December 2025

EDA Project Report

Road Accidents in the United States of the year 2022

Prepared By:
Niharika Agarwal

Introduction

The US Accidents Dataset is a large-scale, countrywide collection of road accident records covering **49** U.S. states. It captures traffic incidents reported between February 2016 and March 2023, resulting in a dataset of approximately **7.7 million** records.

The data were sourced through multiple real-time traffic APIs, which aggregate incident reports from a variety of entities, including:

- U.S. and state Departments of Transportation
- Law enforcement agencies
- Traffic cameras and sensors
- Other road-network monitoring systems

This dataset provides a rich foundation for exploratory data analysis, enabling the study of accident trends across regions, time periods, environmental conditions, and road characteristics.

Purpose of Analysis

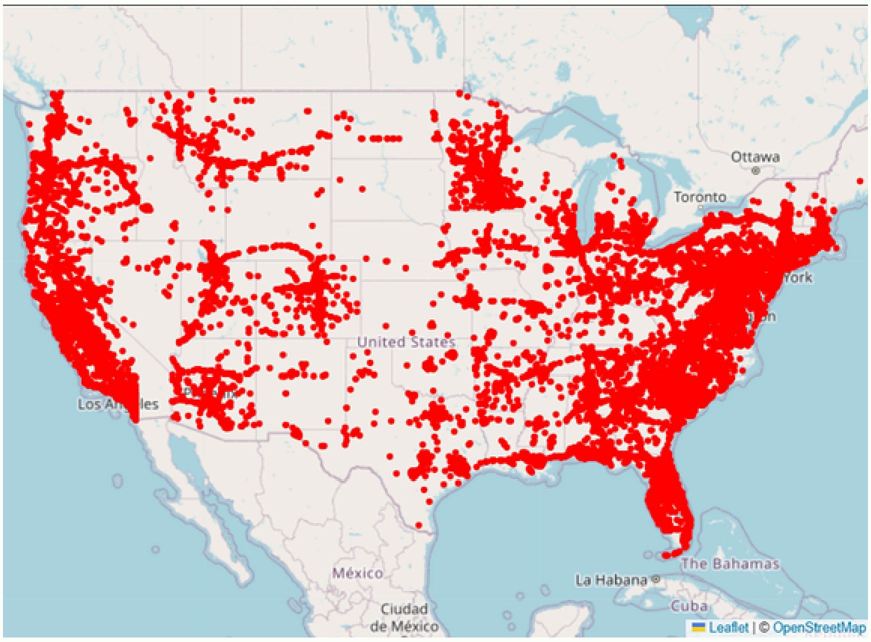
The aim of this analysis is to **understand the patterns, contributing factors, and spatial distribution of road accidents** across the United States using the US Accidents Dataset. Specifically, the analysis seeks to:

- **Normalize** accident counts using population-adjusted metrics, enabling more meaningful comparisons across states with varying population sizes.
- **Identify high-risk locations** and geographic clusters of accidents using spatial visualizations such as maps.
- **Examine** how environmental conditions, road infrastructure, and traffic features (e.g., weather, junctions, signals, crossings) relate to accident occurrence.
- **Provide insights** that can support road safety understanding, policymaking, and future predictive modeling.

Overall, the analysis aims to build a comprehensive, **data-driven** understanding of when, where, and why accidents occur, and to present the findings clearly within an exploratory data analysis framework.

