



# BIG DATA

## PHASE 2

**Prepared By:**

Name	Sec.	B.N	ID
Ahmed Bassem ELKady	1	7	9210048
Daniel Nabil Khalil	1	16	9210386
Mohamed Yasser Mohamed	2	12	9211066
Mustafa Tarek Salah	2	21	9211178

**TEAM\_17**

# TABLE OF CONTENTS

- 01** BRIEF PROBLEM DESCRIPTION
- 02** PROJECT PIPELINE
- 03** DATA PREPROCESSING
- 04** DATA VISUALIZATION & INSIGHTS
- 05** MODEL TRAINING
- 06** MODEL RESULTS & EVALUATION
- 07** UNSUCCESSFUL TRIALS
- 08** ENHANCEMENTS & FUTURE WORK

# BRIEF PROBLEM DESCRIPTION

Car accidents result in significant financial losses and threaten public safety. A small proportion of severe accidents often leads to the greatest impact, making prevention essential. Our project seeks to answer:

**"What patterns contribute to serious car accidents, and how can we use this knowledge to improve road safety?"**

By analyzing millions of US accident records from 2016–2023, we aim to uncover key risk factors and interactions that lead to severe outcomes. Using PySpark for large-scale data analysis, we will develop predictive models and actionable insights to help authorities make informed decisions, enhance road planning, and implement proactive measures to reduce the severity of accidents.

# PROJECT PIPELINE

## 1. Data Loading:

Loaded 7,728,394 rows and 46 columns from Kaggle, including raw, redundant, and missing-value features.

## 2. Data Preprocessing:

Cleaned and transformed the dataset to 6,140,189 rows and 42 columns by removing columns with high missing or low analytical value, filtering outliers, binning continuous variables, grouping weather conditions, and adding Urban/Rural classification. Saved the cleaned dataset for further analysis.

## 3. Exploratory Data Analysis (EDA):

Analyzed and visualized over 20 features using PySpark for data processing and Matplotlib, Seaborn, and Plotly for visualizations.

## 4. Model Development:

Built classifiers using Naive Bayes (MapReduce with RDDs), Logistic Regression, and Random Forest in PySpark. Addressed class imbalance with undersampling/oversampling techniques.

## 5. Model Evaluation & Iteration:

Evaluated model performance, refined approaches, and compared results to select the best solution.

## 6. Future Work:

Identified areas for further enhancement and potential directions for deeper analysis and improved predictive accuracy.

# DATA PREPROCESSING

## 1. Handling Missing Values

- **Initial Dataset:** 7,728,394 rows and 46 columns.
- **Missing Value Analysis:**
  - Calculated the number and percentage of missing (NULL) values for each column.
  - Found high proportions of missing data in:
 

```
=== NULL Count and Percentage by Column ===
End_Lat: 3402762 (44.03%)
End_Lng: 3402762 (44.03%)
Precipitation(in): 2203586 (28.51%)
Wind_Chill(F): 1999019 (25.87%)
```
- **Dropping Columns:**
  - Removed columns with more than 25% missing values:
  - End\_Lat, End\_Lng, Precipitation(in), Wind\_Chill(F)
- **Dropping Rows:**
  - For the remaining columns (with low missing percentages), dropped all rows containing any NULLs.
- **Result:**
  - Dataset reduced to 7,051,556 rows and 42 columns, with no missing values remaining.

## 2. Dropping Irrelevant Columns

- **Column Inspection:**
  - Checked unique values in 'Source', 'Country', and 'Turning\_Loop' columns.
- **Dropped Columns:**
  - ID: Unique identifier, not useful for analysis
  - Country: Only one value ('US')
  - Source: Only three values, not informative
  - Turning\_Loop: Only one value (False)
- **Result:**
  - Dataset now contains 38 relevant columns for further analysis.

```
+-----+
|Country|
+-----+
|US     |
+-----+
```

```
+-----+
|Source |
+-----+
|Source3|
|Source2|
|Source1|
+-----+
```

```
+-----+
|Turning_Loop|
+-----+
|false       |
+-----+
```

# DATA PREPROCESSING

## 3. Removing Redundant Columns

### • Redundancy Check:

- Examined columns with similar meanings and checked for high correlation among continuous variables.

```
Correlation matrix for continuous variables:
Columns: ['Distance(mi)', 'Temperature(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)']
[[ 1.          -0.05598764  0.01140223 -0.09040334 -0.0395247  0.00874711]
 [-0.05598764  1.          -0.33053564  0.12005678  0.22400662  0.03458547]
 [ 0.01140223 -0.33053564  1.          0.10986754 -0.38682745 -0.17264725]
 [-0.09040334  0.12005678  0.10986754  1.          0.04117808 -0.0222837 ]
 [-0.0395247  0.22400662 -0.38682745  0.04117808  1.          0.01454486]
 [ 0.00874711  0.03458547 -0.17264725 -0.0222837  0.01454486  1.          ]]
```

### • Twilight Columns:

- The columns 'Sunrise\_Sunset', 'Civil\_Twilight', 'Nautical\_Twilight', and 'Astronomical\_Twilight' all contained only 'Day' or 'Night' values, making them redundant.

### • Action:

- Dropped 'Civil\_Twilight', 'Nautical\_Twilight', and 'Astronomical\_Twilight', keeping only 'Sunrise\_Sunset' for simplicity.

### • Other Columns:

- Retained 'End\_Time' for accident duration analysis and 'Street' for potential rural/urban analysis.

### • Result:

- Dataset reduced to 35 columns, keeping only the most relevant features.

## 4. Outlier Detection and Removal

### • Method:

- Calculated the 2nd and 98th percentiles for each continuous variable to identify outliers.

### • Action:

- Removed rows where any continuous variable fell outside the [2nd, 98th] percentile range.

### • Effect:

- This approach removed extreme outliers while preserving most of the data.

### • Result:

- Dataset reduced to 6,141,325 rows, improving data quality for further analysis.

```
Distance(mi): 2nd percentile = 0.0, 98th percentile = 4.496
Temperature(F): 2nd percentile = 17.0, 98th percentile = 93.0
Humidity(%): 2nd percentile = 15.0, 98th percentile = 100.0
Pressure(in): 2nd percentile = 25.27, 98th percentile = 30.35
Pressure(in): 2nd percentile = 25.27, 98th percentile = 30.35
Visibility(mi): 2nd percentile = 1.0, 98th percentile = 10.0
Wind_Speed(mph): 2nd percentile = 0.0, 98th percentile = 20.0
```

# DATA PREPROCESSING

## 5. Equal-Width Binning of Continuous Variables

### • Process:

- Discretized the main continuous variables (Distance, Temperature, Humidity, Pressure, Visibility, Wind Speed) using equal-width binning.
  - Created new binned columns (e.g., 'Distance(mi)') and kept the original values as \*\_cont columns.
  - Converted temperature from Fahrenheit to Celsius before binning.

### • Result:

- Each continuous variable now has a corresponding discretized version, making it easier to analyze patterns and trends in the data.

```
+-----+-----+
|Distance(mi)|count |
+-----+-----+
|0.00-1.50   |5674110|
|1.50-3.00   |343977 |
|3.00-4.50   |123238 |
+-----+-----+
```

```
+-----+-----+
|Temperature(C)|count |
+-----+-----+
|-8.33-2.22    |472710 |
|12.78-23.33   |2390579|
|2.22-12.78    |1410634|
|23.33-33.89   |1867402|
+-----+-----+
```

```
+-----+-----+
|Humidity(%) |count |
+-----+-----+
|17.00-37.75  |729271 |
|37.75-58.50  |1532462|
|58.50-79.25  |1923952|
|79.25-100.00|1955640|
+-----+-----+
```

```
+-----+-----+
|Visibility(mi)|count |
+-----+-----+
|1.00-4.00     |309867 |
|4.00-7.00     |323758 |
|7.00-10.00    |5507700|
+-----+-----+
```

```
+-----+-----+
|Pressure(in) |count |
+-----+-----+
|25.46-27.09  |117139 |
|27.09-28.72  |277357 |
|28.72-30.35  |5746829|
+-----+-----+
```

```
+-----+-----+
|Wind_Speed(mph)|count |
+-----+-----+
|0.00-5.00     |1672798|
|10.00-15.00   |1229628|
|15.00-20.00   |498514 |
|5.00-10.00    |2740385|
+-----+-----+
```

## 6. Weather Condition Grouping

### • Process:

- Simplified the 'Weather\_Condition' column by mapping detailed weather descriptions to broader, more meaningful categories (e.g., Cloudy, Snowy, Rainy, Clear, etc.).

### • Result:


- The dataset now focuses on high-level weather patterns, improving interpretability and reducing complexity for analysis.

```
+-----+-----+
|Weather_Condition |
+-----+-----+
|Cloudy            |
|Snowy             |
|Thunderstorm or Hail|
|Clear             |
|Rainy             |
|Freezing Rain & Ice |
|Hazy & Dusty       |
+-----+-----+
```

# DATA PREPROCESSING

## 7. Urban vs. Rural Classification

### • Process:

- Enriched the dataset by classifying each accident as Urban or Rural.
  - Aggregated a new column using external data from the US Census ([link](#)). 
  - Matched each accident's city and state to the census list: if a match was found, labeled as "Urban"; otherwise, labeled as "Rural".

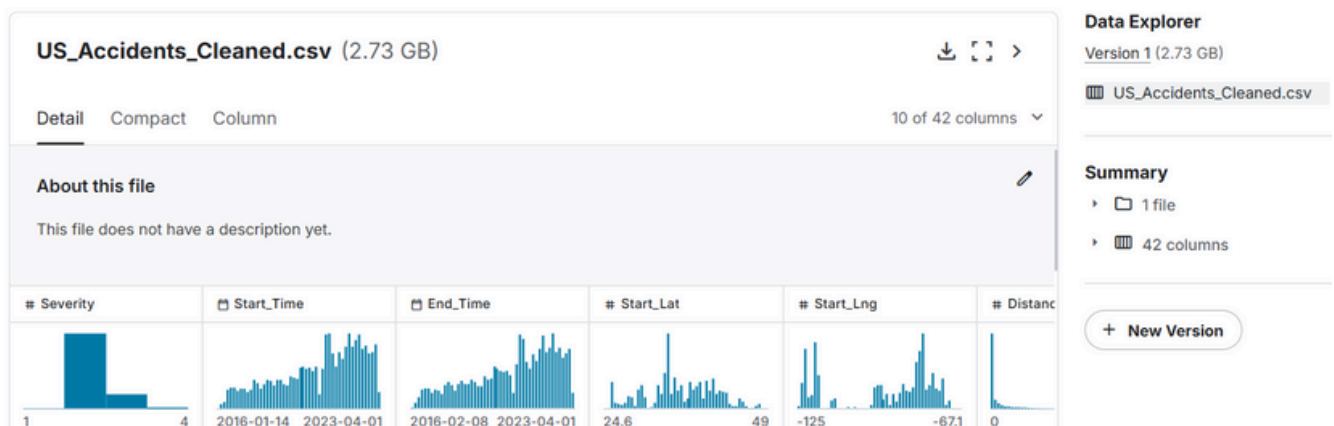
### • Result:

- Added a new 'Urban\_Rural' column, enabling analysis of accident patterns by area type.

```
+-----+-----+
|Urban_Rural| count|
+-----+-----+
|      Urban|1884199|
|      Rural|4255990|
+-----+-----+
```

## 8. Final Dataset Summary

- Before Preprocessing:
  - 7,728,394 rows, 46 columns (raw, with missing values and redundant features)
- After Preprocessing:
  - 6,140,189 rows, 42 columns (fully cleaned, binned, grouped, and enriched)
- Key Changes:
  - Removed columns with high missing or low analytical value
  - Filtered outliers and binned continuous variables
  - Grouped weather conditions
  - Added Urban/Rural classification
- Export:
  - The cleaned dataset was exported and uploaded to Kaggle for further analysis.





