Accident Risk Spatio-Temporal Analysis and Severity Estimation Using U.S. Crash Data

INFO 5810 – Summer 2025 | Group 4

Mary Dallas, Juayl Bukhari, Syam Sai Konakalla, Isagani Hernandez, Breanna Kotary

Abstract

Traffic crashes are a major public safety concern in the United States, resulting in tremendous human and economic losses. In this project, we investigate a large-scale dataset of over 7 million reported traffic accidents across the 49 U.S. states collected from February 2016 to March 2023. Using Python libraries (Pandas, NumPy, Seaborn, Matplotlib, Folium, and scikit-learn) as tools of data mining and predictive modeling, we target identifying the spatiotemporal patterns, discovering the environmental and weather conditions influencing accidents, and constructing a predictive model for the severity of accidents. Through clustering, correlation analysis, and machine learning-based classification models, the project attempts to find answers to the key questions of when and where accidents are most likely to happen and how severe they are likely to be. The output will be communicated through heatmaps and dashboards for policy-making, increasing road safety, and serving he transportation authorities' proactive planning.

Introduction & Background

- Traffic accidents continue to pose a serious threat to public safety across the U.S. With over 7 million reported incidents from 2016 to 2023, our project aims to understand when and where these accidents happen, what factors contribute to their severity, and how machine learning can help anticipate and prevent them.

Research Significance & Questions

- Supports proactive planning for public safety and infrastructure.

Key Questions:

- When and where are accidents most likely to occur?
- What environmental and situational factors increase severity?
- Can ML effectively forecast accident severity?

Project Objectives

- Highlight the high-risk areas and vehicular accident hotspots.
- Predict severity using machine learning models such as RandomForest and XGBoost.
- Enable insights through interactive visualizations and maps.
- Possibly Aid urban planners and transportation authorities in decision making.

Related Work Summary

- Referenced 10+ ML-based studies on accident prediction.
- Hybrid models (CNN-LSTM, SHAP + XGBoost) outperform classical models.
- Clustering + spatial data is crucial to identify hotspots.
- Inspired the use of interpretable ML + dashboards.

Dataset Summary

- Source: Kaggle’s U.S. Accident Dataset.
- Records: 7,728,394 | Features: 46 columns | Time Span: Feb 2016 – Mar 2023.
- Covers road type, weather, location, visibility, etc.

Data Cleaning & Feature Engineering

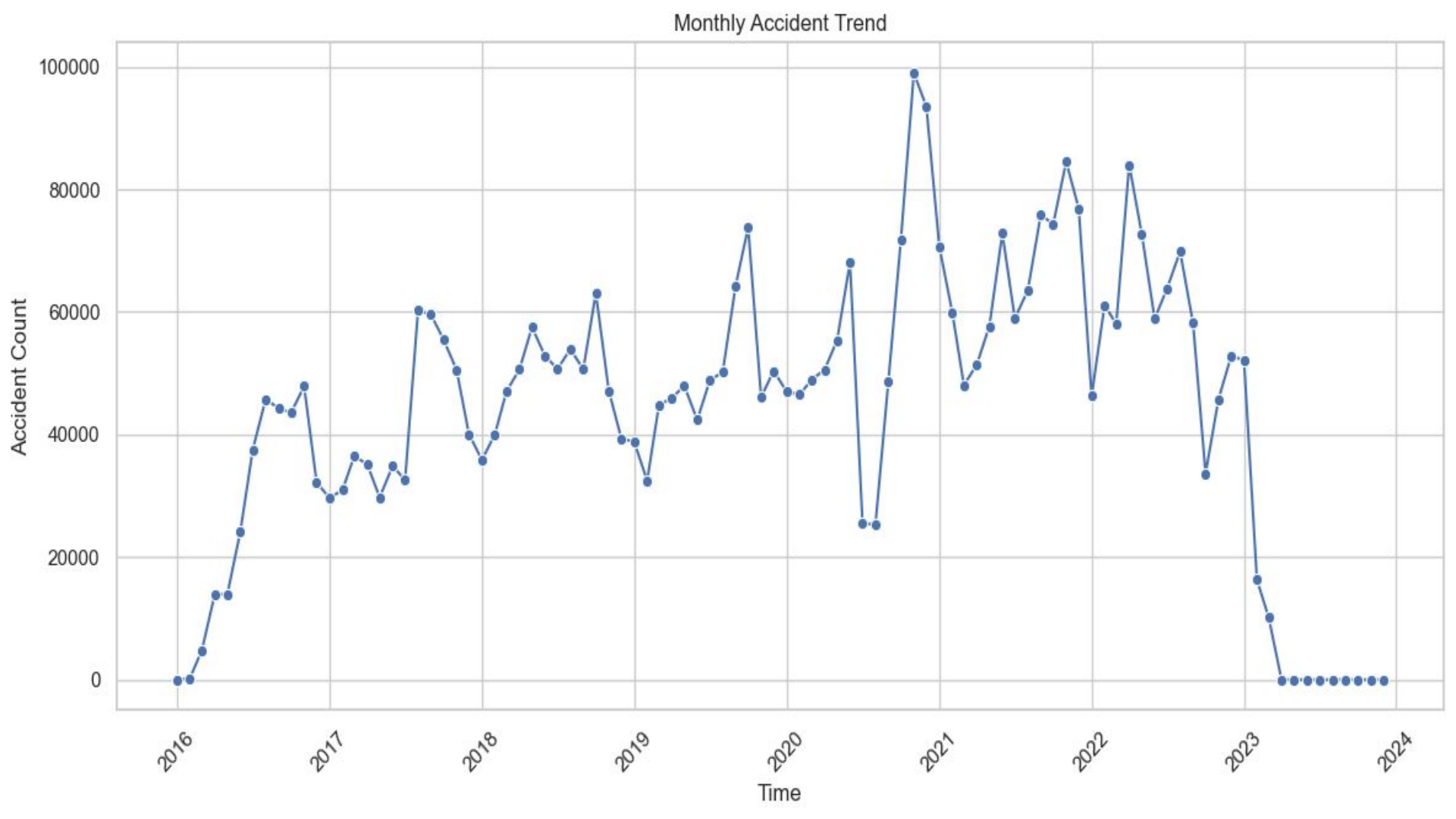
- Removed columns with >40% missing data.
- Imputed remaining missing values using mean/mode.
- Standardized categories (e.g., states, time zones).
- Feature engineering: Hour, DayOfWeek, Month, Weekend, Duration.