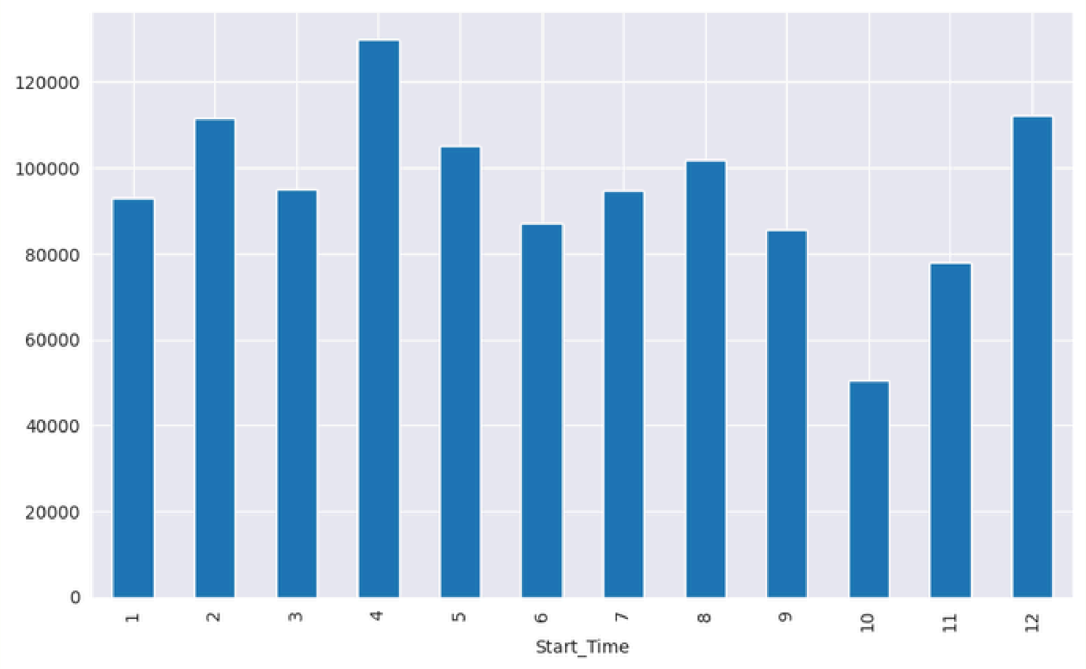
Dataset Overview

- While the original dataset contains 7.7 million records from 2016 to 2023, this analysis focuses specifically on accidents recorded in 2022, as this year provides the most complete and consistent information. It has 1144496 rows and 46 columns
- Some of the key features of the dataset are ‘Accident Severity’, ‘Location of the accident’, ‘information about surrounding infrastructure’, recorded with precise timestamps.
- The ‘End_Lat’ and ‘End_Lng’ fields contained over 50% missing values and were therefore removed during data cleaning.
- The accompanying map visualizes the geographic distribution of the remaining accident records across the country.



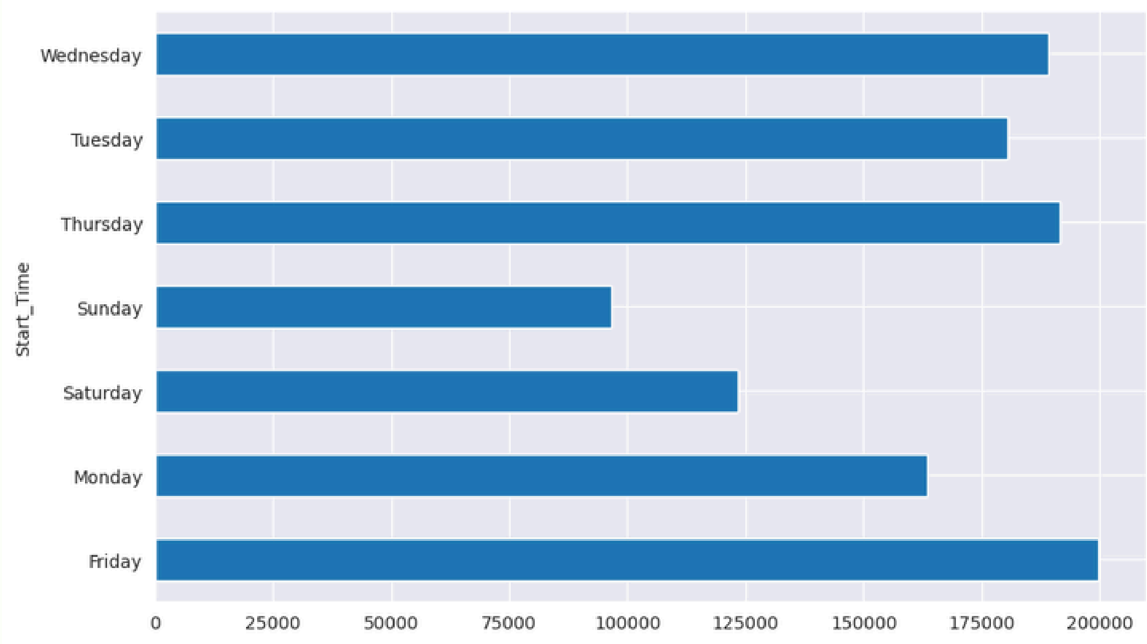
Univariate Analysis

1. Monthwise Plot of No. of Accidents



The chart shows clear monthly variation in accident counts for 2022, with peaks in April, February, and December, and the lowest occurrence in October. Overall, accident frequencies remain relatively high throughout the year, but **seasonal fluctuations are evident**, possibly influenced by weather patterns, travel volumes, or regional conditions.

