Analyzing Road Accidents to Recommend Safer Routes

Group 6 - Pooja Reddy K, Tanvica Samudrala, Yamini Srija K, Jashwant R

1. Introduction and Project Objective

This project aims to analyze patterns in road accidents and recommend safer routes for commuters. With the alarming frequency of traffic incidents and the wide range of contributing factors—such as adverse weather conditions, limited visibility, high-traffic periods, and specific geographic locations—there is a growing need for intelligent, data-driven solutions that can proactively improve road safety. By examining large-scale historical accident data, we seek to uncover underlying trends, identify high-risk zones, and develop a system that can guide users toward less hazardous travel routes. Our initial approach focused on the US Accidents Dataset, a comprehensive compilation of over 7 million recorded accidents across the United States from 2016 to 2023. This dataset provides rich information, including the time, location, severity, environmental context, and infrastructural characteristics of each incident, offering a valuable foundation for predictive modeling and spatial analysis. However, as we delved deeper into the data, we recognized that conducting analyses at a nationwide level introduced considerable computational overhead and visualization challenges due to the dataset's scale and regional variability. To manage these complexities and ensure the feasibility and effectiveness of our modeling efforts, we decided to narrow the project's scope to the state of California. This decision was driven by both practical and analytical considerations. California consistently ranks among the top states in terms of accident frequency, offering a dense and diverse subset of data that is well-suited for focused exploration. Limiting our study to this region enables us to perform more detailed spatial clustering, improve model precision, and create actionable insights that can be directly applied to urban and suburban mobility planning. This refined focus enhances our ability to deliver meaningful safety recommendations tailored to real-world conditions.

2. Datasets and Sources

The primary dataset used in this project is the US Accidents (March 2023) dataset sourced from Kaggle. It includes detailed records of 7.7 million traffic accidents from 2016–2023, with 46 features ranging from geographic coordinates to environmental conditions and temporal factors. The dataset includes:

- Accident ID, Source, and Severity
- Time and location (Start Time, End Time, Start Lat, Start Lng)
- Environmental features (Temperature(F), Wind Speed(mph), Visibility(mi), etc.)
- Road and infrastructure features (Amenity, Junction, Traffic Signal, etc.)

Due to scale and relevance, we filtered the dataset to focus specifically on accidents in California. This decision was informed by exploratory analysis, which showed California has the highest accident density, especially in areas like Los Angeles.

ID	object	Visibility(mi)	float64
Source	object	Wind_Direction	object
Severity	int64	Wind_Speed(mph)	float64
Start_Time	object	Precipitation(in)	float64
End_Time	object	Weather_Condition	object
Start_Lat	float64	Amenity	bool
Start_Lng	float64	Bump	bool
End_Lat	float64	Crossing	bool
End_Lng	float64	Give_Way	bool
Distance(mi)	float64	Junction	bool
Description	object	No_Exit	bool
Street	object	Railway	bool
City	object	Roundabout	bool
County	object	Station	bool
State	object	Stop	bool
Zipcode	object	Traffic_Calming	bool
Country	object	Traffic_Signal	bool
Timezone	object	Turning_Loop	bool
Airport_Code	object	Sunrise_Sunset	object
Weather_Timestamp	object	Civil_Twilight	object
Temperature(F)	float64	Nautical_Twilight	object
Wind_Chill(F)	float64	Astronomical_Twilight	object
Humidity(%)	float64	dtype: object	
Pressure(in)	float64		

	ID	Source	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	...	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
0	A-1	Source2	3	2016-02-08 05:46:00	2016-02-08 11:00:00	39.865147	-84.058723	NaN	NaN	0.01	...	False	False	False	False	False	False
1	A-2	Source2	2	2016-02-08 06:07:59	2016-02-08 06:37:59	39.928059	-82.831184	NaN	NaN	0.01	...	False	False	False	False	False	False
2	A-3	Source2	2	2016-02-08 06:49:27	2016-02-08 07:19:27	39.063148	-84.032608	NaN	NaN	0.01	...	False	False	False	False	True	False
3	A-4	Source2	3	2016-02-08 07:23:34	2016-02-08 07:53:34	39.747753	-84.205582	NaN	NaN	0.01	...	False	False	False	False	False	False
4	A-5	Source2	2	2016-02-08 07:39:07	2016-02-08 08:09:07	39.627781	-84.188354	NaN	NaN	0.01	...	False	False	False	False	True	False