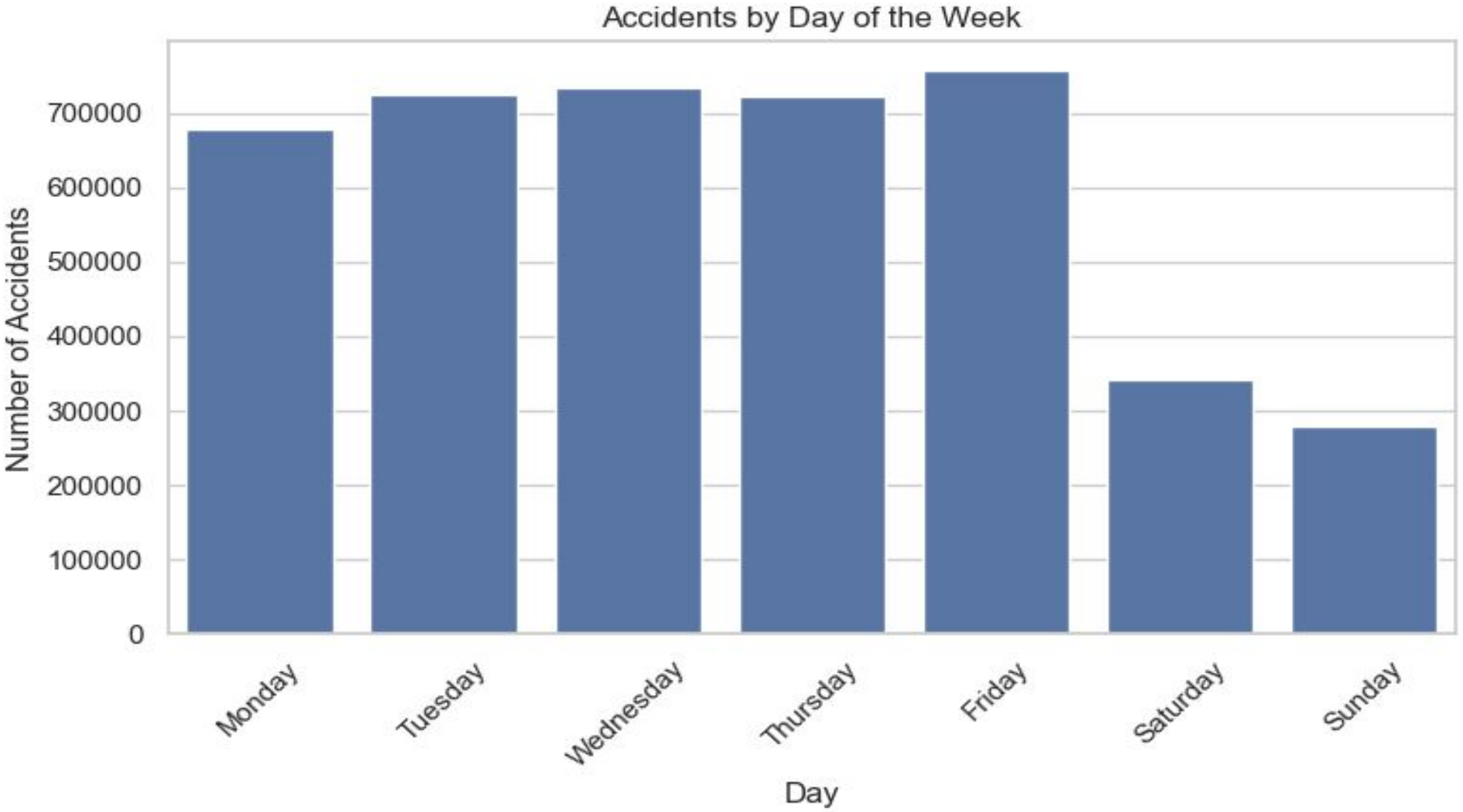
Methodology Overview

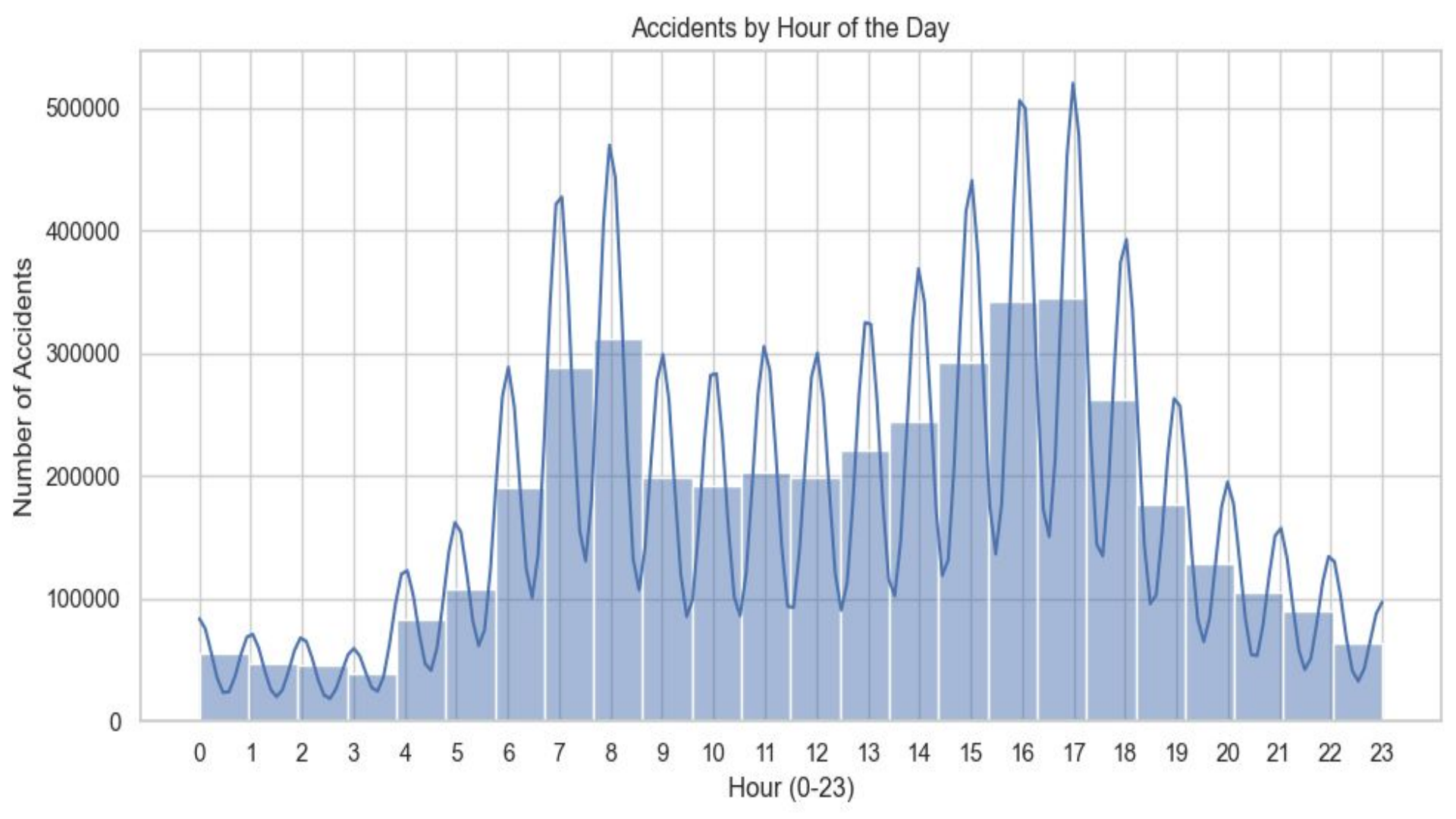
- Hotspot detection: DBSCAN and K-Means.
- Severity prediction models: Logistic Regression, Random Forest, XGBoost.
- Evaluation metrics: Accuracy, Precision, Recall, F1.