2. Daywise Plot of No. of Accidents

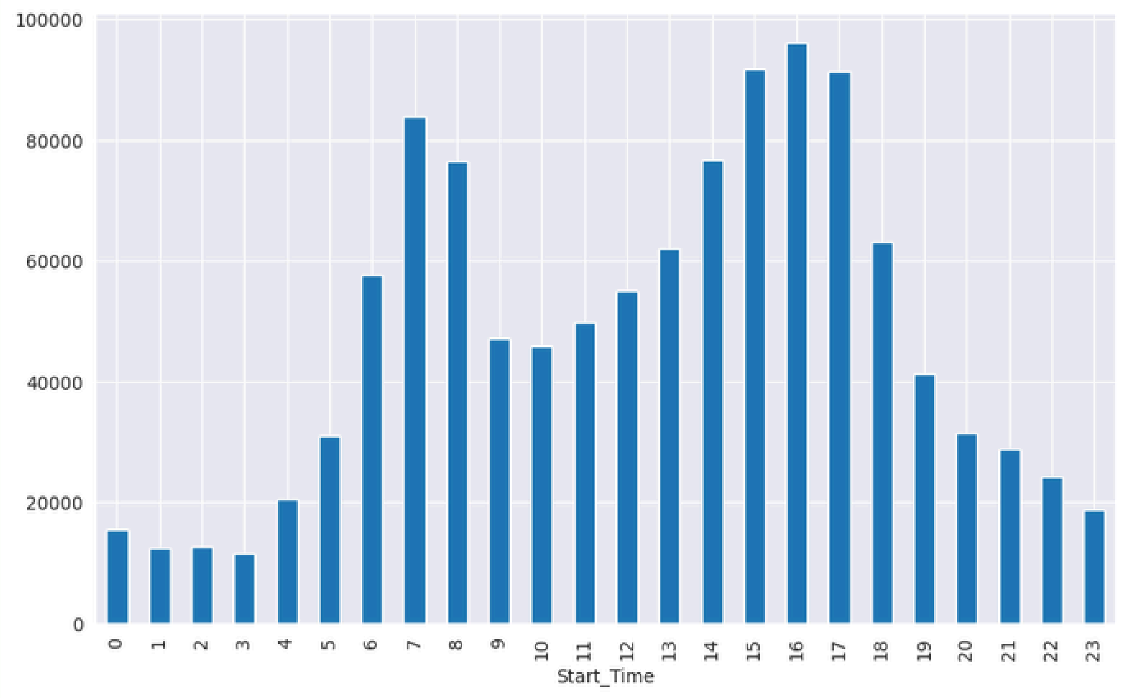


The chart shows that accident frequency varies noticeably across the days of the week. **Friday records the highest** number of accidents, followed by Thursday and Wednesday, while **Sunday has the lowest** count.

- **Higher commuter traffic** from Monday to Friday naturally increases exposure to accidents.
- Lower work-related travel and **lighter overall traffic** likely contribute to the lowest accident count.

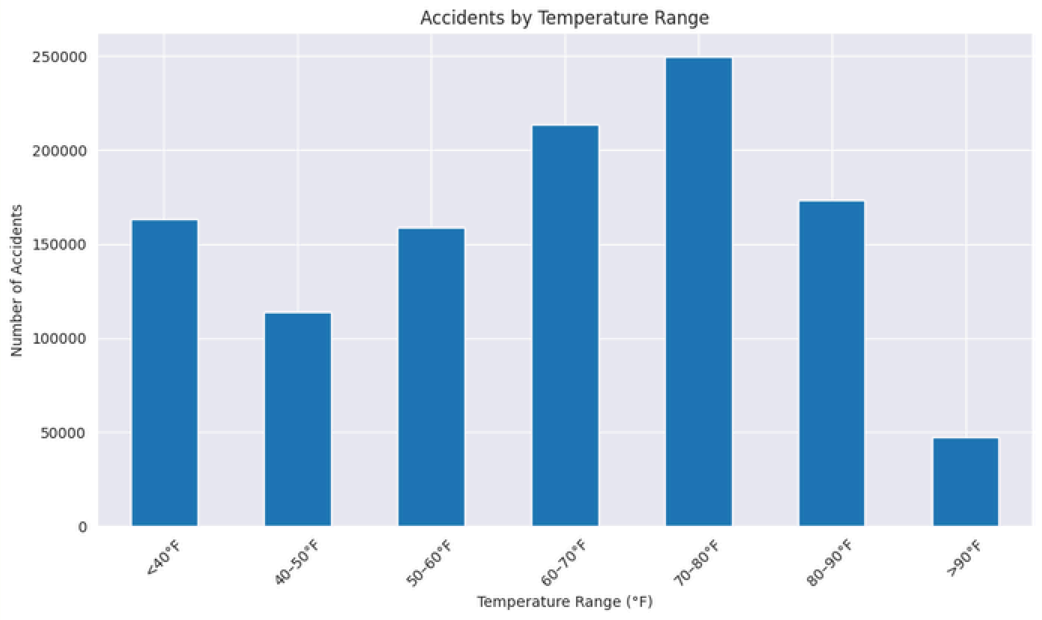
Overall, the trend aligns with typical traffic behavior—more accidents occur on busy weekdays, and fewer on low-traffic weekends.

3. Hourly Plot of the No. of Accidents



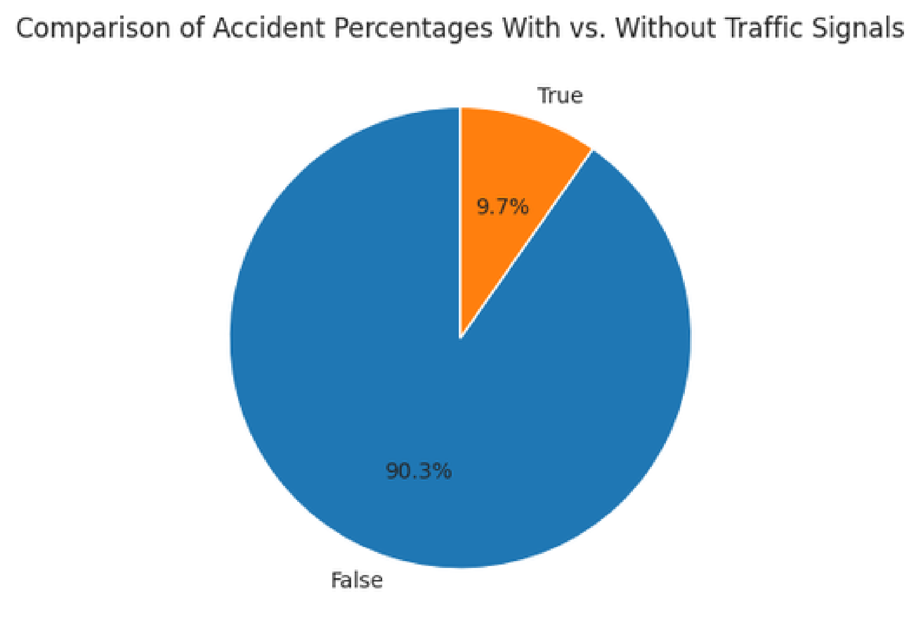
- The chart shows a strong hourly pattern in accident occurrences. Accident counts are lowest during late-night and early-morning hours (0:00–5:00), then **rise sharply with the morning commute** peak around 7:00–8:00.
- After a midday dip, counts climb again and reach the highest peak between 15:00–17:00, **aligning with the evening rush hour**. Following this, accidents gradually decline into the night.
- This pattern reflects typical daily traffic behavior—minimal traffic overnight, heavy volume during commute hours, and the highest congestion in the late afternoon—leading to increased accident frequency during these periods.

