



**University of North Texas**

**ADTA 5940 - Analytics Capstone Experience**

**Dr. Tony Fantasia**

**November 29, 2024**

**Weather Impact on Accident severity**

**Submitted by:**

**Cheelam, Guna Pranay Reddy: 11646048**

**Diddekunta, Siva Teja: 11655074**

**Pidamarthi, Bharath: 11657213**

**Ravi, Meghana :11716212**

**Swathi Sandela :11741222**

## Introduction:

Auto accidents are a public nuisance in highly populated cities such as New York. Snow, fog, rain, and ice are all contributory factors to the seriousness of an accident. The weather contributes to the probability of an accident because of reduced visibility, road friction, and driver reaction time. The impact of weather on motor vehicle accidents will be analyzed for New York City using the Motor Vehicle Collisions - Crashes data set. The data set is used to predict the accident severity by location, weather conditions and by time of the day. The study will also help traffic managers and municipal planners in their quest to make the roads much safer and reduce the number of accidents that relate to bad weather conditions Pisano, Goodwin, & Rossetti, 2008.

## Research Questions:

1. How could the weather conditions impact the severity of accidents in a region?

**H<sub>0</sub>:** Rain, snow, hail or storms and other severe weather conditions doesn't impact the severity of accidents.

**H<sub>a</sub>:** Rain, snow, hail or storms and other severe weather conditions increase the likelihood of severity of accidents.

**Result:** As per the analysis, during weather conditions like rain and snow have tied up with a high number of severe accidents. This confirms that bad weather is the direct cause for severe outcomes.

2. Does the severity of accidents change significantly across boroughs?

**H<sub>0</sub>:** The severity of accidents doesn't change from one borough to the other relatively.

**H<sub>a</sub>:** The severity of accidents changes from one borough to the other relatively, with some boroughs experiencing higher numbers of serious accidents and other boroughs comparatively less.

**Result:** Borough-wise distribution is showing the differences in patterns, and conclusions drawn out of it are that the areas with high traffic density may indeed have higher accident severity. For example, boroughs like Brooklyn might show higher counts, declaring further safety measures.

3. Is there any direct relation between rush hours and the higher number of major accidents?

**H<sub>0</sub>:** There is no relationship between peak hours and the frequency of major accidents.

**H<sub>a</sub>:** Major accidents are more common during peak hours

Result: The accident time analysis in our data shows that during morning and evening hours, majorly during school and office hours, accident frequency is at peak level, which supports our hypothesis that traffic congestion is directly related to the number of major accidents.

4. To what extent do the models Logistic Regression and Random Forests predict severe vs. non-severe accidents?

**H<sub>0</sub>:** Logistic Regression and Random Forest models perform equally well in predicting severe accidents.

**H<sub>a</sub>:** Random Forest performs better than Logistic Regression in predicting severe accidents due to class imbalance.

**Result:** Logistic Regression shows high accuracy of 72% for non-severe accidents but fails to identify severe ones due to imbalance. Random Forest, accuracy is 70%, and it better captured minority class patterns.

## Literature/ Industry Review:

### Current Industry trends:

Properly estimating the severity of road accidents is essential regarding highway security as well as health. For forecasting complex road conditions, considerable traffic, and changing weather, advanced models are helpful. Using machine learning models with basic algorithms to deep learning, one may forecast the severity of accidents by analyzing driving behavior, road conditions, and weather (Celik & Seveli, 2022; Zheng et al., 2018). Evaluated fourteen scholarly papers for the prediction of trends, issues, and potential advancement in accidents severity.

Recent advances in CNNs, Decision Trees, and Logistic Regression can identify the reasons behind mishaps (Çelik & Seveli, 2022; Zheng et al., 2018). Incorporating demographic and real-time meteorological data improves predictive models (Labib et al., 2019; Pińskwar et al., 2024). It's difficult to align the changes in fields in the world (Becker et al., 2022). The research on accident

severity prediction algorithms for road safety is surveyed in this article, along with opportunities for improvement and challenges.

### Machine Learning Techniques for Accident Severity:

Machine learning is used for examining large datasets to detect ways in accident severity because it can support difficult attribute communications (Çelik & Seveli, 2022). (Çelik and Seveli 2022) explored the use of machine learning models (ML), such as XGBoost, Random Forest, K-Nearest Neighbors, and SVM, to classify traffic incidents based on their intensity. The most accurate model they had was Logistic Regression, which showed that even basic models may identify elements that contribute to accident severity, with an accuracy rate of 88% (Çelik & Seveli, 2022).

(Çelik and Seveli 2022) explored group models like XGBoost and Random Forest that productively manage huge datasets and non-linear communications, instead the fact that logistic regression is a direct model. The assumption of these models recommends a increasing trend towards the functioning of difficult machine learning techniques in traffic safety evaluations, although they did not exceed logistic regression in this study (Çelik & Seveli, 2022). Ensemble techniques like Random Forest enhance models by integrating the strengths of decision trees (Labib et al., 2019). The work focuses on testing various machine learning algorithms to identify the optimum dataset-aware accident severity prediction method (Çelik & Seveli, 2022).

### Comparative Analysis of Machine Learning Models:

The road accident data gathered throughout the world helped researchers to design better crash severity prediction methods, by using huge training datasets and test datasets on various machine learning models (Malik et al., 2021). In this paper six different algorithms and their accuracy results were compared for predicting accident severity and analyzed by (Malik et al. 2021): Random Forest, Decision Tree, Logistic Regression, KNN, SVM, and Naïve Bayes. These papers concluded that based on the analysis, Random Forest and Decision Tree techniques outperformed predictions on accident severity evaluations (Malik et al., 2021).

Random Forest and decision trees decreased overfitting and increased accuracy (Malik et al., 2021). For the study of accident data on big datasets with lots of variables, this approach is perfect (Malik et al., 2021). Input variables and catastrophic consequences are explained using decision trees (Malik et al., 2021). Weather data and features were added to the accident dataset, attributed to improving severity forecasts.

(Behboudi et al. 2024) paper has its focus on improving existing machine learning algorithms for better accident severity predictions. Model predictions have improved by including weather data, traffic signals data in cities, and roads layout details (Behboudi et al., 2024). Training with multiple datasets help models in comprehending the severity of traffic accidents (Behboudi et al., 2024). (Behboudi et al. 2024) paper has also discussed that big datasets and the integration of multiple data sources are crucial for better predictions.

### Effect of Weather Conditions on Accident Severity:

Weather has a significant influence on the frequency of accidents and their severity (Pisano et al., 2008). According to Pisano et al., weather influences or causes about a quarter (24%) of accidents. This large share of a quarter indicates that driving is highly impacted by weather (Pisano et al., 2008). Conditions related to weather like visibility, traction control, and vehicle operations add and increase risk and severity of the accident (Pisano et al., 2008). Using real-time data of weather in algorithms for severity prediction can help in improving road safety measures and risk assessment (Pisano et al., 2008).

In another kind of research, (Pińskwar et al., 2024), he investigated the effect of high or low temperatures and wind speed or air pressure on road accidents in Poland. This research clearly suggests that good weather conditions have a great impact on safer driving conditions and road safety of the drivers and passengers (Pińskwar et al., 2024). So, to increase the accuracy of the models he suggested to include the changes or fluctuations in temperatures and air pressure for better results (Pińskwar et al. 2024).

## METHODS:

### Data Collection:

Data is collected from the source below:

<https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes>

This dataset details New York City car accidents. NYCDOT, NYPD, are the principal sources of data. The dataset is updated regularly, ensuring its relevance for ongoing research. The data include detailed information about each collision, helping researchers understand a variety of factors that cause road accidents. This dataset is notional for current metrics, Model prediction, and landscape study of city traffic safety.

### Key Variables include:

**Location Information:** Latitude, longitude, borough, and street intersections are included in each entry. This allows detailed spatial analysis of NYC accident hotspots.

**Time and Date:** The dataset records every crash's exact time and date. This data lets researchers examine accident patterns by time of day, day of the week, or season to determine when they occur most often.

**Contributing Factors:** The data shows human and external factors that cause crashes. Distracted driving, speeding, alcohol use, and road and weather conditions are among these risks. Investigations into accident causes and prevention techniques require this substantial data collecting.

**Crash Severity:** It reports crash intensity, including injuries and deaths. This includes event-related pedestrian, motorist, and cycle counts. Researchers can examine contributing elements to determine significant accident likelihood using severity data.

**Vehicle Information:** The number and types of vehicles in each occurrence allow analysis of how truck, car, and bicycle types affect crash results.

**Data Pre-processing:**

- Imputed missing borough and weather\_description values with the most frequent value, and number\_of\_injuries and number\_of\_deaths with 0 (assuming no data implies no injuries/deaths)
- Created a binary target variable, serious accident, to classify accidents as serious (1) or non-serious (0).
- Defined a serious accident as one with either injuries (number\_of\_injuries > 0) or fatalities (number\_of\_deaths > 0).
- Converted crash\_time to a datetime format to allow grouping by time ranges (e.g., Morning Rush, Midday, Evening Rush, etc.).
- Removed entries in weather\_description that were labeled as "Unspecified" or any other non-informative descriptions, ensuring only meaningful weather data is included.
- Encoded categorical variables borough, weather\_description, and crash\_time\_period to prepare the data for machine learning models.
- Removed columns that were unnecessary or columns which have unstructured data because it is difficult to process in modeling, such as crash\_date and contributing\_factor\_vehicles.
- Ensured that all columns used in modeling were numerical after encoding, as machine learning models require numeric input.

**Model Selection:**

Models used for analysis are Logistic Regression and Random Forest model.

**Why did we choose Logistic Regression and Random Forests models?**

For our project we chose logistic regression model, as our dependent variable is binary, the outcome is to predict whether the accident is (severe/non-severe). As our data is highly linearly separable with the least overlapping of datapoints. And this was our most efficient choice. But still we went ahead and chose Random Forests as we have time series data and to know get the insights of each feature in our data. To understand which feature has contributed more to our predictions, as well as each feature importance score. In both the

model's Logistic regression has outperformed well as our data is linearly separable, and Random Forests helped us in better understanding of features in our data.

#### Logistic Regression Model:

A framework for predicting a binary outcome that fits perfectly for the classification of accidents into a serious and non-serious, which are easy to understand and interpret the specific circumstances-like weather, time of day – impact accident severity. The model's simplicity and computational efficiency make it a valuable baseline for binary classification tasks. However, performance can be limited if the data has complex, non-linear patterns or interactions. In addition, as we have seen in our analysis, it is not a serious situation at the big event, but it is like fighting with unequal information, which mostly supports the majority class, which becomes important.

