COMP 4651 Group Project

Exploratory Data Analysis on Traffic Accidents

LI, Chi Kin CHEUNG, Ching Long LIU Yi Man

ckliam@connect.ust.hk

clcheungag@connect.ust.hk

ymliuaa@connect.ust.hk

I. Introduction

Road safety is a critical concern in the United States, with millions of traffic accidents occurring annually, resulting in significant economic, social, and personal repercussions. This report leverages the **US Accidents Dataset**, which contains detailed records of over 7.7 million traffic accidents spanning from **2016 to March 2023**, to analyze various aspects of road accidents. By exploring trends over time, geographic distributions, and contributing factors such as weather and road conditions, this analysis aims to uncover patterns and insights that could inform strategies to improve road safety and reduce accident frequency.

The report is structured by key sections, including **time-based analysis**, which examines accident trends by year, month, day, and hour; **location analysis**, which focuses on accident hotspots across cities and states; and **road condition and weather analysis**, which investigates the impact of environmental and infrastructural factors on accident severity. Through these analyses, the study provides valuable insights into the dynamics of traffic accidents and highlights areas where interventions can make a meaningful impact in mitigating risks and improving public safety.

II. Dataset Introduction

The US Accidents dataset is a comprehensive collection of 7,728,394 traffic accidents across the United States. Each accident record is uniquely identified and includes temporal information (start and end times), along with a severity rating ranging from 1 to 4. The dataset captures detailed location data, including latitude, longitude, city, state, and time zone of each incident. Weather conditions at the time of accidents are thoroughly documented, incorporating measurements of temperature, humidity, atmospheric pressure, wind chill, wind speed, and visibility. The dataset also features a series of Boolean indicators for various road conditions and infrastructure elements, such as amenities, bumps, crossings, give-way signs, junctions, no-exit zones, railways, roundabouts, stations, stop signs, traffic calming measures, traffic signals, and turning loops. This rich set of attributes makes the dataset valuable for analyzing patterns and factors contributing to traffic accidents in the United States.

III. Development Environment

The analysis is conducted in a Jupyter Notebook environment, leveraging various tools and libraries to perform the required tasks. The **kagglehub** library is used to seamlessly download the dataset directly from Kaggle, ensuring easy access to high-quality data. For distributed data processing, **PySpark** is employed, taking advantage of its robust framework for handling large-scale datasets efficiently. To support visualization and data manipulation, **matplotlib** and **pandas** are utilized. These libraries enable insightful visualizations and facilitate intuitive data exploration and preprocessing.

A key component of this analysis is the use of the **MapReduce paradigm** within PySpark to efficiently count accidents by city and state. By applying the MapReduce model, PySpark processes large volumes of data in a distributed manner, breaking down the task into multiple smaller operations (mapping), and then aggregating the results (reducing). This approach not only improves performance but also ensures scalability when working with extensive datasets.

IV. Location Analysis

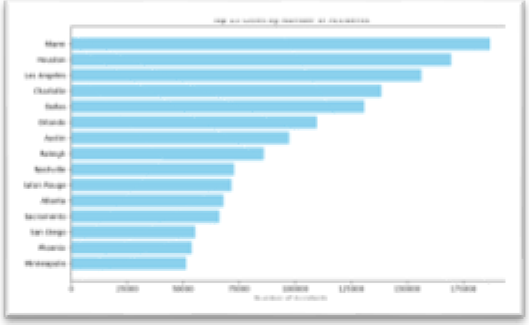


Figure 1. Generated Bar Chart Showing Top 15 Cities by Number of Accidents

From the bar chart, we can see that Miami is the city with the highest number of road accidents in the US (2016-2023). Besides, around 20% of all accident records of these 7 years are from these 15 cities out of 13679 cities in the US.

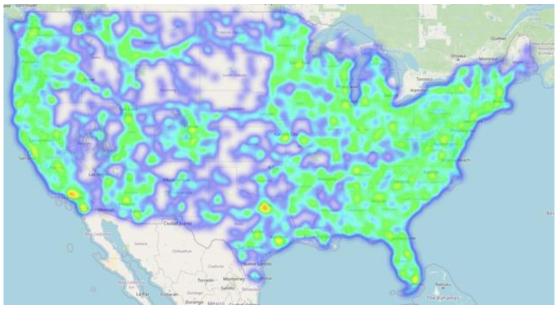
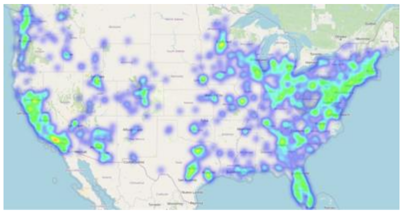
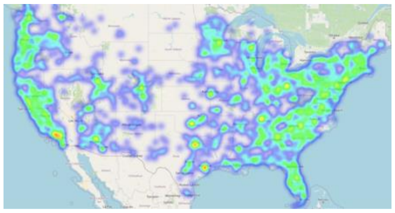


Figure 2. Heat Map Visualization on Road Accident Locations

Also, Road accidents mostly concentrate on metropolitan areas. The heat map on the right visualizes the location and density of road accidents. The green area represents less frequent accidents, and the yellow area represents more frequent accidents. We can see that the distribution of accidents is not even. It concentrates in certain metropolitan areas such as LA city, Dallas and Miami.