4. Relationship Between Temperature and Accident Rates



- A simple comparison of temperature ranges shows that approximately 13.6% of recorded accidents occurred during very low temperature conditions (<40°F), 41.8% in moderate temperature range (60-80°F) and 19.2% at high temperatures (>80°F).
- The distribution of accidents across temperature ranges indicates that most incidents occur in moderate temperatures (60–80°F), which is expected since this range represents the majority of typical driving conditions across the U.S.
- A smaller proportion of accidents happen in very cold conditions (<40°F), largely because such temperatures occur less frequently in many states. Similarly, high-temperature conditions (>80°F) account for around one-fifth of accidents, reflecting regions and seasons where heat is more common.
- **Overall, the percentages align more with how often each temperature range occurs rather than suggesting that any specific temperature band is inherently more dangerous.**

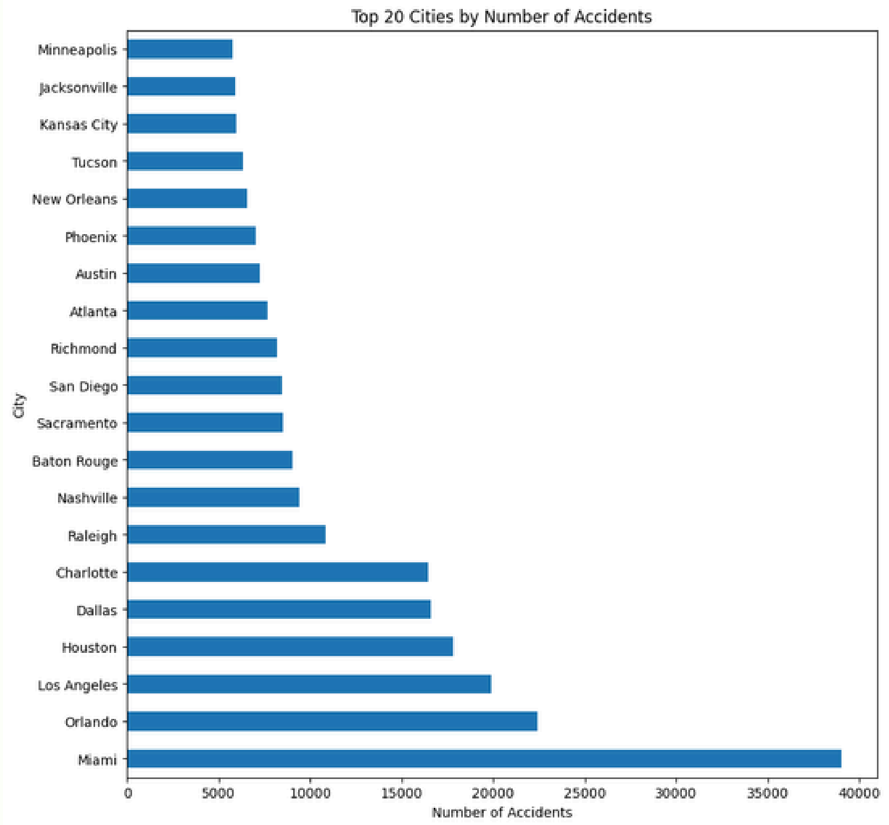
5. Impact of Traffic Signals on Accident Frequency