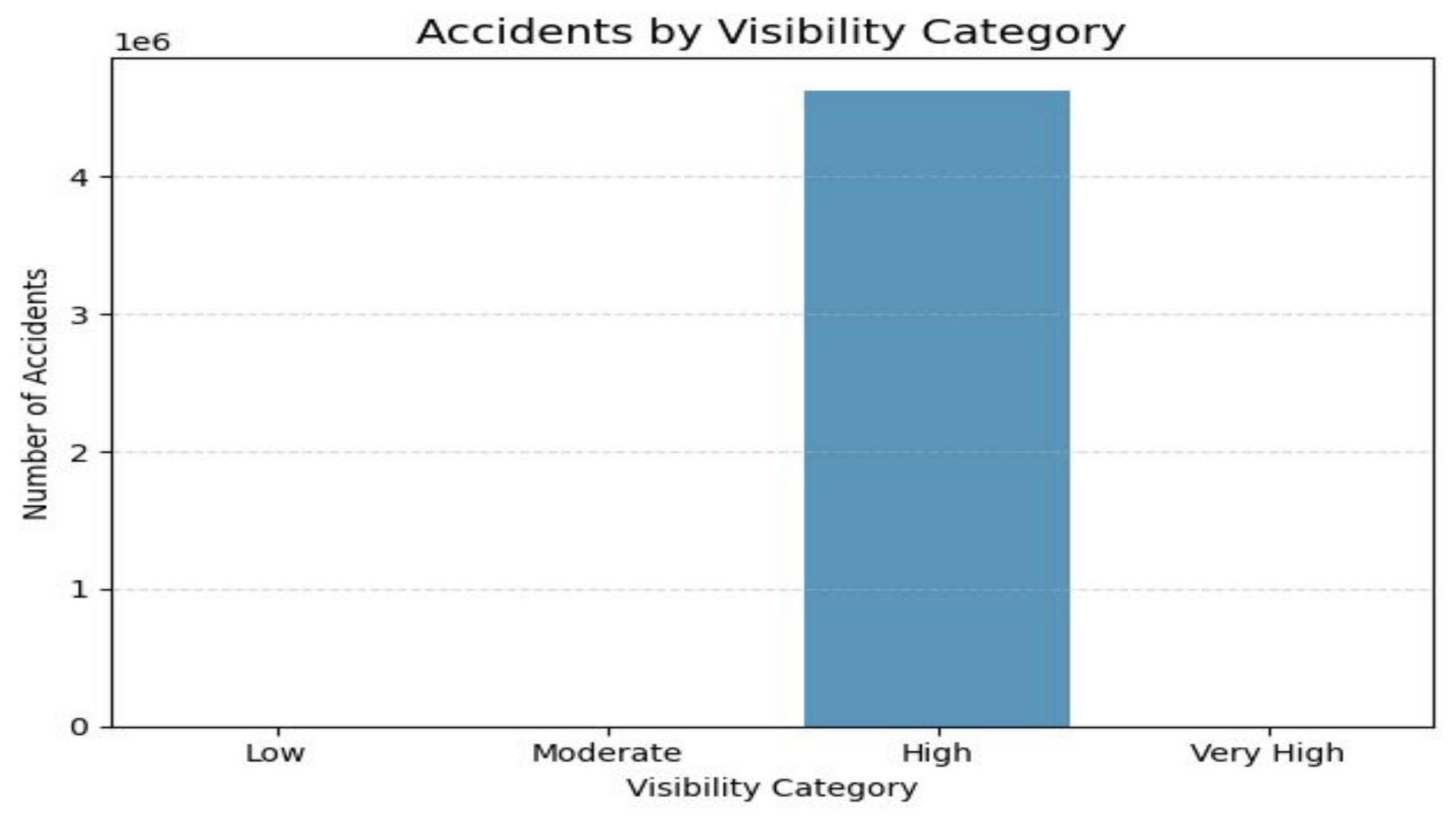
Experimental Design

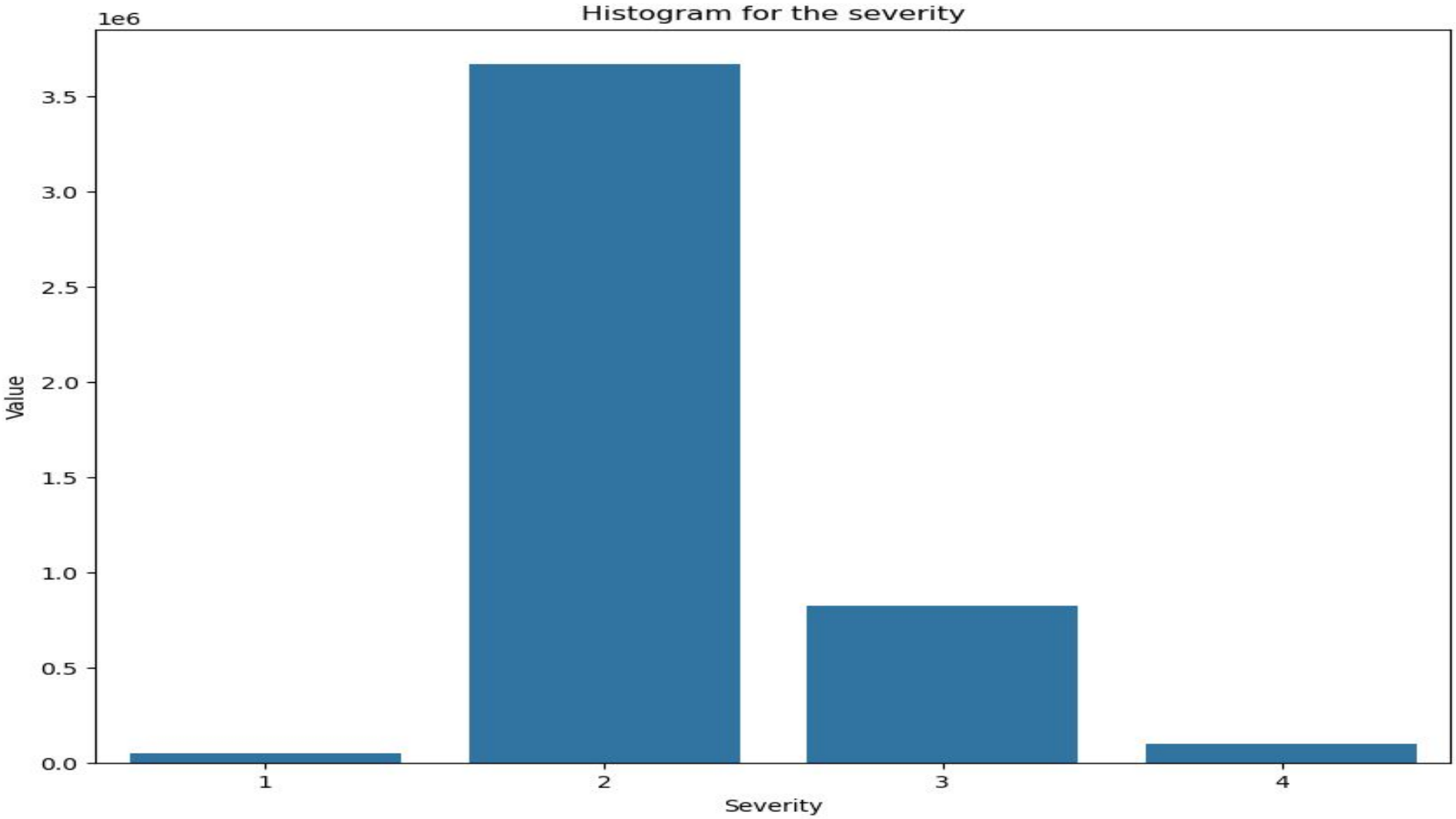
- Split data: 70% train, 15% val, 15% test.
- Undersampling to balance severity classes.
- Trained models on structured features including time and weather.
- Visualized using heatmaps, temporal charts, and dashboards.