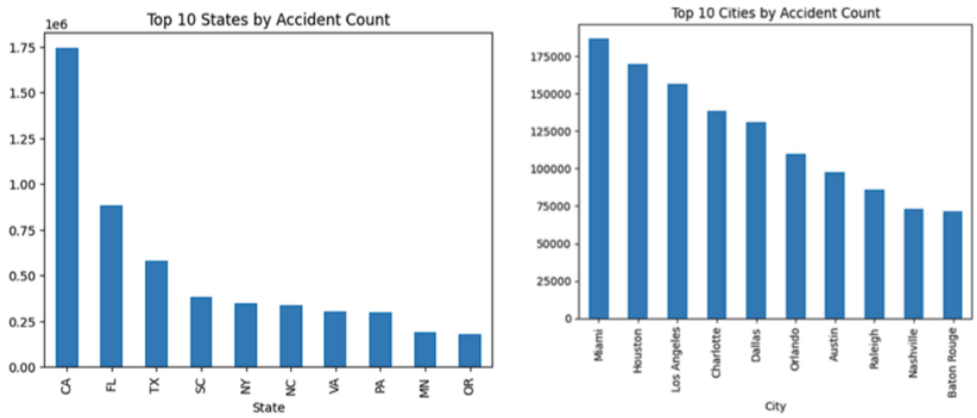
5 rows × 46 columns

3. Exploratory Analysis of the Datasets

Dataset Structure and Feature Selection:

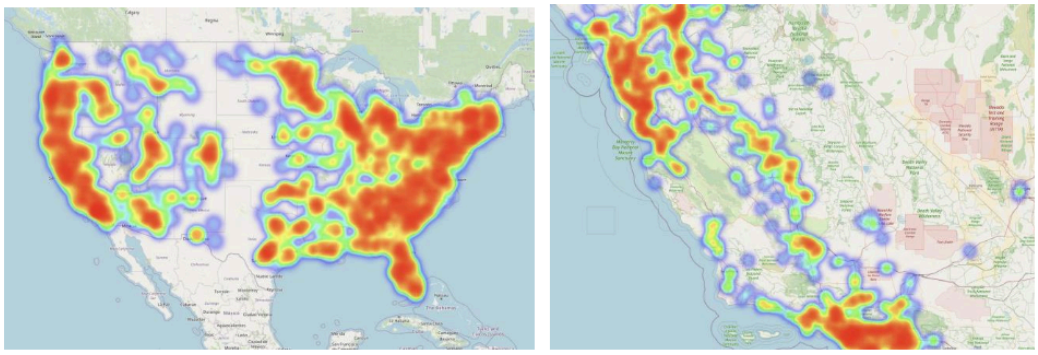
The dataset comprises over 7.7 million records and 46 columns. Features include temporal (Start Time, End Time), environmental (Temperature, Visibility), and spatial attributes (Latitude, Longitude). To improve modeling quality and reduce noise, we dropped columns with more than 40% missing values or low analytic value, such as End Lat, Wind Chill(F), and Airport Code. Bar charts show California and Los Angeles lead the nation in total accidents. This confirmed our decision to focus modeling and recommendations within California due to its high incident volume and data richness.