#### Random Forest Model:

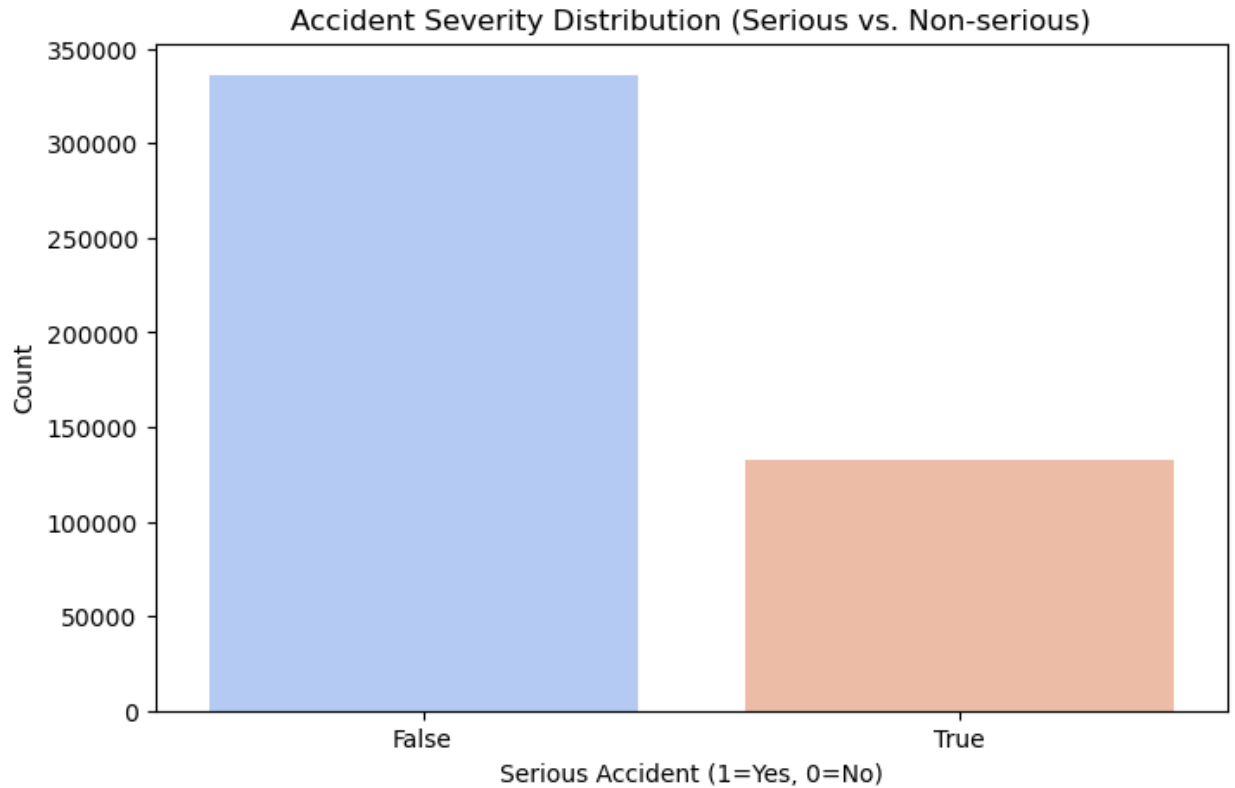
Random Forest is an ensemble learning model that builds multiple decision trees using random subsets of the data, combining their outputs to improve predictive accuracy and capture complex, non-linear relationships within the data. This makes it ideal for datasets where accident severity may be influenced by intricate interactions between variables, like weather, time, and location. Random Forest handles class imbalance more effectively by considering the minority class more robustly, which is particularly beneficial in distinguishing between serious and non-serious accidents. It also provides a critical score that allows us to determine which variables have the greatest impact on the severity of the accident. While random forests are no less expressive than transportation, the ability of random forests to capture relationships makes them a strong contender for accuracy in this assessment.



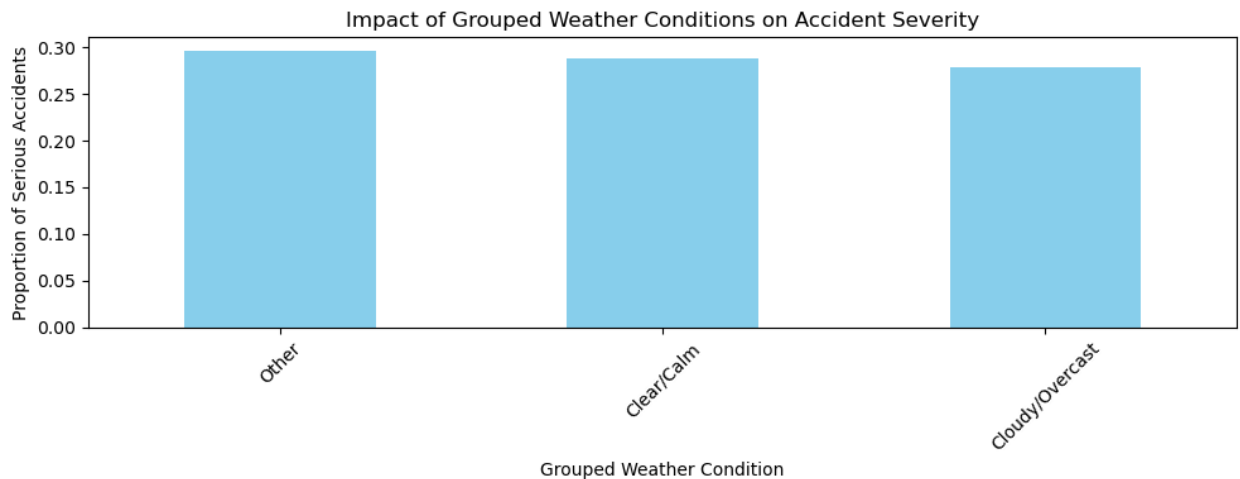
## Results:

### Exploratory Data Analysis (EDA):

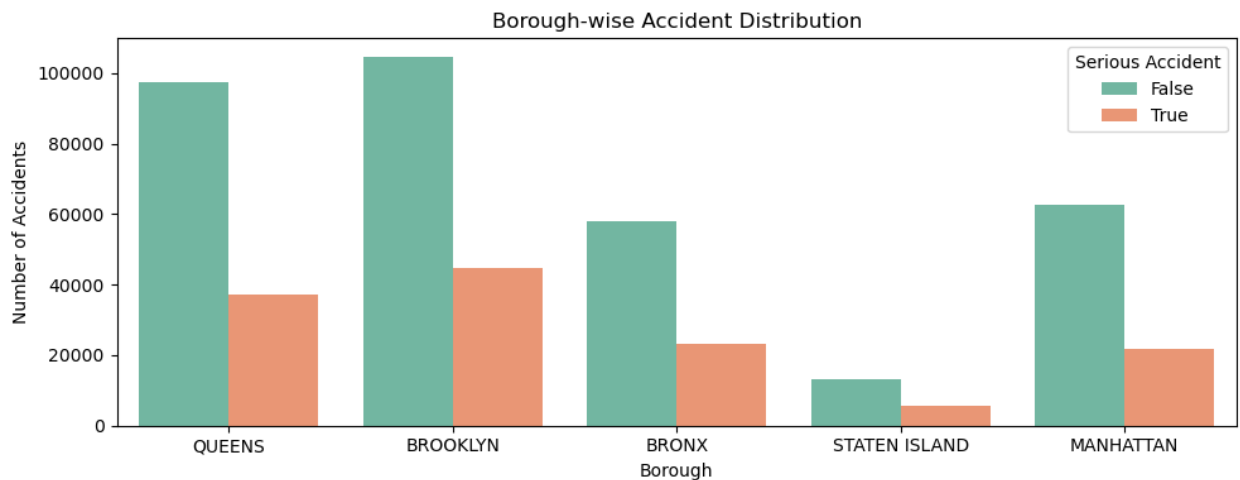
#### Accident Severity Distribution:



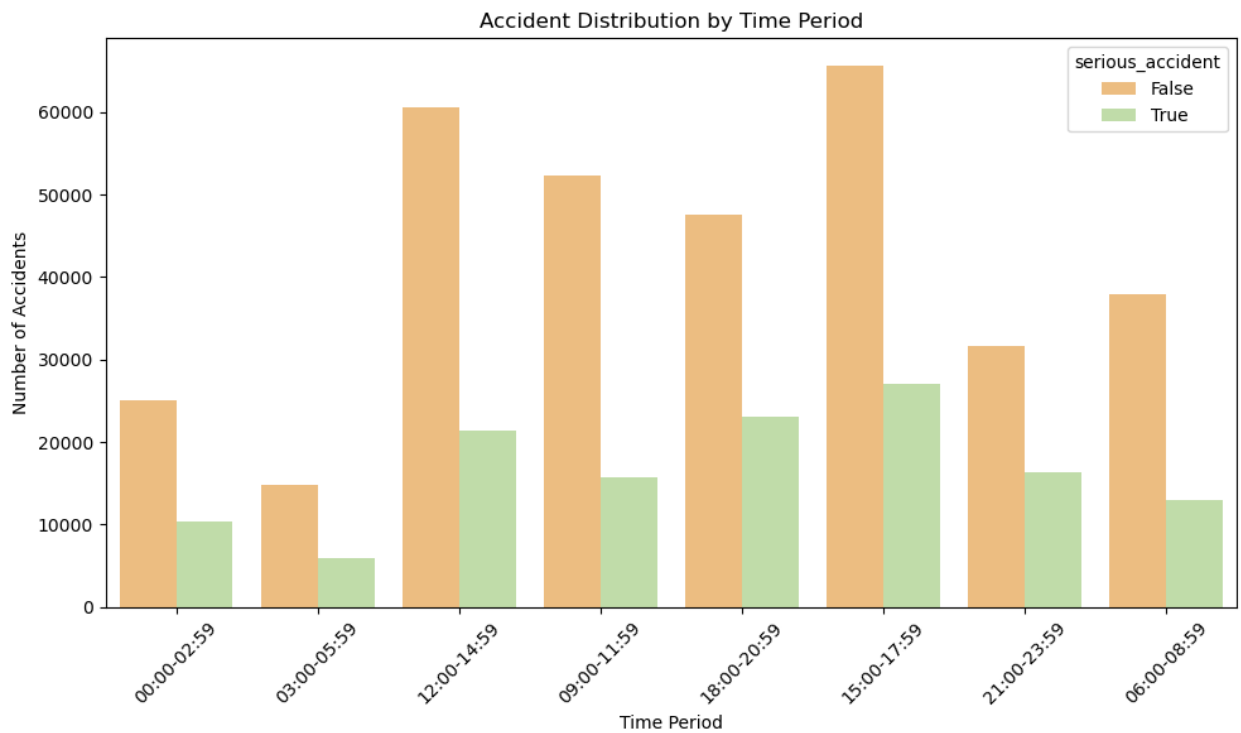
Above Graph shows the distribution of accident severity. It describes the number of serious accidents is less than the non-serious accidents and it indicates the most accidents do not result in serious injuries.

