Exploratory Data Analysis

- StudentID: 22200333

- Name: Harin Park

- 1st Major: ICT 융합

- 2nd Major: Data Science

Brief summary of your proposed project idea.

The project analyzes traffic accident patterns across the United States using a large-scale dataset. It focuses on key factors like time of day, temperature, and location to understand their impact on accident severity and frequency. Through univariate, bivariate, and multivariate analyses, hidden correlations and trends are uncovered. These insights can guide traffic management strategies, enhance road safety, and help prevent accidents, particularly during high-risk times such as rush hours and extreme weather.

1. Data overview

Descriptives statistics on overall data (sample size, number of variables, data type, data range, distribution, etc.)

Total Sample Size (Rows): 7,728,394 Number of Variables (Columns): 45

Key Variables Overview

- **Categorical Virables:**
    - Source, Severity, Street, City, County, State, Zipcode, Timezone, Airport_Code, Wind_Direction, Weather_Condition
- **Geographical Variables:**
    - Start_Lat, Start_Lng, End_Lat, End_Lng, Distance(mi)
- **Date and Time Variables:**
    - Start_Time, End_Time, Weather_Timestamp
- **Weather and Environment-Related Variables:**
    - Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph), Precipitation(in)

- **Boolean Variables (Traffic Features):**
    - Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop
- **Twilight Information:**
    - Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight

2. Univariate analysis

2.1 Severity Analysis

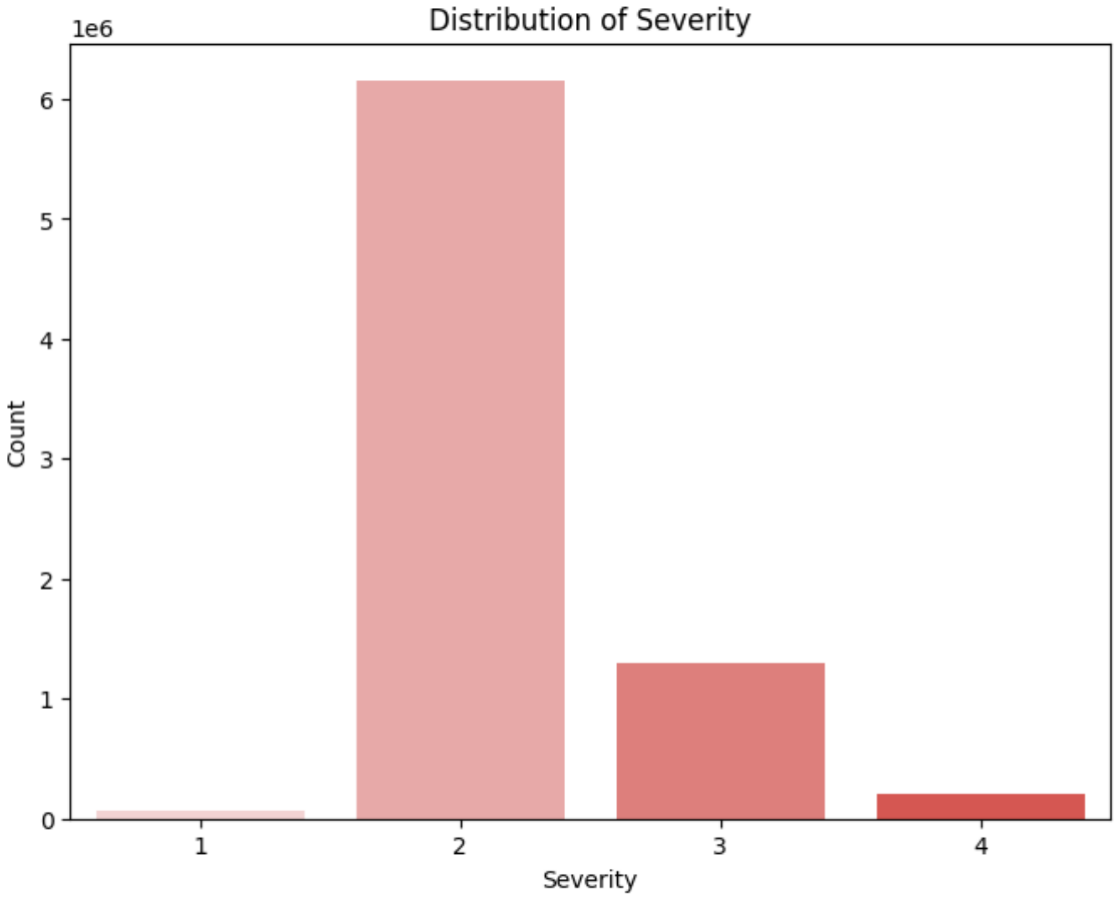

Figure 1. Severity

The majority of accidents are concentrated in Severity 2, indicating that most accidents are of a minor nature. Over 6 million accidents fall into this category, typically causing temporary traffic disruptions. Severity 3 also accounts for a significant portion, representing accidents that can cause major traffic delays.

In contrast, Severity 1 and Severity 4 are rarely observed, indicating very minor and severe accidents, respectively. Based on this distribution, there is a need to enhance rapid response systems for minor accidents and implement monitoring in high-risk areas to prevent severe accidents.

2.2 Start_Lat and Start_Lng Analysis

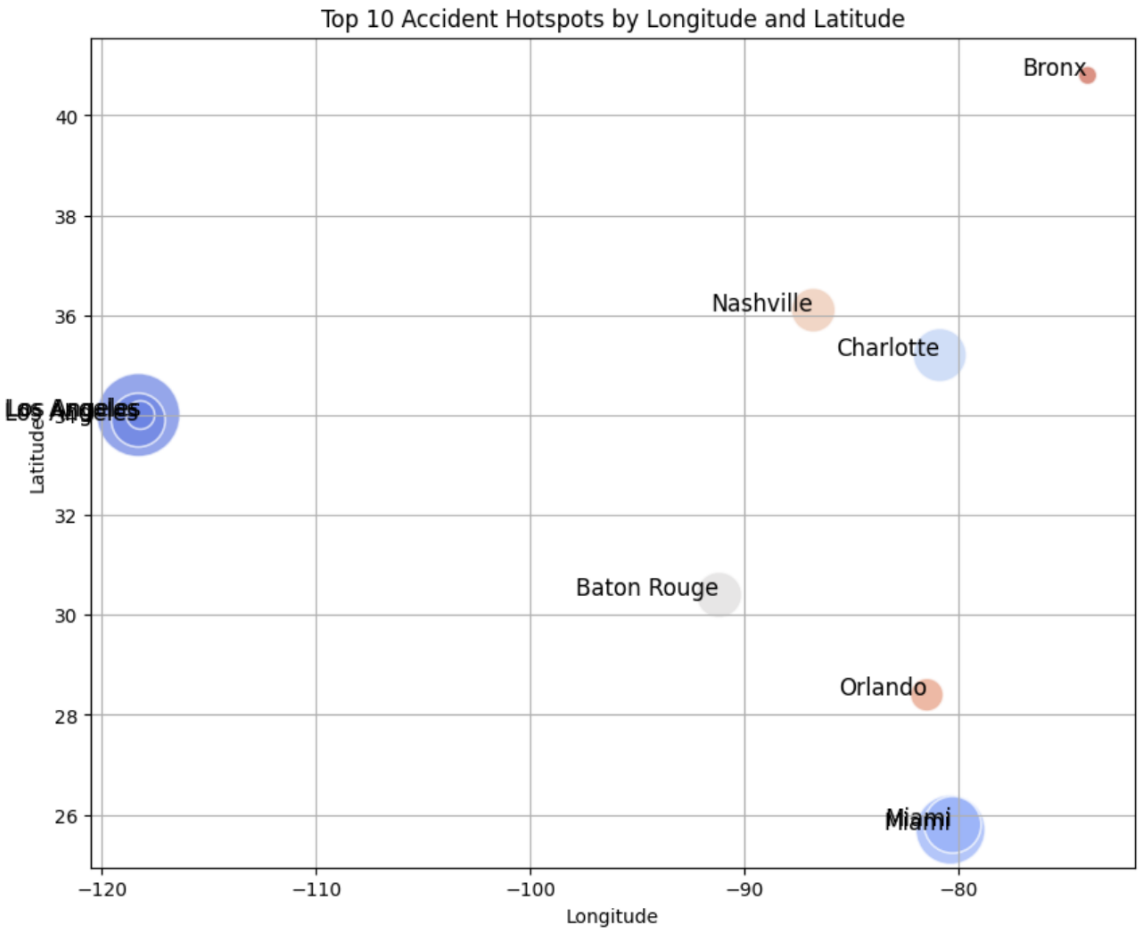

Figure 2. Accident Hotspot Areas

Los Angeles and Miami are represented by the largest circles, indicating that these two cities have the highest frequency of accidents. This suggests that accidents tend to occur more frequently in areas with high traffic volumes and population density.