Top 10 States and Cities by Accident Count

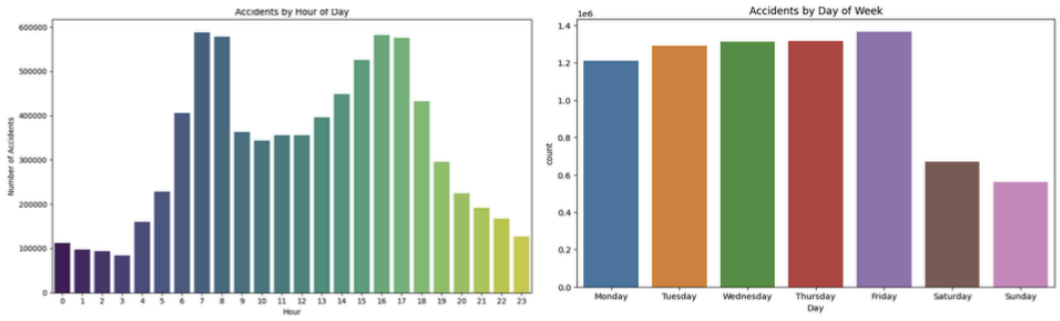


Bar charts show California and Los Angeles lead the nation in total accidents. This confirmed our decision to focus modeling and recommendations within California due to its high incident volume and data richness.

Interactive U.S. Accident Heatmap



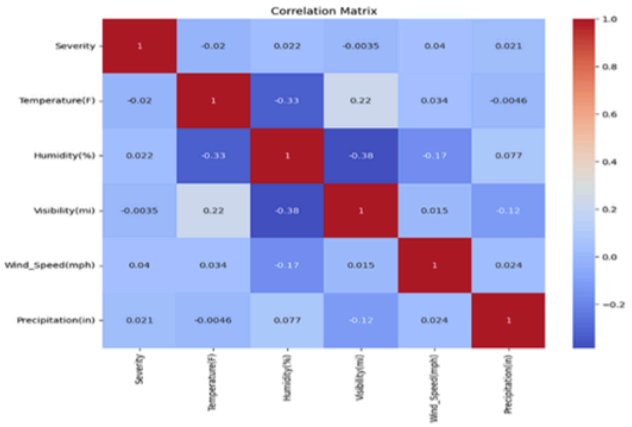
Accidents by Hour of Day and by Day of Week



The first histogram displays two peaks during morning and evening commute hours. The pattern highlights the importance of hour-of-day as a factor influencing accident likelihood. The second graph shows that accidents steadily increase from Monday to Friday and decline over the

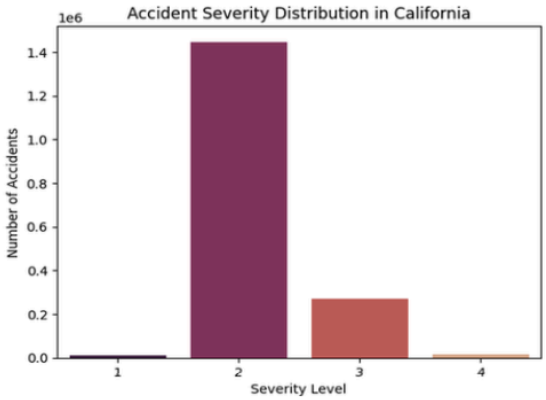
weekend. This trend indicates weekday commuting is a key risk period and should inform temporal scoring.

Environmental Feature Summary and Correlation Matrix



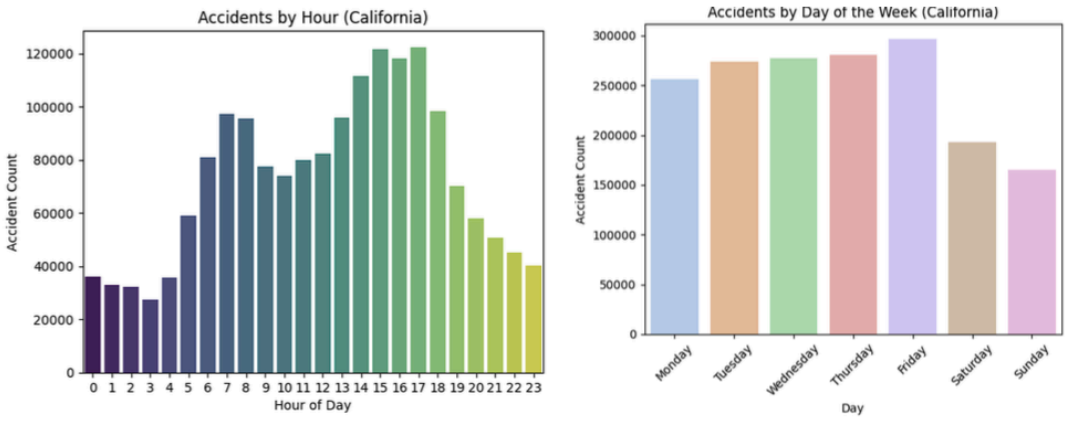
Summary statistics and correlations show weak relationships between environmental variables and severity. However, features like low visibility and humidity remain relevant for inclusion.