**Weather Condition Impact on Accidents:**

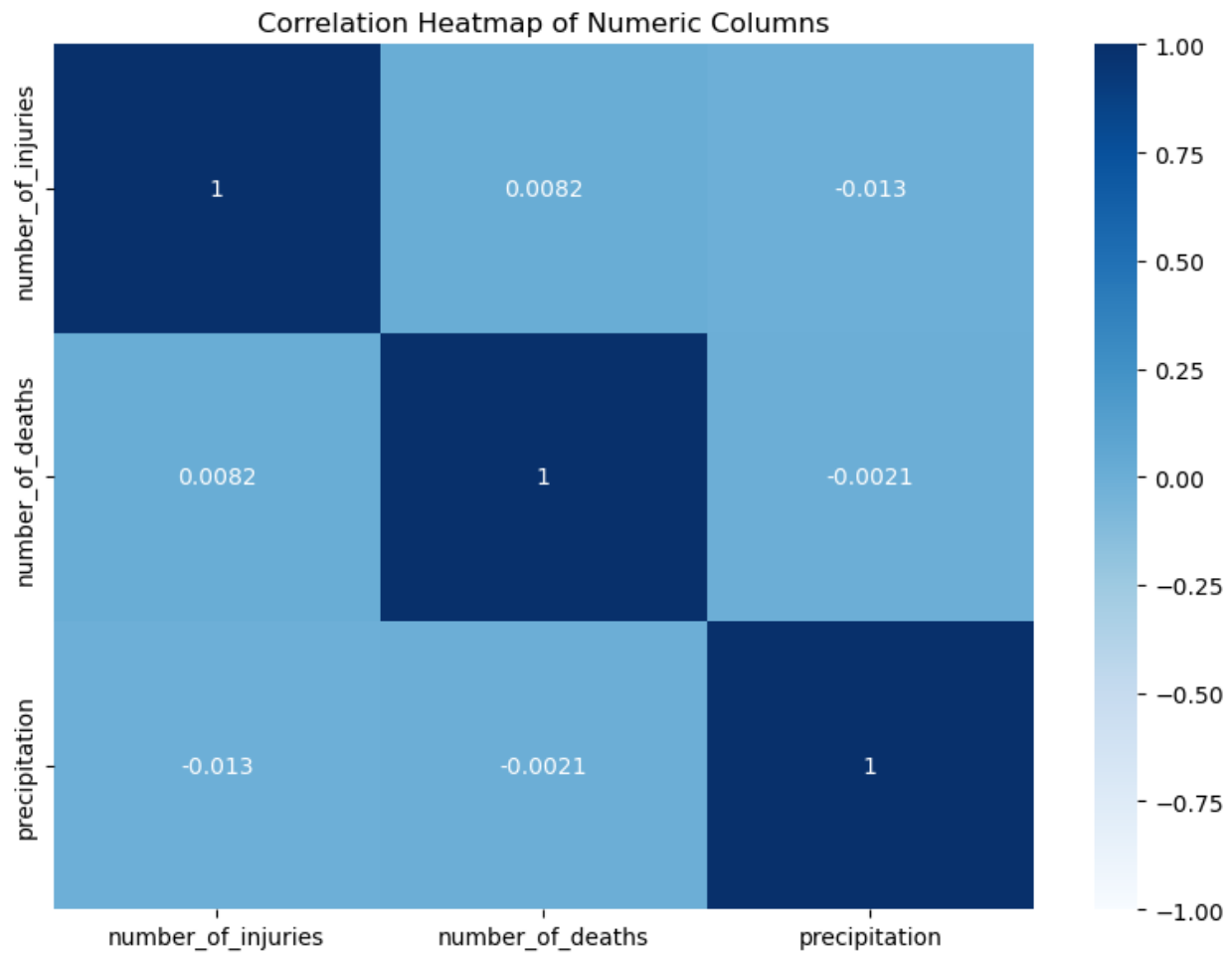
As per the graph there no difference in variation in the proportion of serious accidents across the weather conditions in the map This implies that factors other than weather, such driving habits or road conditions, can be quite important in determining how serious accidents are in this dataset.

**Borough-wise Accident Distribution**

The Above Graph clearly explains the accident frequency in each borough, broken down by severity. For the improvement of traffic safety there needs to be serious attention required to avoid the high number of serious accidents. Let's say for example, most accidents happened in Brooklyn because of higher traffic density/ road conditions that affect the number of accidents rate. Variations among boroughs may indicate the necessity for specific interventions or tailored policies aimed at enhancing road safety.

**Accident Distribution by Time Period:**

This graph illustrates the frequency of accidents over time, highlighting accident hotspots. In peak hours (morning and evening) with huge accidents, it will give the idea that congestion and higher traffic volumes during these times. Illustrating severe and non-severe accidents over time can also help conclude whether those hours are more lead to serious accidents, which could suggest measures such as increased patrols or road safety campaigns during high-risk hours.

**Correlation Heatmap:**

Heat map showing the correlation between numerical variables:

Number of injuries and number of deaths shows that there is a high positive correlation between these two and simulating more injuries likely to be occurred because of these accidents. This shows a relation between injury count and severity. In a linear way it shows that the precipitation does not directly influence accident severity, but it may indirectly affect the factors that influence the visibility of road slipperiness.

## Interpretations

Based on our results we have interpreted that bad weather and bad road constructions increase severity in accidents. The differing patterns from one borough to the other, tells that road constructions, steep angles, position of traffic signals majorly contribute to higher accident rates in specific regions. Lastly, peak hour traffic increases accident risks, based on the driver's state of mind whether alcoholic or not, visually not impaired, intensity of fog, slippery conditions on the roads due to snow. Logistic Regression model has outperformed in predictions of severity accidents considering the attributes like weather conditions, time, road construction, traffic signals, driver's behavior.

## Summary:

This study majorly focuses on important factors leading to major accidents in New York city. Mainly concentrating on bad weather, road construction, steep angles and peak hours. However, limitations include a lack of real-time traffic data and limited details in weather conditions. Future studies could incorporate additional variables, like real-time road monitoring and driver behavior data, to improve prediction accuracy. Despite these limitations, the findings provide a robust foundation for implementing safety measures in high-risk urban areas in New York.

## References:

- Çelik, A., & Sevlı, O. (2022). Predicting traffic accident severity using machine learning techniques. *Türk Doğa ve Fen Dergisi*, 11(3), 79-83.
- Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, 90, 128-139.
- Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques. *arXiv preprint arXiv:2406.13968*.

- Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019, June). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In *2019 7th international conference on smart computing & communications (ICSCC)* (pp. 1-5). IEEE.
- Malik, S., El Sayed, H., Khan, M. A., & Khan, M. J. (2021, December). Road accident severity prediction—a comparative analysis of machine learning algorithms. In *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)* (pp. 69-74). IEEE.
- Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., ... & Wang, Z. (2019). Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access*, 7, 39897-39910.
- Pińskwar, I., Choryński, A., & Graczyk, D. (2024). Good weather for a ride (or not?): how weather conditions impact road accidents—a case study from Wielkopolska (Poland). *International journal of biometeorology*, 68(2), 317-331.
- Nazif-Munoz, J. I., Martínez, P., Williams, A., & Spengler, J. (2021). The risks of warm nights and wet days in the context of climate change: assessing road safety outcomes in Boston, USA and Santo Domingo, Dominican Republic. *Injury epidemiology*, 8, 1-9.
- Eisenberg, D., & Warner, K. E. (2005). Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. *American journal of public health*, 95(1), 120-124.

**Note:**

- Used suggestions from auto-correction tool in word.
- Referred ChatGPT to get syntaxes for analysis part.

# Final\_Project\_code\_group5

December 2, 2024

## 0.1 Importing required packages:

```
[1]: import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,
    accuracy_score
import matplotlib.pyplot as plt
```

```
C:\Users\megha\anaconda\Lib\site-packages\pandas\core\arrays\masked.py:60:
UserWarning: Pandas requires version '1.3.6' or newer of 'bottleneck' (version
'1.3.5' currently installed).
from pandas.core import (
```

## 0.2 Loading Dataset

```
[2]: accident_data = pd.read_csv('C:\\Users\\megha\\OneDrive - UNT\\
    System\\Desktop\\Capstone_project\\Capstone_finaldataset.csv', encoding='latin1')
```

```
[3]: print(accident_data.columns)
```

```
Index(['crash_date', 'borough', 'zip_code', 'latitude', 'longitude',
      'collision_id', 'crash_time_period', 'contributing_factor_vehicles',
      'vehicle_types', 'number_of_injuries', 'number_of_deaths',
      'street_name', 'street_type', 'weather_description', 'precipitation',
      'precipitation_type', 'temp_max', 'temp_min'],
      dtype='object')
```

## 0.3 Data Cleaning

### 0.3.1 1) Formatting crash\_date to datetime

```
[4]: accident_data['crash_date'] = pd.to_datetime(accident_data['crash_date'],
    errors='coerce')
```

### 0.3.2 2) Handling missing values in columns (weather, severity, location)

#### 1) Imputing missing 'borough' values with the most frequent value

```
[5]: from sklearn.impute import SimpleImputer

borough_imputer = SimpleImputer(strategy='most_frequent')
accident_data['borough'] = borough_imputer.
↳fit_transform(accident_data[['borough']]).ravel()
```

## 2)Imputing missing ‘weather\_description’ values with the most frequent value

```
[6]: weather_imputer = SimpleImputer(strategy='most_frequent')
accident_data['weather_description'] = weather_imputer.
↳fit_transform(accident_data[['weather_description']]).ravel()
```

