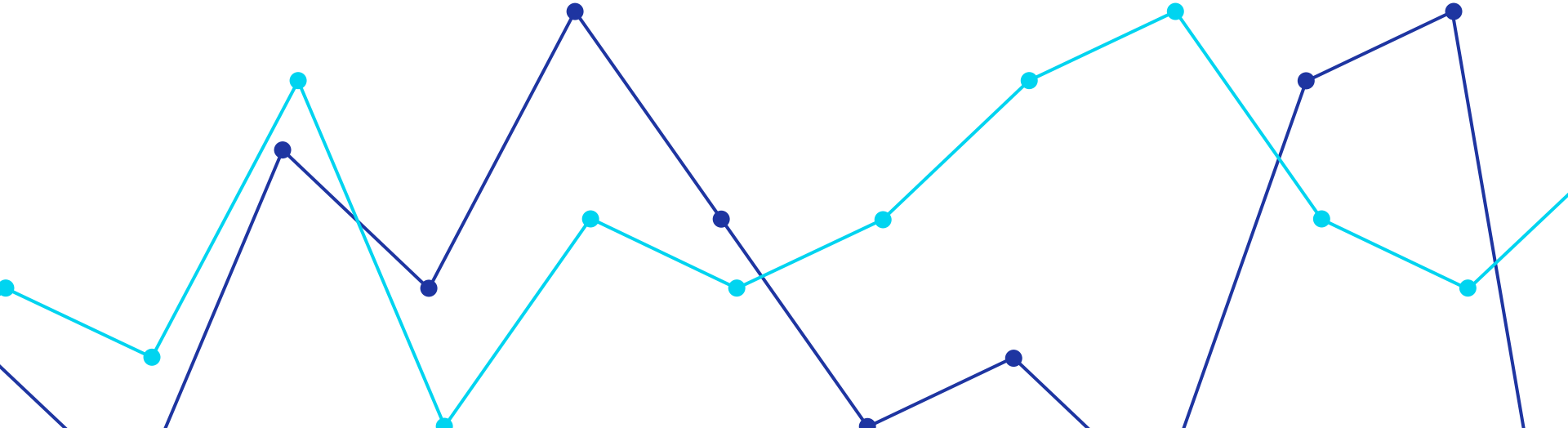
US Accidents Analysis

Ahmed Osama Helmy
Aliaa Gheis
Abdallah Ahmed
Omar Mahmoud



Problem Definition

- Road safety is a critical concern, and understanding accident patterns can help cities improve traffic management and reduce accident rates.
- This project aims to analyze accident data to identify high-risk locations, contributing factors, and potential mitigation strategies.
- By leveraging big data processing, we will extract valuable insights for transportation authorities and urban planners.

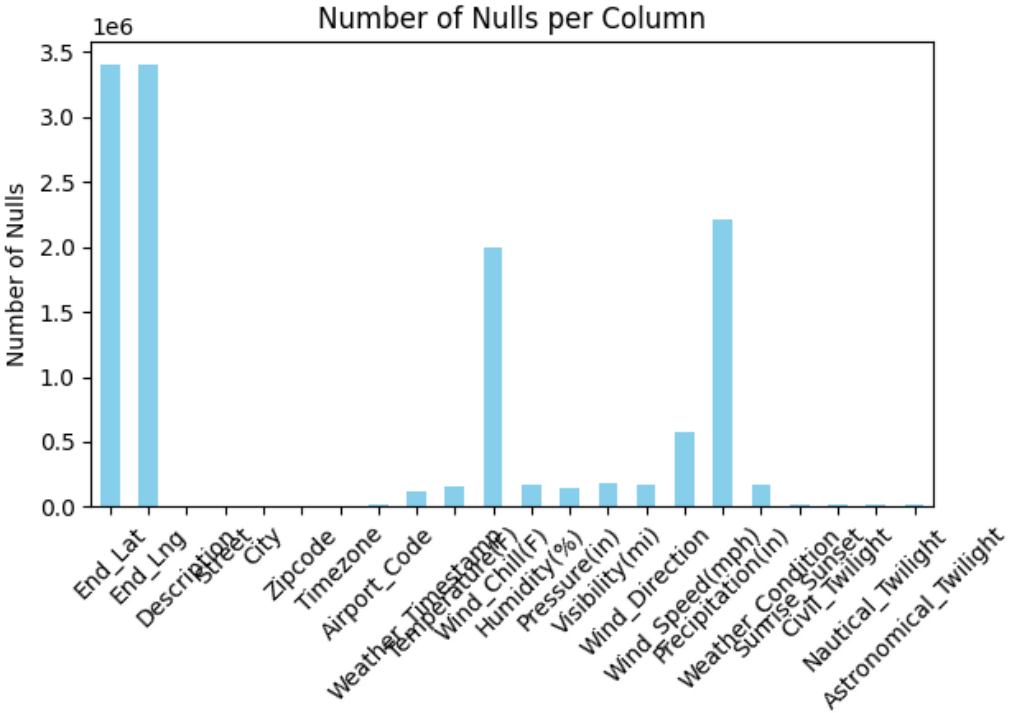
From Data to Insights

- **Data Exploration**
- **Data Visualization**
- **Data Cleaning**
- **Feature Engineering**
- **Aggregation or using models**
- **Visualizing & Extracting Insights**