Accident Severity Distribution in California



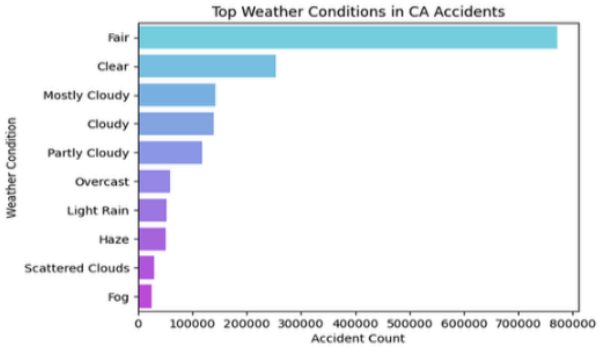
Most California accidents are Severity 2, with a few critical cases. This imbalance will require an adjustment in model training to improve prediction for rare but severe incidents.

Accidents by Hour of Day and Day of the Week in California



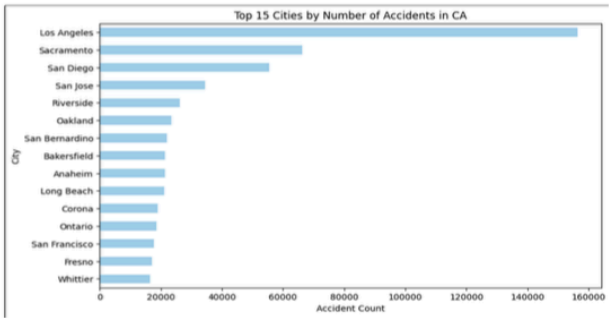
Accidents by hour in California mirror national patterns, with accident peaks during typical rush hours. These findings confirm the temporal consistency of traffic risk. Accidents by day of the week show Fridays have the highest accident count, while weekends see fewer incidents. Day-of-week remains a reliable temporal feature for safety scoring.

Top Weather Conditions in California Accidents



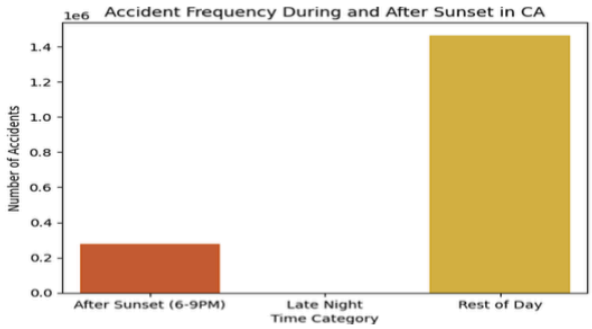
Accidents mostly occur under clear conditions, though fog and rain are associated with greater severity. Weather remains an important supporting factor in assessing risk.

Top 15 Cities by Number of Accidents in California



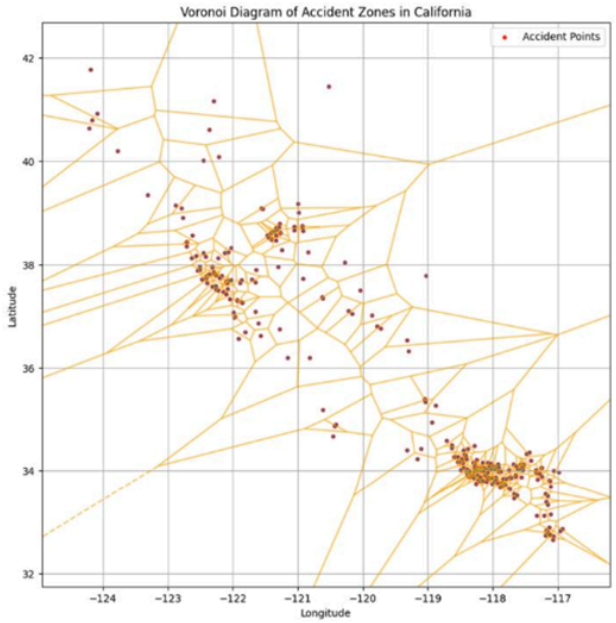
Los Angeles, Sacramento, and San Diego rank highest in accident volume. These cities are focal points for future modeling and recommendation outputs.