## 3)Impute missing ‘number\_of\_injuries’ and ‘number\_of\_deaths’ with 0 (assuming no data implies no injuries/deaths)

```
[7]: injuries_deaths_imputer = SimpleImputer(strategy='constant', fill_value=0)
accident_data[['number_of_injuries', 'number_of_deaths']] =_
↳injuries_deaths_imputer.fit_transform(
    accident_data[['number_of_injuries', 'number_of_deaths']]
)
```

## 4)Filter rows to exclude ‘Unspecified’ weather\_description and create a binary target variable

```
[8]: accident_data_cleaned = accident_data[~accident_data['weather_description'].str.
↳contains('Unspecified', na=False)].copy()
```

## 5)Create a binary target variable for accident severity (1: Serious accident, 0: Not serious)

```
[9]: accident_data_cleaned['serious_accident'] =_
↳(accident_data_cleaned['number_of_injuries'] > 0) |_
↳(accident_data_cleaned['number_of_deaths'] > 0)
```

## 0.3.3 3)Numerical summary of the key columns

```
[10]: accident_data_cleaned.describe()
```

```
[10]:
```

	crash_date	zip_code	latitude \
count	504520	472235.000000	463077.000000
mean	2020-07-19 23:05:34.397843712	10877.324584	40.483561
min	2019-01-01 00:00:00	7002.000000	0.000000
25%	2019-08-02 00:00:00	10454.000000	40.667114
50%	2020-04-13 00:00:00	11208.000000	40.717648
75%	2021-07-05 00:00:00	11354.000000	40.780150
max	2024-01-27 00:00:00	11697.000000	40.912884
std	NaN	542.606136	3.134080



	longitude	collision_id	number_of_injuries	number_of_deaths	\
count	463077.000000	5.045200e+05	504520.000000	504520.000000	
mean	-73.473821	4.309164e+06	0.754388	0.003839	
min	-74.254845	3.822228e+06	0.000000	0.000000	
25%	-73.962960	4.182927e+06	0.000000	0.000000	
50%	-73.919235	4.309168e+06	0.000000	0.000000	
75%	-73.863010	4.435342e+06	2.000000	0.000000	
max	0.000000	4.698710e+06	80.000000	8.000000	
std	5.686799	1.458659e+05	1.483648	0.091582	

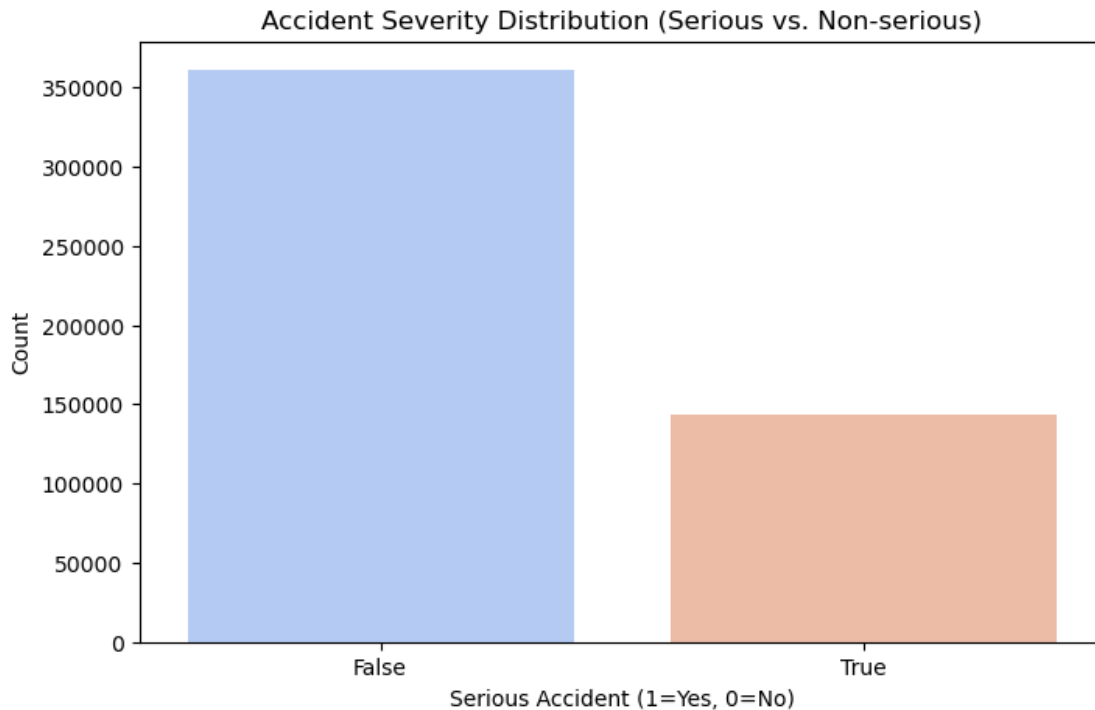
  

	precipitation	temp_max	temp_min
count	504492.000000	504492.000000	504492.000000
mean	1.780470	18.348564	10.952365
min	0.000000	-7.300000	-14.500000
25%	0.000000	9.400000	3.300000
50%	0.230000	19.200000	11.200000
75%	1.168000	27.100000	19.300000
max	71.630000	36.700000	26.900000
std	4.515698	9.929415	9.153330

## 0.4 Exploratory Data Analysis(EDA):

### 0.4.1 Accident Severity Distribution-shows the distribution of accidents by severity.

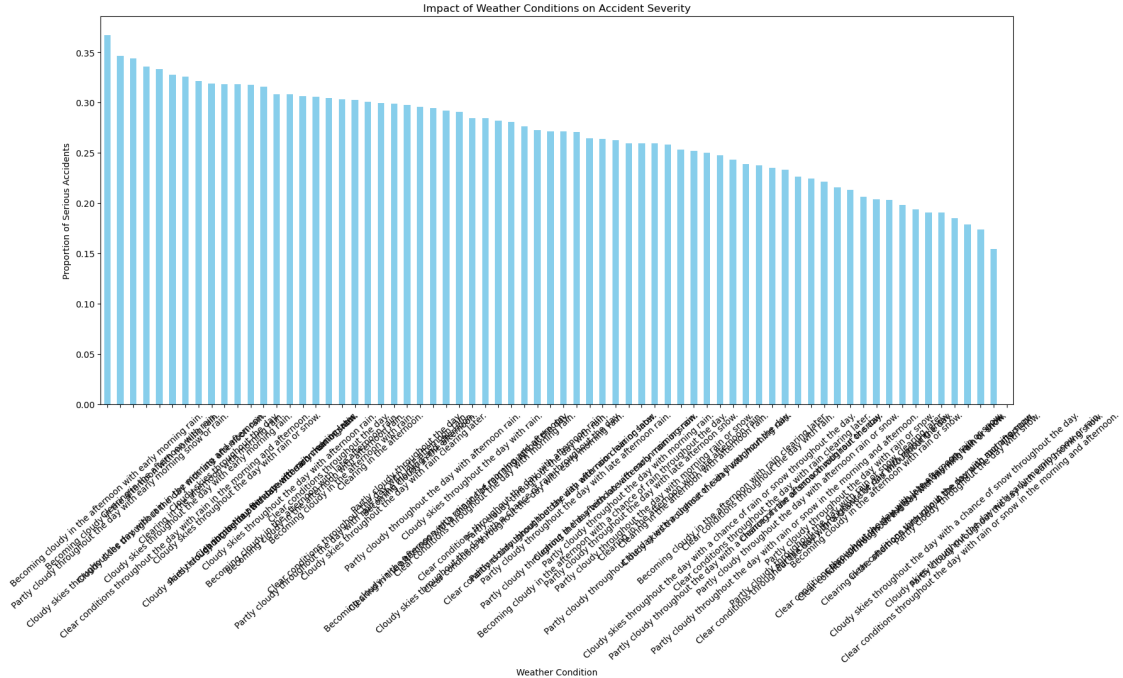
```
[11]: import seaborn as sns
plt.figure(figsize=(8, 5))
sns.countplot(x='serious_accident', data=accident_data_cleaned,
             palette='coolwarm')
plt.title('Accident Severity Distribution (Serious vs. Non-serious)')
plt.xlabel('Serious Accident (1=Yes, 0=No)')
plt.ylabel('Count')
plt.show()
```



If the count of serious accidents (injuries or fatalities) is significantly lower than non-serious accidents, it indicates that most accidents do not result in serious injuries or fatalities. This distribution is typical for accident datasets and suggests that further analysis or model adjustments might be necessary to account for the imbalance if predictive modeling is intended.

**Weather Condition Impact on Accidents** helps identifying which weather conditions lead to more severe accidents.

```
[12]: plt.figure(figsize=(18, 11))
weather_impact = accident_data_cleaned.
    ↳groupby('weather_description')['serious_accident'].mean().
    ↳sort_values(ascending=False)
weather_impact.plot(kind='bar', color='skyblue')
plt.title('Impact of Weather Conditions on Accident Severity')
plt.xlabel('Weather Condition')
plt.ylabel('Proportion of Serious Accidents')
plt.xticks(rotation=42)
plt.tight_layout()
plt.show()
```



Higher bars indicate weather conditions where a greater proportion of accidents are severe. For example, if conditions like heavy rain or snow show a higher proportion of serious accidents, this would indicate that such conditions pose a higher risk. Conversely, if clear weather has a low proportion, it could mean that these conditions are safer or lead to fewer severe outcomes.

```
[13]: # Define a dictionary to map detailed descriptions to the grouped categories
grouped_weather_mapping = {
    'Clear/Calm': ['Clear', 'Clear throughout the day', 'Clear sky', 'Sunny'],
    'Cloudy/Overcast': ['Cloudy', 'Partly cloudy', 'Mostly cloudy', 'Overcast'],
    'Rainy': ['Rain', 'Light rain', 'Heavy rain', 'Showers', 'Rain throughout the day'],
    'Snowy/Icy': ['Snow', 'Light snow', 'Heavy snow', 'Snow showers', 'Sleet', 'Icy conditions'],
    'Foggy/Hazy': ['Fog', 'Haze', 'Mist', 'Foggy conditions'],
    'Windy/Stormy': ['Windy', 'Stormy', 'Thunderstorms', 'High wind']
}

# Function to assign each weather description to a grouped category
def assign_grouped_weather(description):
    for group, keywords in grouped_weather_mapping.items():
        if any(keyword in description for keyword in keywords):
            return group
    return 'Other' # Assign 'Other' for descriptions that don't fit any group

# Apply the grouping function
```