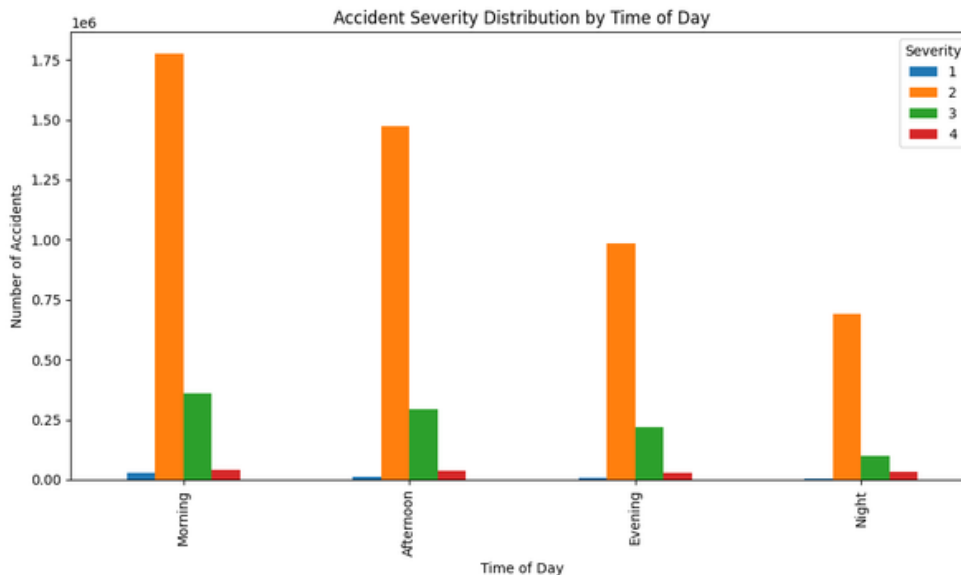
# DATA VISUALIZATION & INSIGHTS

## Accident Severity by Time of Day

- Severity 2 accidents are most common in all time segments, peaking in the morning and decreasing towards night.
- Severity 3 accidents follow a similar but lower trend.
- Severity 1 (least severe) and Severity 4 (most severe) are rare but consistent across all times.
- Total accidents are highest in the morning and afternoon, lowest at night.

## Business Takeaways:

- Prioritize emergency and traffic resources during morning and afternoon.
- Most incidents are moderate (Severity 2), but severe cases (Severity 4) occur steadily—constant readiness is needed.
- Further analysis should explore what factors drive higher severity at certain times.



# DATA VISUALIZATION & INSIGHTS

**How does the presence of different road objects (e.g., Junctions, Crossings, Traffic Signals) relate to accident severity?**

## Analysis:

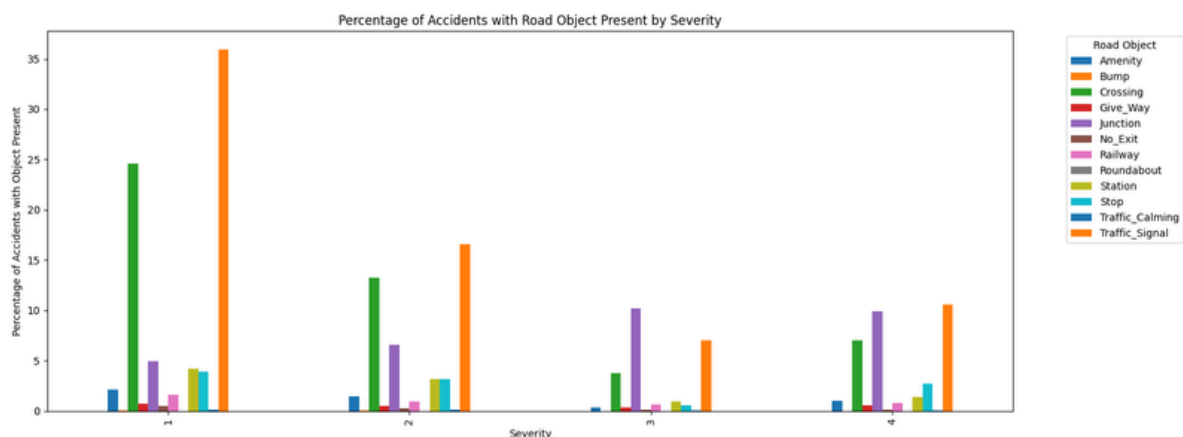
- Calculated the percentage of accidents at each severity level that occurred in the presence of various road objects.
- Used PySpark for aggregation and Matplotlib for visualization.

## Insight:

- Junctions and Traffic Signals are more common in severe accidents (Severity 3 & 4), indicating intersections and signalized areas are riskier for serious incidents.
- Crossings are linked to less severe accidents (Severity 1 & 2); their presence drops as severity increases.
- Traffic calming features (Bumps, Roundabouts, etc.) are rarely present in severe accidents, suggesting they may help reduce risk.
- Railway, Station, Stop signs, and Amenities show low or consistent presence across all severities.

## Business Implications:

- Prioritize intersection safety improvements (signage, signal timing, visibility) to reduce severe accidents.
- Expand traffic calming measures to more locations.
- Enhance crosswalk safety to reduce minor accidents.



# DATA VISUALIZATION & INSIGHTS

## 1. How do weather conditions, time of day, and road features (e.g., Traffic\_Signal, Junction) interact to influence accident severity across different states?

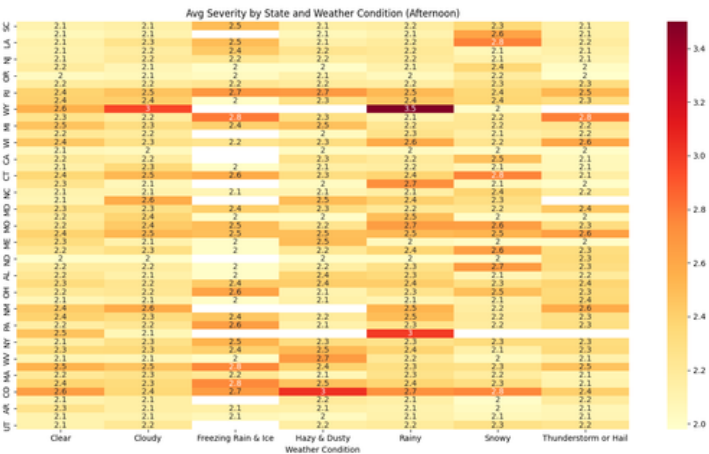
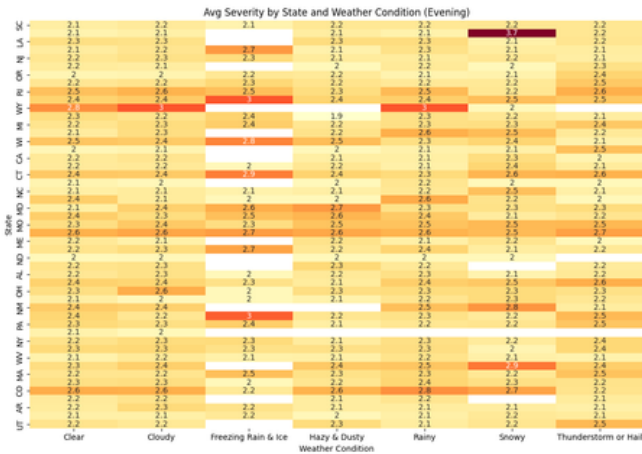
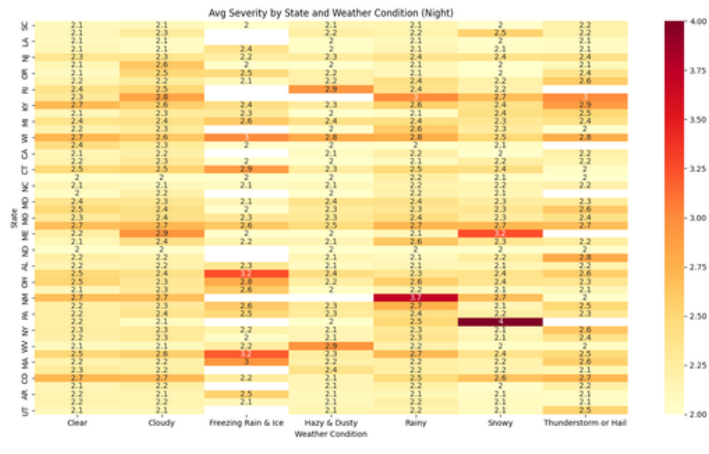
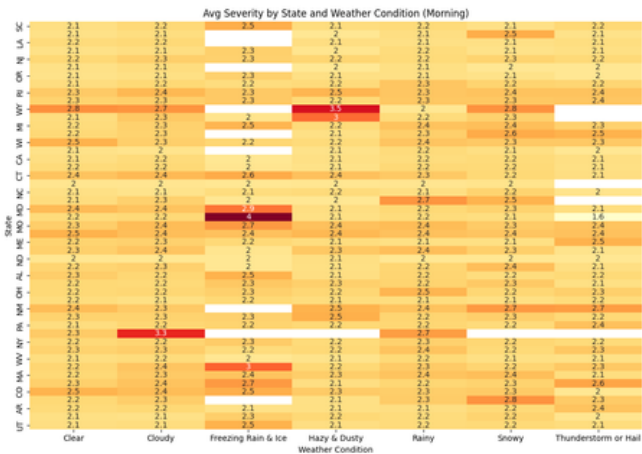
### Business Value:

Enables urban planners and law enforcement to identify high-risk periods and locations, optimize traffic signals, and prioritize safety improvements based on when and where severe accidents are most likely.

### Insights:

The four heatmaps show average accident severity by state and weather condition for each time of day (morning, afternoon, evening, night):

- **Adverse Weather Impact:** Severity spikes during freezing rain, snow, and storms, especially in certain states and time periods.
- **Time-of-Day Variation:** Some states (e.g., MO, MI, NY) experience higher severity in the evening or night under poor weather, highlighting increased risk after dark.
- **Consistent Patterns:** Clear and cloudy weather generally have lower severity, but hazardous conditions consistently elevate risk across all times.
- **Actionable Focus:** Targeted interventions—such as improved signage, signal timing, or patrols—should focus on high-risk weather and time combinations, especially at intersections and near traffic signals.



# DATA VISUALIZATION & INSIGHTS

## 2. Are longer accidents (by distance) more severe in certain states or cities?

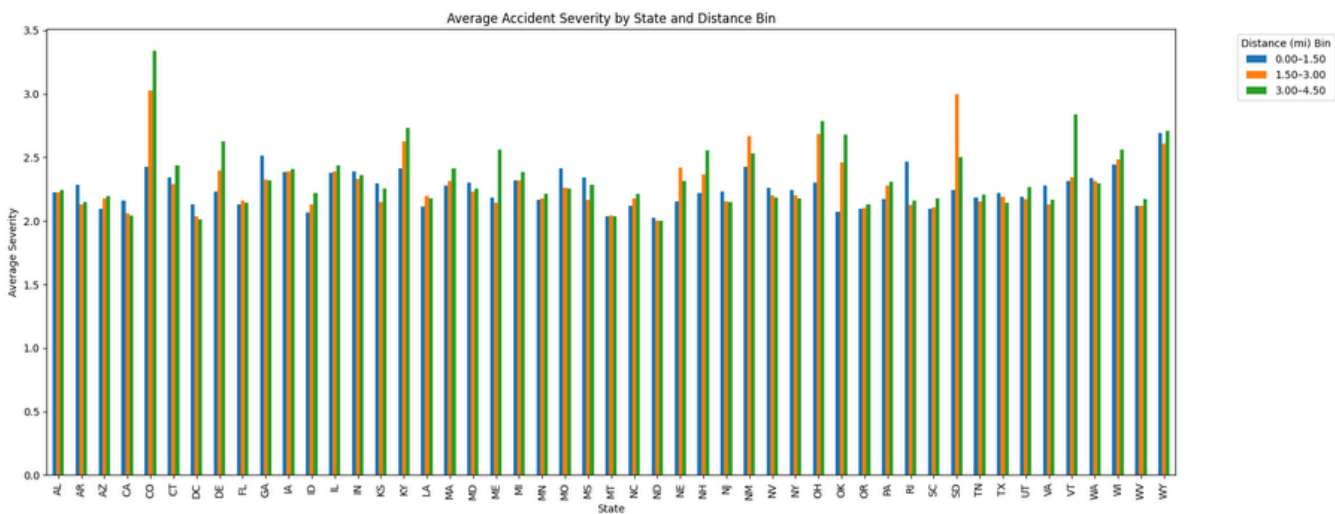
**Business Value:**

Helps local authorities and urban planners pinpoint states where longer accidents are more severe, guiding targeted interventions such as speed regulation, improved road design, or enhanced emergency response in high-risk zones.

**Insights:**

The grouped bar chart compares average accident severity across states for different accident distance bins:

- **Severity Trends:** In most states, severity remains relatively stable across distance bins, but some states (e.g., CA, NM, SD, WA) show a noticeable increase in severity for longer accidents.
- **High-Risk States:** States with higher severity for longer accidents may have unique road conditions, higher speeds, or delayed emergency response, indicating a need for focused safety measures.
- **Actionable Focus:** Authorities in these states should investigate contributing factors and consider targeted interventions—such as speed enforcement or infrastructure improvements—on routes where long, severe accidents are more common.