Accident Frequency During and After Sunset in California



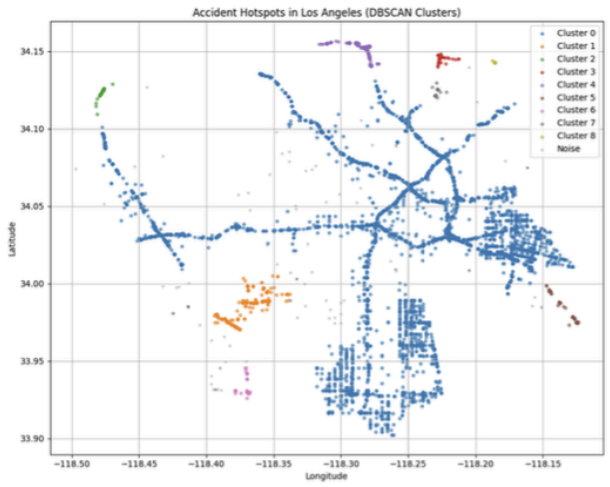
A comparative bar chart shows elevated accident counts during and shortly after sunset, suggesting visibility may play a role in increased risk.

Voronoi Diagram of Accident Zones in California



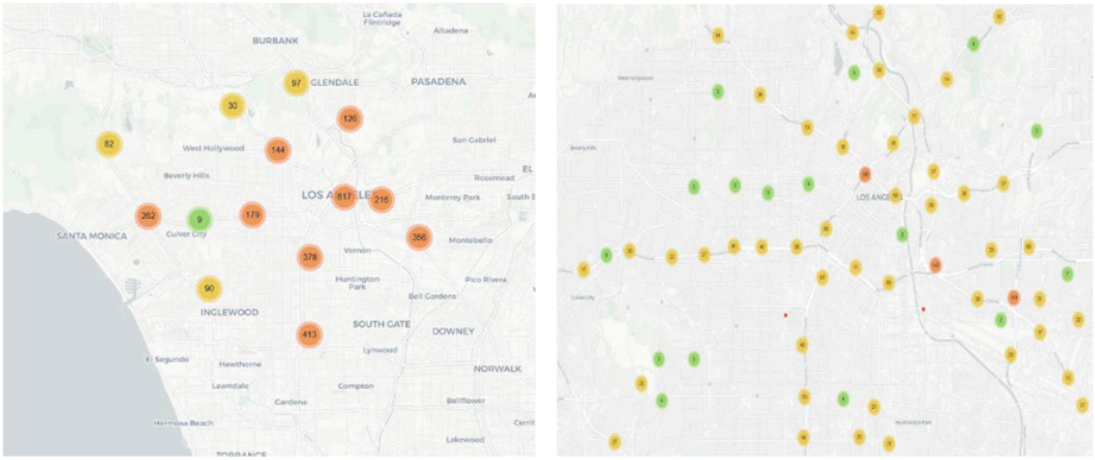
A Voronoi map of 300 sampled coordinates visually partitions California into proximity zones. Clusters form tightly around urban cores, useful for spatial segmentation.

Accident Hotspots in Los Angeles (DBSCAN Clusters)



Using DBSCAN, nine spatial clusters were detected in Los Angeles. These hotspots will guide route avoidance logic and regional risk prioritization.

Interactive Clustering of Accident Hotspots in Los Angeles



An interactive Folium map visualizes DBSCAN clusters in LA with color-coded markers. Zoom levels reveal fine-grained risk concentrations useful for route-based feedback.

4. Missing Value Analysis:

We analyzed missing data and found that columns like End Lat, Precipitation(in), and Wind Chill(F) had the highest null percentages, up to 44%. These were either imputed or excluded based on downstream use cases.