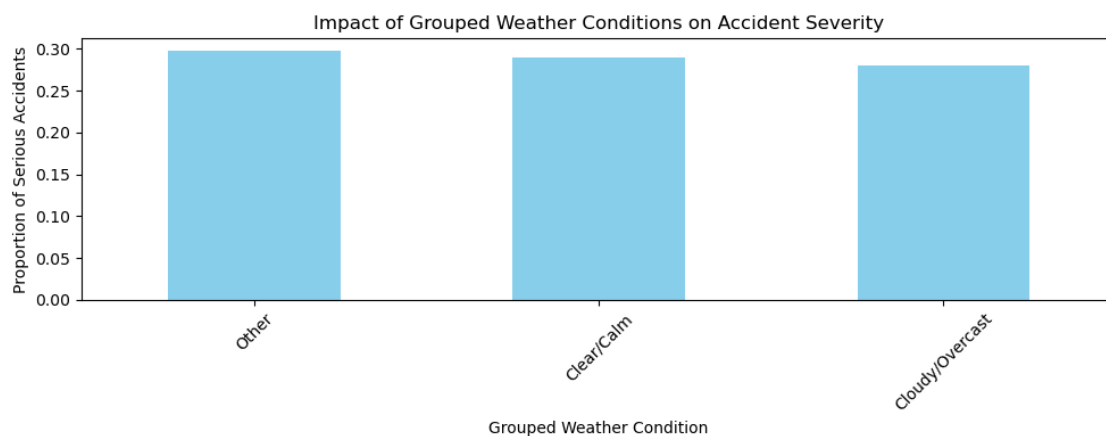
```

accident_data_cleaned['grouped_weather'] =
    ↪accident_data_cleaned['weather_description'].apply(assign_grouped_weather)

# Calculate the proportion of serious accidents by grouped weather condition
grouped_weather_impact = accident_data_cleaned.
    ↪groupby('grouped_weather')['serious_accident'].mean().
    ↪sort_values(ascending=False)

# Plot the grouped weather impact on accident severity
plt.figure(figsize=(10, 4))
grouped_weather_impact.plot(kind='bar', color='skyblue')
plt.title('Impact of Grouped Weather Conditions on Accident Severity')
plt.xlabel('Grouped Weather Condition')
plt.ylabel('Proportion of Serious Accidents')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

```

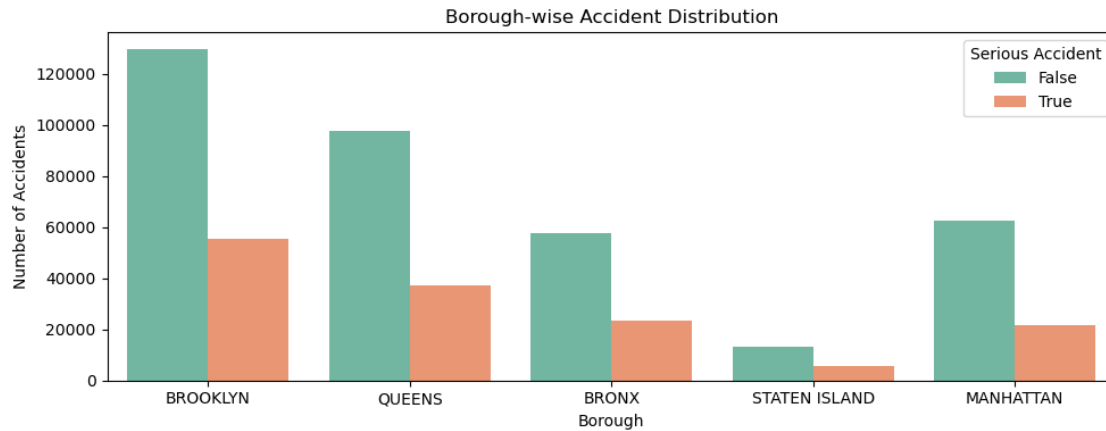


Borough-wise Accident Distribution-Count plot of accidents for each borough, segmented by severity (serious or not serious)

```

[14]: import seaborn as sns
plt.figure(figsize=(10, 4))
sns.countplot(x='borough', hue='serious_accident', data=accident_data_cleaned,
    ↪palette='Set2')
plt.title('Borough-wise Accident Distribution')
plt.xlabel('Borough')
plt.ylabel('Number of Accidents')
plt.legend(title='Serious Accident')
plt.tight_layout()
plt.show()

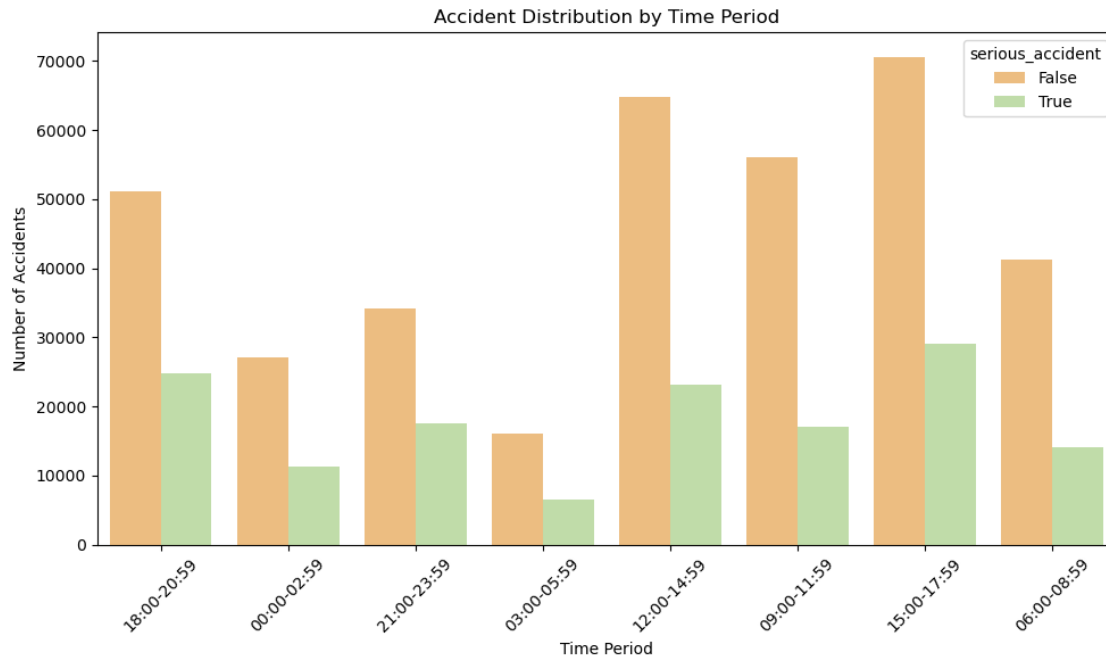
```



This graph provides insights into the accident frequency in each borough, broken down by severity. A borough with high counts of serious accidents may require closer attention for traffic safety improvements. For instance, if Brooklyn has the most accidents, it might indicate higher traffic density or specific road conditions contributing to accident rates. Differences between boroughs might suggest the need for targeted interventions or localized policies to improve road safety.

**Time Period Analysis**-shows how accidents are distributed across different time periods (rush hours, off-hours)

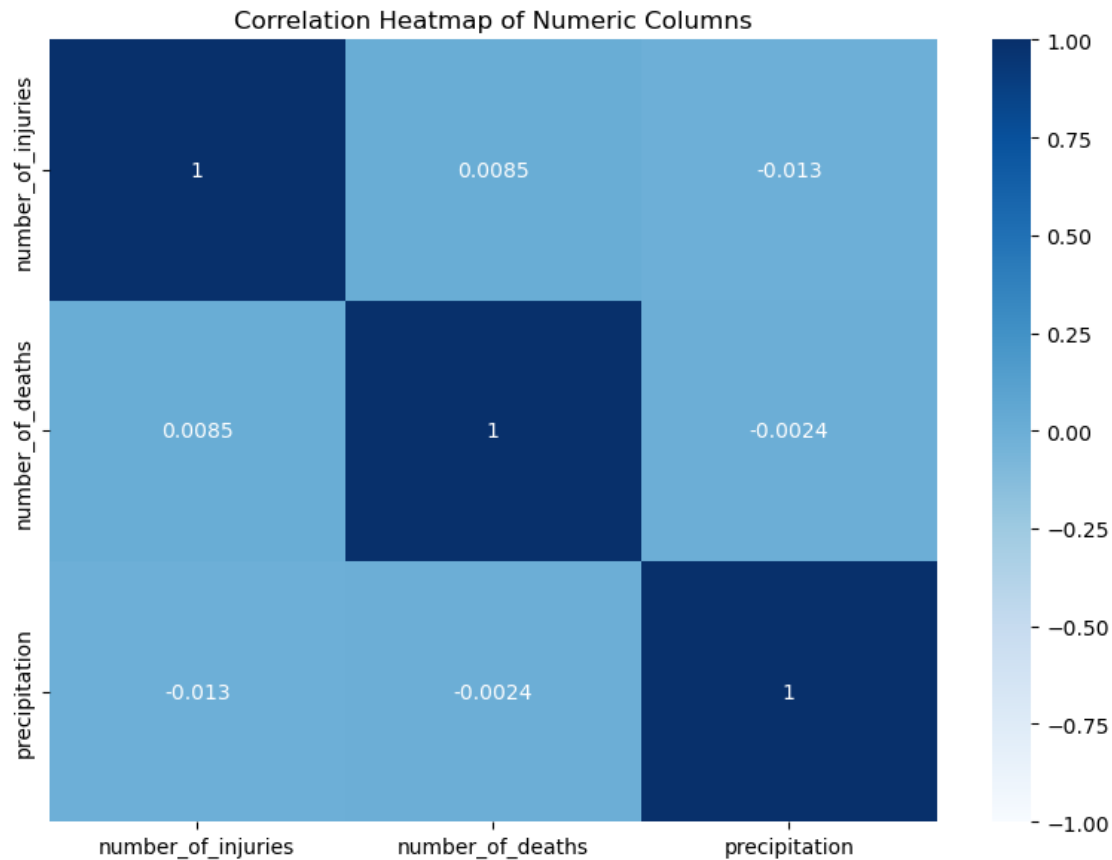
```
[15]: plt.figure(figsize=(10, 6))
sns.countplot(x='crash_time_period', hue='serious_accident',
             data=accident_data_cleaned, palette='Spectral')
plt.title('Accident Distribution by Time Period')
plt.xlabel('Time Period')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



This chart shows accident frequency across different time periods, indicating peak times for accidents. If rush hours (e.g., morning and evening peaks) have more accidents, it suggests that congestion and higher traffic volumes during these times contribute to accident rates. Comparing severe and non-severe accidents by time period can also help in understanding whether certain hours are more prone to serious accidents, potentially guiding measures like increased patrols or road safety campaigns during high-risk hours.

**Correlation Heatmap (for numeric columns)**-shows the correlation between numeric columns like `number_of_injuries`, `number_of_deaths`, and `precipitation`.