Major Cities with High Accident Frequency:

- Bronx: The accident frequency is notably high in the New York metropolitan area, particularly on urban roads with heavy traffic.
- Charlotte and Orlando: Frequent accidents can be observed on major roads and urban areas in the southern regions, as highlighted on the map.

Geographic Distribution:

- Accidents are concentrated around major metropolitan areas such as California, Florida, and New York. This demonstrates that traffic congestion in large cities is a key factor contributing to accidents.
- Major cities with high accident rates are primarily concentrated along the East Coast, West Coast, and Southern regions, highlighting the need for accident prevention measures in these areas.

2.3 Temperature Analysis

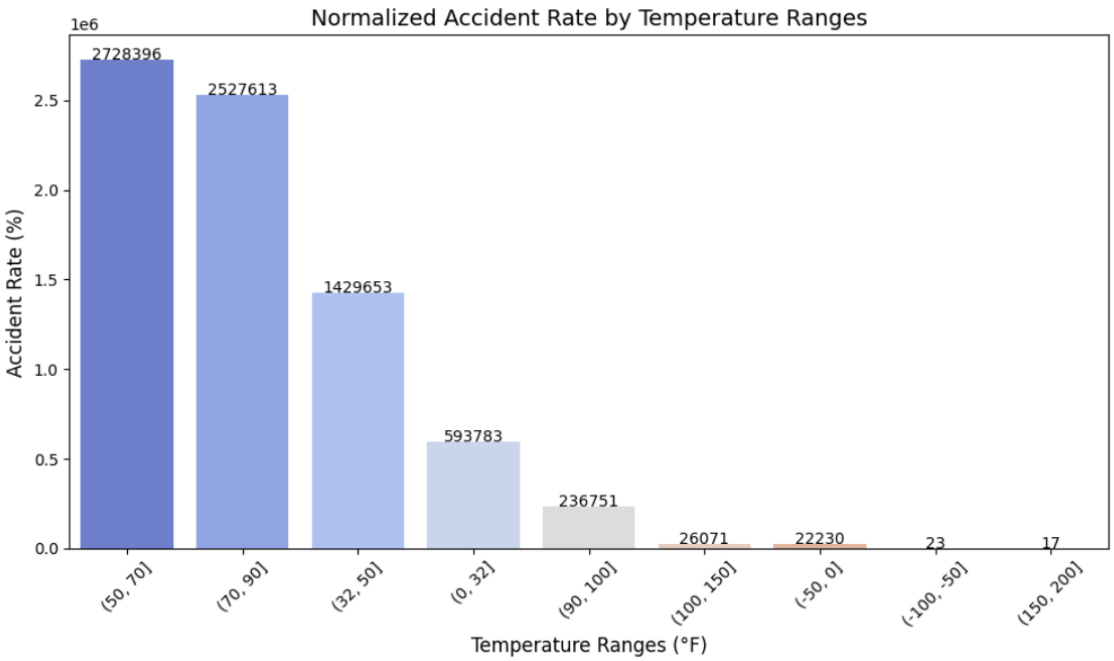

Figure 3. Accident Frequency by Temperature Range

Highest Accident Rate:

- The temperature range between 50°F and 70°F had the highest number of accidents, with over 2.72 million incidents. This range represents moderate temperatures commonly experienced across most parts of the United States, likely corresponding to periods of high vehicle activity.

Accidents in Extreme Heat and Cold:

- Accidents occur far less frequently in extreme temperatures, both in extremely high heat (90°F ~ 100°F) and extremely low cold (-50°F ~ 0°F). Around 230,000 accidents were recorded in the 90°F ~ 100°F range, while almost no accidents occurred at temperatures below -50°F.
- This may be because in extreme temperature conditions, fewer vehicles are on the road, as people tend to stay indoors, leading to lower driving activity.