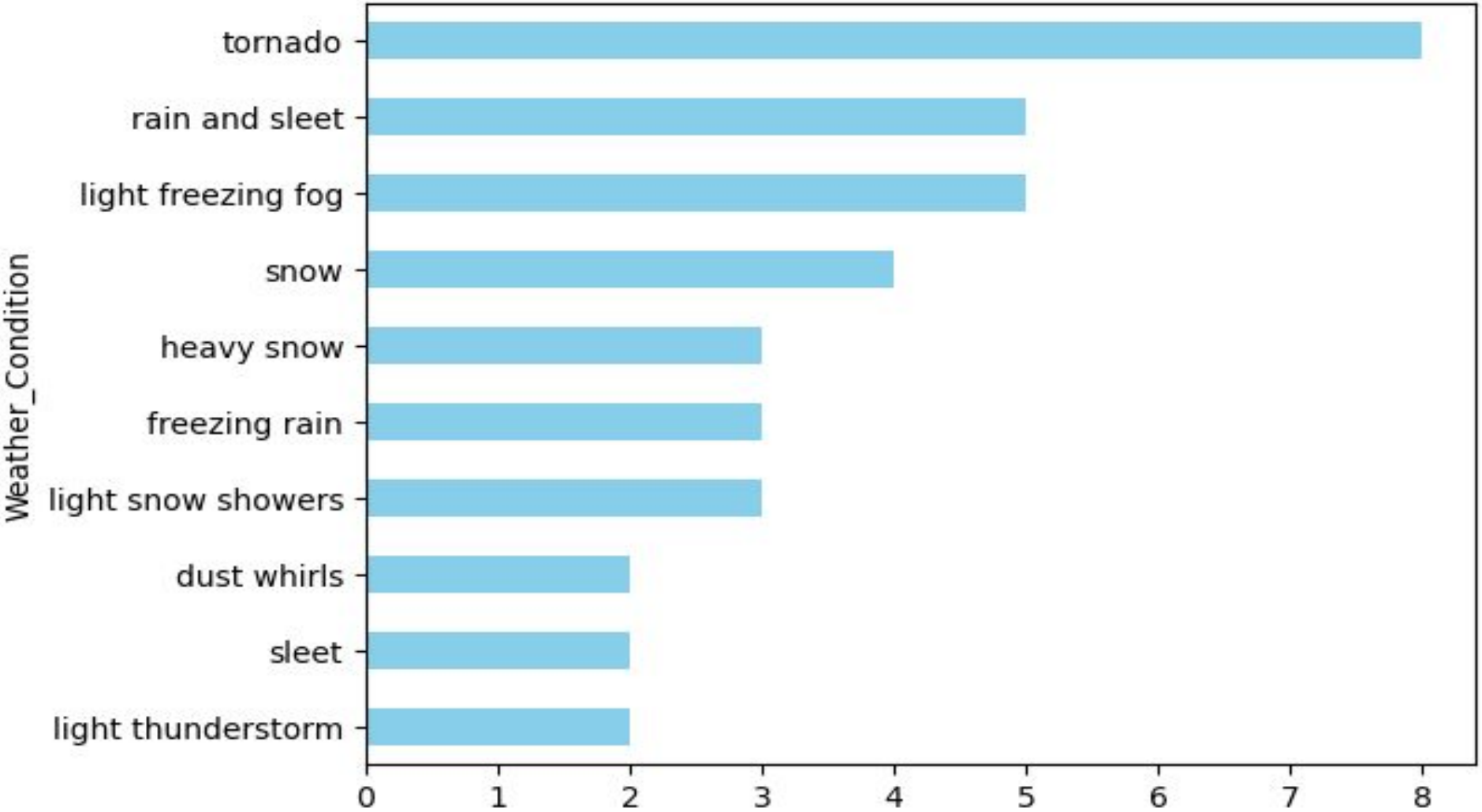

Monthly Accident Trend
Accident Count
100000
80000
60000
40000
20000
0
2016
2017
2018
2019
2020
2021
2022
2023
2024
Time

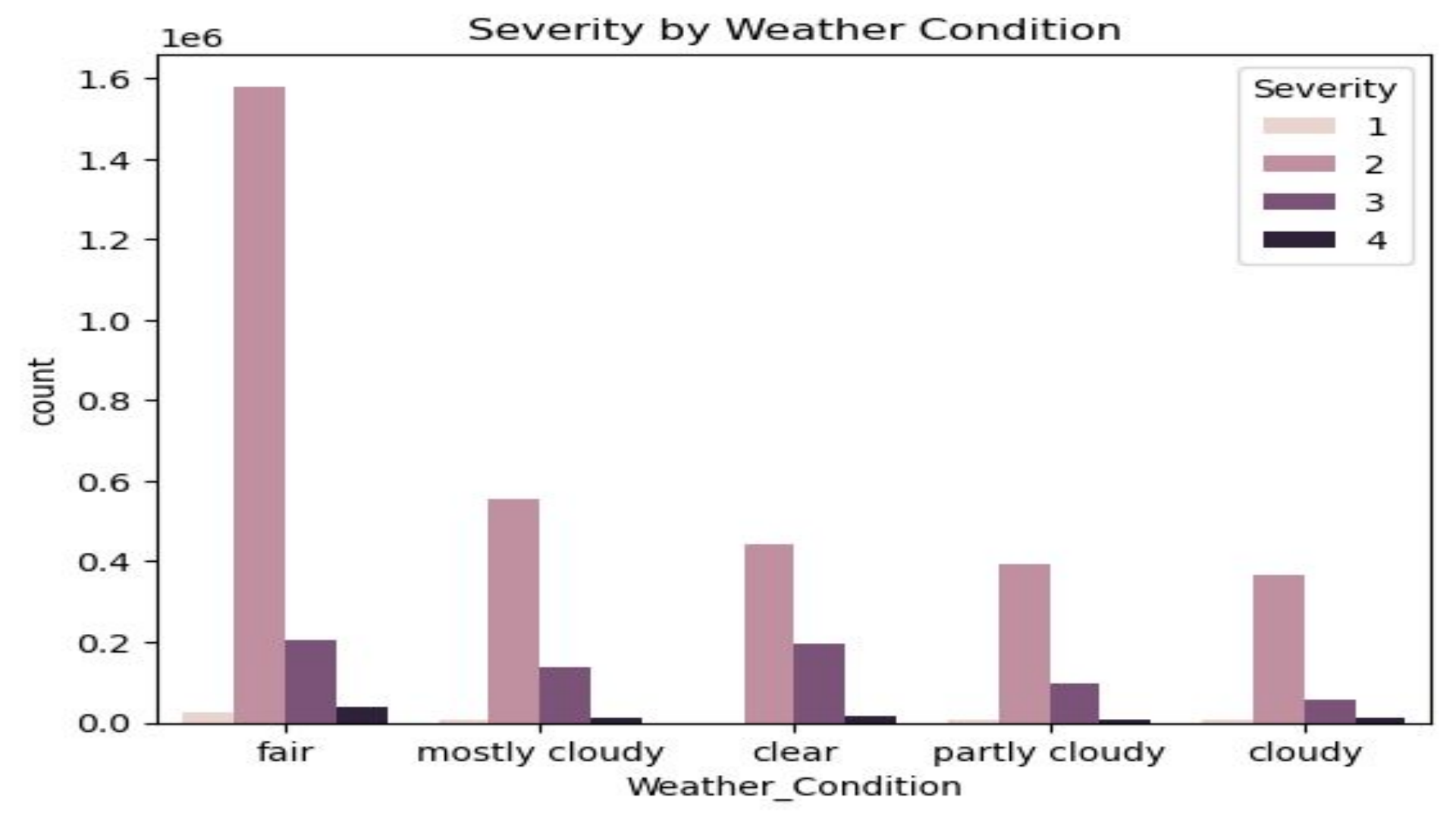

Accidents by Day of the Week
Number of Accidents
0
100000
200000
300000
400000
500000
600000
700000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day