```
[16]: plt.figure(figsize=(8, 6))
sns.heatmap(
    accident_data_cleaned[['number_of_injuries', 'number_of_deaths', 'precipitation']].corr(),
    annot=True, cmap='Blues', vmin=-1, vmax=1)
plt.title('Correlation Heatmap of Numeric Columns')
plt.tight_layout()
plt.show()
```



The correlation heatmap shows the relationships between numerical variables:

A high positive correlation between number\_of\_injuries and number\_of\_deaths suggests that accidents with more injuries are likely to have fatalities, too, indicating a relationship between injury count and severity.

Weak or no correlation between precipitation and injury/death counts suggests that precipitation alone does not directly influence accident severity in a linear way, though it may still have an indirect impact through other factors like visibility or road slipperiness.

## 0.5 Step-1

### 0.5.1 Extract relevant features and drop 'contributing\_factor\_vehicles' since it contains problematic strings

```
[17]: accident_data_cleaned = accident_data_cleaned[['crash_date', 'borough', 'crash_time_period', 'weather_description', 'precipitation', 'serious_accident']]
```

### 0.5.2 Encode categorical columns: borough, weather\_description, and crash\_time\_period

```
[18]: accident_data_cleaned_encoded = pd.get_dummies(accident_data_cleaned,
↳ columns=['borough', 'weather_description', 'crash_time_period'])
```

### 0.5.3 Drop the datetime column (crash\_date) because it is not numeric and can't be used in Logistic Regression

```
[19]: accident_data_cleaned_encoded = accident_data_cleaned_encoded.
↳ drop(columns=['crash_date'])
```

### 0.5.4 Split the dataset into features and target

```
[20]: X = accident_data_cleaned_encoded.drop(columns=['serious_accident'])
y = accident_data_cleaned_encoded['serious_accident']
```

## 0.6 Logistic Regression

### 0.6.1 Step 2: Train-Test Split

```
[21]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
↳ random_state=42)
```

## 0.7 Step 3: Logistic Regression Model

### 1) Impute Missing Values:

```
[31]: from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='mean') # You can also use 'median' or
↳ 'most_frequent'
X_train = imputer.fit_transform(X_train)
X_test = imputer.fit_transform(X_test)
```

### 2) Train a Logistic Regression classifier

```
[27]: log_reg = LogisticRegression(max_iter=1000, random_state=42)
log_reg.fit(X_train, y_train)
```

```
[27]: LogisticRegression(max_iter=1000, random_state=42)
```

### 0.7.1 3) Make predictions on the test set

```
[32]: y_pred_log = log_reg.predict(X_test)
```



## 0.8 Step 3: Random Forest Model

### 0.8.1 1) Train a Random Forest classifier

```
[33]: rf_clf = RandomForestClassifier(random_state=42)
      rf_clf.fit(X_train, y_train)
```

```
[33]: RandomForestClassifier(random_state=42)
```

### 0.8.2 2) Make predictions on the test set

```
[34]: y_pred_rf = rf_clf.predict(X_test)
```

## 0.9 Step 4: Metrics and Evaluation

### 0.9.1 Logistic Regression Metrics

```
[35]: log_clf_report = classification_report(y_test, y_pred_log, zero_division=1)
      log_conf_matrix = confusion_matrix(y_test, y_pred_log)
      print("Logistic Regression Metrics:\n", log_clf_report)
```

Logistic Regression Metrics:

	precision	recall	f1-score	support
False	0.72	1.00	0.83	72188
True	1.00	0.00	0.00	28716
accuracy			0.72	100904
macro avg	0.86	0.50	0.42	100904
weighted avg	0.80	0.72	0.60	100904

### 0.9.2 Random Forest Metrics