2.4 Time of Day Analysis

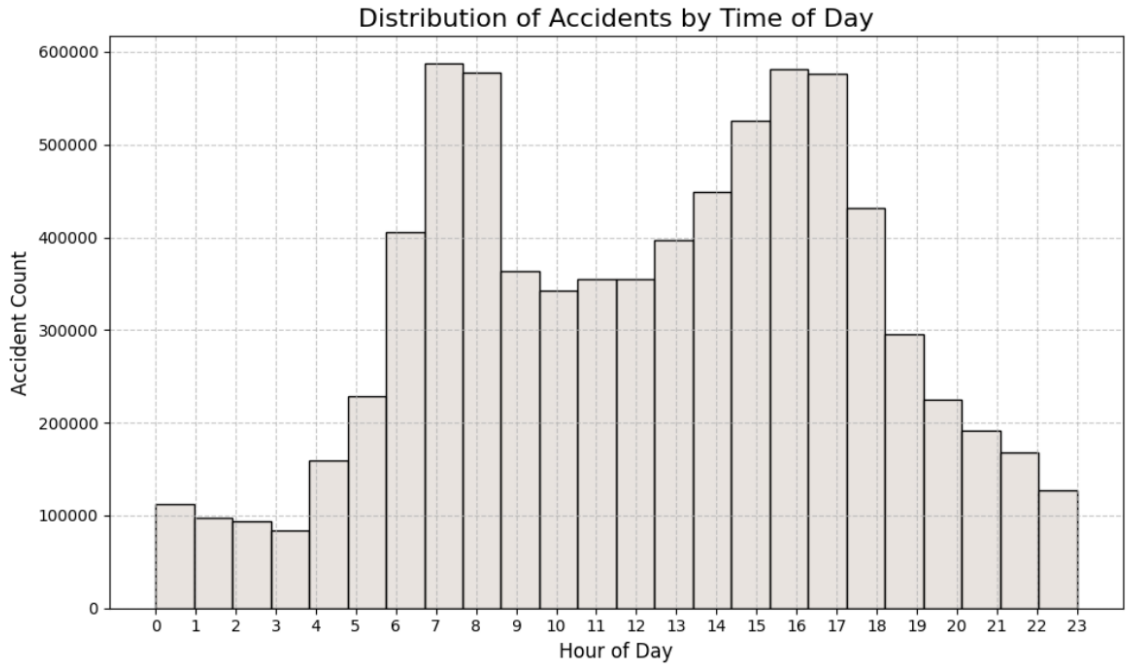

Figure 4. Distribution of Accidents by Time of Day

Accident Frequency during Rush Hours:

- There is a significant increase in accident frequency between 7 AM and 9 AM, and again between 4 PM and 6 PM. This reflects the times when many people are commuting to and from work, indicating that accidents occur more frequently during times of heavy traffic.

Daytime Accidents:

- A relatively high frequency of accidents is also observed between 10 AM and 3 PM. While this time frame does not coincide with rush hours, it is still a period with significant traffic flow, potentially contributing to the higher accident rate.

Early Morning and Nighttime Accidents:

- Accident frequency is considerably lower between midnight and 6 AM. This is a period of reduced traffic volume as fewer vehicles are on the road. Accident frequency also tends to decrease after 8 PM in the evening.

Overall Patterns:

- Double Peak Pattern: The accident distribution shows two distinct peaks during the morning and evening rush hours, while accident occurrences are more evenly distributed throughout the rest of the day.
- This pattern highlights how traffic flow is concentrated during specific times of the day, with a higher risk of accidents when traffic volume increases.

3. Multivariate analysis

Presenation of hidden patterns between variables (correlation, clustering, etc.)