Data Exploration/Visualization

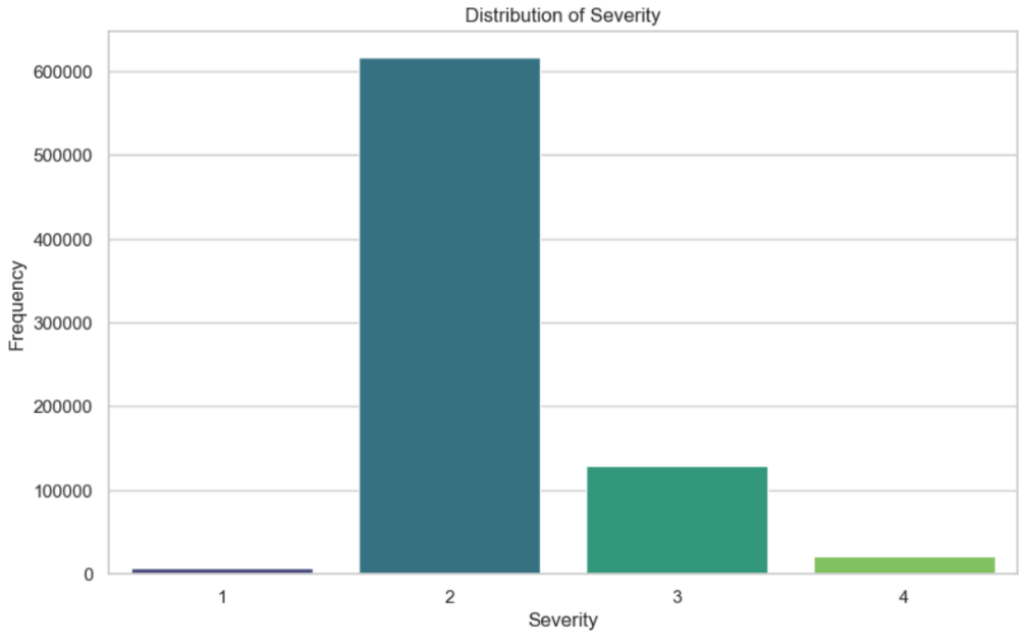
Nulls in Data

Some Columns have a lot of missing values



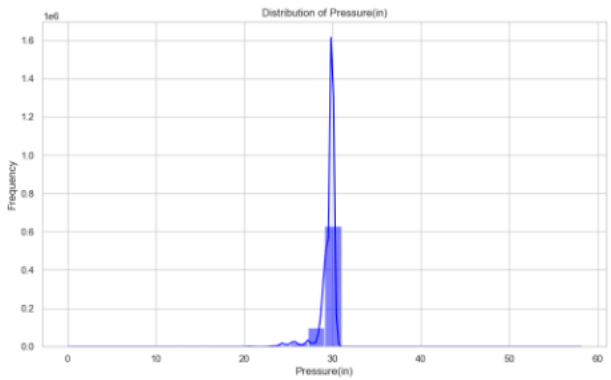
Severity Distribution

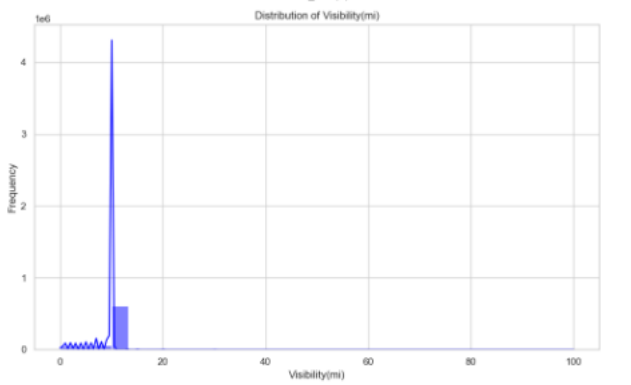
Severity 2 is dominating

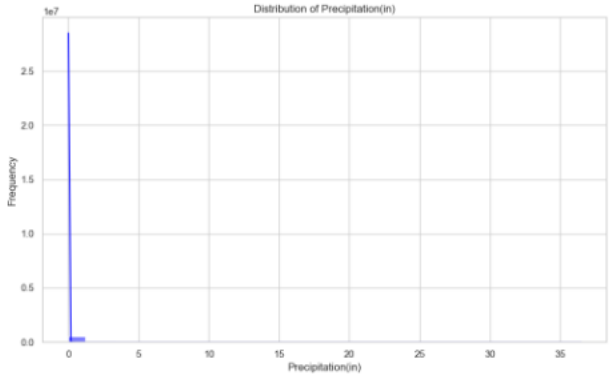


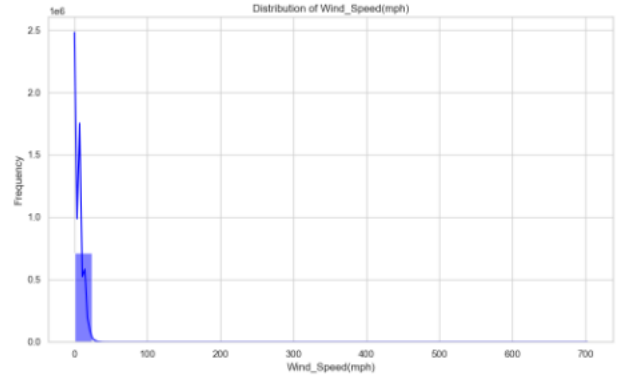
Outliers

Some columns have outliers



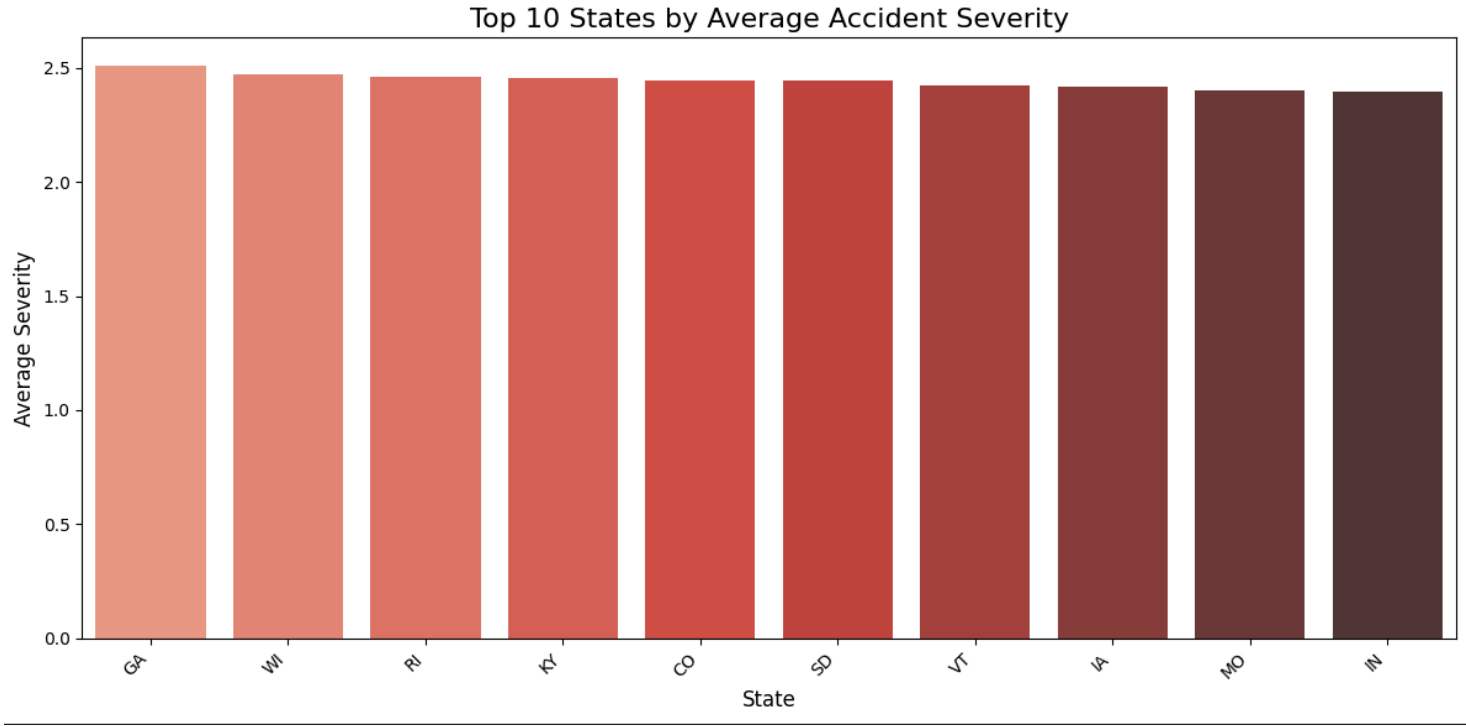






Avg Severity per state

Uniform Distribution



Data Cleaning

Irrelevant Columns

Eliminates columns that don’t provide meaningful data for modeling or analysis.

Columns such as ID, Source, Description, Street, City, Zipcode, Airport_Code, etc., are dropped as they are not useful for analysis or prediction tasks.

Handling Missing Values

Dropping Columns like End_Lat & End_Lng as the percentage of missing values was greater than 40%

Imputing missing values in numeric columns by inserting the mean value

Imputing missing values in categorical columns by inserting the mode value

Outliers

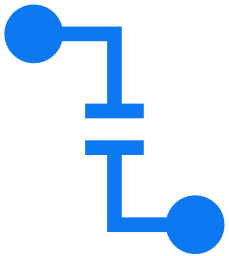
Eliminating records with temperature higher than 56.7 C as reported in this article that the maximum US temperature was 134.4°F (56.7°C)

For Wind Speed values, we identified outliers by considering the maximum observed wind speeds. According to the World Meteorological Organization, the highest recorded wind speed was 254 mph (408 km/h). We decided to remove any records with wind speeds exceeding this threshold to eliminate extreme outliers.

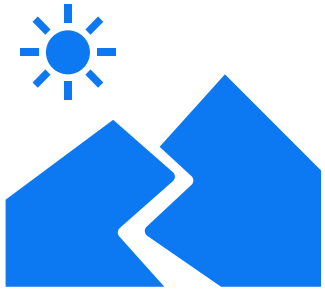
We used IQR in other columns

Feature Engineering

Adding Road-Related Features



A Boolean variable Is_Complex_Road was added to interpret whether the road is complex by utilizing the other variables like (Junction, Railway, Crossing)



This will help in giving insights into the effect of complexity of roads.

Risk Score

The dataset was aggregated at the state level to simplify the analysis and provide actionable insights at a regional scale.

Key metrics such as average accident severity, total accident count, and risk score were computed for each state.

The **Risk score** was calculated as the product of average severity and accident count, capturing both the frequency and severity of accidents.

The **Risk score** was then normalized to be from 0 to 1.

Also, A Boolean variable **Is_High_Risk** was added to detect if a state was high risky or not by using the 75th quartile.

Time Related Features

- **Hour of the Day**: Captures the time of day when accidents occur (e.g., morning rush hours, nighttime)
- **Day of the Week**: Identifies whether accidents are more frequent on weekdays or weekends.
- **Month**: Highlights seasonal trends in accident occurrences (e.g., higher rates during winter months due to adverse weather conditions).
- **Year**: Tracks long-term trends in accident frequency over multiple years.
- **Duration**: Tracks the duration of the accident in minutes by subtracting the start time from the end time
- **Season**: Determines the Season when the accident happened (Summer, ...)