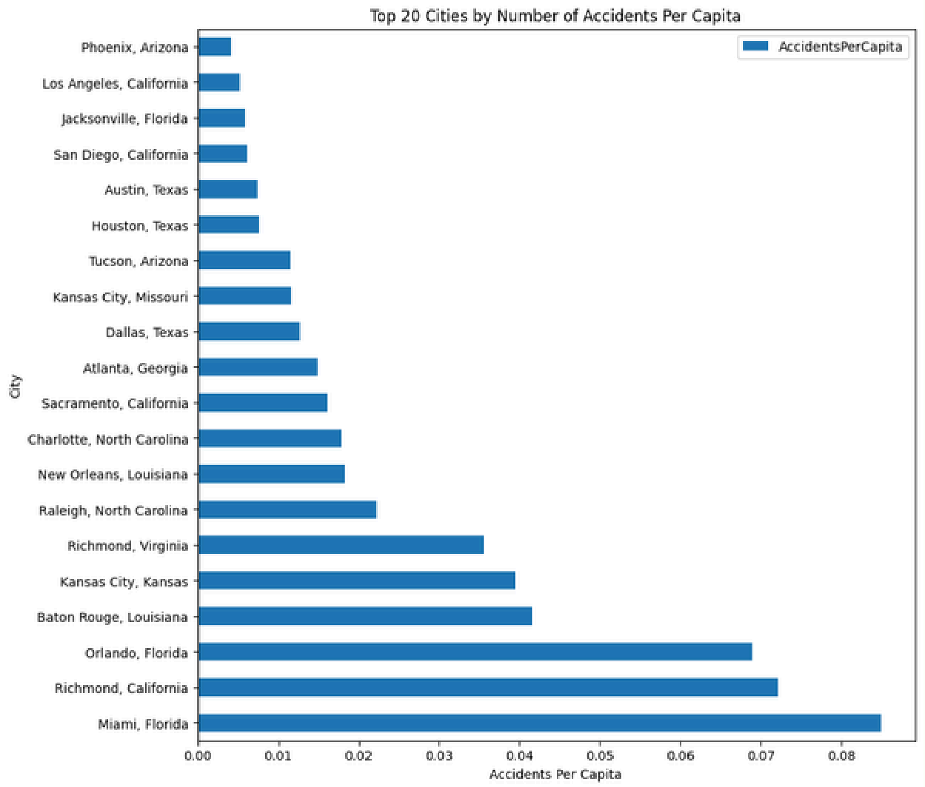


The majority of accidents (around 90%) occur at locations without traffic signals, while only about 10% occur at signalized intersections. This suggests that the presence of traffic signals is not associated with a higher overall accident count and may help regulate traffic at busy intersections. However, this **does not account for exposure as many roads simply do not have signals**. So, further analysis would be needed to determine their true impact on accident risk.

Multivariate Analysis

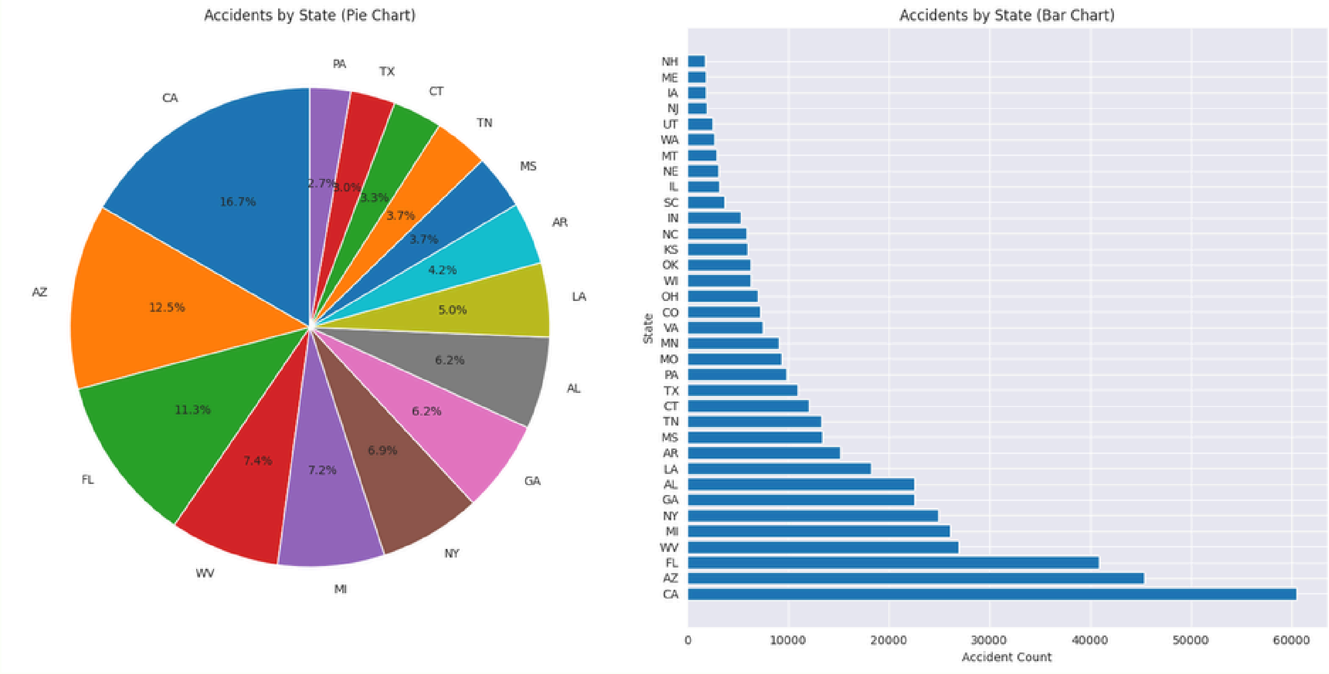
1. Which Metric Is More Informative: Total Accidents or Per-Capita Rates?