3.1 Correlation Analysis of Incident Severity by Time and Month

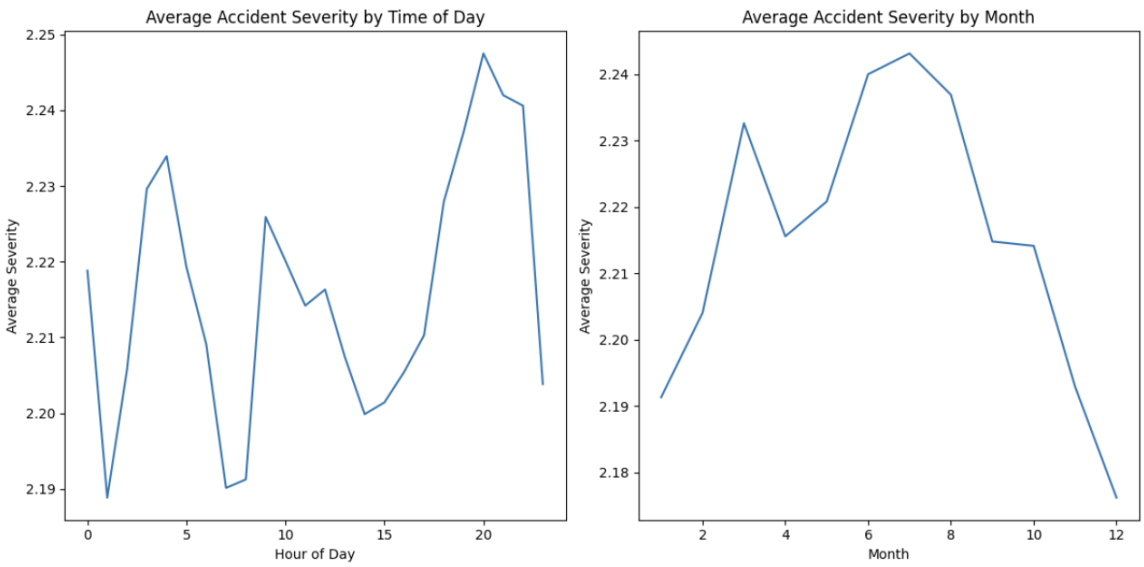

Figure 5. Analyzing Incident Severity by Time of Day & Month

Analyzing Incident Severity by Time of Day

The graph on the left shows the average incident severity by time of day.

- It is evident that the average severity of accidents increases during rush hour, specifically between 7:00 am and 9:00 am, and again between 4:00 pm and 6:00 pm. These times coincide with high traffic volume and road congestion, which can lead to more severe accidents when they occur. Notably, the average severity peaks between 4:00 pm and 6:00 pm, likely reflecting the fatigue of drivers or the intensified rush hour congestion at this time. Conversely, the post-midnight and early morning hours (from midnight to 6:00 am) exhibit lower average severity, indicating that these are periods of lighter traffic, during which minor accidents are more likely to occur.

Incident Severity Analysis by Month The graph on the right shows the average incident severity by month.

- Accident severity tends to rise gradually from early spring (February) to summer (June) and then declines during the summer months. During the summer period (June through August), although road conditions are typically favorable, the increase in traffic volume due to the vacation season may lead to more accidents, thereby increasing accident severity. In the fall (September through November), accident severity tends to be relatively lower, which could be attributed to more stable weather conditions and potentially lower traffic volumes. In December, accident severity drops significantly, which may be due to drivers being more cautious and reducing their speed in response to poor winter road conditions.

This analysis shows that accident severity tends to rise during periods of high traffic volume and in certain seasons (summer). This information can be valuable when developing traffic management strategies and accident prevention policies.