Severity Distribution and Temporal Patterns

The majority of accidents are categorized as Severity 2. Histogram plots revealed peaks in accident occurrence during morning and evening rush hours. Weekly trends showed an increase from Monday through Friday, with a decline over the weekend, highlighting the importance of time in assessing accident risk.

Environmental and Geographic Trends

Environmental conditions like fog and rain were less frequent but correlated with higher severity. A Folium-based heatmap and Voronoi diagrams helped visualize high-density clusters and segment California into accident zones. DBSCAN clustering further identified 9 hotspots in the Los Angeles region.

Feature Engineering for Safety Scoring

We constructed a custom Risk Score using:

- **Base Score:** Proportional to severity
- **Environmental Score:** Penalizing unsafe weather/visibility conditions
- **Infrastructure Score:** Based on the presence of safety features (e.g., traffic signals)

- **Time Score:** Penalizing rush hour periods (before 6 AM and after 8 PM)

The final score was normalized and visualized using interactive maps to guide route-based decision-making.

5. Routing Analysis and Implementation

To complement the exploratory data analysis of accident trends, we designed a routing system aimed at identifying safer alternative routes in accident-prone areas. Using the OpenStreetMap road network and California's traffic accident data, we explored how data-driven routing can provide actionable insights for safer travel.

We selected five key cities in California—Los Angeles, San Francisco, San Diego, Sacramento, and Fresno—as case studies. Each city's road network was analyzed within a 500-meter radius around its center to balance performance and relevance.

5.1 Initial Routing: Simple Risk-Based Logic

Methodology

In the initial implementation, we defined risk purely based on accident density. Each road segment in the downloaded street network was checked for the number of nearby accidents within approximately 30 meters.

The formula used for each road segment was:

$$\text{risk\_weight} = \text{length} \times (1 + 5 \text{accident\_count})$$

This ensures that road segments with no nearby accidents keep their weight as their true length (no penalty), while segments with more accidents get penalized proportionally, discouraging the routing algorithm from selecting them.

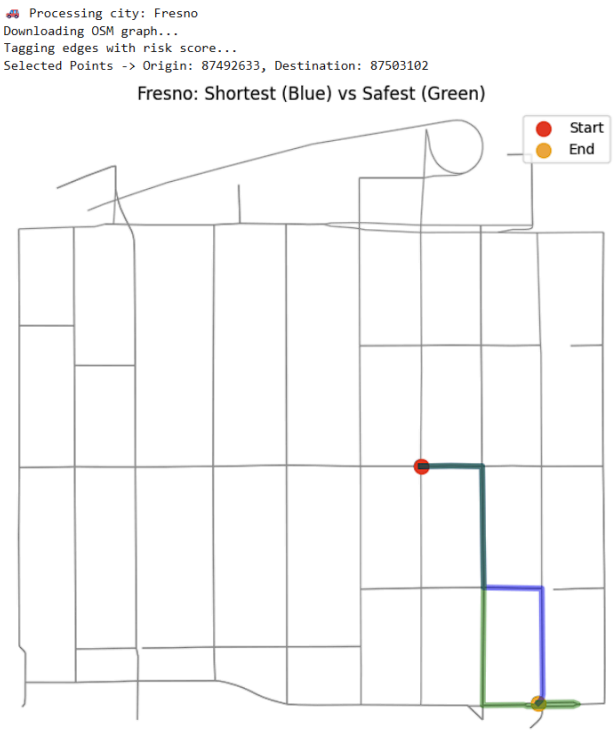
Visualization and Results

For each city, we selected two random points within the graph and computed:

- The shortest route (blue): the minimum-distance path.
- Safest route (green): the path minimizing the total risk_weight.

In most cases, we observed that the safest route diverged slightly from the shortest path, skirting areas with clustered accidents. However, in sparse areas or small graphs, both routes occasionally overlapped due to limited alternative paths.

Example:



5.2 Enhanced Routing: Complex Risk Logic:

Methodology

To improve upon the simple accident count method, we introduced a more nuanced risk formula that integrates additional risk factors:

- Severity: Each accident's severity (scale 1–4) directly contributes to its risk impact.
- Weather Conditions: Accidents occurring in adverse weather (Rain, Fog, Snow, etc.) were considered more dangerous and multiplied by a 1.5× factor.