The comparison between total accidents and per-capita accident rates **shows a clear difference in interpretation.** Cities like Miami, Orlando, Los Angeles, and Houston appear at the top when considering total accidents, reflecting their large populations and heavy traffic volumes. However, when adjusted for population, a different set of cities such as Miami, Richmond (CA), Orlando, and Baton Rouge show much higher accident rates per resident, indicating greater relative risk. **This demonstrates that per-capita metrics provide a more accurate and fair measure of accident risk than raw accident counts.**

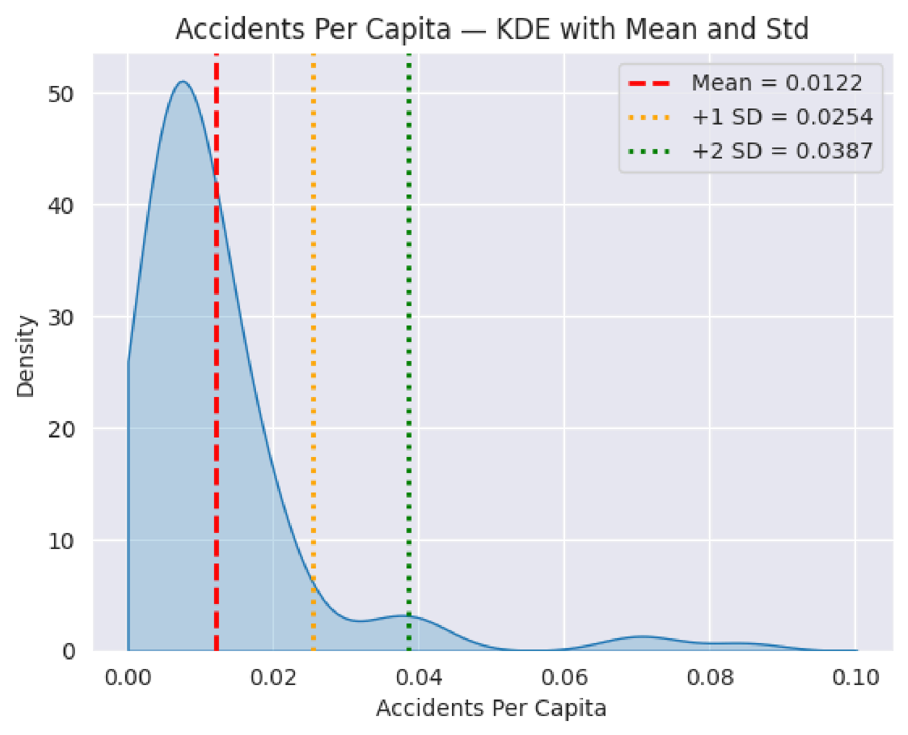
2. State Distribution of the Top 100 Most Accident-Prone Cities



- The state-wise breakdown of accidents shows a **clear concentration in a few high-traffic states**. California records the highest number of accidents, followed by Arizona and Florida, reflecting their large populations, extensive highway systems, and significant daily travel volumes. States such as West Virginia, Michigan, Georgia, Alabama, and Louisiana also contribute substantial shares, although at lower levels than the top three.

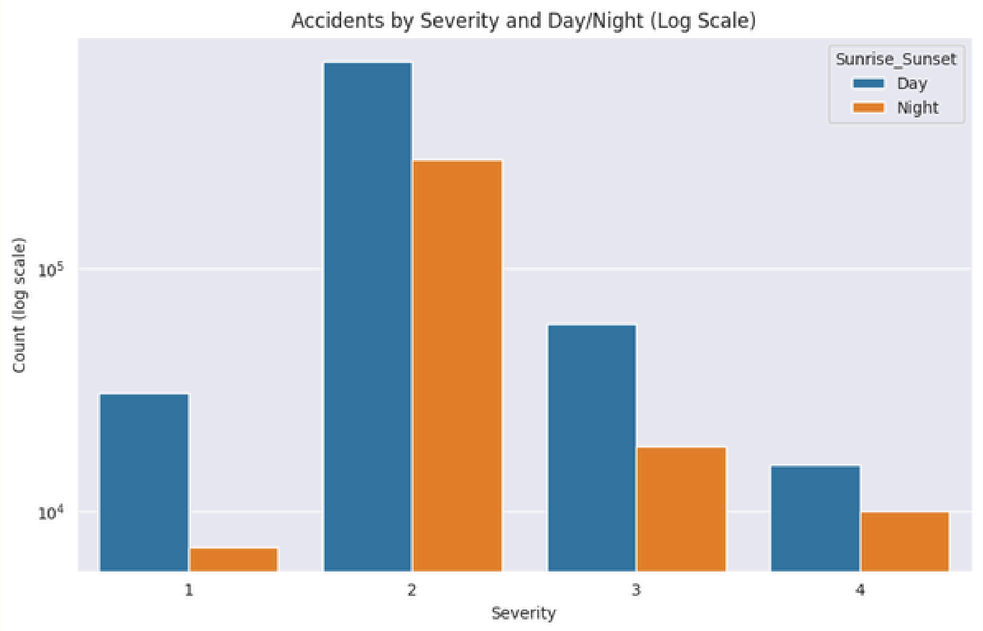
- In contrast, **many states show comparatively low accident counts**, indicating less dense traffic conditions or smaller populations. Overall, the distribution highlights that accident frequency is strongly influenced by state-level factors such as **population size, urbanization, traffic density**, and regional driving patterns, with a **handful of states accounting for a major portion of total incidents.**