# DATA VISUALIZATION & INSIGHTS

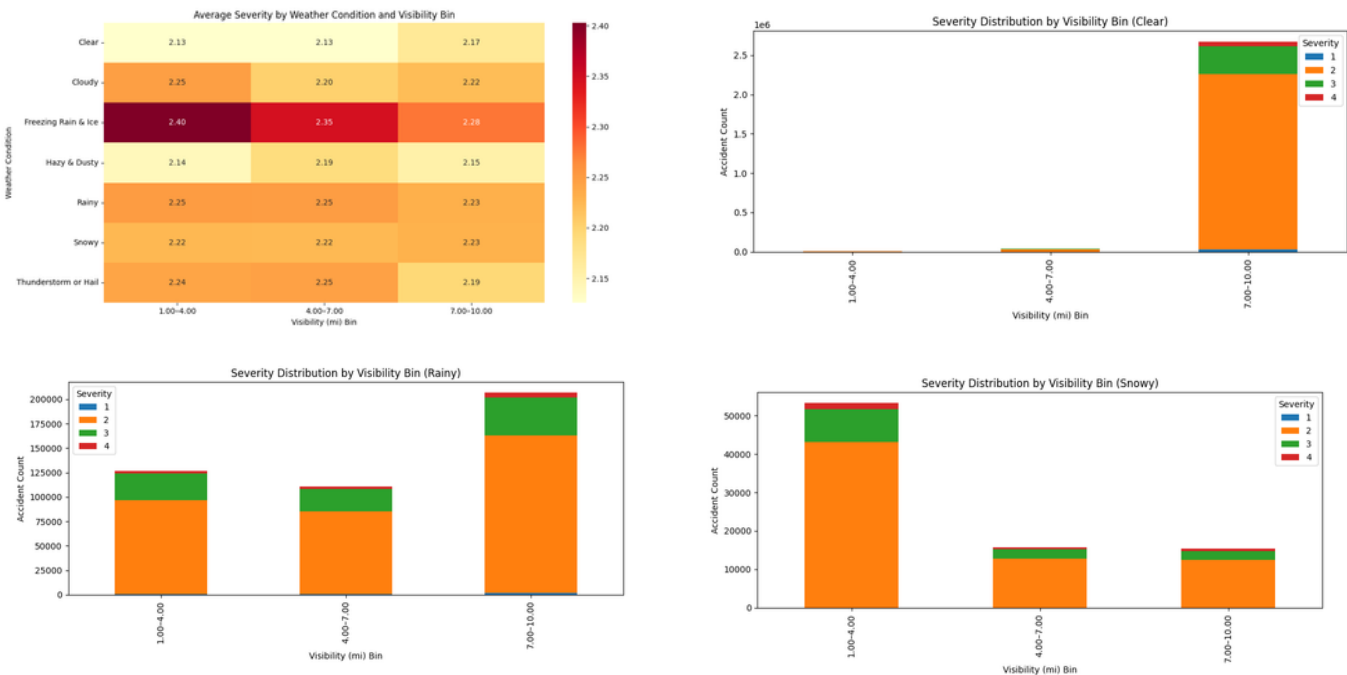
## 3. What is the relationship between visibility conditions and accident frequency/severity during different weather types?

**Business Value:**

Helps agencies develop weather-specific safety policies and dynamic warning systems by identifying visibility thresholds that correlate with higher accident rates and severity under various weather conditions.

**Insights:**

- **Accident Frequency:**
  - Most accidents occur at higher visibility in clear and rainy weather, but low-visibility bins (1–4 mi) see a higher proportion of severe accidents, especially in snowy conditions.
- **Severity Patterns:**
  - The heatmap shows that average severity is highest during freezing rain & ice, especially at low visibility (2.40), and remains elevated for snowy and rainy conditions.
  - Severity is generally lower in clear weather, regardless of visibility.
- **Weather-Specific Risks:**
  - In snowy and rainy weather, low visibility increases both accident count and severity, highlighting the need for targeted warnings and interventions during such conditions.
- **Actionable Focus:**
  - Dynamic warning systems and stricter controls should be considered when visibility drops below 4 miles, especially during adverse weather.



# DATA VISUALIZATION & INSIGHTS

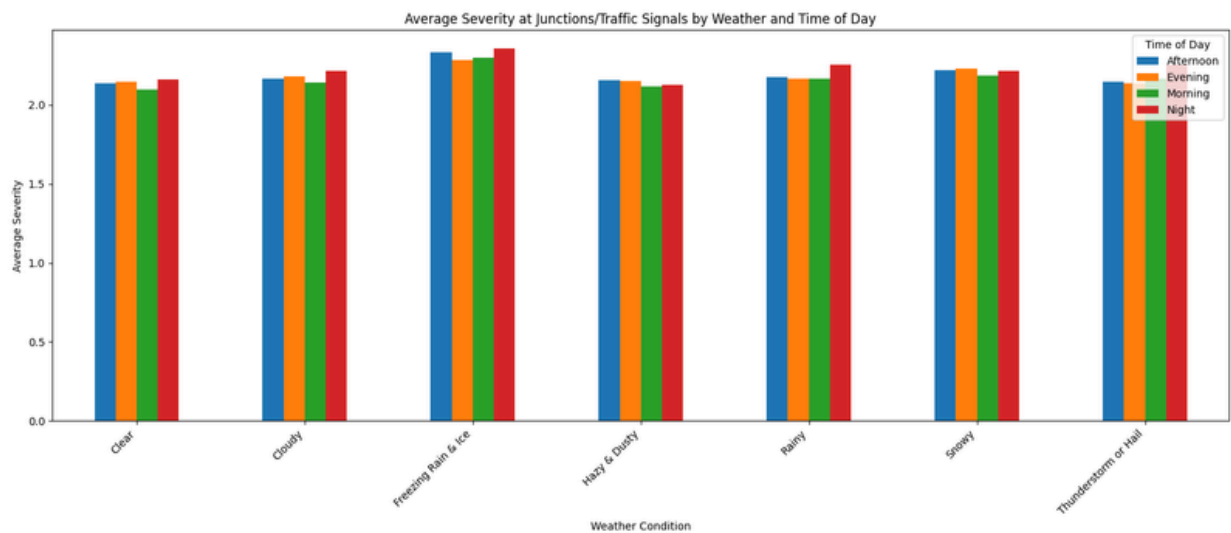
## 4. Do certain road features (e.g., Junction, Traffic\_Signal) correlate with higher accident severity under specific weather or time conditions?

### Business Value:

Enables city planners and traffic engineers to improve infrastructure design and optimize signal placement by understanding how weather and time of day amplify accident severity at critical road features.

### Insights:

- **Consistent Elevation:** Accident severity at junctions and traffic signals is generally higher during adverse weather, especially freezing rain & ice, across all times of day.
- **Time-of-Day Effect:** Nighttime and evening hours tend to show slightly higher severity, particularly under hazardous weather conditions.
- **Weather Amplification:** The combination of poor weather (e.g., freezing rain, rain, snow) and critical road features increases risk, suggesting targeted interventions (e.g., better lighting, advanced warnings, or adaptive signal timing) during these periods.
- **Actionable Focus:** Prioritizing safety improvements at intersections and signals during adverse weather and nighttime can help reduce severe accidents.



# DATA VISUALIZATION & INSIGHTS

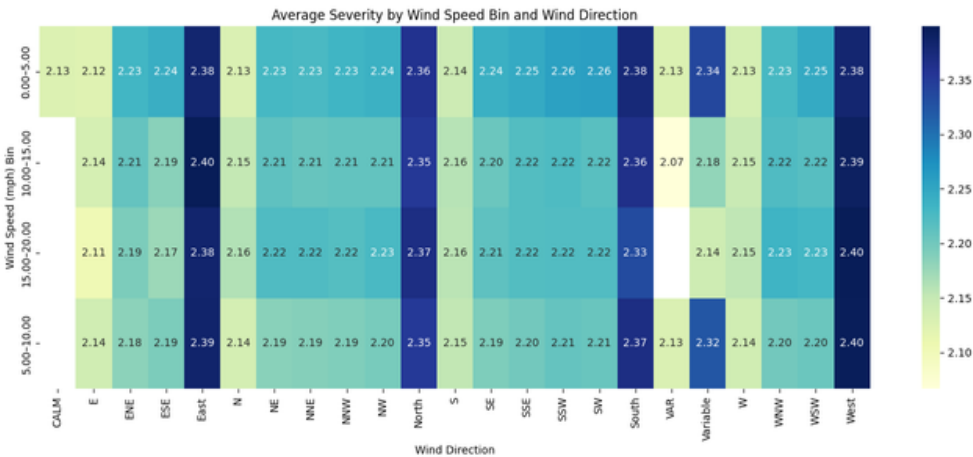
## 5. How does accident severity change with wind speed and direction?

**Business Value:**

Supports the development of weather-aware traffic management systems by identifying wind conditions that may increase accident severity, especially in exposed or high-speed travel areas.

**Insights:**

- **Severity Patterns:** Accident severity is slightly higher for certain wind directions (e.g., East, South, West) and at moderate wind speeds (10–20 mph), but overall differences are modest.
- **Data Distribution:** Most data is concentrated in the cardinal directions (N, E, S, W), which may bias the results and limit the ability to detect strong directional effects.
- **Actionable Focus:** While no extreme spikes are observed, monitoring and issuing warnings during moderate winds—especially from the East or West—could help mitigate risk in vulnerable areas.



# DATA VISUALIZATION & INSIGHTS

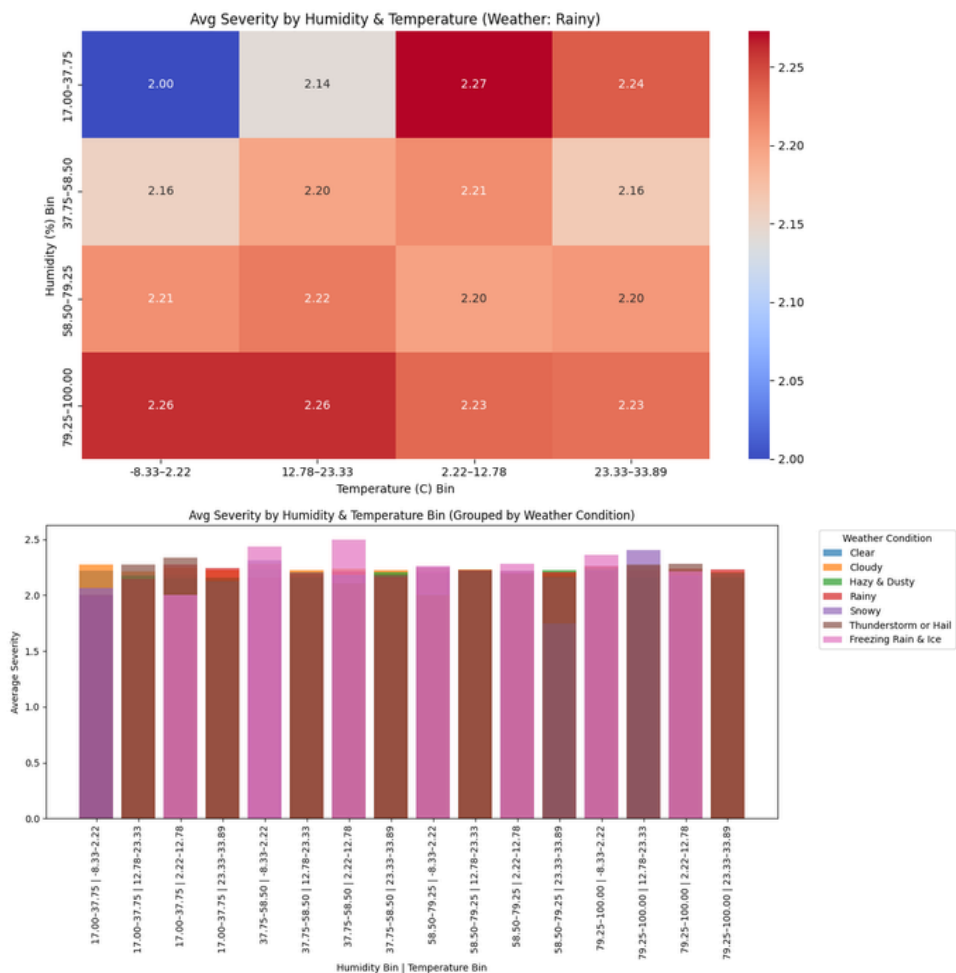
## 6. Are there patterns in accident severity based on humidity, temperature, and weather condition combinations?

### Business Value:

Enables the creation of more nuanced, condition-aware traffic alerts and infrastructure planning by revealing how specific combinations of humidity, temperature, and weather impact accident severity.

### Insights:

- **Combined Effects:** Accident severity tends to increase with higher humidity, especially when paired with extreme temperatures, across most weather conditions.
- **Weather-Specific Patterns:** Under rainy conditions, severity is lowest at low humidity and temperature, but rises steadily with both higher humidity and temperature, peaking at the most humid and warmest bins.
- **Consistent Trends:** Freezing rain & ice and snowy conditions also show elevated severity at higher humidity and temperature combinations.
- **Actionable Focus:** These patterns suggest that dynamic warnings and targeted interventions should consider not just single weather variables, but their combinations—especially during humid, warm, or stormy periods.