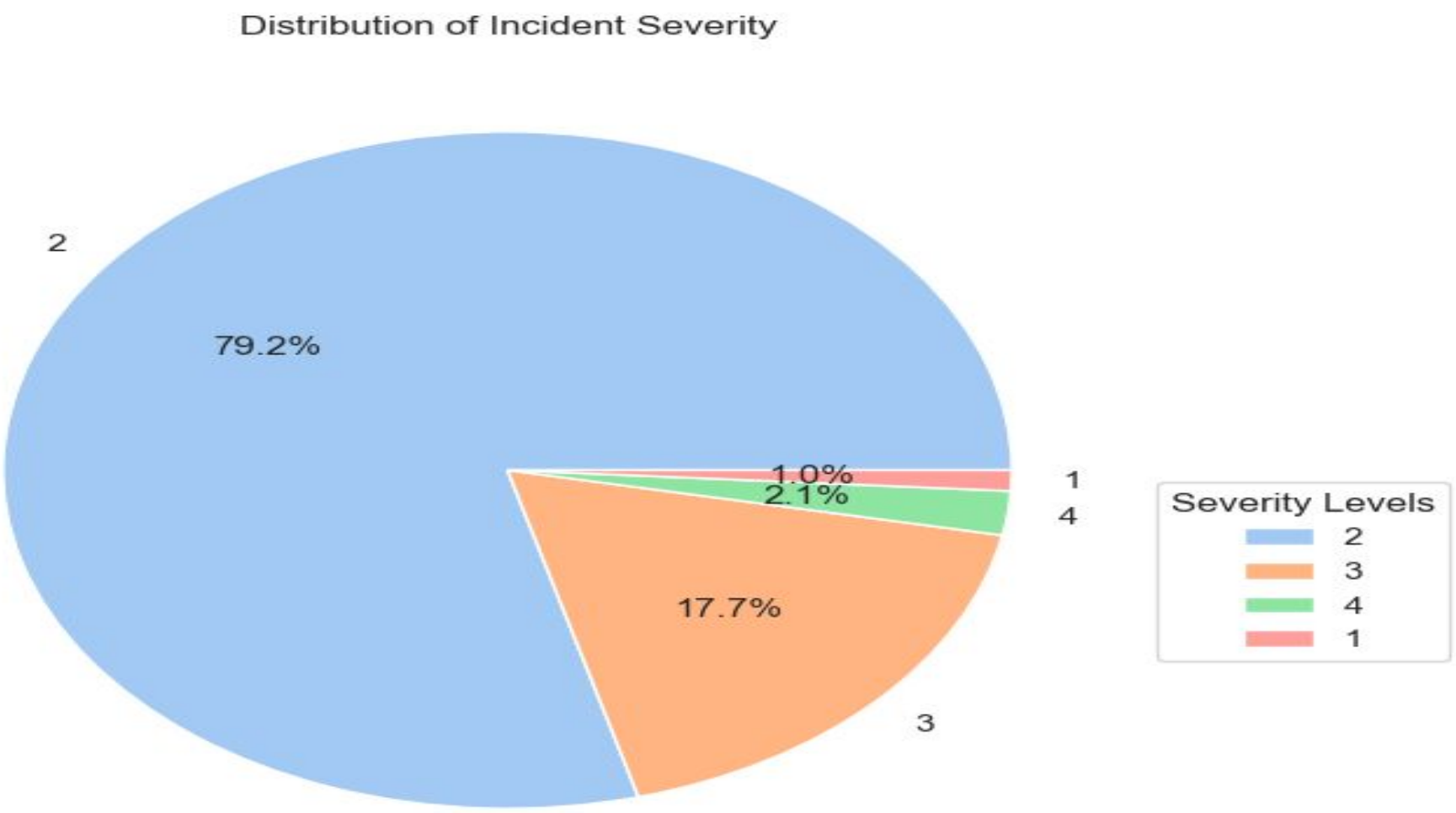

Accidents by Hour of the Day
Number of Accidents
0
100000
200000
300000
400000
500000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour (0-23)


Accidents by Visibility Category
1e6
4
3
2
1
0
Number of Accidents
Low
Moderate
High
Very High
Visibility Category


Histogram for the severity
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Value
1
2
3
4
Severity

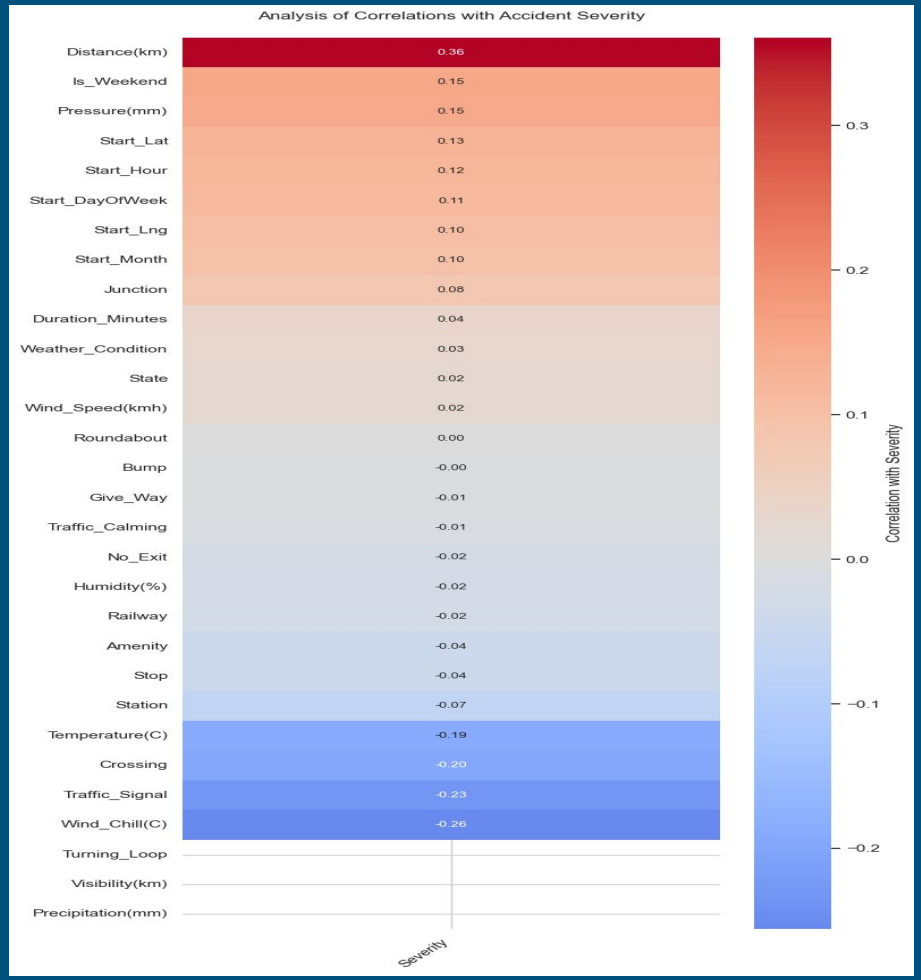

tornado
rain and sleet
light freezing fog
snow
heavy snow
freezing rain
light snow showers
dust whirls
sleet
light thunderstorm
Weather_Condition
0
1
2
3
4
5
6
7
8