3.2 Clustering Analysis of Accident Patterns by Temperature and Distance

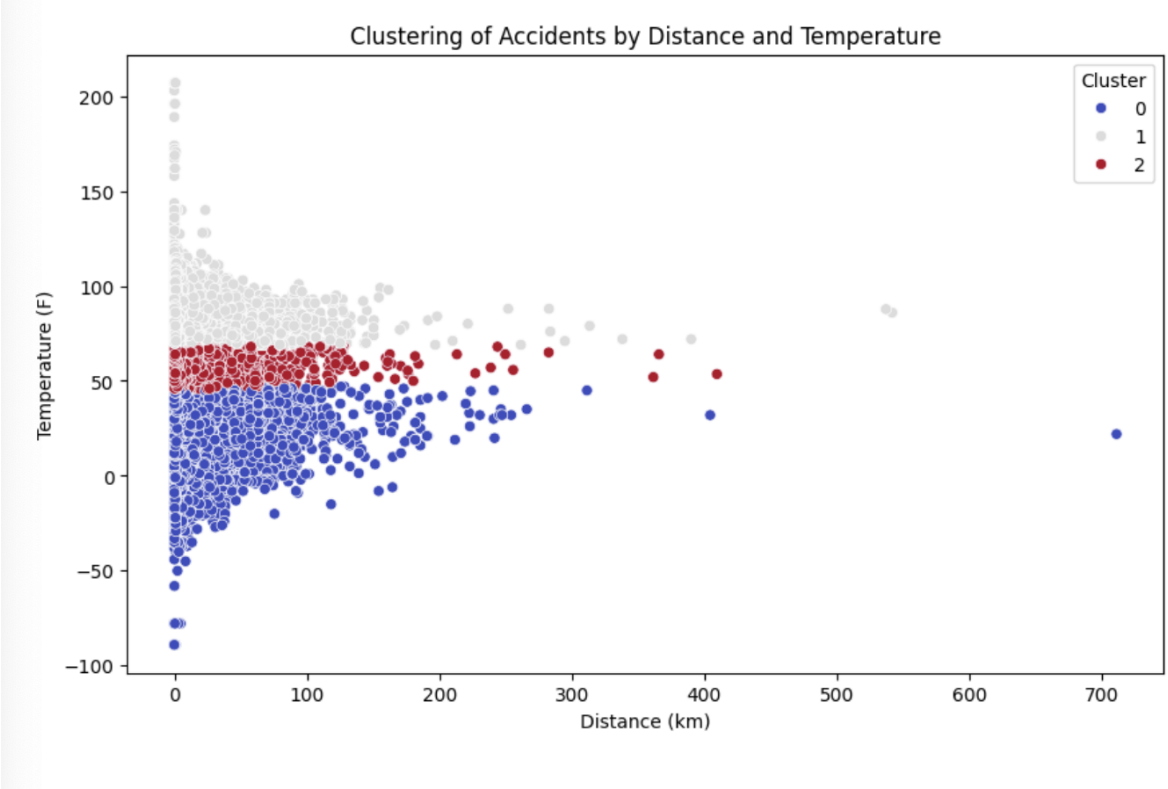

Figure 6. Clustering of Accident Patterns by Temperature and Distance

Blue Cluster (Cluster 0)

- Accidents primarily occurred in cold temperatures (-50°F to 50°F).
- Most accidents happened within distances of less than 100 km, though some occurred during long-distance driving, exceeding 300 km.
- This cluster indicates accidents that occurred in cold weather, where icy roads or reduced visibility may increase accident risk.

Gray Cluster (Cluster 1)

- Accidents occurred in moderate temperatures (50°F to 100°F).
- This cluster mostly involves accidents within short distances of less than 50 km.
- These accidents occurred under relatively stable road conditions, but with heavy traffic flow or congestion, leading to a higher likelihood of accidents.

Red Cluster (Cluster 2)

- Accidents in moderate temperatures (50°F to 100°F) that occurred during long-distance driving.
- Many accidents occurred at distances over 100 km, which suggests that driver fatigue or long-distance travel contributed to these incidents.
- Although road conditions were generally favorable in this temperature range, long-distance driving accidents are more common in this cluster.

This clustering analysis shows a link between accidents, temperature, and driving distance. In colder temperatures, accidents are more frequent on shorter trips, while in moderate temperatures, they tend to occur during long-distance driving. Factors like driver fatigue and weather conditions impact accident risk.

With these insights, targeted measures can be implemented to reduce risks associated with winter driving and long-distance travel.

4. Suggestion

Based on the insights obtained from the previous analyses, a targeted traffic management system can be proposed. This system would focus on high-risk times, locations, and weather conditions that significantly contribute to accidents. The project's key suggestions are:

Enhanced Traffic Monitoring during Rush Hours: Since accidents peak during morning and evening rush hours, implementing advanced traffic monitoring and accident prevention measures such as congestion control, traffic light adjustments, and increased police presence could reduce accident severity during these times.

Seasonal Safety Campaigns: The analysis showed that accident severity tends to rise during the summer months. A targeted campaign during vacation seasons to raise awareness of accident risks, promote cautious driving, and enforce stricter traffic regulations could be highly effective.

Weather-Based Road Management: Cold weather and icy road conditions were identified as critical factors for accident occurrence in shorter distances. A system that provides real-time weather updates, coupled with winter road treatments (such as de-icing) and timely alerts to drivers about hazardous road conditions, could significantly reduce accident risks in colder regions.

Long-Distance Driving Alerts: Accidents linked to long-distance driving, especially in moderate temperatures, indicate the role of driver fatigue. Introducing mandatory rest breaks or fatigue monitoring systems for long-distance drivers could help reduce these types of accidents.

By addressing these specific areas, the proposed project will not only contribute to safer road conditions but also provide a framework for proactive accident prevention policies.