# DATA VISUALIZATION & INSIGHTS

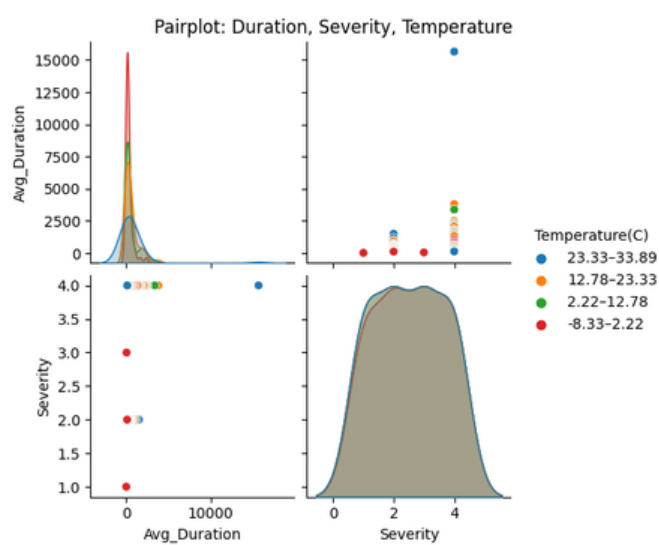
## 7. Which combinations of environmental factors and accident characteristics are most associated with prolonged accidents?

**Business Value:**

Enables traffic authorities to optimize emergency response and reduce congestion by identifying environmental and situational patterns that contribute to longer accident durations.

**Insights:**

- **Correlation:** There is a moderate positive correlation (0.35) between accident severity and duration, indicating that more severe accidents tend to last longer.
- **Environmental Factors:** The pairplot shows that while most accidents are short in duration, higher severity and certain temperature ranges are linked to longer-lasting incidents.
- **Multi-factor Influence:** The 3D scatter (not shown here) further highlights that prolonged accidents often occur when severity is high, visibility is low, and temperatures are at the extremes.
- **Actionable Focus:** Emergency response strategies should prioritize severe accidents, especially under challenging environmental conditions (e.g., low visibility, extreme temperatures), to minimize traffic disruption.



	Avg_Duration	Severity
Avg_Duration	1.000000	0.347782
Severity	0.347782	1.000000

# DATA VISUALIZATION & INSIGHTS

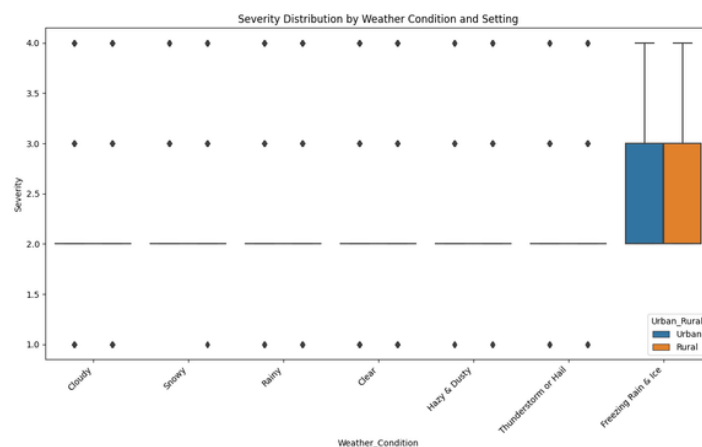
## 8. What is the combined impact of weather and urban/rural settings on accident severity and frequency?

### Business Value:

Supports insurance risk assessment and guides infrastructure policymaking by revealing how environmental and geographic contexts jointly affect accident outcomes.

### Insights:

- **Statistically Significant Differences:** The chi-square test ( $\chi^2 = 2090.14$ ,  $p < 0.00001$ ) confirms a strong association between weather, setting (urban/rural), and accident frequency.
- **Severity Patterns:** The boxplot shows that severe accidents (higher severity scores) are more common during hazardous weather (e.g., freezing rain, snow) in both urban and rural areas, but the distribution is broader in rural settings.
- **Frequency Trends:** Urban areas generally experience more accidents across all weather types, but rural settings see a higher proportion of severe outcomes during adverse weather.
- **Actionable Focus:** Insurers can refine risk models by factoring in both weather and location, while policymakers should prioritize rural infrastructure upgrades and targeted weather-related safety campaigns.



# DATA VISUALIZATION & INSIGHTS

**9. How does the combination of discretized Distance(mi), Visibility(mi), and Wind\_Speed(mph) affect accident severity during specific weather conditions (e.g., rain, snow)?**

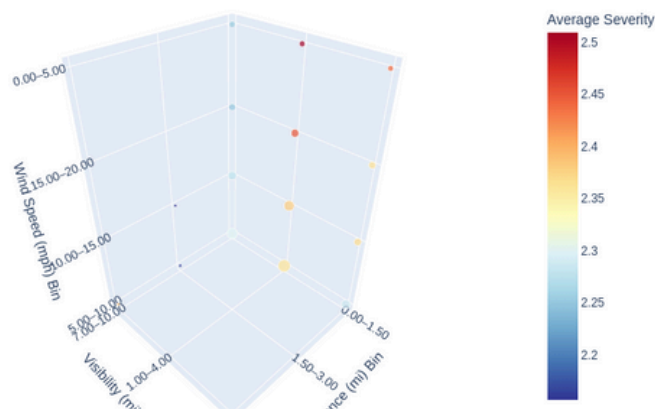
## Business Value:

Enables dynamic traffic management (e.g., speed limits, alerts) during adverse weather by identifying how combinations of road and weather factors influence severity.

## Insights:

- **Multi-Factor Risk:** 3D scatter plots reveal that accident severity increases when low visibility, high wind speed, and longer travel distances coincide, especially during adverse weather like rain and snow.
- **Weather-Specific Patterns:** In snowy and rainy conditions, clusters of high-severity accidents are most common in bins with poor visibility and higher wind speeds, regardless of distance.
- **Targeted Interventions:** These findings suggest that dynamic warnings and speed reductions should be prioritized when multiple risk factors are present, not just during bad weather alone.
- **Actionable Focus:** Transportation agencies can use these insights to deploy real-time alerts and adjust traffic controls based on combined environmental risks, reducing the likelihood of severe accidents.

Severity by Distance, Visibility, Wind Speed (Freezing Rain & Ice)



# DATA VISUALIZATION & INSIGHTS

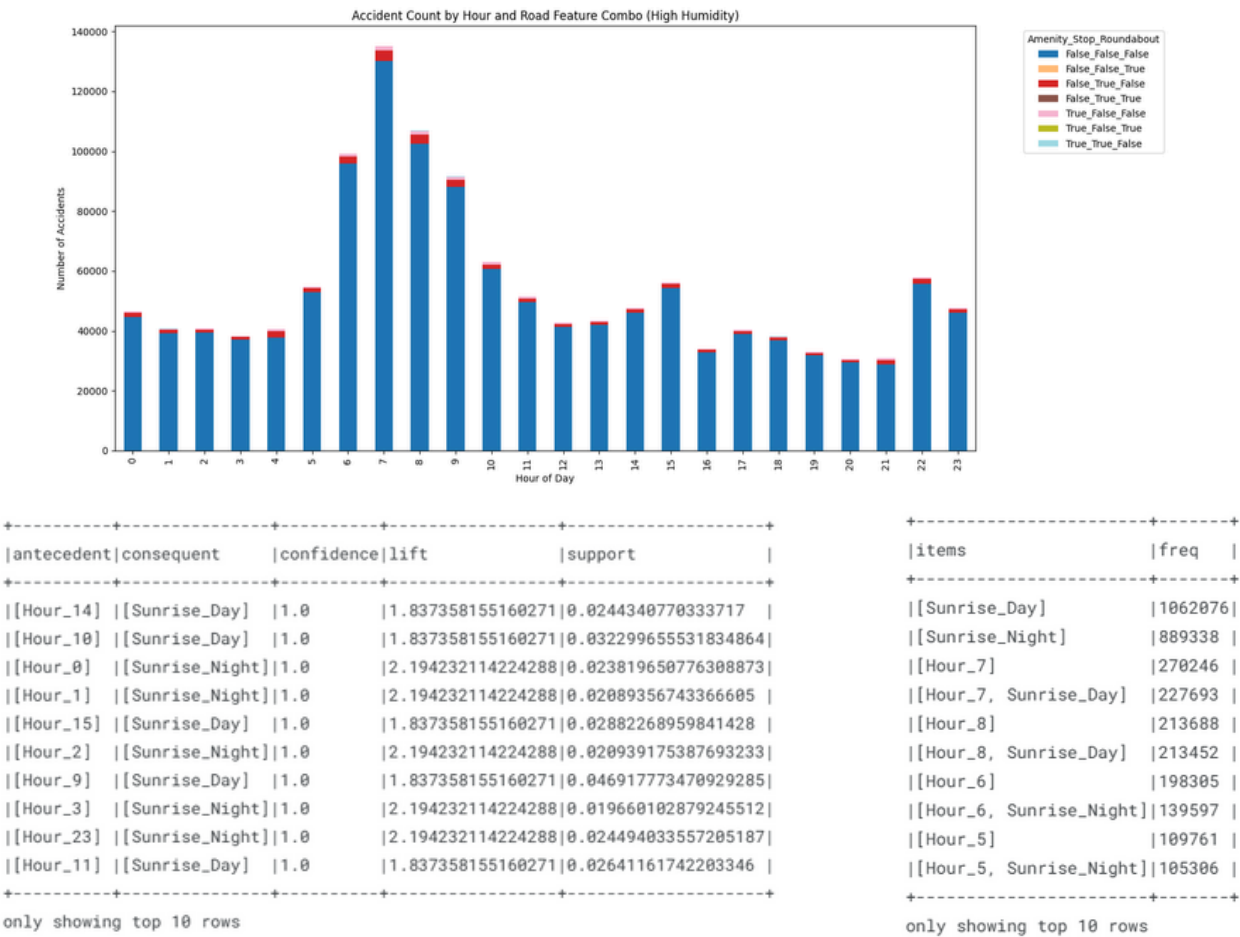
10. Which combinations of road features (e.g., Amenity, Stop, Roundabout) and time-based factors (e.g., hour of Start\_Time, Sunrise\_Sunset) are most associated with accidents in high-humidity conditions (Humidity(%) ≥ 79.25)?

**Business Value:**

Guides infrastructure upgrades and patrol scheduling by identifying high-risk combinations under humid conditions.

**Insights:**

- **Peak Risk Hours:** Accident counts spike during morning rush hours (6–9 AM), especially around sunrise, as shown by the stacked bar chart.
- **Feature Combinations:** Most accidents occur where none of the features (Amenity, Stop, Roundabout) are present, but combinations involving stops and roundabouts, though less frequent, may indicate specific risk points.
- **Frequent Patterns:** FP-Growth analysis reveals strong associations between certain hours (e.g., Hour\_7, Hour\_14) and sunrise/sunset periods, with high confidence and lift values for these time-feature combinations.
- **Actionable Focus:** Targeted patrols and infrastructure improvements (e.g., better signage or lighting) should be prioritized at high-risk hours and locations, especially during humid, low-visibility conditions.



# DATA VISUALIZATION & INSIGHTS

## 11. Where are the most dangerous accident hotspots?

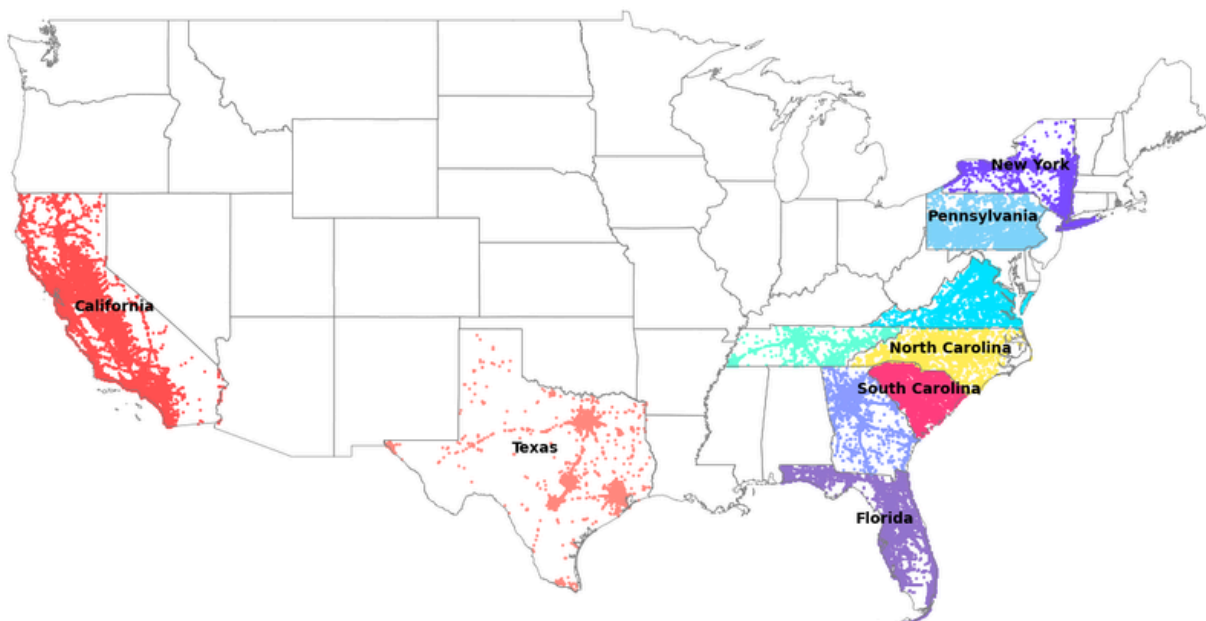
### Business Value:

Helps transportation departments and insurance companies identify high-risk areas for targeted safety interventions and risk assessment.