For each road segment, the risk formula became:

$$\text{risk\_weight} = \text{length} \times (1 + 5 \sum(\text{severity} \times \text{weather\_factor}))$$

Where:

- $\text{weather\_factor} = 1.5$ if bad weather, otherwise 1.0.

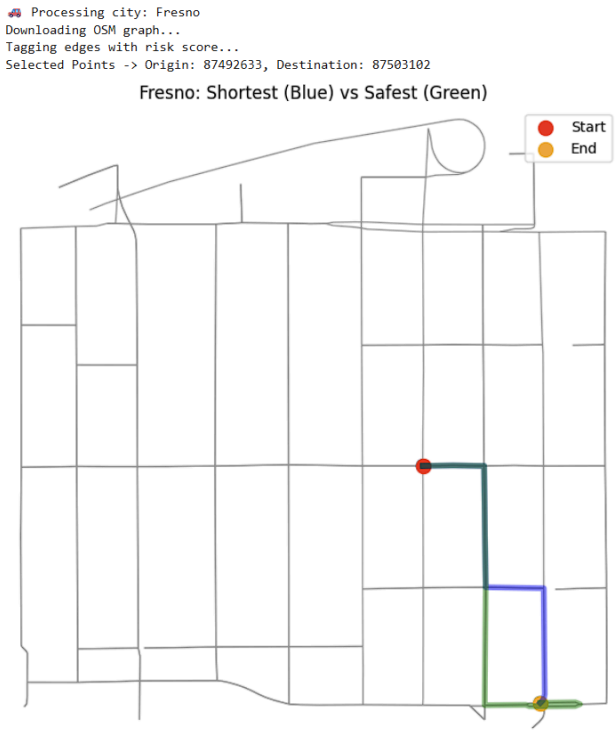
Visualization and Results

By incorporating severity and weather conditions, the routing algorithm prioritized avoidance of not just accident-dense areas but specifically those with more severe and hazardous incidents. This resulted in:

- More pronounced differences between the shortest and safest routes.
- In cities like Los Angeles, the safest path diverged further, especially in areas with multiple severe accidents.

This enhancement demonstrates that accident data's depth matters: not all accidents pose an equal risk, and safer routing should reflect this.

Example:

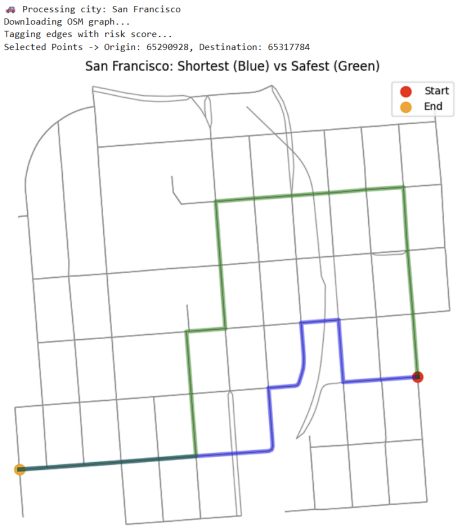


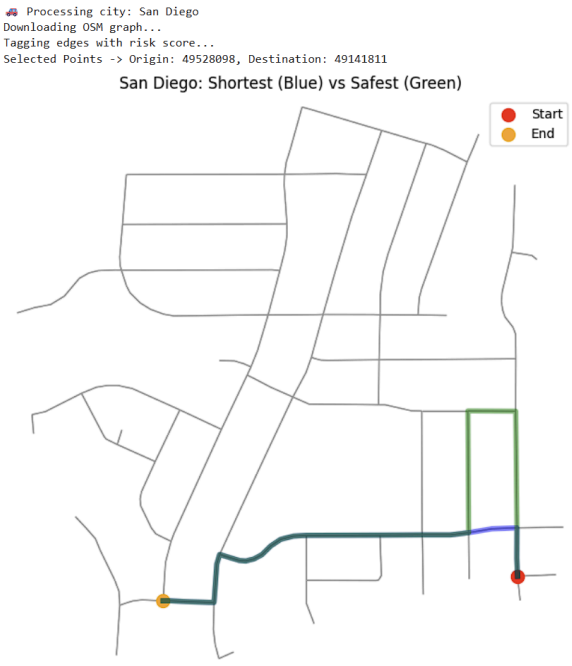
5.3 Comparative Observations

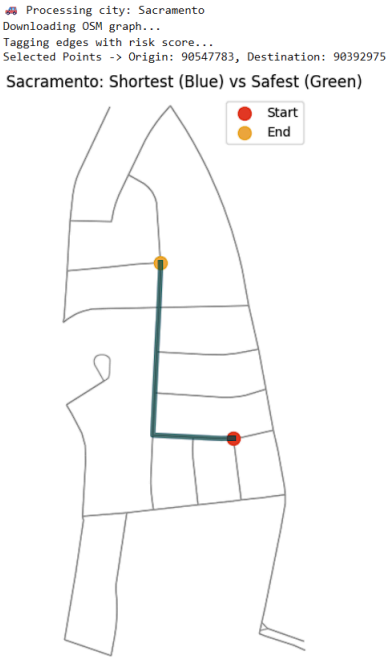
An important observation from the analysis is that, due to the limited 500-meter radius, the randomly selected origin–destination points are often located in close proximity to each other. As a result, the computed shortest and safest routes frequently overlap or are nearly identical, since the constrained spatial area offers fewer alternative paths. Additionally, within such short distances, environmental factors like weather conditions have minimal impact on route differentiation. Consequently, both the basic accident-scaling risk model and the more complex risk logic incorporating severity and weather tend to produce similar outcomes. This reinforces the notion that to fully capture meaningful variation in safety-aware routing, a larger spatial context is necessary.

- Simple Risk vs. Complex Risk:
  The complex model will often result in more effective avoidance of truly high-risk areas, compared to the simple model which only considered count.

- City Differences:
  Larger cities (LA, SF) had denser networks and more accidents, allowing the safest route to meaningfully diverge. Smaller cities (Fresno) saw less difference due to fewer available paths.







6. Limitations and Next Steps

For this analysis, the radius was limited to 500 meters around each city center to ensure computational feasibility, which naturally constrained the available route options and may have excluded broader path variations. Additionally, the dataset was downsampled to 500 rows for testing purposes, which, while practical for rapid prototyping, reduced the overall granularity and real-world accuracy of the risk assessment. Another limitation is that the current model does not account for time-based risk factors such as rush hour traffic patterns or temporal spikes in accident frequency. As part of future work, we plan to address these limitations by leveraging the full dataset, expanding the spatial radius to cover more comprehensive areas, and integrating dynamic variables such as rush hour indicators. We also aim to deploy the routing system as a live, interactive application—potentially using a tool like Streamlit—to support real-time, safety-aware navigation for end users.

7. Conclusion

This project aimed to enhance routing safety by analyzing California crash data and integrating multiple risk dimensions into route planning. We focused on five key cities, enriching the accident dataset with features like weather-based risk factors and severity scores. Initially, we applied a simple risk model that penalized road segments based purely on nearby accident density, using a risk formula that scaled linearly with accident count. Enhanced routing incorporated additional complexity, factoring in accident severity (on a 1–4 scale) and adverse weather conditions, which increased risk weighting by 1.5× when present.

Our routing experiments, conducted within a 500-meter radius around each city center for efficiency, revealed that while the simple and complex risk models performed similarly in constrained areas, the enhanced model more effectively prioritized the avoidance of high-severity, high-risk zones. Visualizations highlighted meaningful route divergence in larger cities like Los Angeles and San Francisco, whereas smaller cities with sparser networks saw limited variation between the shortest and safest paths.

A key takeaway is that both the spatial radius and data granularity significantly impact the potential for meaningful risk-aware routing. Our current setup, while practical for prototyping, limited alternative routing options and muted the influence of environmental risk factors such as weather and severity. To address these constraints, future work will expand the radius of analysis, leverage the full dataset, and incorporate dynamic variables like rush-hour traffic patterns. Ultimately, we aim to deploy this system as a live, interactive routing application—empowering users with real-time, safety-informed navigation choices.

Github : https://github.com/Pooja1819/UsableAI