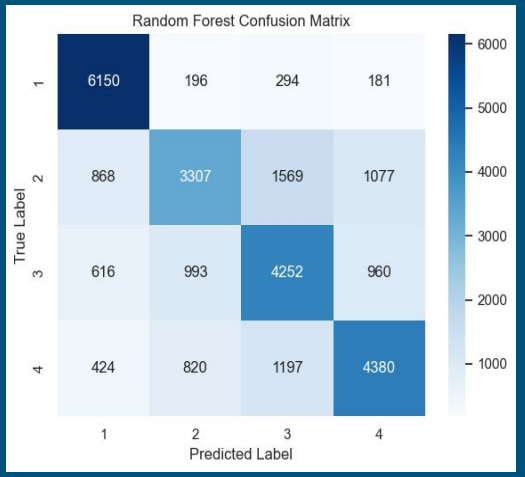

Severity by Weather Condition
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
count
Severity
1
2
3
4
fair
mostly cloudy
clear
partly cloudy
cloudy
Weather_Condition

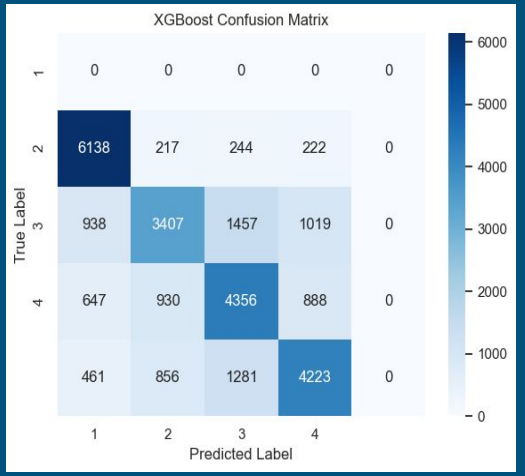

Distribution of Incident Severity
2
79.2%
1.0%
2.1%
1
4
17.7%
3
Severity Levels
2
3
4
1

Exploratory Analysis Highlights

- Rush hours (7–9AM, 4–6PM) the had most accidents.
- Visibility & weather significantly affected severity.
- Top cities and highways identified for repeated risk.


Analysis of Correlations with Accident Severity
Distance(km) 0.36
Is_Weekend 0.15
Pressure(mm) 0.15
Start_Lat 0.13
Start_Hour 0.12
Start_DayOfWeek 0.11
Start_Lng 0.10
Start_Month 0.10
Junction 0.08
Duration_Minutes 0.04
Weather_Condition 0.03
State 0.02
Wind_Speed(kmh) 0.02
Roundabout 0.00
Bump -0.00
Give_Way -0.01
Traffic_Calming -0.01
No_Exit -0.02
Humidity(%) -0.02
Railway -0.02
Amenity -0.04
Stop -0.04
Station -0.07
Temperature(C) -0.19
Crossing -0.20
Traffic_Signal -0.23
Wind_Chill(C) -0.26
Turning_Loop
Visibility(km)
Precipitation(mm)
Severity
Correlation with Severity
0.3
0.2
0.1
0.0
−0.1
−0.2

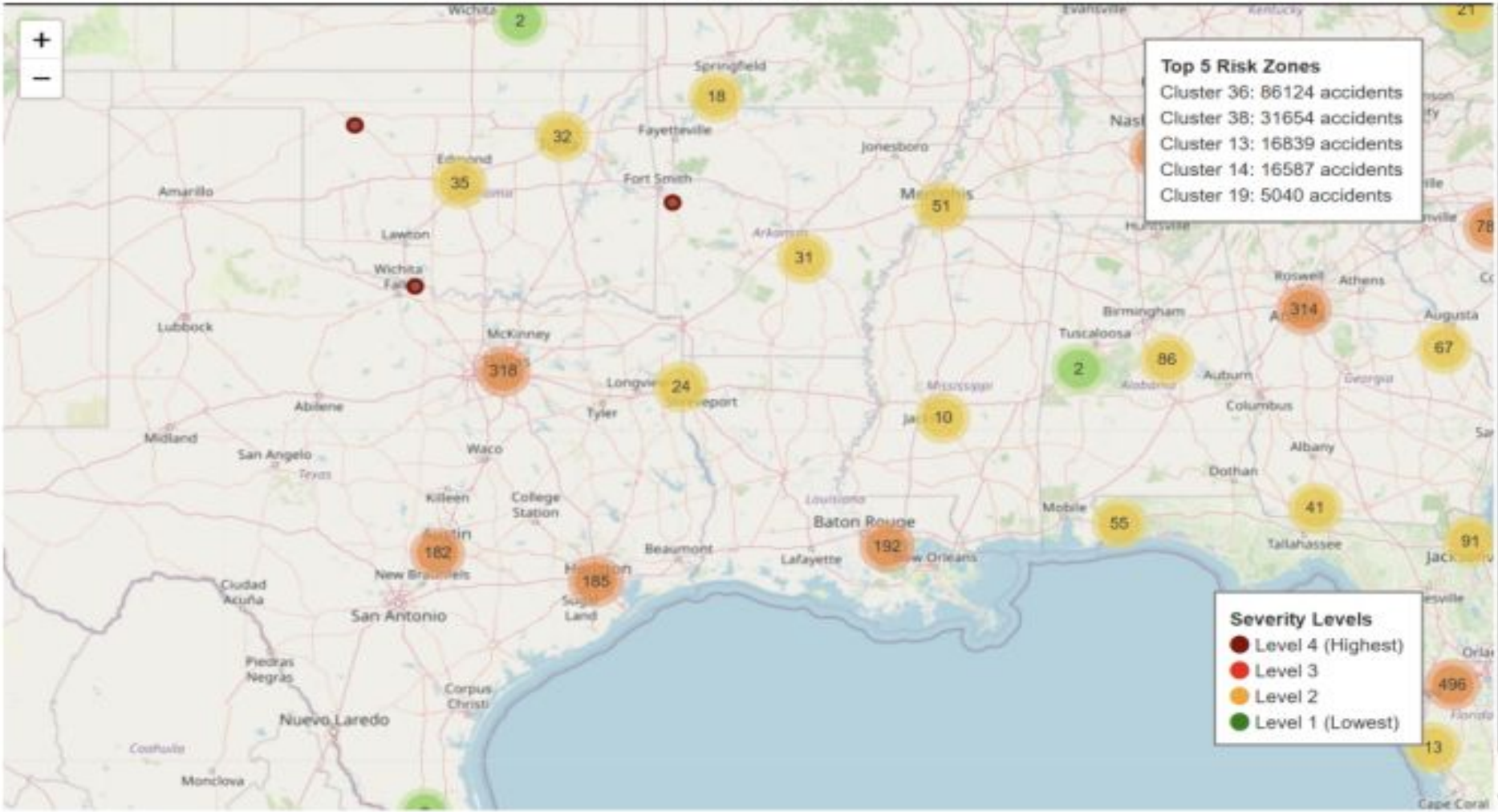

Random Forest Confusion Matrix
True Label
Predicted Label
1 6150 196 294 181
2 868 3307 1569 1077
3 616 993 4252 960
4 424 820 1197 4380
1 2 3 4
6000
5000
4000
3000
2000
1000


XGBoost Confusion Matrix
True Label
Predicted Label
1 0 0 0 0 0
2 6138 217 244 222 0
3 938 3407 1457 1019 0
4 647 930 4356 888 0
461 856 1281 4223 0
1 2 3 4
6000
5000
4000
3000
2000
1000
0

Clustering Insights

- 5 major clusters identified with different risk levels.
- Cluster 36: Most severe and accident-dense zone (avg severity 2.64).
- Low-signal or poorly lit areas had higher severity ratings.
- Urban areas had more volume, but rural zones had higher severity.

Traffic Accidents Map



Model Evaluation Summary

- Random Forest: Accuracy = 66.3%, strong recall on minor accidents.
- XGBoost: Better diagonal separation and F1 score across all classes.
- Random Forest more interpretable; XGBoost more precise.