3. Do Most Cities Show Moderate Accident Rates or Outliers With Very High Rates?



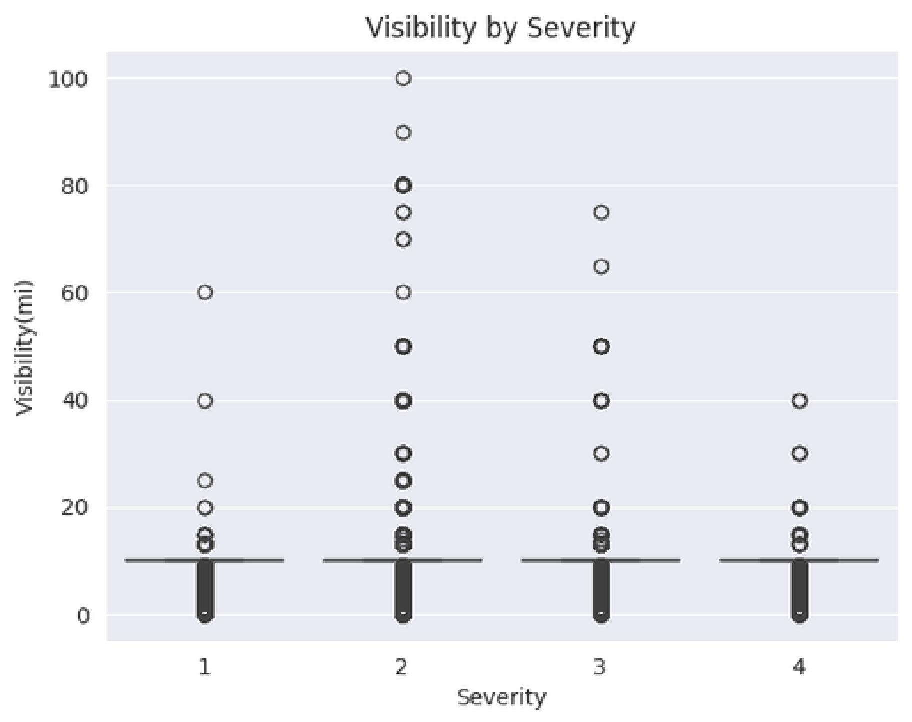
- The distribution of accidents per capita across U.S. cities is **highly right-skewed**. Most cities have relatively low accident rates, clustered around the mean of 0.0122 accidents per resident.
- When examining accident frequency in absolute terms, 21.2% of cities have more than 5,000 accidents, indicating that only a **small fraction of cities generate very high accident volumes**—primarily due to their large populations and heavy traffic.
- However, when adjusting for population size, 39.8% of cities have an accident rate greater than 0.01 (i.e., more than 1 accident per 100 residents). This suggests that a substantial share of cities experience **elevated accident rates relative to their population**, even if their total accident count is not very high. This reiterates the importance of using per-capita metrics to capture underlying risk that raw accident counts may obscure.
- Overall, the KDE analysis shows that while most cities fall near the average accident rate, **a non-trivial tail of cities exhibit significantly higher risk**, especially those exceeding the +1 SD or +2 SD thresholds.

4. Comparing Accident Severity Between Daytime and Nighttime



The chart shows that accidents of all severities occur more often during the day, with Severity 2 accidents dominating both day and night. However, the gap between day and night narrows as severity increases. For example, while daytime Severity 1 and 2 accidents far exceed nighttime counts, the difference becomes smaller for Severity 3 and especially Severity 4 incidents. This suggests that although daytime accidents are more common overall, **nighttime conditions are associated with a greater proportion of higher-severity crashes.**

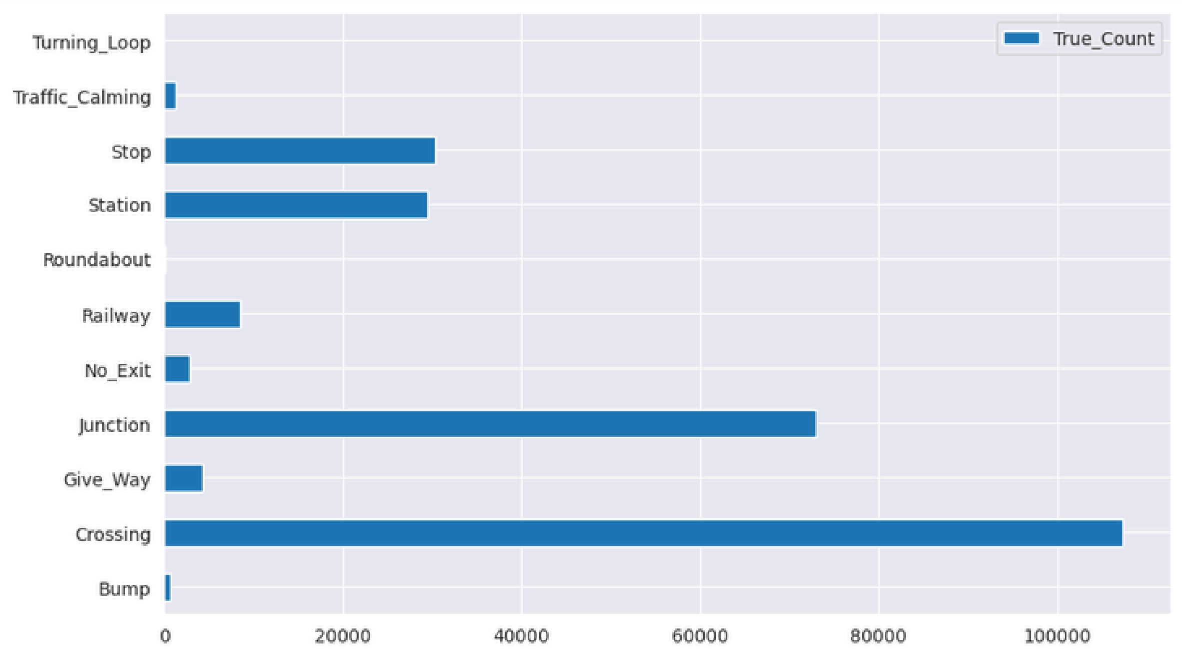
5. Impact of Visibility on Accident Severity



The visibility distribution appears **very similar** across all severity levels. The **boxplots** show that the median visibility is consistently 10 miles for every severity category, and the interquartile ranges almost completely overlap. The summary statistics confirm this pattern: mean visibility ranges only slightly—from about 9.45 miles for Severity 1 to 9.08 miles for Severity 4—indicating no meaningful downward trend with increased severity. **Although there are occasional low-visibility outliers**, they occur across all severity groups rather than being concentrated in the higher-severity categories.