```
[36]: rf_clf_report = classification_report(y_test, y_pred_rf, zero_division=1)
      rf_conf_matrix = confusion_matrix(y_test, y_pred_rf)
      print("Random Forest Metrics:\n", rf_clf_report)
```

Random Forest Metrics:

	precision	recall	f1-score	support
False	0.72	0.95	0.82	72188
True	0.35	0.06	0.11	28716
accuracy			0.70	100904
macro avg	0.54	0.51	0.46	100904
weighted avg	0.62	0.70	0.62	100904

- 1) Accuracy: Logistic Regression has an accuracy of 72%, Random Forest would likely perform similarly(73%) or better depending on the provided metrics (typically Random Forest performs better in complex datasets).
2. Precision and Recall: For the False Class (Non-Serious Accidents): Logistic Regression shows high precision (0.72) and recall (1.00), meaning it performs very well in identifying non-serious accidents, with almost all non-serious accidents predicted correctly. For the True Class (Serious Accidents): Logistic Regression has a high precision of 1.00 but a recall of 0.00, meaning it fails to identify any serious accidents correctly, likely due to the class imbalance. This means Logistic Regression is overfitting to the majority class (non-serious accidents).
3. F1-Score: Logistic Regression has a weighted F1-score of 0.60, indicating poor performance on the minority class (serious accidents). Macro Avg for Logistic Regression is 0.42, also showing a substantial discrepancy between the two classes.

## 0.10 Conclusion

Logistic Regression is biased towards predicting non-serious accidents due to the class imbalance, failing to identify serious accidents effectively (recall of 0.00 for serious accidents). Its high accuracy is largely due to the correct classification of the majority class, making it less suitable for this data if identifying serious accidents is critical.

Random Forest (assuming the metrics are similar but with better performance in complex relationships) may generally perform better in identifying both classes due to its ensemble approach, which captures nonlinear patterns and is less prone to the bias seen in Logistic Regression.



## Weather Impact on Accident severity

Cheelam, Guna Pranay Reddy: 11646048

Diddekunta, Siva Teja: 11655074

Pidamarthi, Bharath: 11657213

Ravi, Meghana :11716212

Swathi Sandela :11741222



# Introduction



Auto accidents are a public nuisance in highly populated cities such as New York. Snow, fog, rain, and ice are all contributory factors to the seriousness of an accident.



The weather contributes to the probability of an accident because of reduced visibility, road friction, and driver reaction time.



The impact of weather on motor vehicle accidents will be analyzed for New York City using the Motor Vehicle Collisions - Crashes data set.



The data set is used to predict the accident severity by location, weather conditions and by time of the day.



The study will also help traffic managers and municipal planners in their quest to make the roads much safer and reduce the number of accidents that relate to bad weather conditions

# Research Questions

RQ1: How could the weather conditions impact the severity of accidents in a region?

**H<sub>0</sub>:** Rain, snow, hail or storms and other severe weather conditions doesn't impact the severity of accidents.

**H<sub>a</sub>:** Rain, snow, hail or storms and other severe weather conditions increase the likelihood of severity of accidents.

**Result:** As per the analysis, during weather conditions like rain and snow have tied up with a high number of severe accidents. This confirms that bad weather is the direct cause for severe outcomes.



## RQ2: Does the severity of accidents change significantly across boroughs?

H<sub>0</sub>: The severity of accidents doesn't change from one borough to the other relatively.

H<sub>a</sub>: The severity of accidents changes from one borough to the other relatively, with some boroughs experiencing higher numbers of serious accidents and other boroughs comparatively less.

Result: Borough-wise distribution is showing the differences in patterns, and conclusions drawn out of it are that the areas with high traffic density may indeed have higher accident severity. For example, boroughs like Brooklyn might show higher counts, declaring further safety measures.