### Insights:

- **Top Hotspot States:** California, Florida, and Texas have the highest accident counts, with California alone reporting over 1.3 million cases.
- **Geographic Clustering:** The map visualization highlights dense accident clusters in major urban and highway regions within these states.
- **Regional Patterns:** The top 10 states are concentrated along the coasts and the Southeast, indicating regional factors such as population density, weather, and infrastructure may contribute to higher accident rates.
- **Actionable Focus:** Prioritizing safety campaigns, infrastructure upgrades, and insurance risk models in these hotspot states can yield the greatest impact in reducing accident frequency and severity.

Visualization of Top 10 Accident Prone States in US



# DATA VISUALIZATION & INSIGHTS

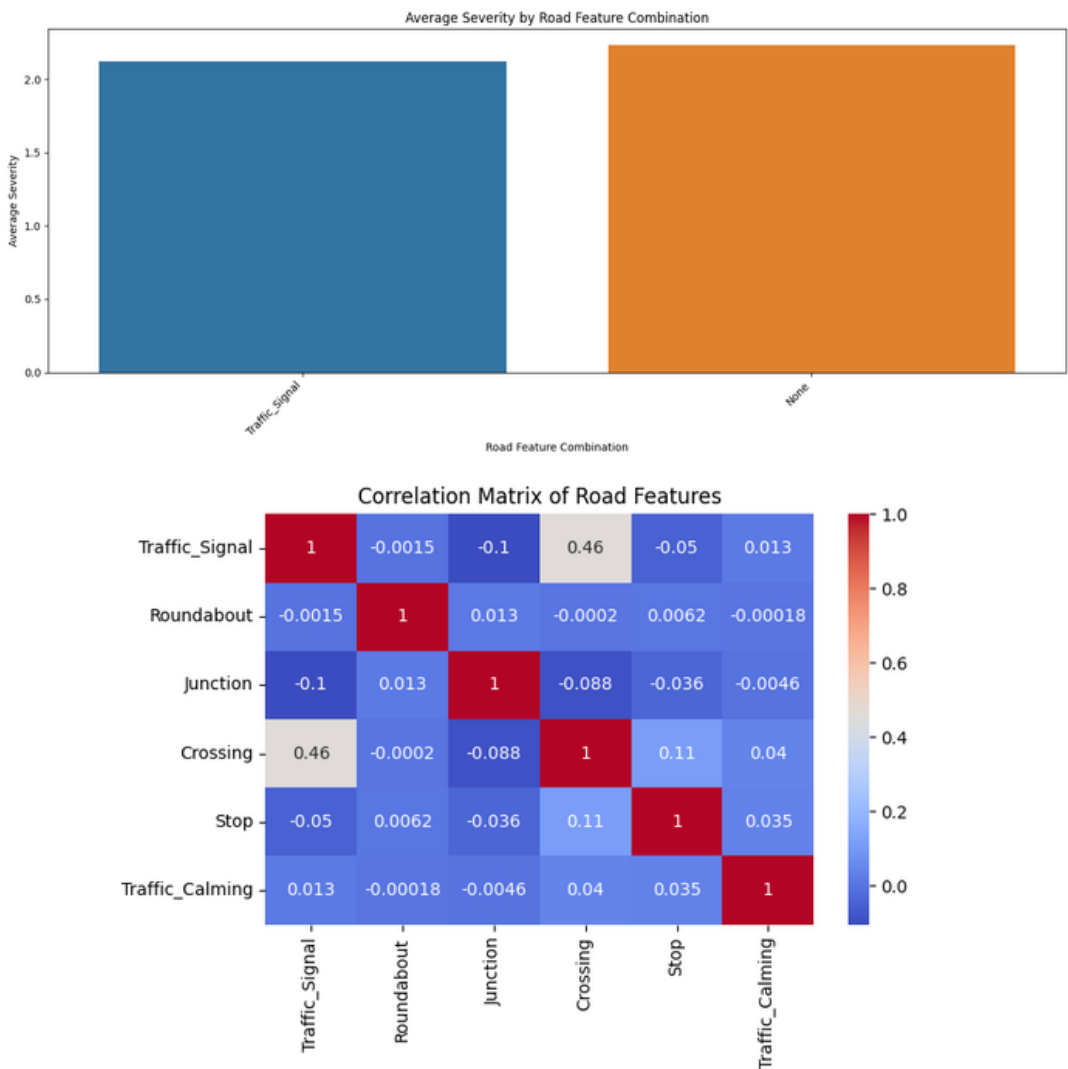
## 12. Which combinations of road features best reduce accident severity?

**Business Value:**

Informs infrastructure investment decisions by identifying the most effective safety features.

**Insights:**

- **Feature Correlation:** The correlation matrix shows that most road features are weakly correlated, except for a moderate positive relationship between Traffic\_Signal and Crossing (0.46), suggesting these features often co-occur.
- **Severity Reduction:** The bar plot indicates that locations with Traffic\_Signal have a lower average accident severity compared to locations with no safety features. This suggests that traffic signals are effective at reducing the severity of accidents.
- **Actionable Focus:** Prioritizing the installation of traffic signals—especially at crossings—can help lower accident severity. Further analysis of more complex feature combinations is needed to identify additional synergies for safety improvements.



# DATA VISUALIZATION & INSIGHTS

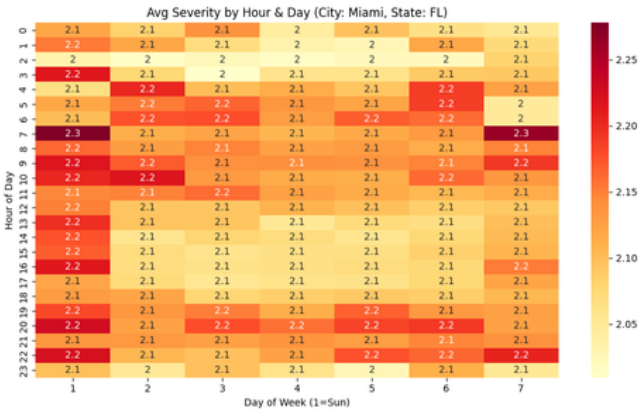
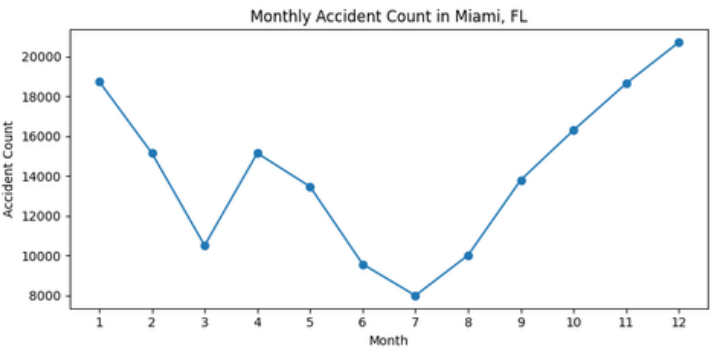
## 13. How do accident characteristics vary by time and location?

**Business Value:**

Optimizes emergency response resource allocation and helps develop targeted safety campaigns.

**Insights:**

- **Danger Hours by City:** The table shows that the most dangerous hours (highest average severity) differ by city, with some cities like Visalia, CA, and Evansville, IN, experiencing peak severity at specific hours (e.g., 9 AM, 8 PM).
- **Seasonal Trends:** In Miami, FL, accident counts are lowest in summer (July) and peak in winter (December), indicating a strong seasonal pattern.
- **Hourly & Weekly Patterns:** The heatmap for Miami, FL, reveals that average severity varies by both hour and day of the week, with certain early morning and late-night hours showing slightly higher severity.
- **Actionable Focus:** Emergency services and safety campaigns should be tailored to local danger hours and seasonal trends, focusing resources during high-risk times and in cities with consistently high severity.



Top danger hour for each city (highest avg severity):				
	State	City	Hour	Avg_Severity
452	CA	Visalia	9	4.000000
962	IN	Evansville	20	3.862069
1194	MO	West Plains	21	3.809524
18	AL	Montgomery	12	3.724138
424	CA	Santa Barbara	8	3.703704
1324	NY	Lake George	7	3.217391
1479	PA	Carlisle	20	3.000000
1265	NJ	Brick	7	3.000000
1179	MO	Kansas City	19	2.961538
1499	PA	Mercer	15	2.933333

# DATA VISUALIZATION & INSIGHTS

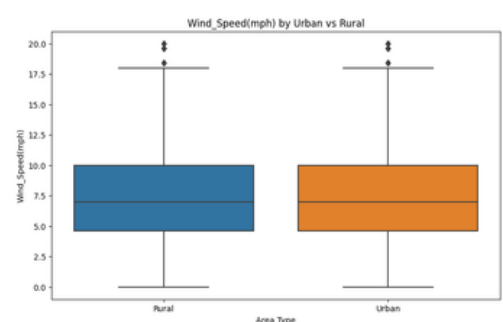
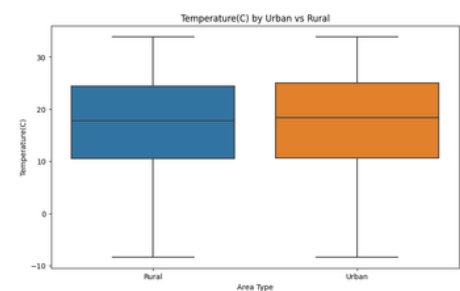
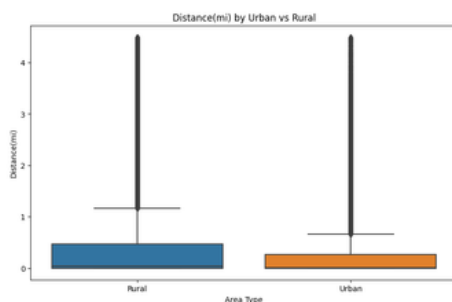
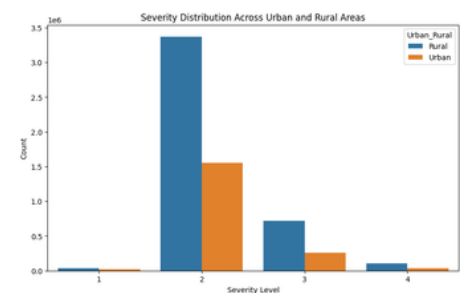
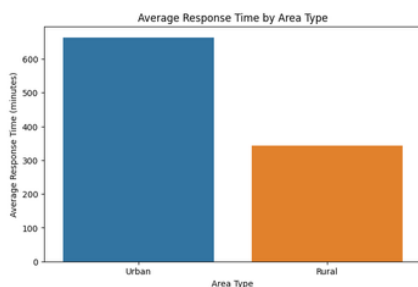
## 14. How do accident patterns differ between population densities (Urban vs. Rural)?

### Business Value:

Helps tailor safety initiatives to the unique challenges of different environment types.

### Insights:

- **Severity Distribution:** Rural areas have a higher proportion of severe accidents (levels 3 and 4), while urban areas see more low-severity cases.
- **Response Time:** Average emergency response time is significantly longer in urban areas, possibly due to congestion or higher call volumes.
- **Environmental Factors:** Boxplots show that wind speed, temperature, visibility, and accident distance distributions are similar between urban and rural areas, but rural accidents tend to involve slightly longer distances.
- **Actionable Focus:** Rural areas need targeted interventions for high-severity accidents and improved emergency access, while urban areas should focus on reducing response delays and managing high accident volumes.