Evaluation

```
Validation Set Performance:
              precision    recall  f1-score   support

           1       0.77      0.90      0.83      6821
           2       0.61      0.47      0.53      6820
           3       0.59      0.63      0.61      6821
           4       0.66      0.64      0.65      6821

    accuracy                           0.66     27283
   macro avg       0.66      0.66      0.65     27283
weighted avg       0.66      0.66      0.65     27283


Test Set Performance:
              precision    recall  f1-score   support

           1       0.76      0.90      0.83      6821
           2       0.62      0.48      0.54      6821
           3       0.58      0.62      0.60      6821
           4       0.66      0.64      0.65      6821

    accuracy                           0.66     27284
   macro avg       0.66      0.66      0.66     27284
weighted avg       0.66      0.66      0.66     27284
```

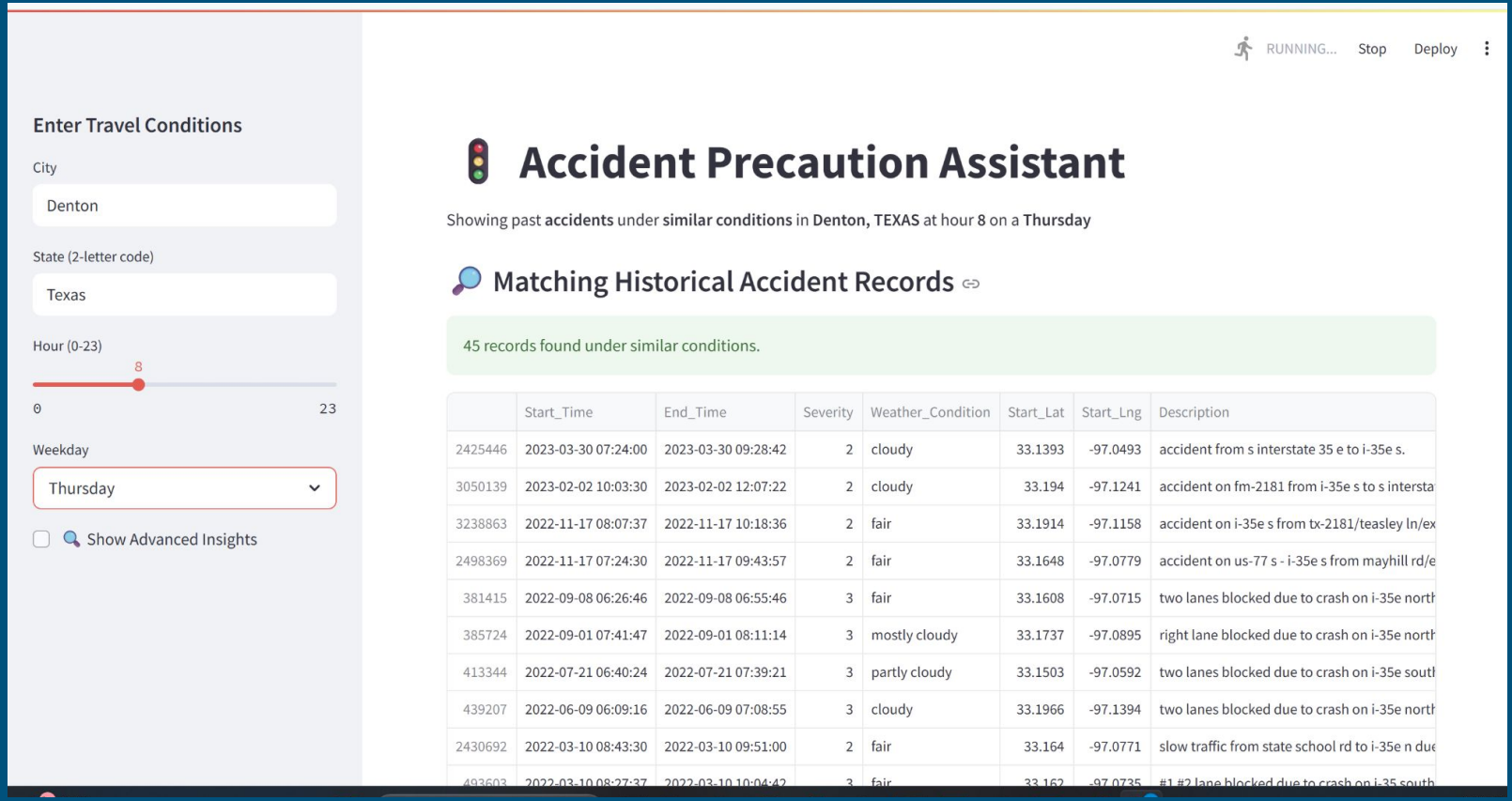
Model	Accuracy	Precision	Recall	F1
RandomForest	0.663	0.658	0.663	0.657
XGBoost	0.664	0.660	0.664	0.658

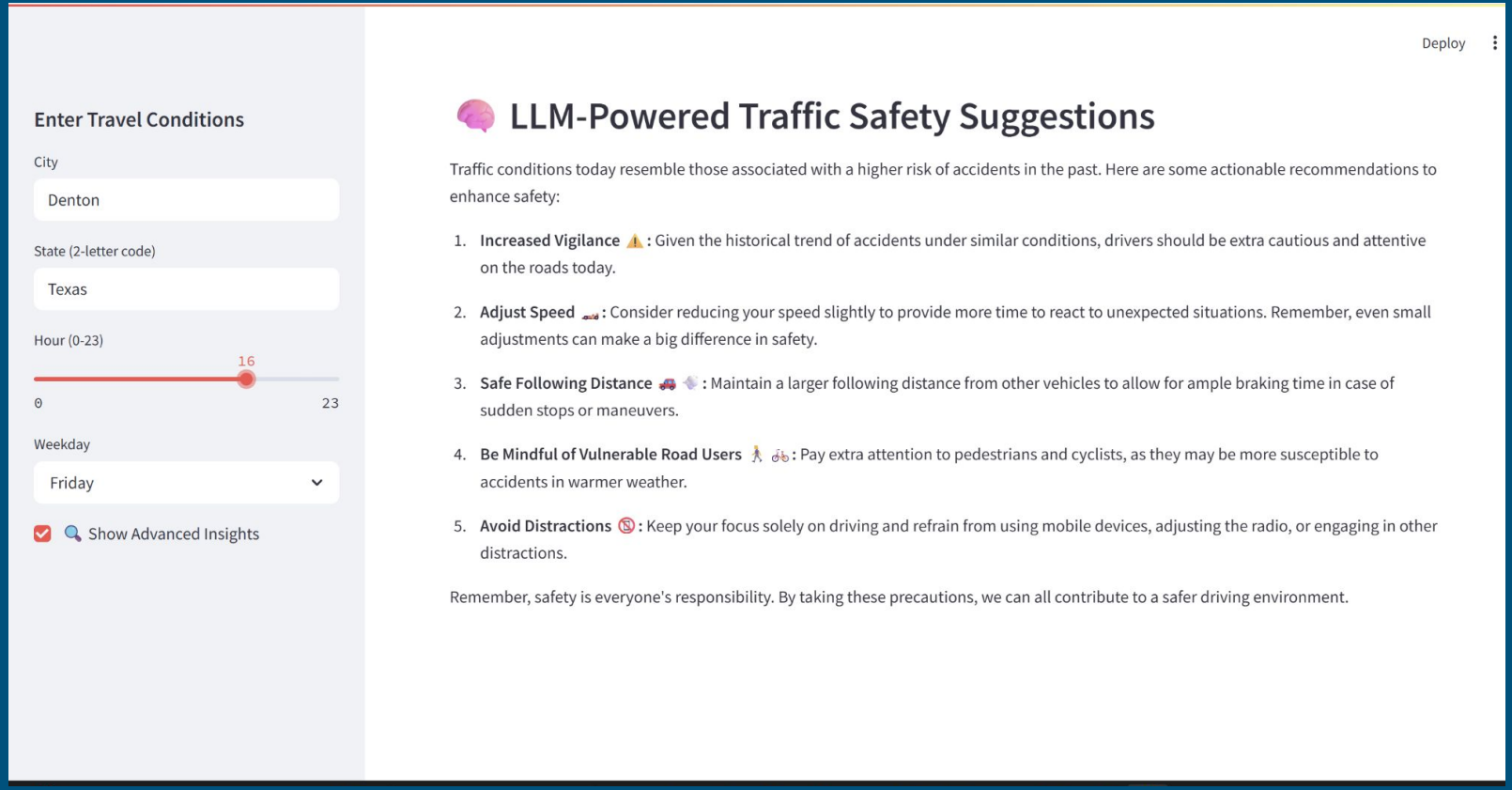
Random Forest Performance

```
Accuracy: 0.6642721008649758
              precision    recall  f1-score   support

           0       0.75      0.90      0.82      6821
           1       0.63      0.50      0.56      6821
           2       0.59      0.64      0.62      6821
           3       0.66      0.62      0.64      6821

    accuracy                           0.66     27284
   macro avg       0.66      0.66      0.66     27284
weighted avg       0.66      0.66      0.66     27284
```

