support

        High       1.00      0.97      0.99       108
         Low       1.00      0.99      0.99     64917
    Moderate       0.99      0.98      0.99       553
        None       1.00      1.00      1.00    220535

    accuracy                           1.00    286113
   macro avg       1.00      0.98      0.99    286113
weighted avg       1.00      1.00      1.00    286113
```

The confusion matrix and classification report further illustrate the model's effectiveness. In particular, the precision and recall for all categories (High, Low, Moderate, None) are nearly perfect, reflecting very few false positives and negatives. This suggests that the model is highly reliable in predicting severity categories in traffic incidents, with substantial support (number of samples) across categories, especially in the 'None' category. This performance indicates that the model can be a critical tool in understanding and predicting traffic incident severity, aiding in more focused and effective interventions.

Model Evaluations and Findings:

Logistic Regression and SVM: Both models demonstrated exceptionally high accuracy, near or at 100%. These results, though outstanding, raise concerns about overfitting, given that perfect or near-perfect performance is rare in practical, real-world applications.

Random Forest: This model also showed near-perfect accuracy but displayed a slightly more nuanced understanding of class distinctions, especially among less frequent categories. Although misclassifications were minimal, they provided a more realistic performance scenario than the absolute scores of the other models.

Cross-Validation: Conducting k-fold cross-validation on the SVM model further confirmed the high accuracy across different data splits, with the mean accuracy consistently close to 1.00 and a very low standard deviation. This suggests strong model stability and generalizability across the data used.

Conclusion

Summary of project

1. Yearly Accident Trends:

A sharp rise in accidents was observed from 2012 to 2014, with a peak in 2014. Post-2014, accident numbers stabilized and then declined significantly after 2018, further influenced by reduced mobility due to the COVID-19 pandemic.

2. Monthly and Daily Patterns:

Accidents typically peak in the summer months and decrease towards the end of the year.

Daily trends show the highest number of accidents during morning (8 AM) and evening (5 PM) rush hours, with a notable decrease after these peak times.

3. Impact of Environmental Factors:

Nighttime driving significantly increases accident severity, likely due to reduced visibility and higher instances of impaired driving.

4. Accident Frequency vs. Severity:

Although accidents are frequent during rush hours, the severity is higher later in the evening, around 10 PM, suggesting increased risks due to factors such as fatigue and reduced visibility.

5. Trends by Borough:

Brooklyn records the highest number of collisions and severity, indicative of high traffic and potentially hazardous conditions. Staten Island, with fewer collisions, shows lower traffic volume and density.

6. Contributing Factors to Accidents:

Major contributors include driver inexperience and distraction, highlighting the need for enhanced driver education and enforcement.

Other significant factors are failure to yield and unsafe maneuvers, suggesting improvements in road design and traffic signals are necessary.

7. Demographic Impact and Vehicle Types:

Pedestrians and cyclists suffer higher severities and fatalities, emphasizing the need for focused safety measures.

Two-wheelers are involved in more severe accidents compared to other vehicles, indicating a need for targeted safety interventions for these groups.

What I Learned from This Project:

From this project, I learned the importance of data in understanding and addressing public safety issues. The detailed examination of accident data by time, lighting conditions, vehicle type, and contributing factors allows for a nuanced approach to improving traffic safety. It also underscores the role of environmental and behavioral factors in accident severity, suggesting targeted strategies for prevention and policy enhancement. This project has shown how comprehensive data analysis can lead to actionable insights for better decision-making in public safety initiatives.

Future Directions

Upon successful completion of this project, several future directions could enhance the impact and scope of the analysis:

1. **Expanding Data Sources**: Integrating additional datasets, such as real-time traffic data, weather patterns, and road conditions, would provide a more comprehensive view of the factors influencing collision severity, strengthening the predictive model and enabling more precise insights.
2. **Advanced Modeling Techniques**: Future work could explore more complex modeling approaches, including deep learning and ensemble methods, to improve the accuracy of severity classification and uncover more intricate relationships among factors.

3. **Policy and Public Safety Collaboration**: Insights gained from this project could support collaboration with policymakers and public safety advocates, informing campaigns focused on high-risk behaviors and targeted interventions in identified hotspots.

4. **Real-Time Monitoring and Alerts**: With further development, the predictive model could be adapted for real-time analysis, enabling proactive measures to mitigate risks during high-collision periods or in high-risk locations.

References:

https://ieeexplore.ieee.org/abstract/document/9672012

https://storymaps.arcgis.com/stories/7e2494cb05d547a6b01550be0b5e2ac5

https://nycdatascience.com/blog/student-works/new-york-city-motor-vehicle-collision-data-visualization/

OpenAI. ChatGPT. November 2024 version, OpenAI, 2024, www.openai.com.