Outliers

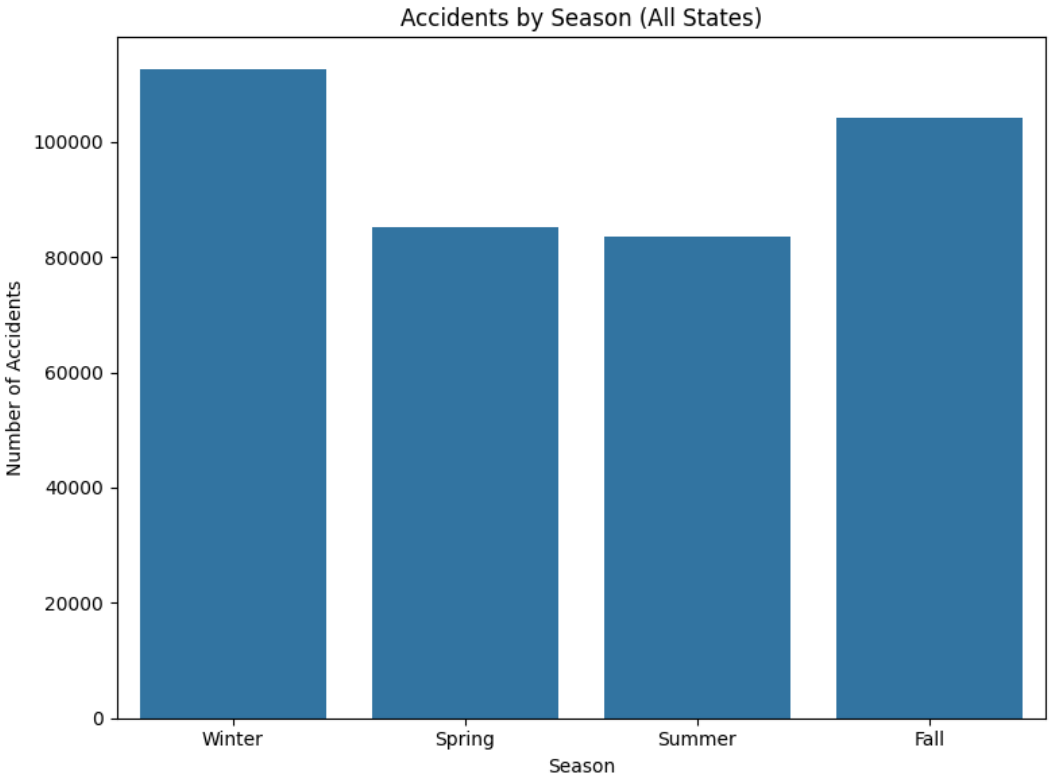
- Eliminating records with temperature higher than 56.7 C as reported in this article that the maximum US temperature was 134.4°F (56.7°C)
- For Wind Speed values, we identified outliers by considering the maximum observed wind speeds. According to the World Meteorological Organization, the highest recorded wind speed was 254 mph (408 km/h). We decided to remove any records with wind speeds exceeding this threshold to eliminate extreme outliers.
- We used IQR in other columns

Outliers

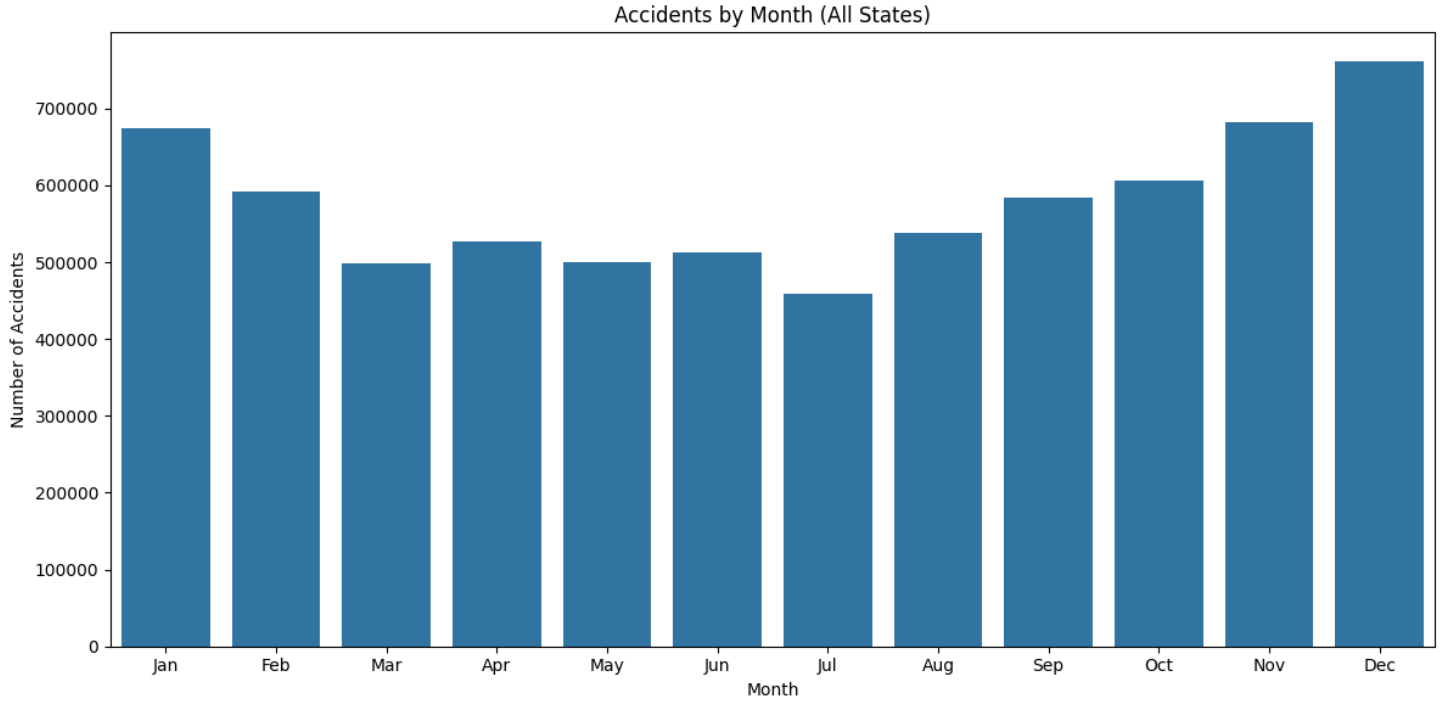
- Eliminating records with temperature higher than 56.7 C as reported in this article that the maximum US temperature was 134.4°F (56.7°C)
- For Wind Speed values, we identified outliers by considering the maximum observed wind speeds. According to the World Meteorological Organization, the highest recorded wind speed was 254 mph (408 km/h). We decided to remove any records with wind speeds exceeding this threshold to eliminate extreme outliers.
- We used IQR in other columns

Accidents based on Time

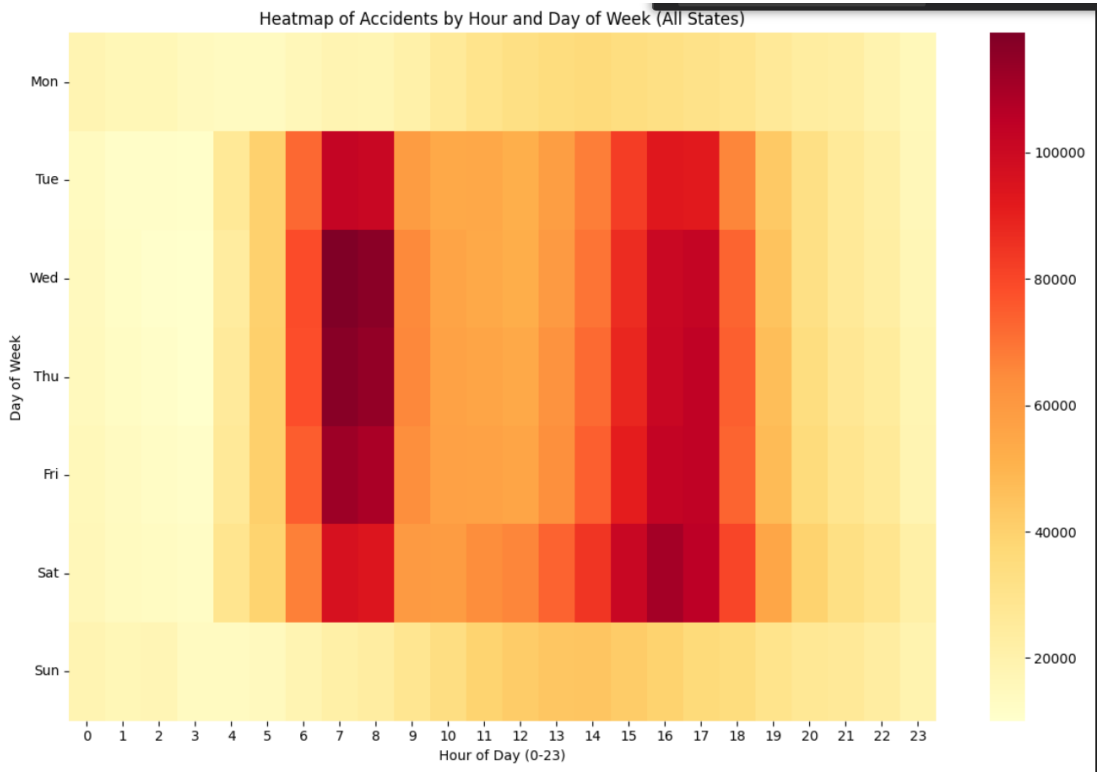
Accident Trends by Season.