# DATA VISUALIZATION & INSIGHTS

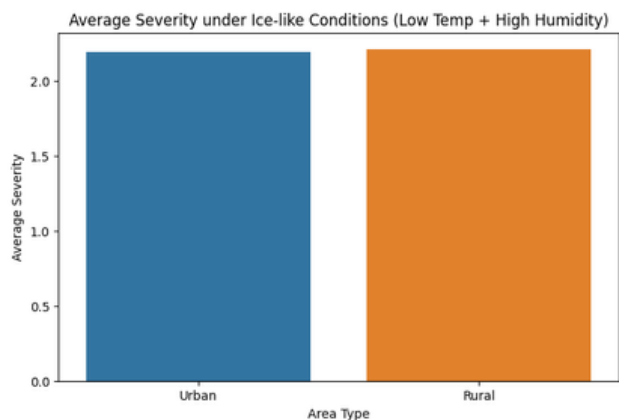
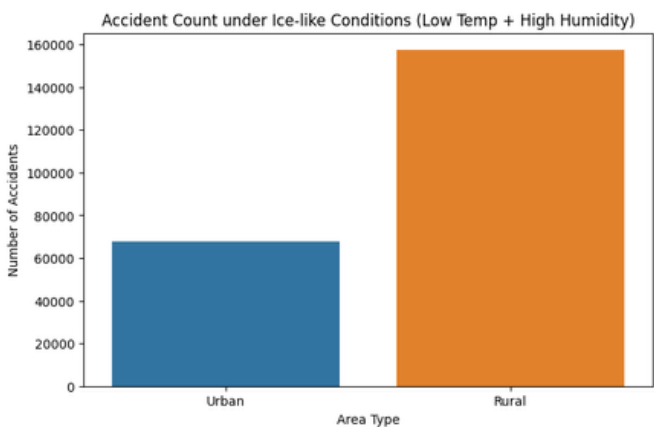
15. Does the impact of combined environmental factors (e.g., low Temperature(C) bin AND high Humidity(%) bin, suggesting potential ice) on accident frequency and severity differ between urban (high-density City/County) and rural areas?

**Business Value:**

Informs infrastructure planning (e.g., road materials resilient to specific conditions in certain area types), targeted de-icing/maintenance schedules, and public safety campaigns tailored to urban vs. rural drivers under specific environmental threats.

**Insights:**

- **Accident Frequency:** Rural areas experience more than twice as many accidents as urban areas under ice-like conditions (low temperature + high humidity).
- **Severity:** The average severity of accidents is high in both urban and rural areas, but slightly higher in rural settings.
- **Actionable Focus:** Rural regions require prioritized de-icing, maintenance, and targeted safety messaging during icy conditions, while urban areas should maintain robust response plans due to consistently high severity.



# DATA VISUALIZATION & INSIGHTS

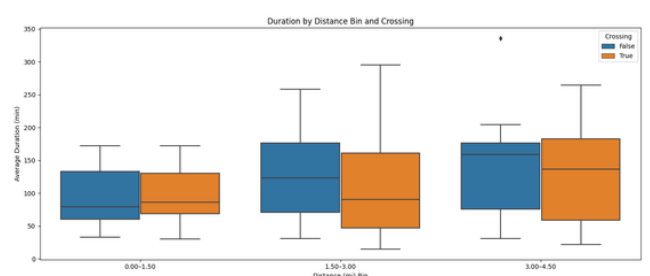
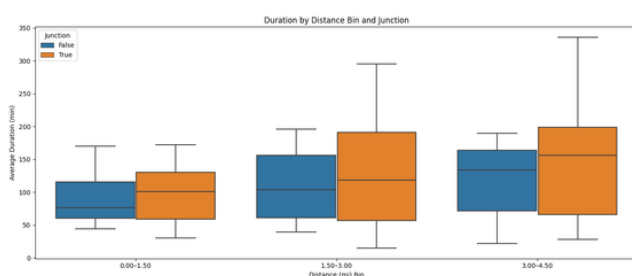
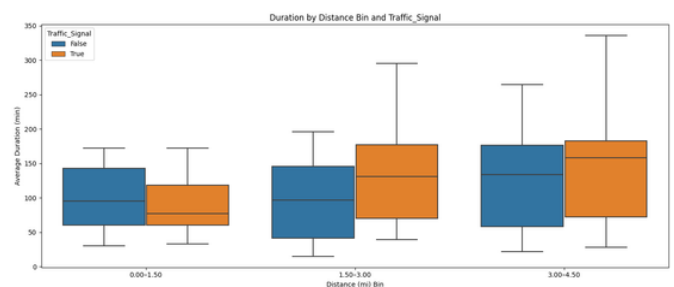
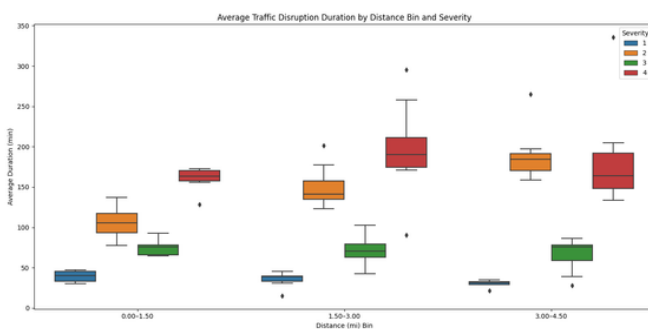
**16. How does the duration of traffic disruption (calculated from Start\_Time, End\_Time) vary based on the combination of Severity, Distance(mi) (discretized), and the presence of key road features (Traffic\_Signal, Junction, Crossing)?**

## Business Value:

Helps emergency response teams, traffic management centers, and towing services estimate clearance times more accurately based on initial accident reports, improving resource deployment and traffic flow management post-accident.

## Insights:

- **Severity & Distance:** Higher severity and longer accident distances are strongly associated with longer traffic disruption durations. Severe accidents (Severity 4) in the largest distance bins can last over 3 hours on average.
- **Road Features Impact:** The presence of traffic signals, junctions, or crossings generally increases average disruption duration for similar distance bins, likely due to increased traffic complexity and safety protocols at these locations.
- **Consistent Patterns:** For all road features, the relationship between distance and duration holds—larger accidents take longer to clear, but the effect is amplified at intersections and crossings.
- **Operational Use:** These patterns enable more precise, data-driven estimates for clearance times, allowing agencies to prioritize resources for severe, large-scale accidents at complex road features.



# DATA VISUALIZATION & INSIGHTS

**17. How do accidents in the United States cluster based on environmental factors such as temperature, humidity, and accident severity?**

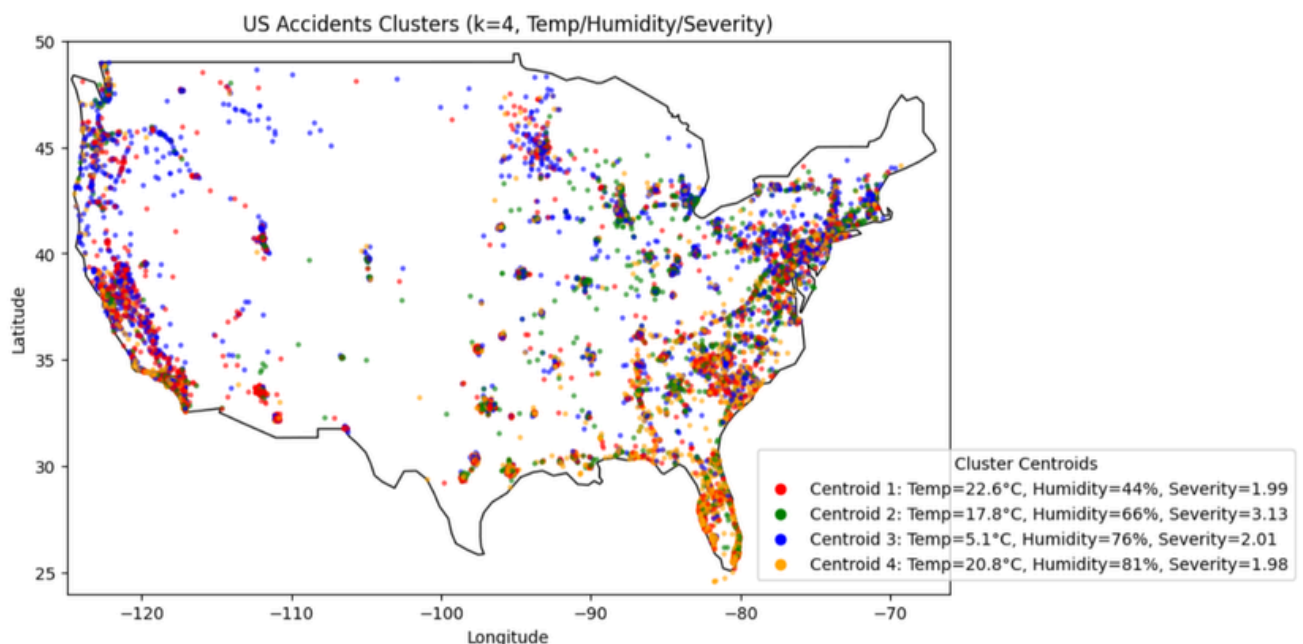
**What patterns can we observe regarding weather conditions and the severity of accidents?**

## Business Value:

Understanding how environmental factors influence accident clustering helps insurance companies, city planners, and emergency services anticipate risk hotspots, tailor safety campaigns, and allocate resources more effectively based on local weather patterns.

## Insights:

- **Cluster 1 (Red):** Warm, dry regions (e.g., Southern California, Texas) see moderate-severity accidents at high temperatures and low humidity.
- **Cluster 2 (Green):** The most severe accidents occur in moderate temperatures and humidity, often in urban or high-traffic areas.
- **Cluster 3 (Blue):** Cold, humid conditions (Northeast, Midwest) are linked to winter-related accidents of moderate severity.
- **Cluster 4 (Orange):** Humid, tropical areas (Florida, Gulf Coast) experience frequent but less severe accidents.
- **Overall:** Accident severity and frequency are strongly influenced by local weather. Notably, moderate weather conditions are associated with the highest severity, suggesting that risk is not limited to extreme weather. This insight can guide targeted interventions and preparedness strategies.



# MODEL TRAINING

In this section, we describe the model training process.

We experimented with Logistic Regression, Random Forest, and Naive Bayes classifiers. The detailed performance results will be discussed in the Model Results section.

Based on the exploration conducted in pre-processing, we observed that accidents with severity level 4 are significantly more serious than accidents at other severity levels, between which the distinction is far less clear. Therefore, we decided to focus specifically on level 4 accidents and regroup the severity levels into a binary classification problem: severity = 4 versus severity  $\neq$  4.

Additionally, for the Logistic Regression model, we performed hyperparameter tuning using cross-validation to optimize key parameters such as the regularization strength (**regParam**), the elastic net mixing parameter (**elasticNetParam**), and the maximum number of iterations (**maxIter**).

## Preprocessing

We performed a small amount of preprocessing where we dropped the following redundant columns:

```
redundant_cols = [  
    "Distance(mi)", "Temperature(C)", "Humidity(%)", "Pressure(in)",  
    "Visibility(mi)", "Wind_Speed(mph)", "Severity", "Weather_Timestamp",  
    "Description", "End_Time", "Distance(mi)_cont",  
]
```

After this preprocessing step, we found that our data was highly imbalanced: The imbalance ratio (majority/minority) was approximately 41.7, indicating a severe skew in class distribution.

is_high_severity	count
0	5,955,314
1	142,813

## Handling Imbalance

Initially, we attempted **downsampling** the majority class using various factors:

- Keeping 10x more majority samples than minority
- Keeping 5x more majority samples than minority
- Keeping 3x more majority samples than minority

Although higher sampling ratios (e.g., 10x and 5x) resulted in high overall accuracy, the models suffered from **severe overfitting**.

In particular, precision and recall for class 1 (serious accidents) were nearly zero. When reducing the ratio to 3x, the models began to exhibit some precision and recall for the minority class.

However, this approach resulted in a training dataset that fell below our desired minimum size of 1 million records.

# MODEL TRAINING

## Further Feature Exploration

Following these findings, we conducted additional exploration to determine if dropping certain features could help.

We resampled 20,000 rows from the dataset and analyzed relationships between Points of Interest (POIs) and accident severity.

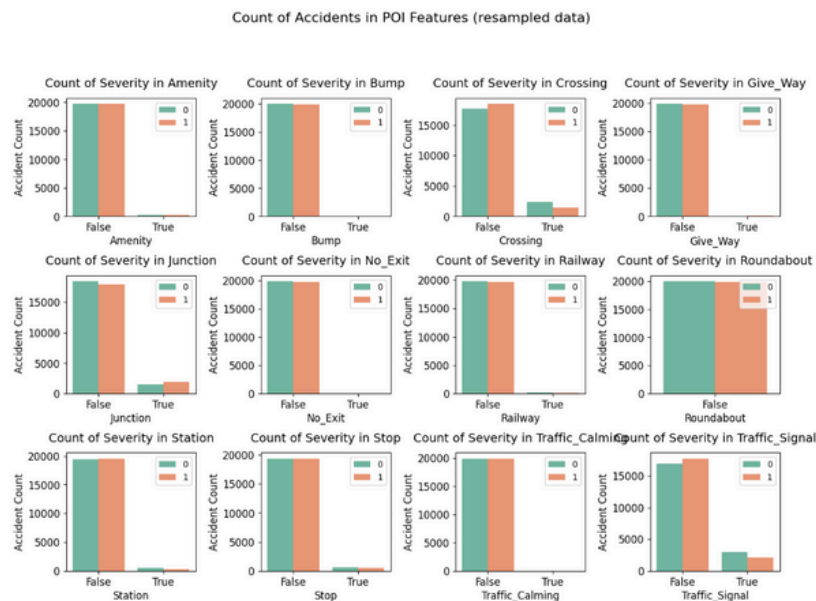
We observed the following:

- Accidents near traffic signals and crossings were less likely to be serious.
- Accidents near junctions were more likely to be serious, possibly because speed is a more critical factor at junctions, whereas drivers usually slow down at crossings and traffic signals.