Dashboard & Application

- Streamlit dashboard allows users to:
- Filter by location, time, weather.
- View geospatial accident patterns.
- Provide AI-driven safety suggestions.
- LLM module provides real-time travel tips.


RUNNING...
Stop
Deploy
Enter Travel Conditions
City
Denton
State (2-letter code)
Texas
Hour (0-23)
8
0
23
Weekday
Thursday
Show Advanced Insights
Accident Precaution Assistant
Showing past accidents under similar conditions in Denton, TEXAS at hour 8 on a Thursday
Matching Historical Accident Records
45 records found under similar conditions.
Start_Time
End_Time
Severity
Weather_Condition
Start_Lat
Start_Lng
Description
2425446 2023-03-30 07:24:00 2023-03-30 09:28:42 2 cloudy 33.1393 -97.0493 accident from s interstate 35 e to i-35e s.
3050139 2023-02-02 10:03:30 2023-02-02 12:07:22 2 cloudy 33.194 -97.1241 accident on fm-2181 from i-35e s to s intersta
3238863 2022-11-17 08:07:37 2022-11-17 10:18:36 2 fair 33.1914 -97.1158 accident on i-35e s from tx-2181/teasley ln/ex
2498369 2022-11-17 07:24:30 2022-11-17 09:43:57 2 fair 33.1648 -97.0779 accident on us-77 s - i-35e s from mayhill rd/e
381415 2022-09-08 06:26:46 2022-09-08 06:55:46 3 fair 33.1608 -97.0715 two lanes blocked due to crash on i-35e north
385724 2022-09-01 07:41:47 2022-09-01 08:11:14 3 mostly cloudy 33.1737 -97.0895 right lane blocked due to crash on i-35e north
413344 2022-07-21 06:40:24 2022-07-21 07:39:21 3 partly cloudy 33.1503 -97.0592 two lanes blocked due to crash on i-35e south
439207 2022-06-09 06:09:16 2022-06-09 07:08:55 3 cloudy 33.1966 -97.1394 two lanes blocked due to crash on i-35e north
2430692 2022-03-10 08:43:30 2022-03-10 09:51:00 2 fair 33.164 -97.0771 slow traffic from state school rd to i-35e n due


Deploy
Enter Travel Conditions
City
Denton
State (2-letter code)
Texas
Hour (0-23)
16
0
23
Weekday
Friday
Show Advanced Insights
LLM-Powered Traffic Safety Suggestions
Traffic conditions today resemble those associated with a higher risk of accidents in the past. Here are some actionable recommendations to enhance safety:
1. Increased Vigilance ⚠️: Given the historical trend of accidents under similar conditions, drivers should be extra cautious and attentive on the roads today.
2. Adjust Speed 🏎️: Consider reducing your speed slightly to provide more time to react to unexpected situations. Remember, even small adjustments can make a big difference in safety.
3. Safe Following Distance 🚗💨: Maintain a larger following distance from other vehicles to allow for ample braking time in case of sudden stops or maneuvers.
4. Be Mindful of Vulnerable Road Users 🚶🚲: Pay extra attention to pedestrians and cyclists, as they may be more susceptible to accidents in warmer weather.
5. Avoid Distractions 🚫: Keep your focus solely on driving and refrain from using mobile devices, adjusting the radio, or engaging in other distractions.
Remember, safety is everyone's responsibility. By taking these precautions, we can all contribute to a safer driving environment.

Conclusion

- Our project created a smart system that uses U.S. crash data to analyze accident risks and suggest safety precautions. We combined crash data with real-time information, machine learning, and AI to turn raw numbers into useful, actionable advice. Essentially, we built a bridge between complex data and practical road safety.

Potential Impact

A huge potential to make our roads safer:

City Planners: It can assist the cities in recognizing hazardous roads and deciding where to install new signs or traffic lights, or execute other safety interventions.

Drivers: It gives drivers the information needed to make intelligent travel decisions and helps them avoid risky areas and conditions.

For Public Safety: Our research supports the formulation of more effective public safety campaigns based on data.

Future Integration: This might be embedded in popular GPS apps for instant warnings or even in a city dashboard for real-time monitoring.

Key Findings

Risky Hours of Commuting: Accidents tend to occur primarily during weekday rush hour traffic, with the highest recorded incidents around 8 AM and 5 PM.

Severity Prediction: The severity of accidents was often correctly predicted by our models, with an accuracy of almost 66 percent. It was particularly good at predicting minor and fatal accidents.

Danger Zones: Clustering of accident data enabled us to detect particular "hotspots" where a high number of crashes is registered, and the average severity is high. These spots most probably require infrastructure interventions.

User-friendly App: A fully working prototype was developed, so now one can view historic accidents around and receive personalized driving tips powered by an Artificial Intelligence that takes into account local conditions and current weather.

Limitations

Some of our limitations are:

Data Gaps: It is a great challenge for the model to predict some middle levels of accident severity because we have few samples of those crashes in our dataset.

Limited Weather Data: The weather data used was not very detailed and may affect some predictions.

Static Hotspots: According to the system, "Danger zones" don't alter over time.

No Live Feeds: The AI must use historical data rather than real-time data from traffic cameras or other live sensors, so it could never take into account new construction or a temporary road closure.

Future Work

There are so many ideas lined up to make this system even better:

Live Updates: The objective now is to feed real-time traffic and camera feeds into our model so that dynamic predictions can be made.

Smarter AI: We want to use better Machine Learning to better comprehend complex accident patterns.

Take It Mobile: We would love to develop a mobile application that provides voice-guided assistance and safety alerts.

Test on Actual Drivers: The next stage is to investigate how these AI-driven suggestions genuinely alter driver behavior and increase safety.

Team Contributions

- Mary Dallas – Intro, Purpose, Significance.
- Juayl Bukhari – Related Work.
- Syam Sai Konakalla – Methodology, Modeling.
- Isagani Hernandez – Data Collection & Cleaning.
- Breanna Kotary – Experimentation, Visualizations.

Thank You