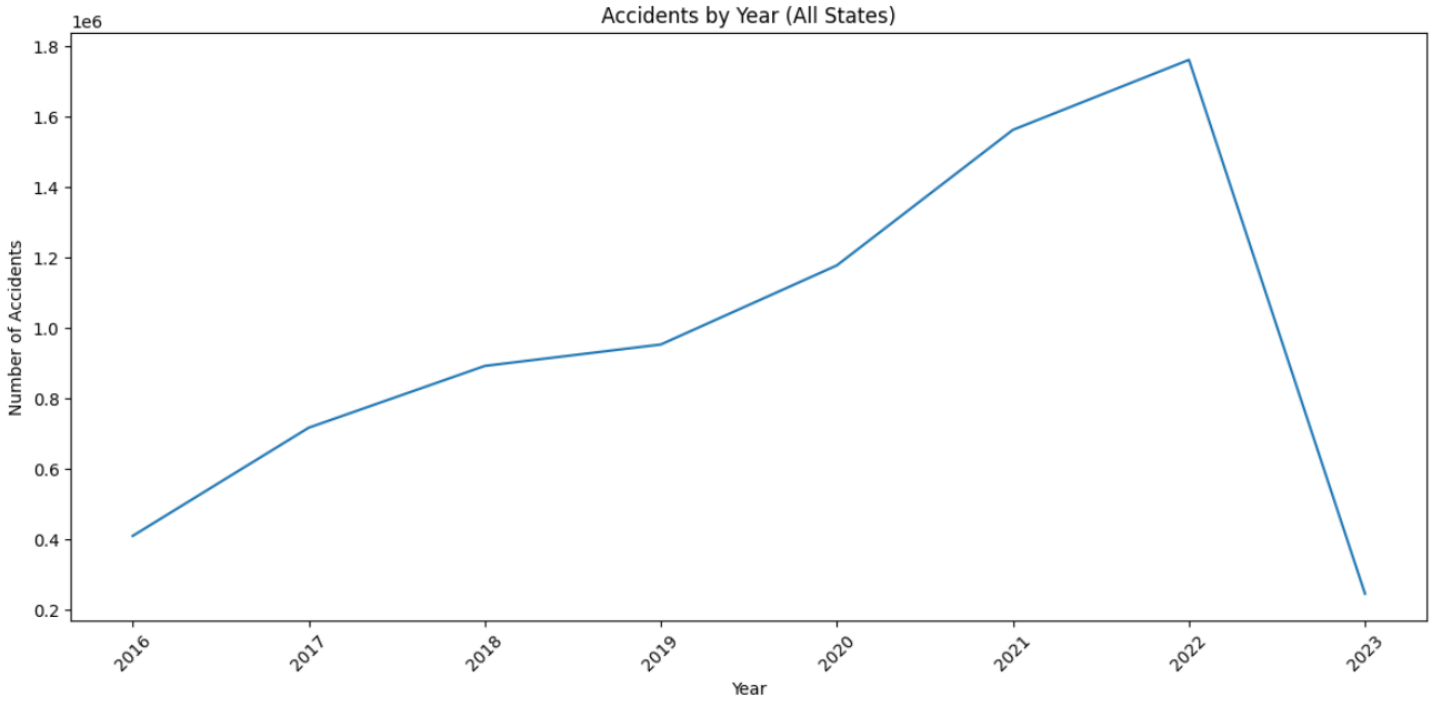
Accident Trends by Month.



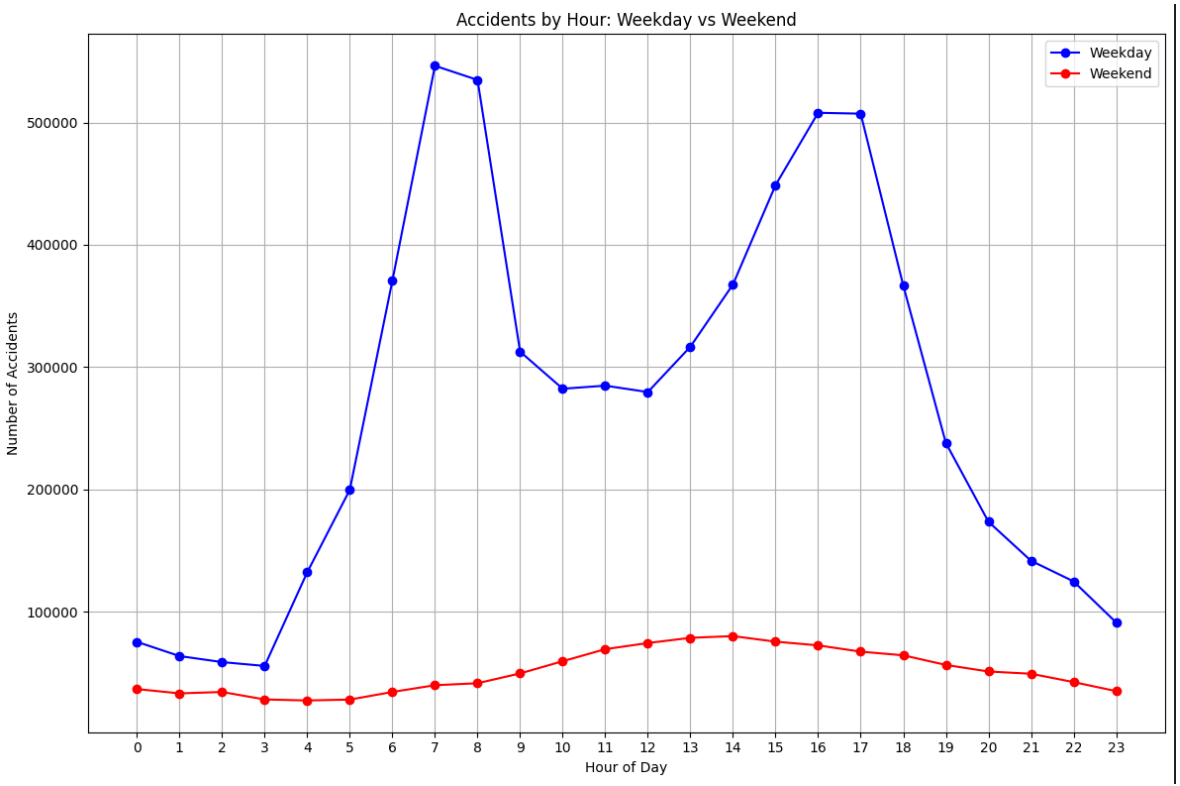
Accident Trends by Hour.



Accident Trends by Year.