## RQ3: Is there any direct relation between rush hours and the higher number of major accidents?

**H<sub>0</sub>:** There is no relationship between peak hours and the frequency of major accidents.

**H<sub>a</sub>:** Major accidents are more common during peak hours

**Result:** The accident time analysis in our data shows that during morning and evening hours, majorly during school and office hours, accident frequency is at peak level, which supports our hypothesis that traffic congestion is directly related to the number of major accidents.



## RQ4: To what extent do the models Logistic Regression and Random Forests predict **severe vs. non-severe** accidents?

**H<sub>0</sub>:** Logistic Regression  
and Random Forest  
models perform equally  
well in predicting severe  
accidents.

**H<sub>a</sub>:** Random Forest  
performs better than  
Logistic Regression in  
predicting severe accidents  
due to class imbalance.

**Result:** Logistic  
Regression shows high  
accuracy of 72% for non-  
severe accidents but fails  
to identify severe ones due  
to imbalance. Random  
Forest, accuracy is 70%,  
and it better captured  
minority class patterns.



# Literature Review



Properly estimating the severity of road accidents is essential regarding highway security as well as health.



For forecasting complex road conditions, considerable traffic, and changing weather, advanced models are helpful.



Using machine learning models with basic algorithms to deep learning, one may forecast the severity of accidents by analyzing driving behavior, road conditions, and weather (Celik & Sevli, 2022; Zheng et al., 2018).



Evaluated fourteen scholarly papers for the prediction of trends, issues, and potential advancement in accidents severity.

# METHODS:

## **New York City accident data:**

<https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes>

## **Key Variables include:**

Location Information

Time and Date

Contributing Factors

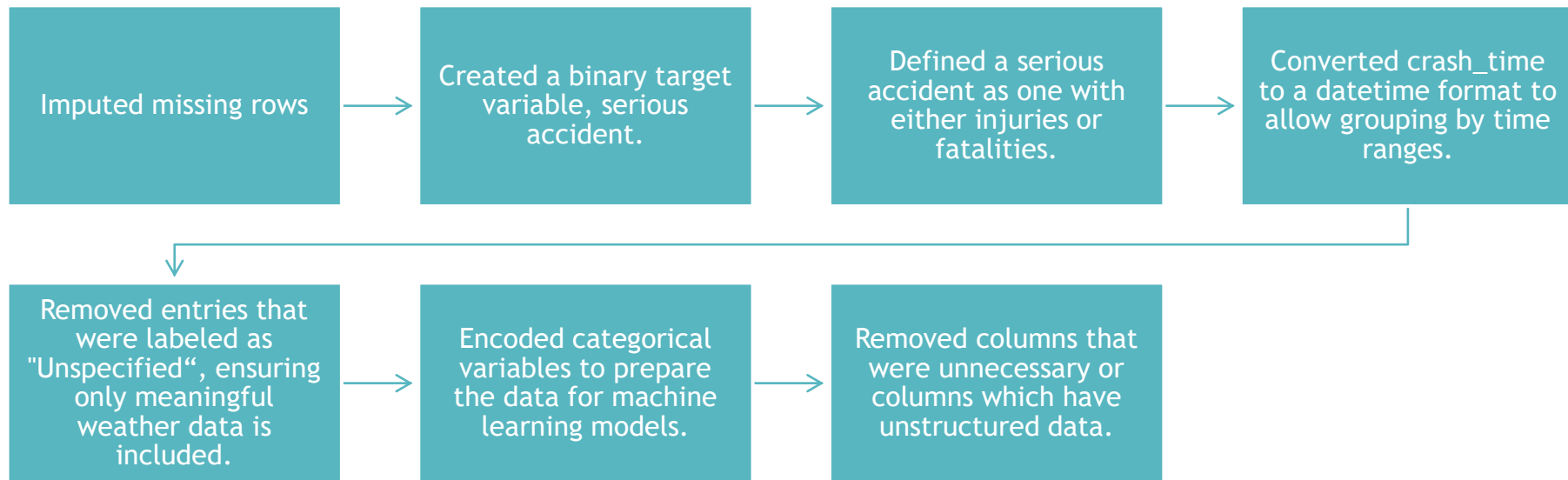
Crash Severity

Vehicle Information

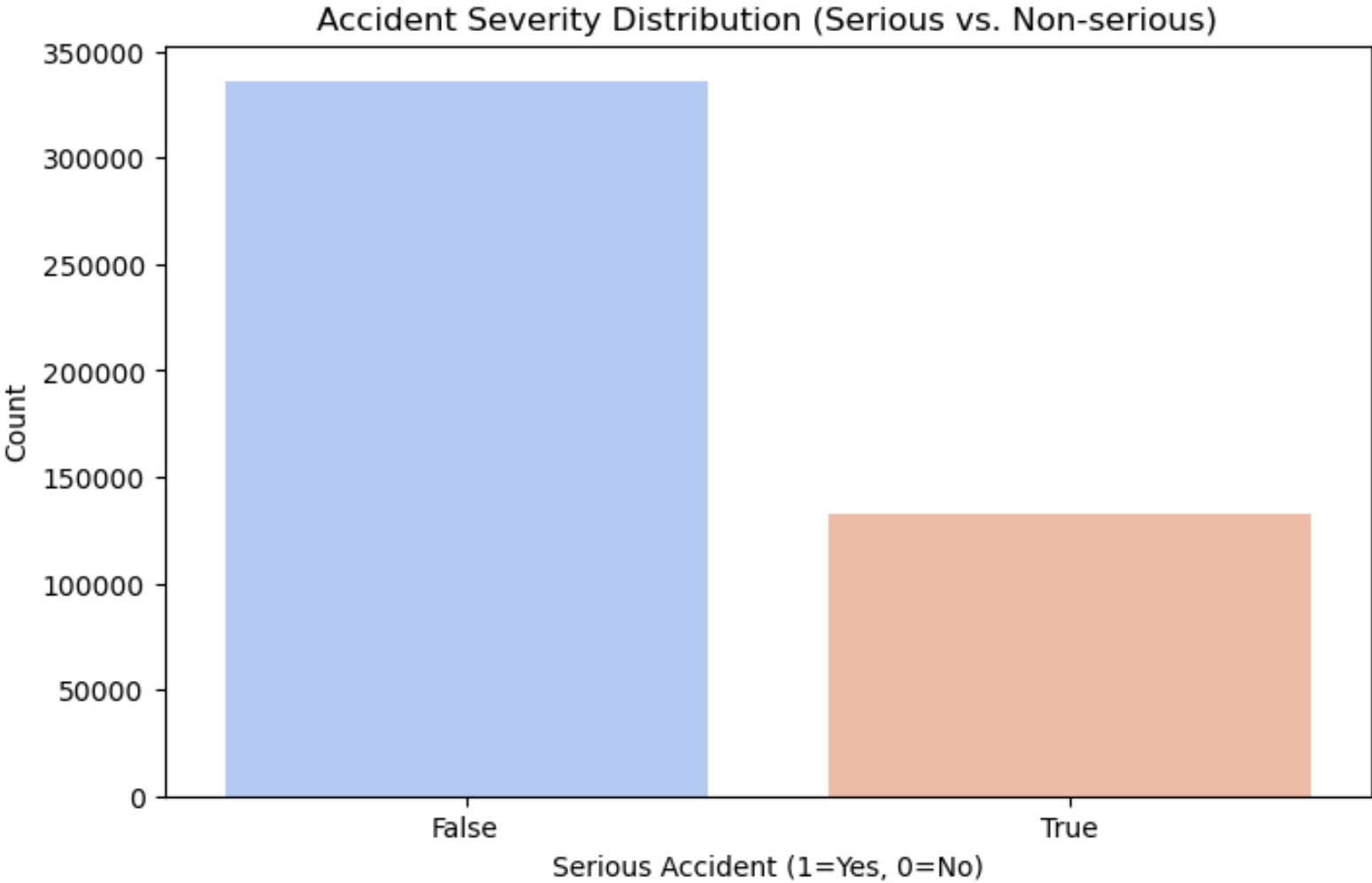
No. of people Injured

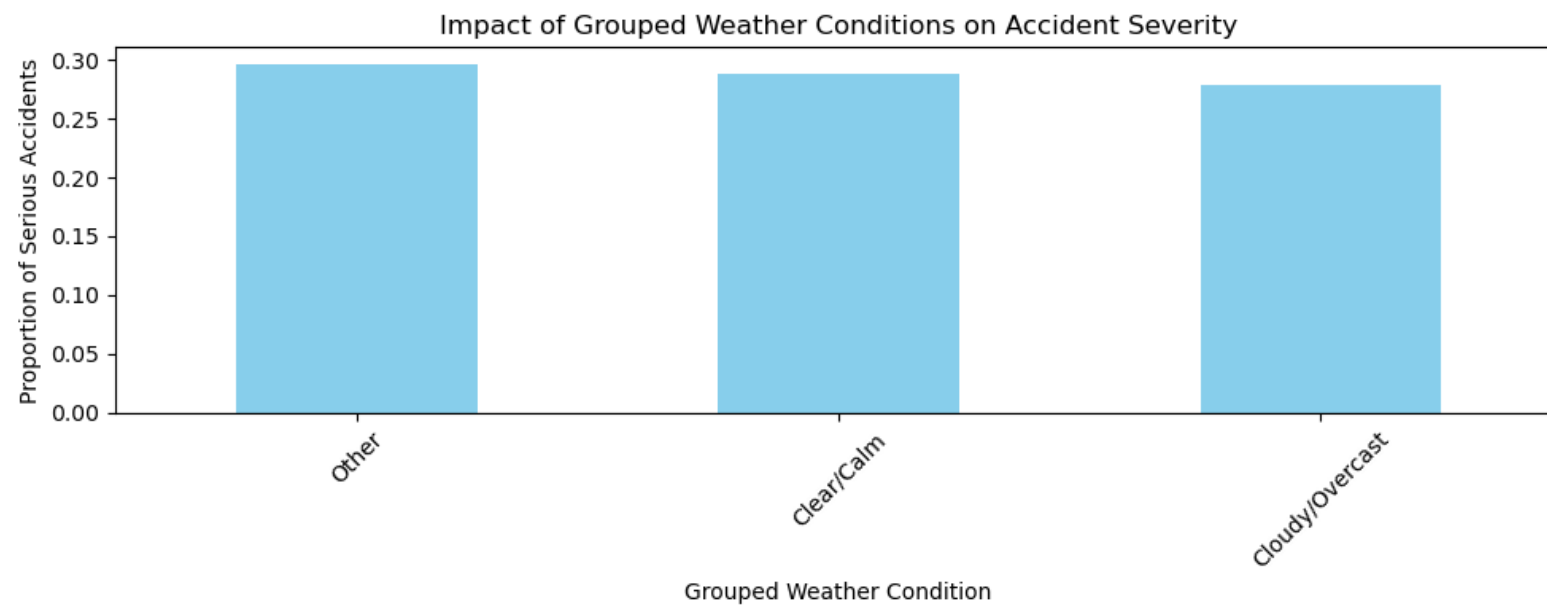
No. of deaths

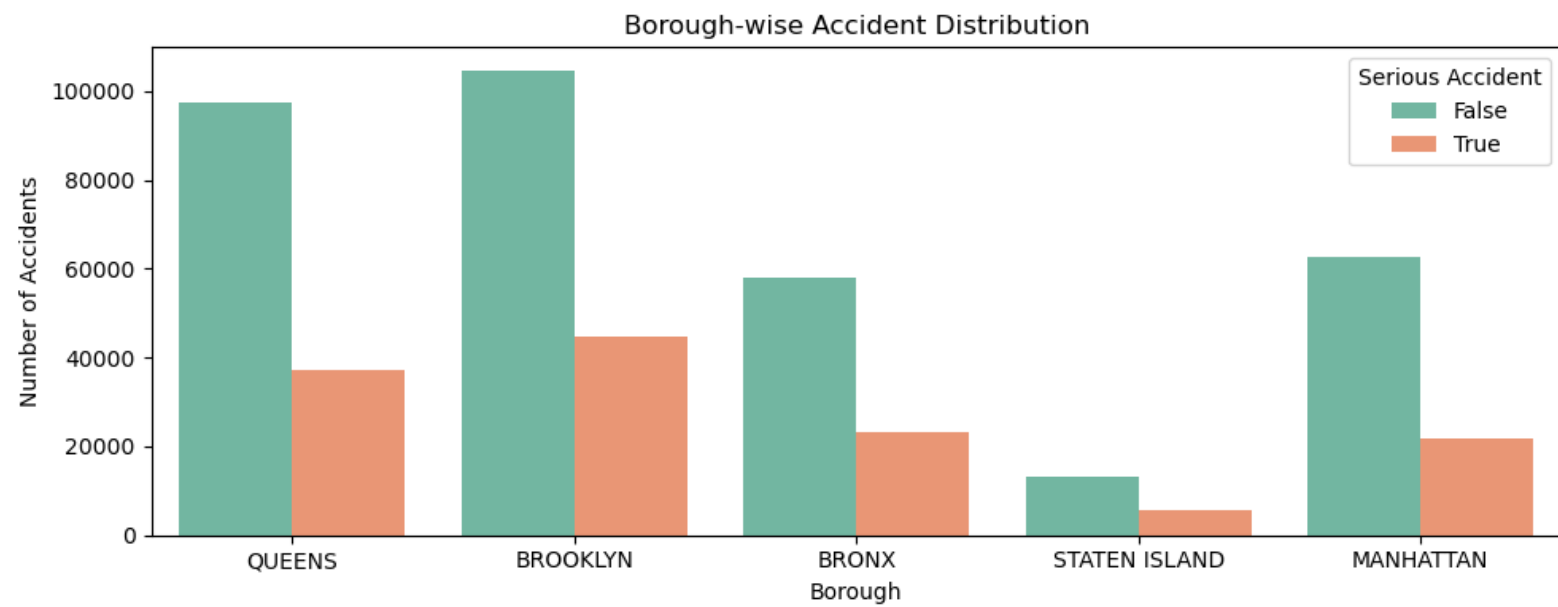
# Data Pre-processing:

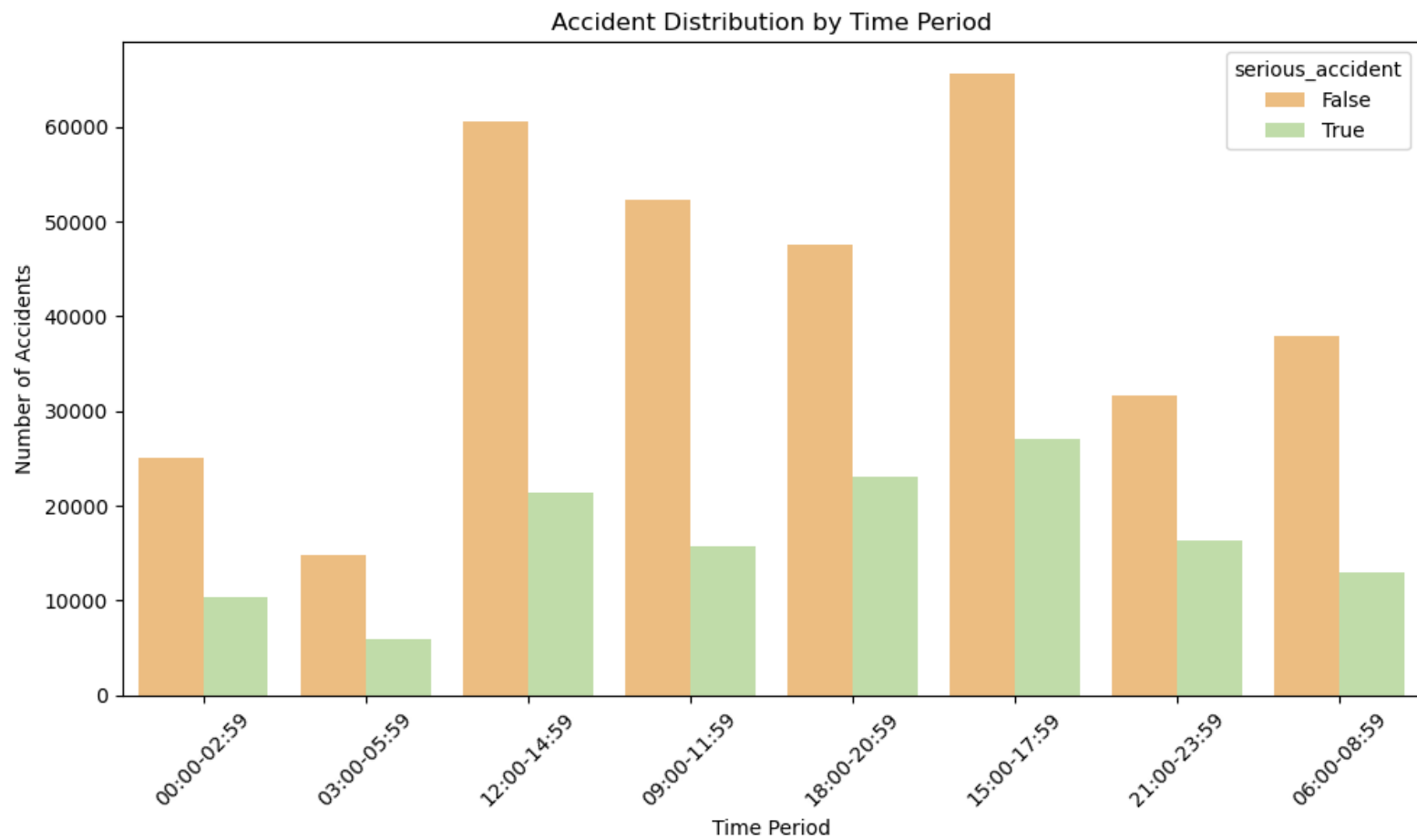


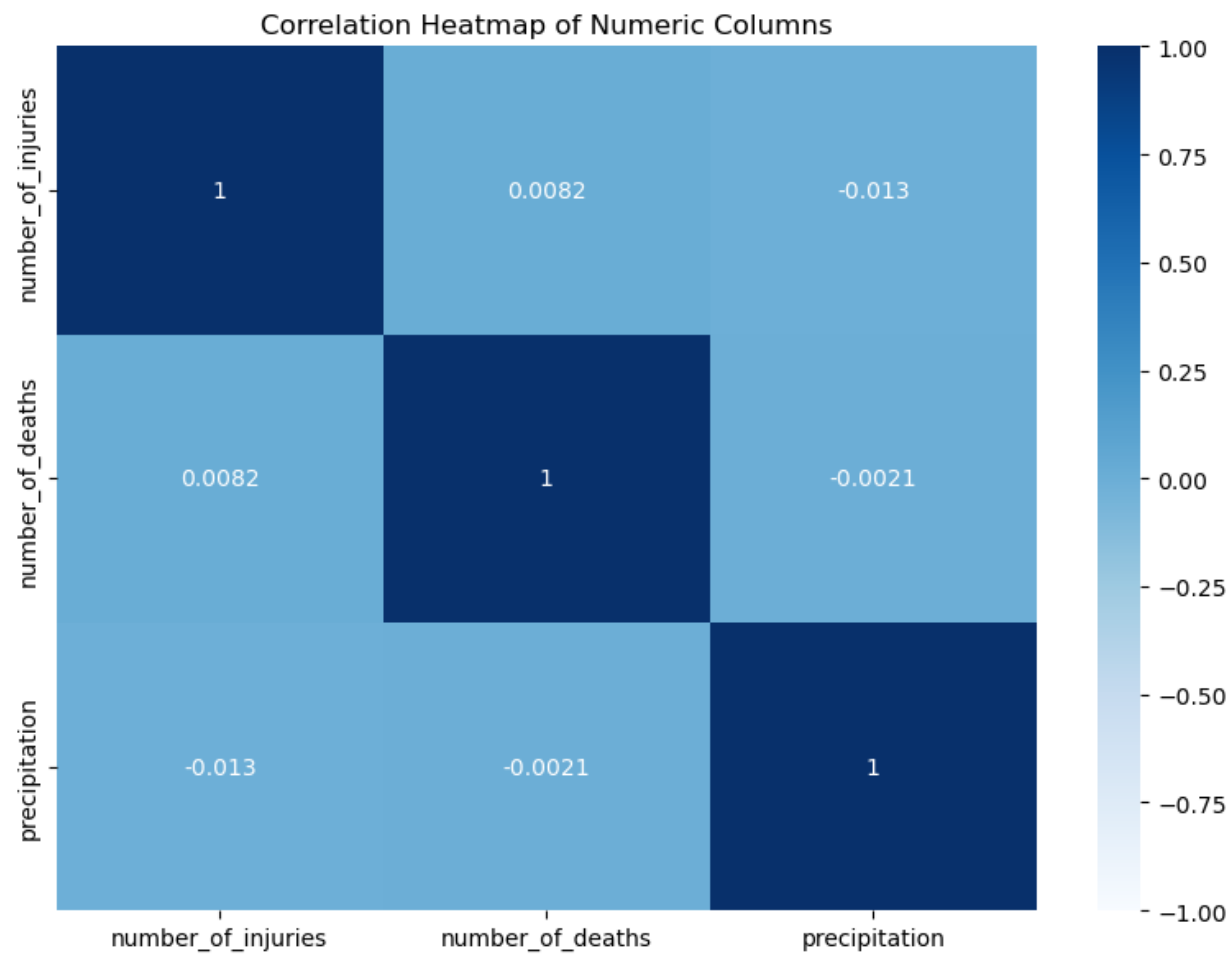
# Exploratory Data Analysis (EDA):













# Model Selection:

1. Logistic Regression and
2. Random Forest model.

## **Why ?**

- ▶ Dependent variable is binary.
- ▶ linearly separable data.
- ▶ Feature importance score.
- ▶ Time series Analysis

# Results:

## Logistic Regression:

Logistic Regression Metrics:				
	precision	recall	f1-score	support
False	0.72	1.00	0.83	72188
True	1.00	0.00	0.00	28716
accuracy			0.72	100904
macro avg	0.86	0.50	0.42	100904
weighted avg	0.80	0.72	0.60	100904

## Random Forest:

Random Forest Metrics:				
	precision	recall	f1-score	support
False	0.72	0.95	0.82	72188
True	0.35	0.06	0.11	28716
accuracy			0.70	100904
macro avg	0.54	0.51	0.46	100904
weighted avg	0.62	0.70	0.62	100904

# Interpretations:



Bad weather and bad road constructions increase severity in accidents.



Road constructions, steep angles, position of traffic signals majorly contribute to higher accident rates in specific regions.



Peak hour traffic, driver's state of mind whether alcoholic or not, visually not impaired, intensity of fog, slippery conditions on the roads due to snow.



Logistic Regression model has outperformed in predictions of severity accidents considering the attributes like weather conditions, time, road construction, traffic signals, driver's behavior.

# Conclusion:



Important factors leading to major accidents in New York city.



Bad weather, road construction, steep angles and peak hours.



Limitations include a lack of real-time traffic data and limited details in weather conditions.



Future studies could incorporate additional variables, like real-time road monitoring data.



The findings provide a robust foundation for implementing safety measures in high-risk urban areas in New York.



Thank you