Additionally, some POI features were so unbalanced that they did not meaningfully contribute to severity prediction.

Thus, we decided to drop the following features:

['Bump', 'Give\_Way', 'No\_Exit', 'Roundabout', 'Traffic\_Calming']



## Upsampling

After feature selection, we attempted **upsampling** to balance the dataset more effectively.

We applied a **1:1 ratio** between majority and minority classes, creating approximately **700,000 records** for each class.

To make the synthetic minority data more realistic, we introduced small amounts of noise into the numerical features.

This approach led to much more sensible model behavior, with improved precision and recall, especially for the minority class.

The full evaluation results will be presented in the **Model Results** section.

# MODEL RESULTS & EVALUATION

For up sampling techniques we discussed before with ratio 1:1 and around 700K row for each class and using 24 features.

## Logistic Regression

precision recall f1-score support

0 0.697 0.673 0.685 143100

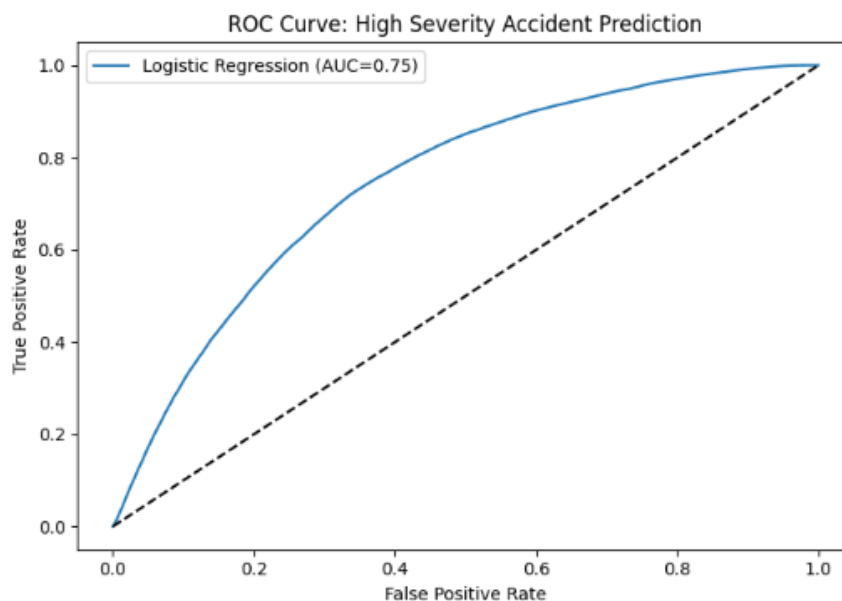
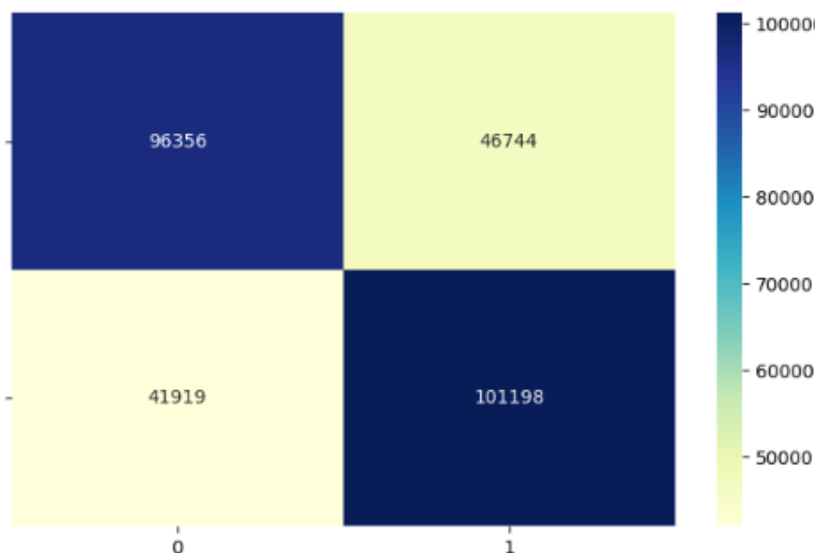
1 0.684 0.707 0.695 143117

accuracy 0.690 286217

macro avg 0.690 0.690 0.690 286217

weighted avg 0.690 0.690 0.690 286217

Confusion Matrix



## Random Forest

Didn't proceed with it as it takes a lot of time and was hard to fit these data and features on it

## Naive Bias

Didn't proceed with it as it takes a lot of time and was hard to fit these data and features on it

# MODEL RESULTS & EVALUATION

For up sampling techniques we discussed before with ratio 1:1 and around 700K row for each class and using 18 features instead of 24.

## Logistic Regression

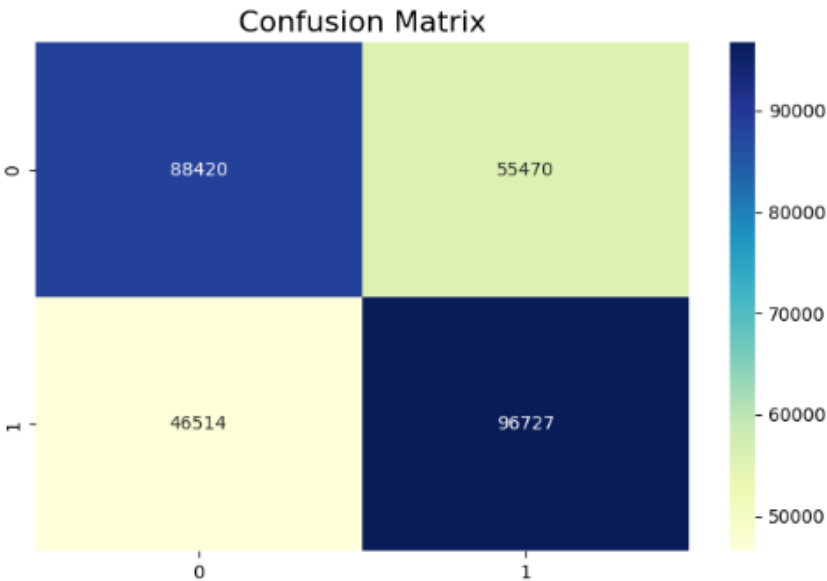
Train set balance

is_high_severity	count
0	574639
1	574179

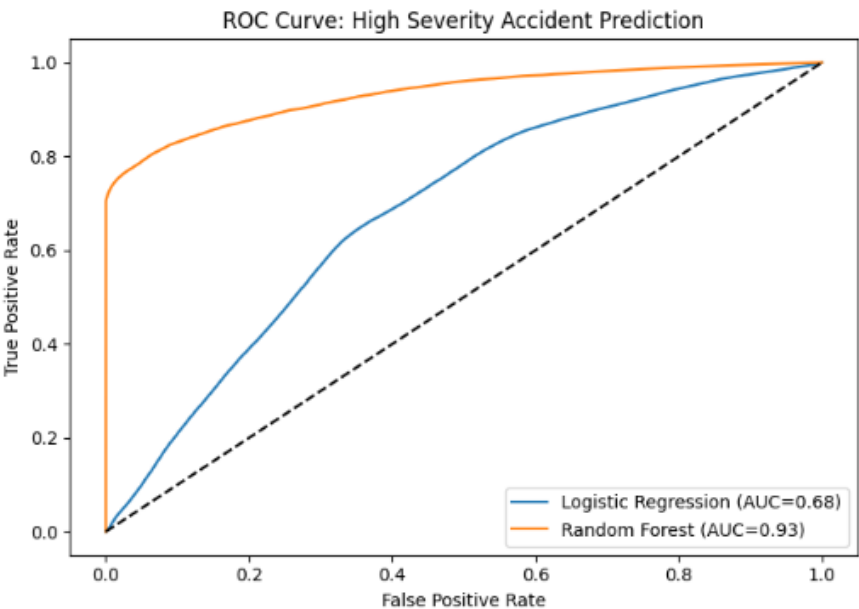
precision    recall    f1-score    support

0	0.655	0.614	0.634	143890
1	0.636	0.675	0.655	143241

accuracy			0.645	287131
macro avg	0.645	0.645	0.645	287131
weighted avg	0.645	0.645	0.644	287131



Logistic Regression AUC: 0.684  
Random Forest AUC: 0.933



# MODEL RESULTS & EVALUATION

For up sampling techniques we discussed before with ratio 1:1 and around 700K row for each class and using 18 features instead of 24.

## Random Forest

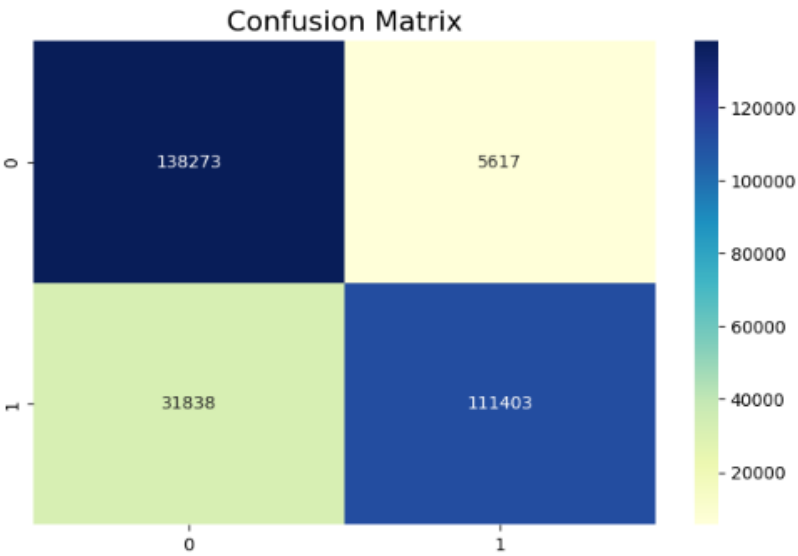
precision recall f1-score support

0	0.813	0.961	0.881	143890
1	0.952	0.778	0.856	143241

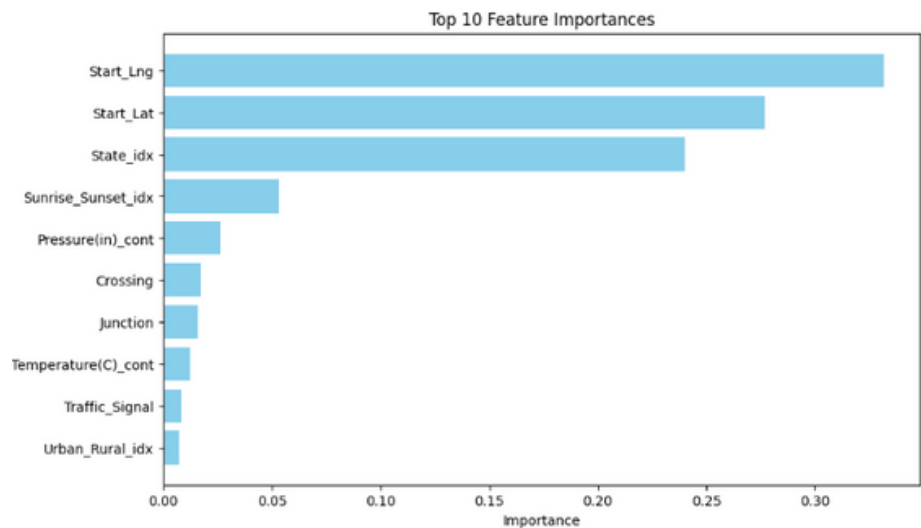
accuracy 0.870 287131

macro avg 0.882 0.869 0.868 287131

weighted avg 0.882 0.870 0.868 287131



- The feature importance plot indicates that high-resolution spatio-temporal accident patterns are the most predictive features for severity, followed by pressure, population, and road type as other key factors.





# MODEL RESULTS & EVALUATION

For up sampling techniques we discussed before with ratio 1:1 and around 700K row for each class and using 18 features instead of 24.