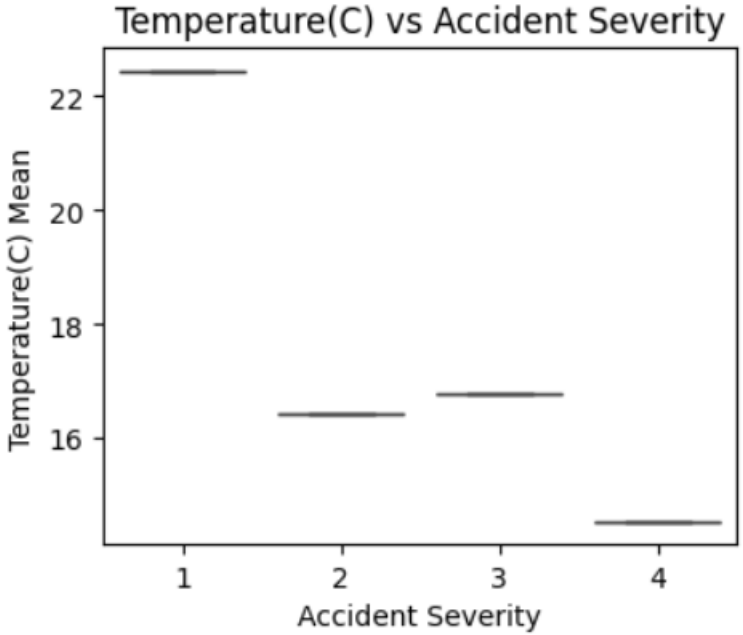


Weekends and Weekdays behavior

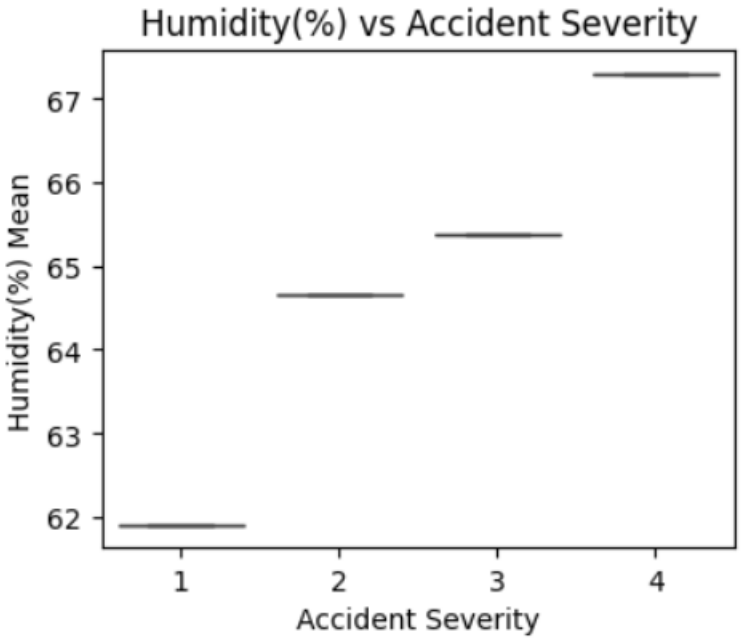


Weather Effect on Accidents

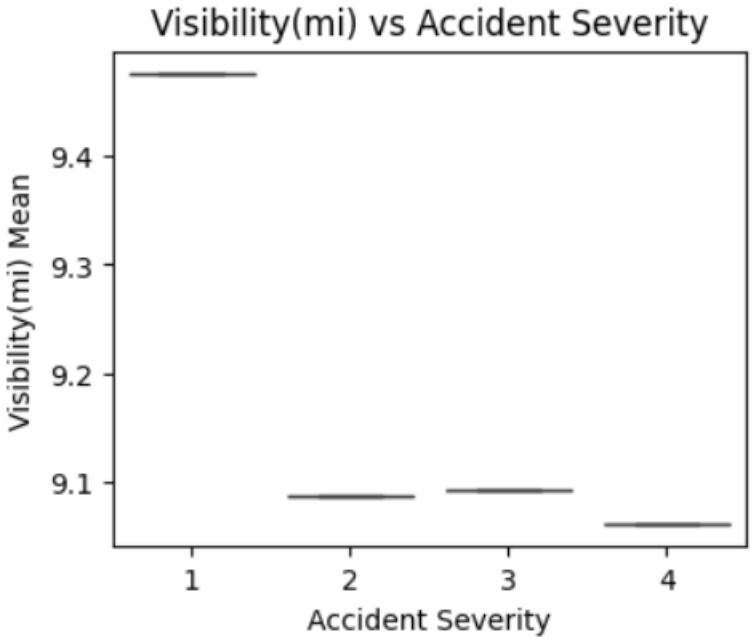
Temperature & Severity



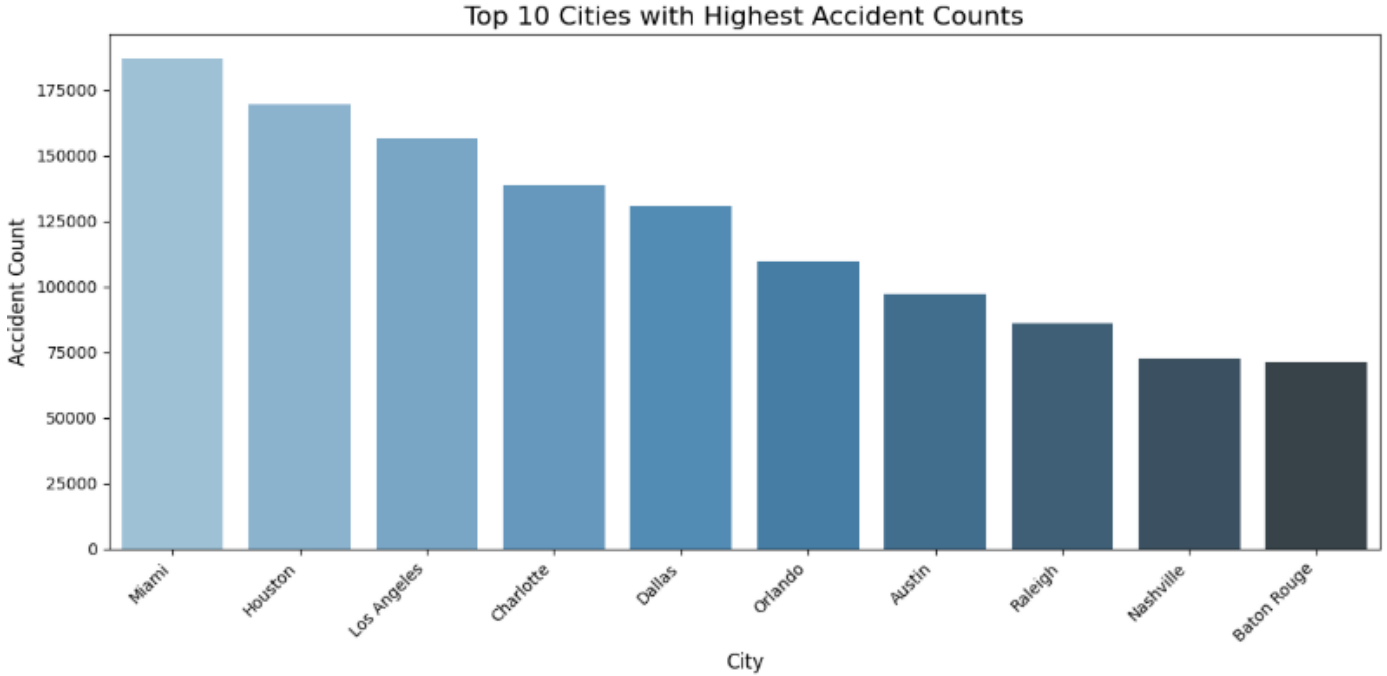
Humidity & Severity

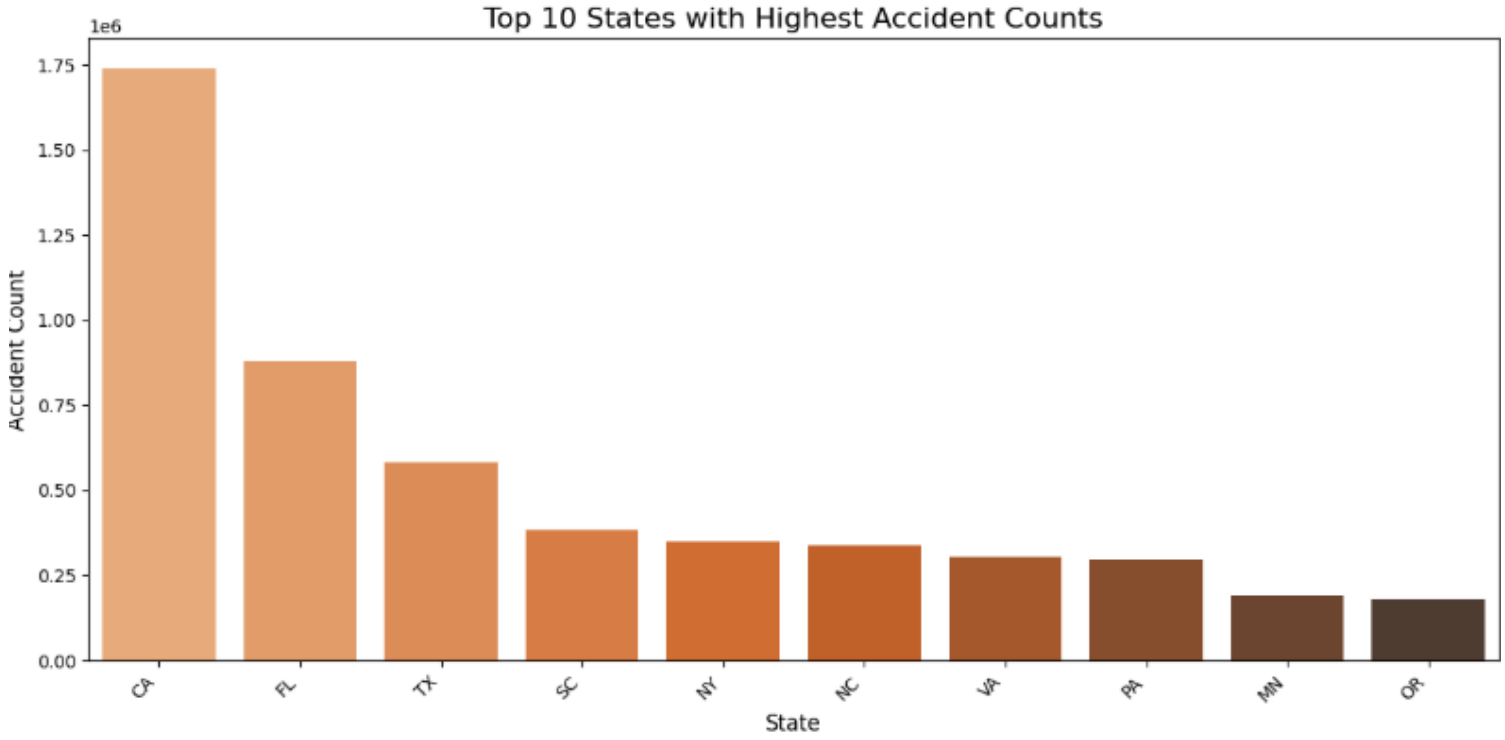