Overall, the data suggests that **low visibility does not show a strong association** with higher accident severity, as most accidents—regardless of severity—occur under normal visibility conditions.

6. Do Certain Road Features Contribute to More Severe Accidents?



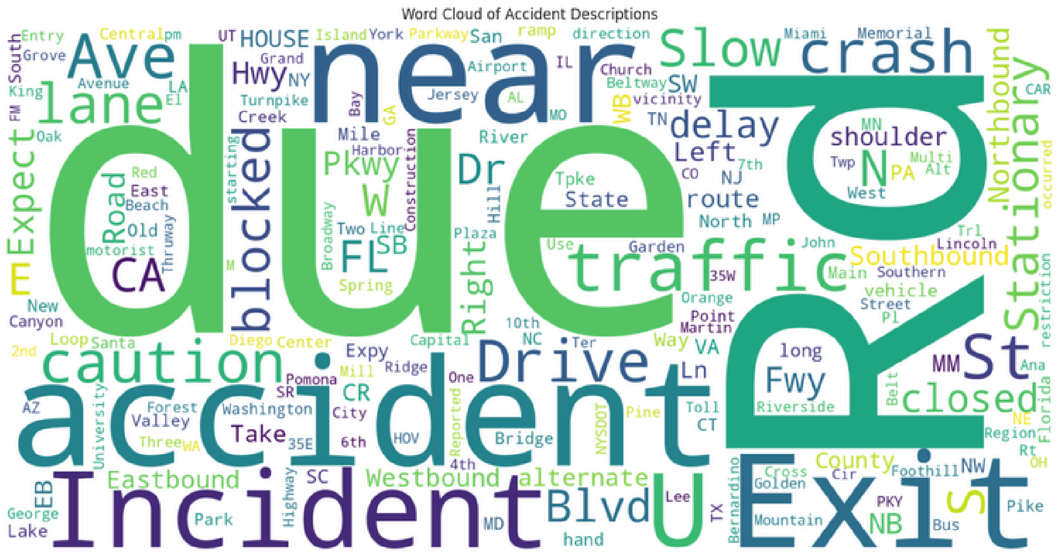
Among the infrastructure features recorded, **accidents occur most frequently near crossings and junctions**, followed by stop signs and stations. Features like roundabouts, bumps, and turning loops show very low accident counts. This suggests that locations involving intersecting traffic flows—such as crossings and junctions—are associated with a higher likelihood of accidents, likely due to increased vehicle interactions and conflict points.

Key Insights & Summary

- **Temporal patterns**: Accidents peak during morning and evening rush hours (7–8 AM and 3–5 PM) and are highest on Fridays and in the months of February, April, and December, reflecting strong traffic-volume effects.
- **Weather impacts**: Most accidents occur under moderate temperatures (60–80°F) and normal visibility conditions, indicating that temperature and visibility have limited direct influence on accident likelihood or severity.
- **Severity drivers**: Although daytime accidents dominate overall counts, night-time crashes represent a higher proportion of severe incidents, suggesting roles for reduced visibility, fatigue, and higher nighttime speeds.
- **Infrastructure correlations**: Accidents cluster around crossings and junctions, where traffic flows intersect, while features like roundabouts, bumps, and turning loops show minimal association with crashes.
- **High-risk cities (normalized)**: Population-adjusted accident rates reveal that cities such as Miami, Orlando, Baton Rouge, Richmond (CA), and Kansas City (KS) exhibit disproportionately high accident risk despite not always having the highest raw counts.
- **Spatial concentration**: A small set of states—especially California, Arizona, and Florida—account for a large share of total accidents, reflecting dense road networks and high travel demand.
- **Day vs night dynamics**: Total daytime accidents far exceed nighttime counts, but severity increases disproportionately at night, highlighting different underlying risk factors.
- **Temperature distribution caveat**: Temperature-based conclusions reflect exposure frequency rather than hazard; the U.S. spends far more hours in moderate ranges, inflating accident counts there.

Limitations & Further Work

- **Uneven dataset coverage:** Accident reporting varies across states and cities, leading to potential geographic biases in the analysis.
- **Missing crash-level detail:** The dataset lacks information on fatalities, vehicle type, driver behavior, and speed, limiting severity interpretation.
- **Incomplete weather attributes:** Some weather and visibility fields contain missing or inconsistent values.
- **Population data merged manually:** Per-capita calculations depend on external population sources, which may not perfectly align with the accident year.
- **Future work:** Incorporating vehicle-level data, traffic volume (VMT), and more consistent reporting would enable deeper risk modeling and predictive analysis.



A word cloud of the description