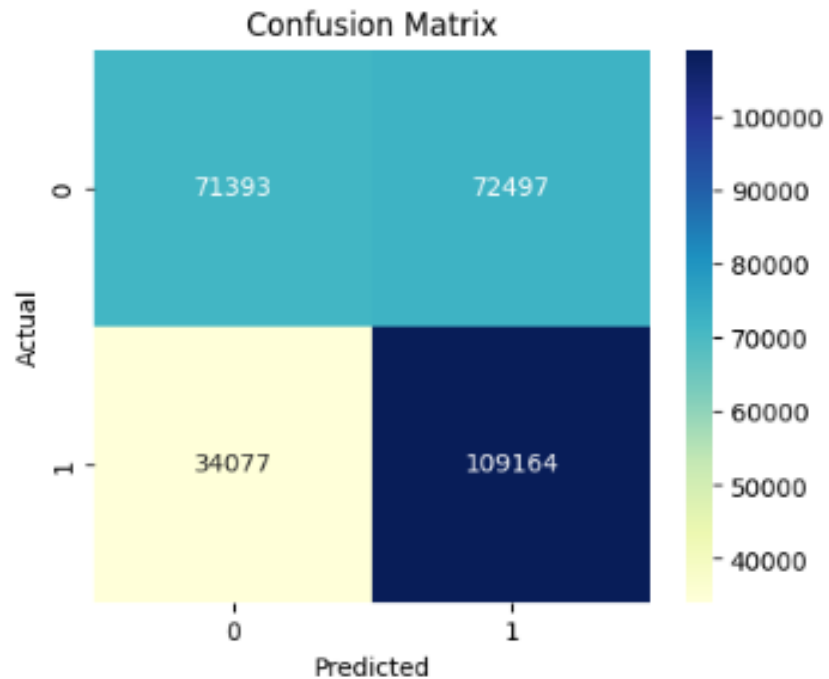
## Naive Bias

**Precision: 0.6009**

**Recall: 0.7621**

**F1 Score: 0.6720**

**Accuracy: 0.6288**



**Class priors: {0: 0.5002002057767201, 1: 0.4997997942232799}**

Indicating we did a good balancing

# UNSUCCESSFUL TRIALS

Here we did a down sampling with 10x , 5x

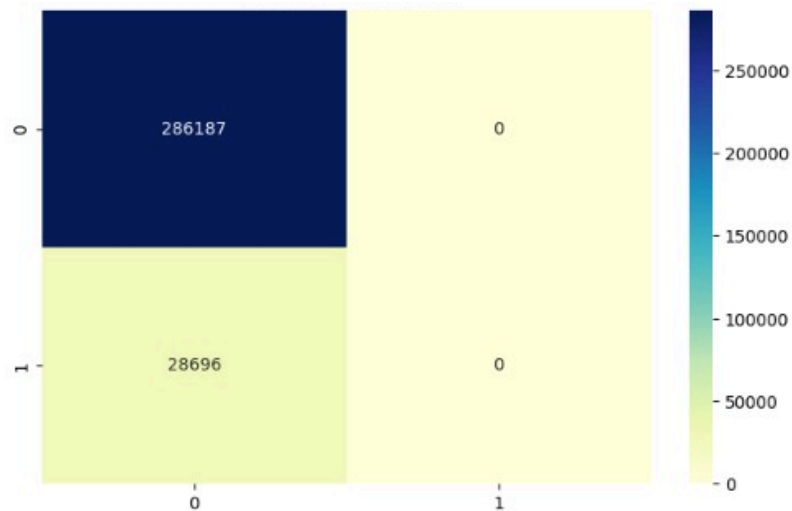
## Logistic Regression

	precision	recall	f1-score	support
0	0.909	1.000	0.952	286187
1	0.000	0.000	0.000	143117

accuracy			0.909	314883
macro avg	0.454	0.500	0.476	314883
weighted avg	0.826	0.909	0.909	314883

Confusion Matrix

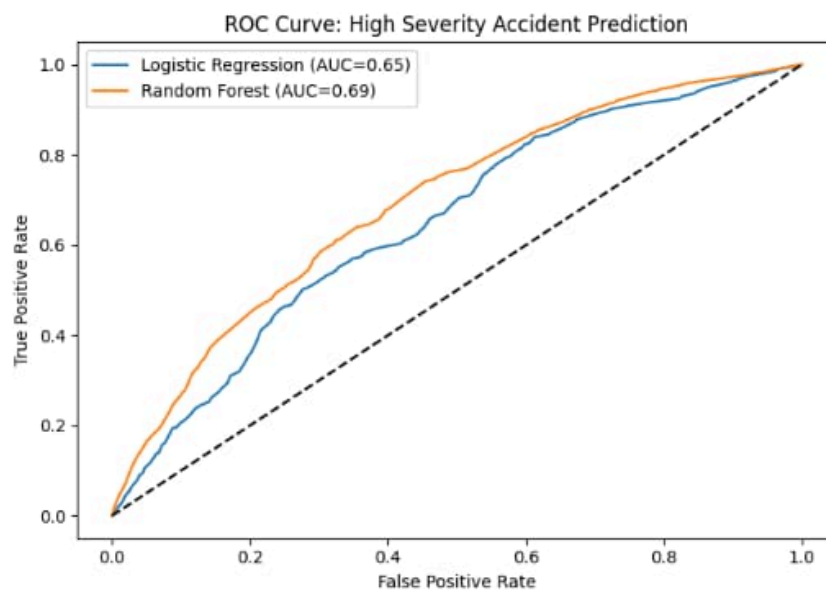


## Random Forest

Just exactly the same results from the logistic regression

## Naive Bias

Didn't proceed with it



# UNSUCCESSFUL TRIALS

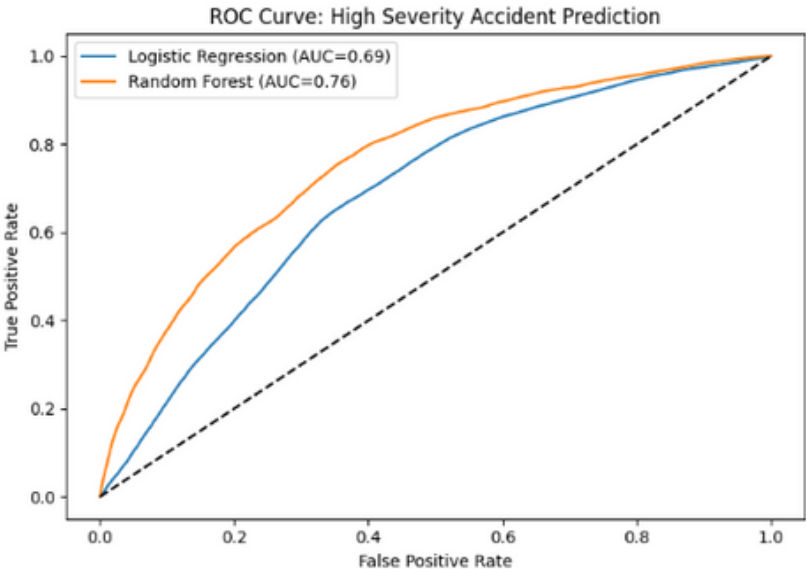
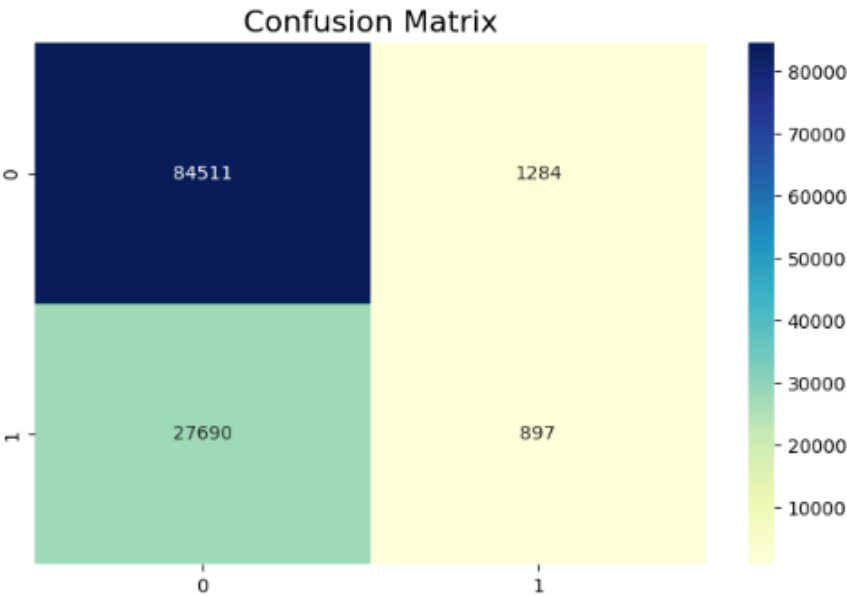
Here we did a down sampling with 3x

## Logistic Regression

	precision	recall	f1-score	support
0	0.753	0.985	0.854	85795
1	0.411	0.031	0.058	28587
accuracy	0.747 114382			
macro avg	0.582	0.508	0.456	114382
weighted avg	0.668	0.747	0.655	114382

Train set balance	
is_high_severity	count
0	345033
1	114897

- Best Regularization Parameter (regParam): 0.01
- Best ElasticNet Parameter (elasticNetParam): 0.0(ridge)
- Best MaxIter: 50

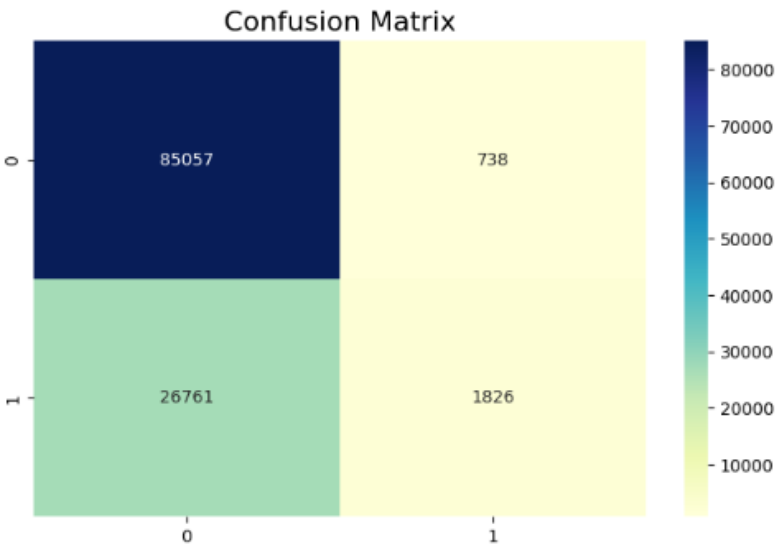


# UNSUCCESSFUL TRIALS

Here we did a down sampling with 3x

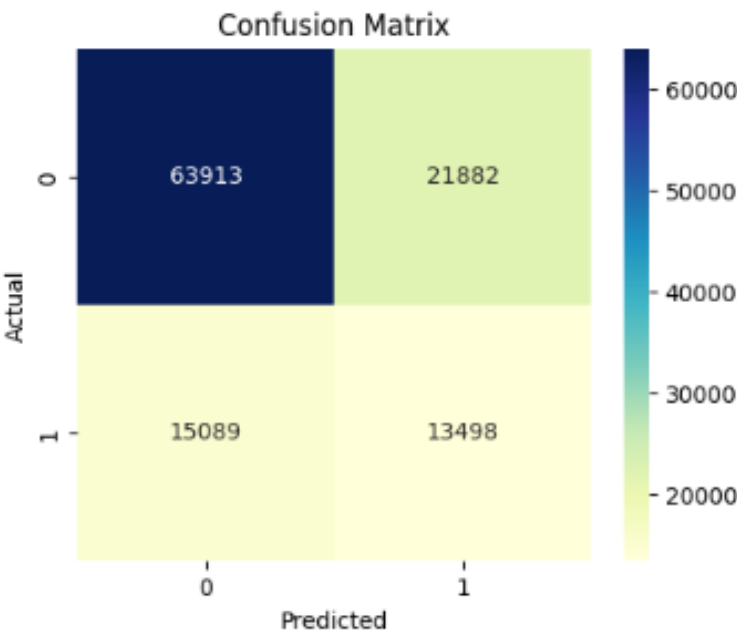
## Random Forest

	precision	recall	f1-score	support
0	0.761	0.991	0.861	85795
1	0.712	0.064	0.117	28587
accuracy				0.760 114382
macro avg				0.736 0.528 0.489 114382
weighted avg				0.749 0.760 0.675 114382



## Guassian Naive Bias(Map-Reduce)

Precision: 0.3815  
Recall: 0.4722  
F1 Score: 0.4220  
Accuracy: 0.6768



# ENHANCEMENTS & FUTURE WORK

1. **Enhanced Handling of Class Imbalance**
2. **Testing other classification algorithms like Gradient Boosting Machines (e.g., XGBoost, LightGBM) or deep learning models.**
3. **Expanding Data Sources:**
  - a. Incorporating additional data sources could greatly improve the model's predictive capabilities.
  - b. One opportunity is to integrate demographic information. For instance, we can leverage census data: we obtained several key variables — including total population, the percentage of commuters by driving, public transit, or walking, and median household income — from the ACS 5-Year Estimates (2018) for all counties. County names were then extracted to link this information to the accident data.
4. **Integration into a Real-Time Prediction System:** A natural next step would be to integrate the model into a real-time accident risk prediction system. Such a system could predict the risk of severe accidents dynamically, possibly by mapping real-time traffic, weather, and location