Visibility & Severity

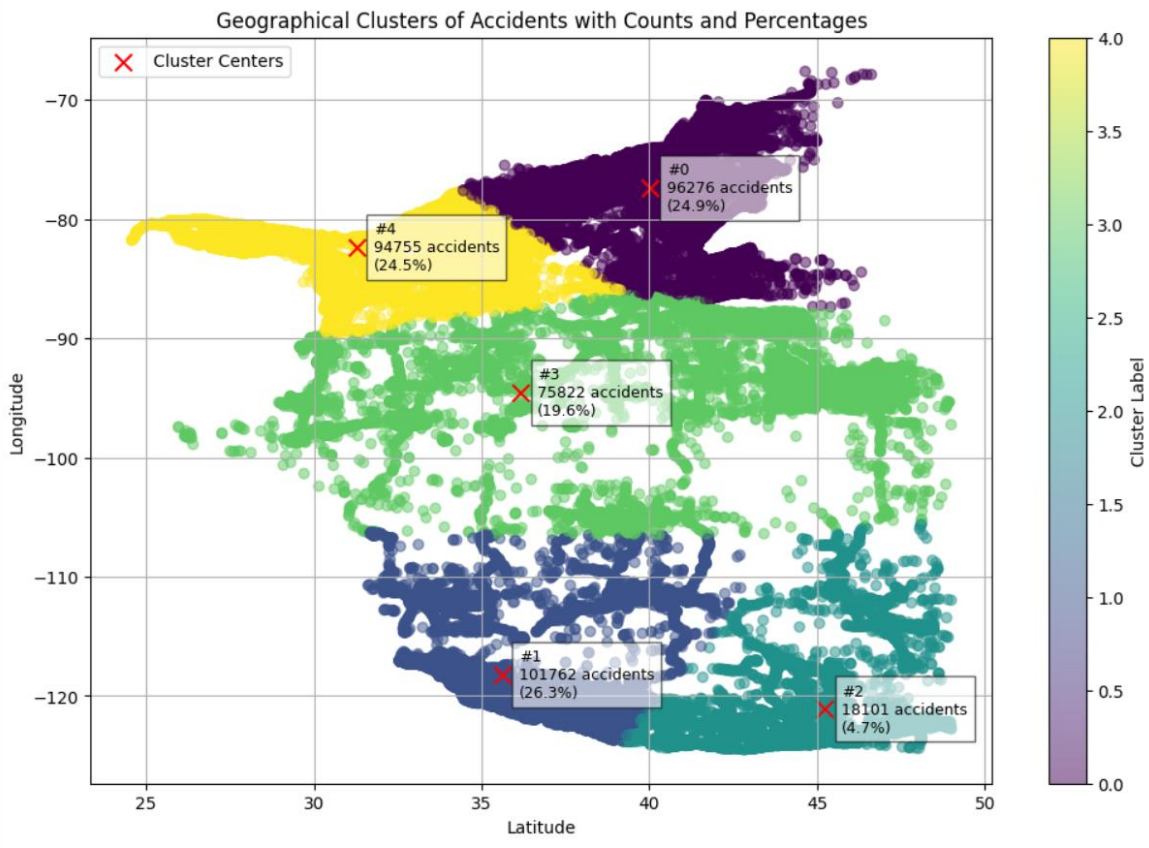


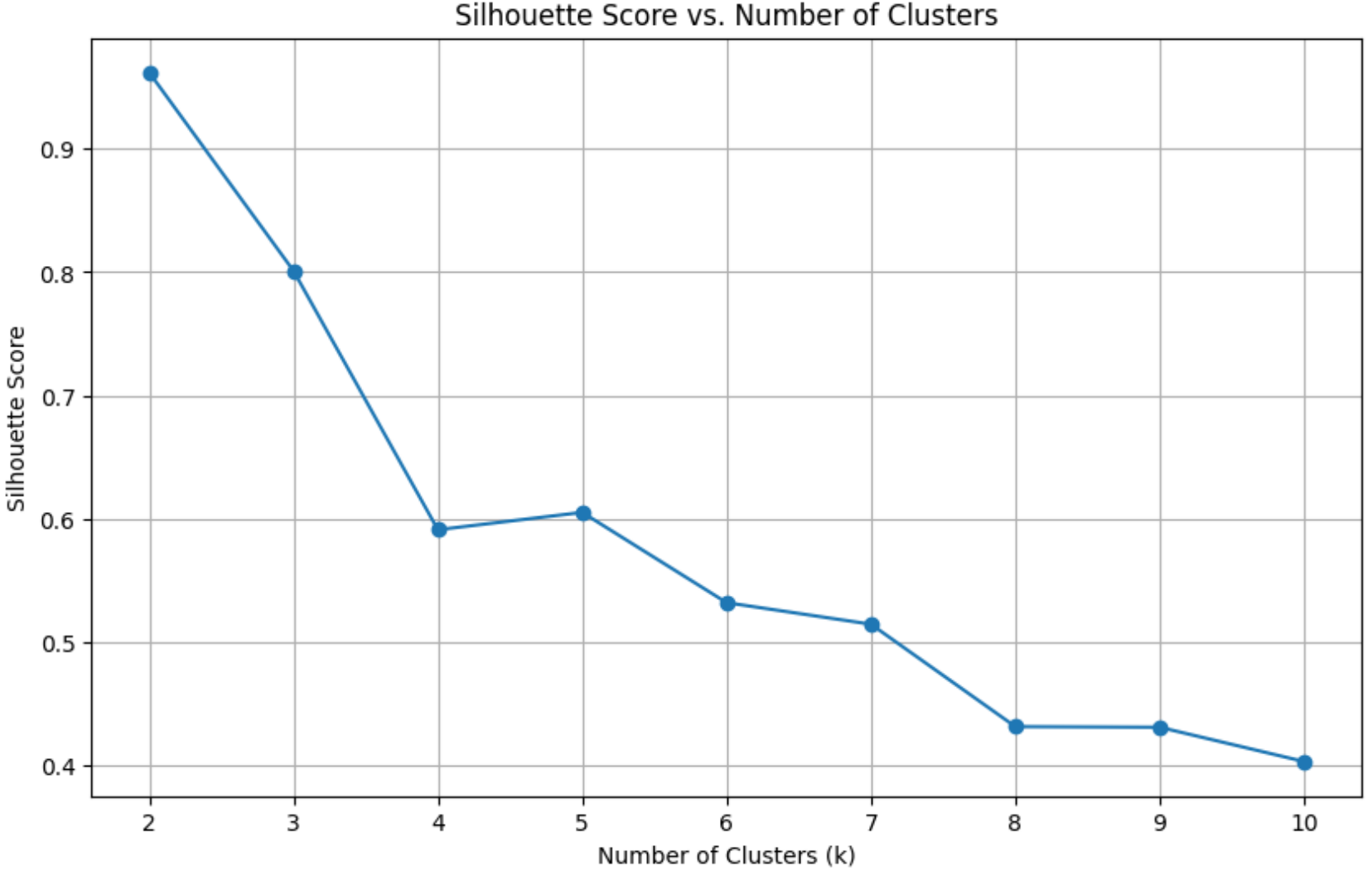
Accidents based on Location


Top 10 Cities with Highest Accident Counts
Accident Count
0
25000
50000
75000
100000
125000
150000
175000
Miami
Houston
Los Angeles
Charlotte
Dallas
Orlando
Austin
Raleigh
Nashville
Baton Rouge
City

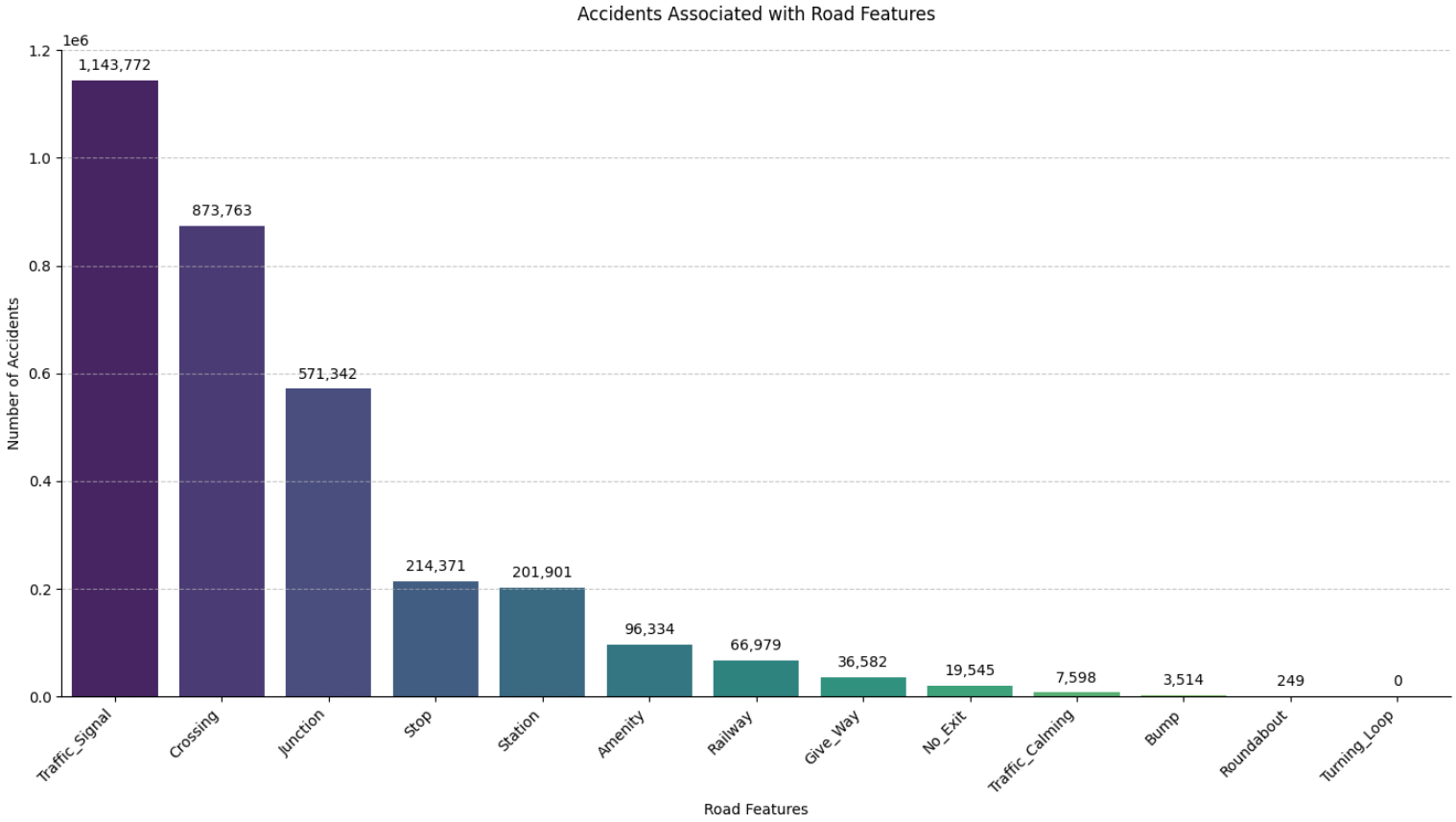

Top 10 States with Highest Accident Counts
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Accident Count
CA
FL
TX
SC
NY
NC
VA
PA
MN
OR
State

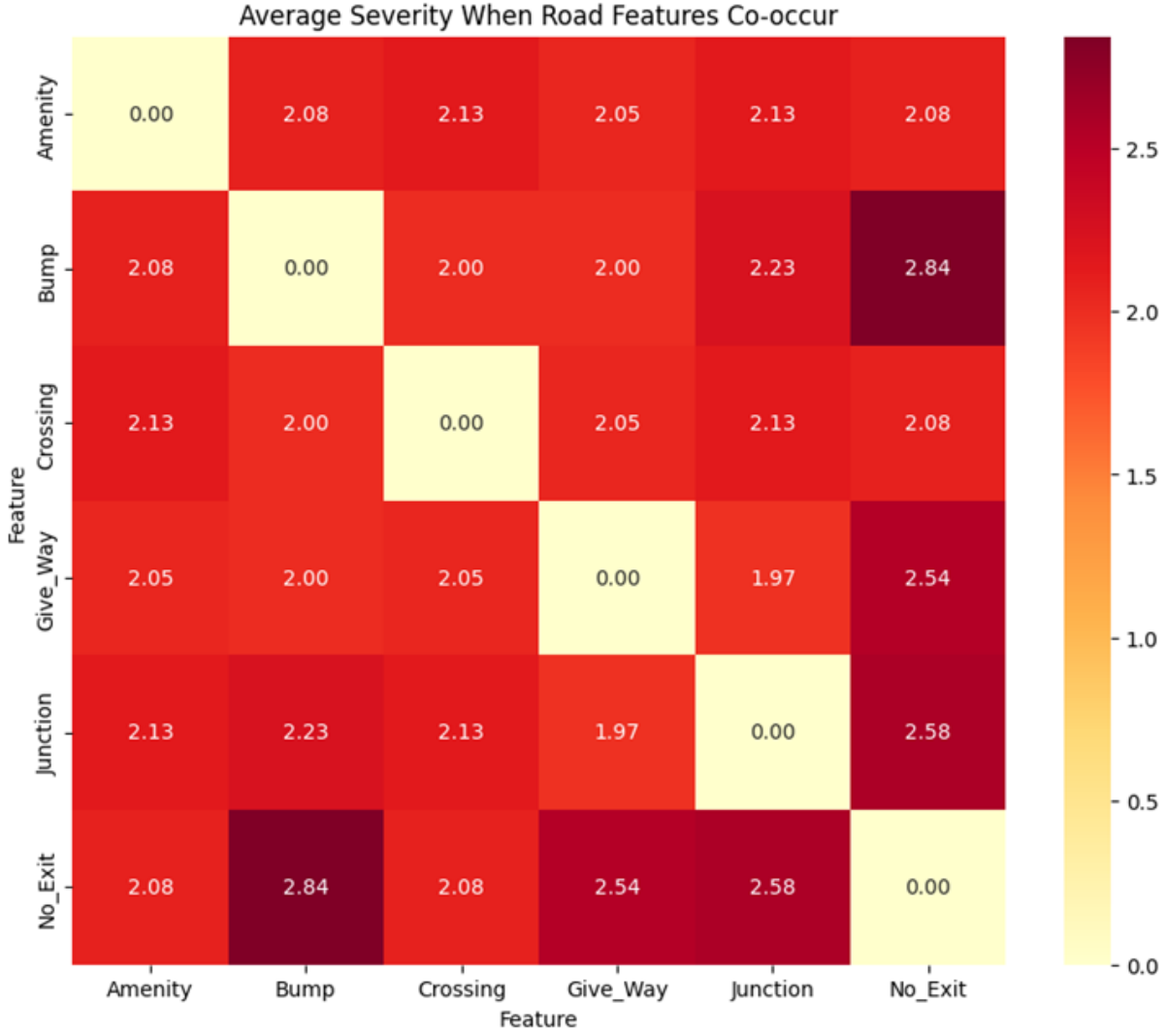
Top 5 Clusters in US with Accidents




Silhouette Score vs. Number of Clusters
Silhouette Score
0.9
0.8
0.7
0.6
0.5
0.4
2
3
4
5
6
7
8
9
10
Number of Clusters (k)

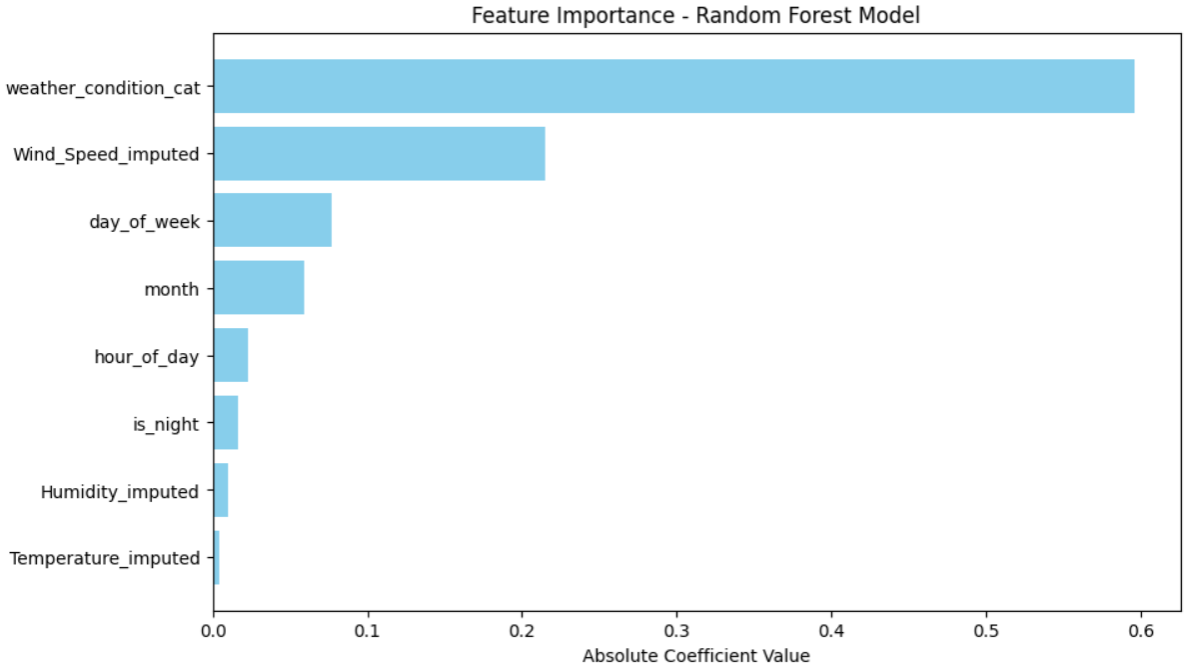
Road Conditions & Accidents


Accidents Associated with Road Features
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Number of Accidents
1,143,772
873,763
571,342
214,371
201,901
96,334
66,979
36,582
19,545
7,598
3,514
249
0
Traffic_Signal
Crossing
Junction
Stop
Station
Amenity
Railway
Give_Way
No_Exit
Traffic_Calming
Bump
Roundabout
Turning_Loop
Road Features


Average Severity When Road Features Co-occur
Feature
Amenity
Bump
Crossing
Give_Way
Junction
No_Exit
0.00 2.08 2.13 2.05 2.13 2.08
2.08 0.00 2.00 2.00 2.23 2.84
2.13 2.00 0.00 2.05 2.13 2.08
2.05 2.00 2.05 0.00 1.97 2.54
2.13 2.23 2.13 1.97 0.00 2.58
2.08 2.84 2.08 2.54 2.58 0.00
Amenity
Bump
Crossing
Give_Way
Junction
No_Exit
Feature
2.5
2.0
1.5
1.0
0.5
0.0

Predicting Severity & Risk

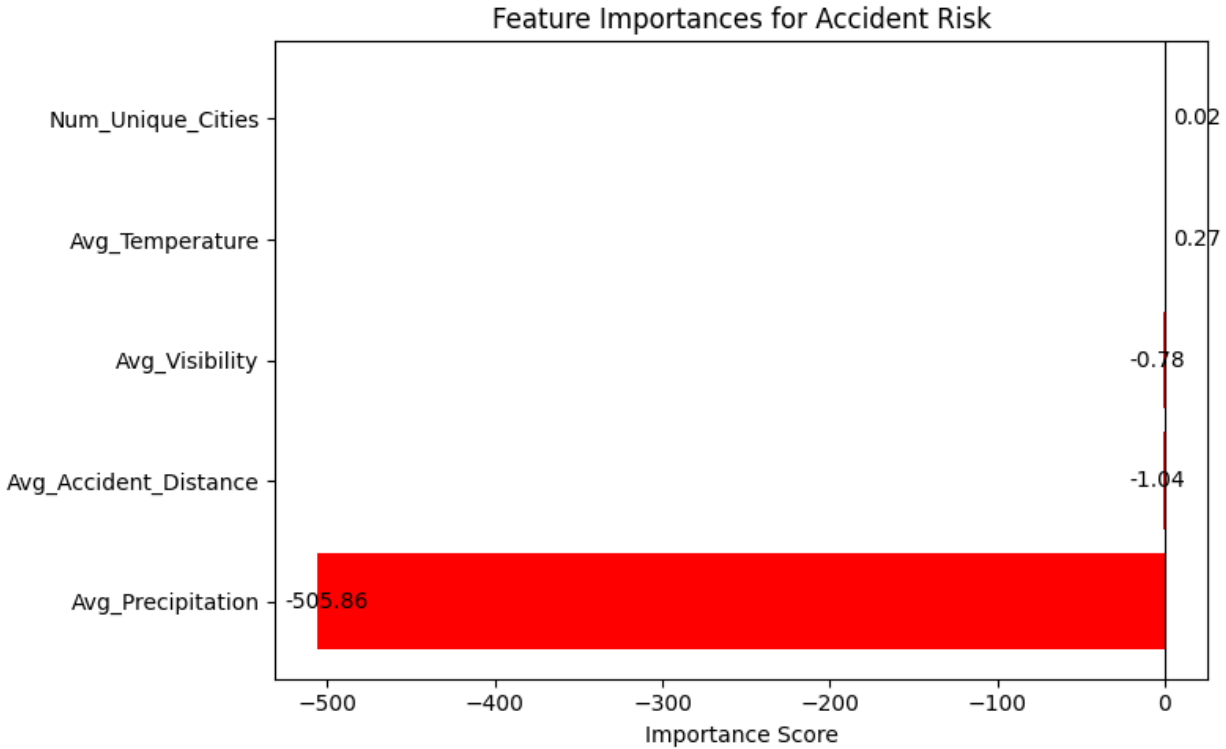
Multiclass Classification



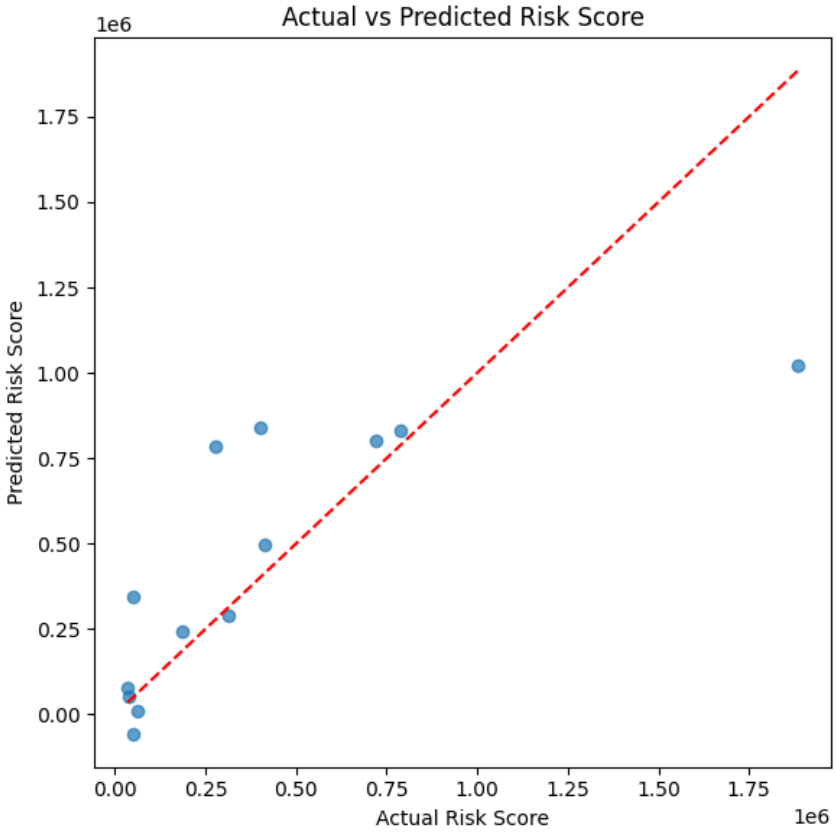
Binary Classification for Severity

Model	AUC	F1	Accuracy
Logistic Regression	0.7297	0.7492	0.8077
Random Forest	0.7509	0.7184	0.8053
Gradient-Boosted Trees	0.8089	0.7660	0.8192
Decision Tree	0.6452	0.7438	0.8112

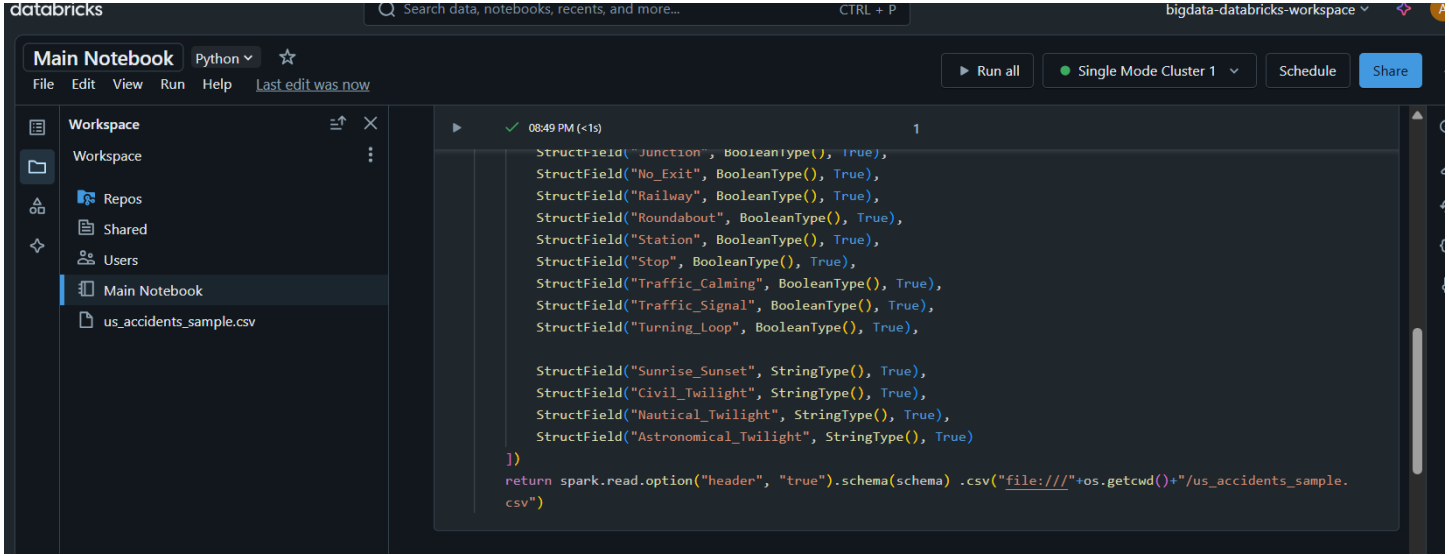
Risk Classification for States

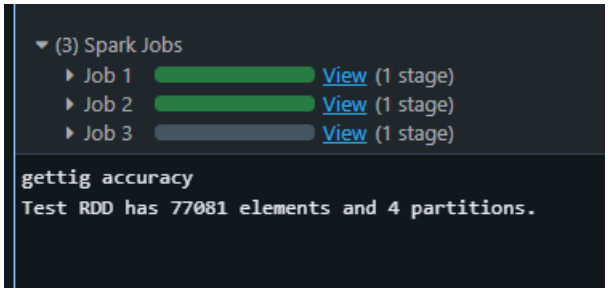


Risk Prediction using Regression



Azure Databricks





Fully Distributed Mode