 2016	 2017	 2018

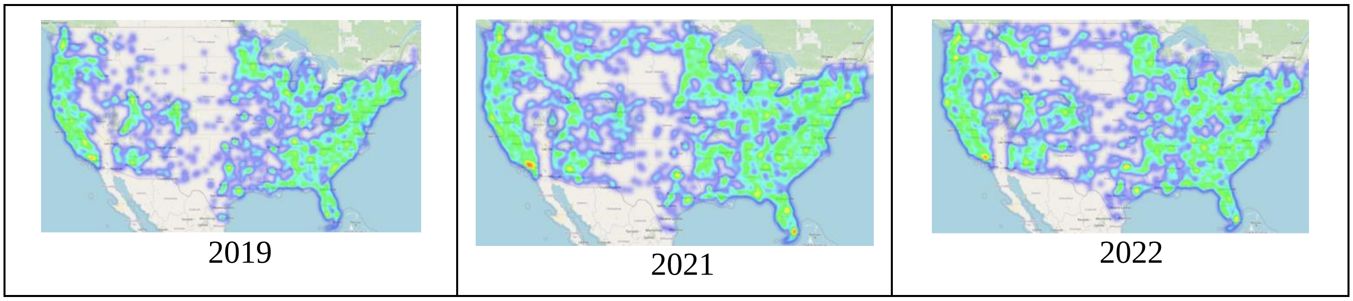


Figure 3. Heat Map Illustration on Geographic Distribution of Road Accidents in United States During Year 2016-2022

The heat maps illustrate the geographic distribution of road accidents across the United States from 2016 to 2022. Over the years, accident density has increased in metropolitan and high-traffic regions, with consistent hotspots in areas like the Northeast, California, and the Southeast. This trend indicates growing traffic and urbanization in these regions. The patterns also show a substantial spread of accidents over time, highlighting the expanding scope of road safety concerns nationwide.

V. Time Analysis

A. Year Analysis

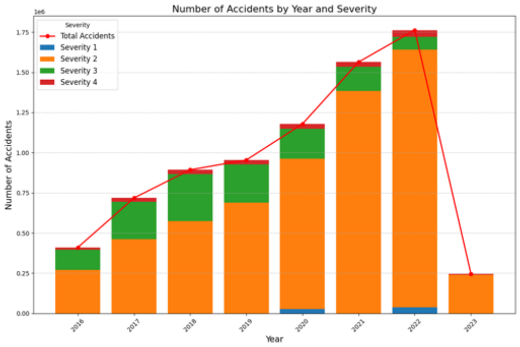


Figure 4. Bar Chart Showing Number of Accidents by Year and Severity

Road accidents have shown a **consistent upward trend** from 2016 to 2022, with an average increase of 226,787 cases per year. This steady rise indicates a worsening road safety situation. Note: The 2023 data is partial, covering only January through March of that year. Besides, most of the accidents happen are at **Severity level 2**, which refers to minor injuries and no casualty.

B. Month Analysis

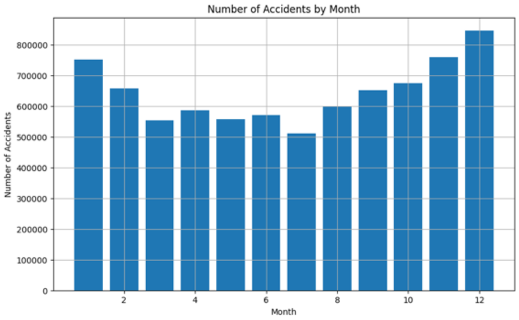


Figure 5. Bar Chart Showing Number of Accidents by Month

The bar chart on the right side shows the distribution of accidents across months. Accidents occur more frequently in **winter and fall** than in **spring and summer**. This gives us an insight that season might also be an important factor to occurrence of road accidents

C. Day Analysis

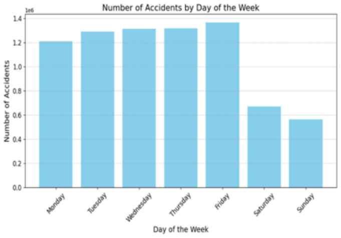
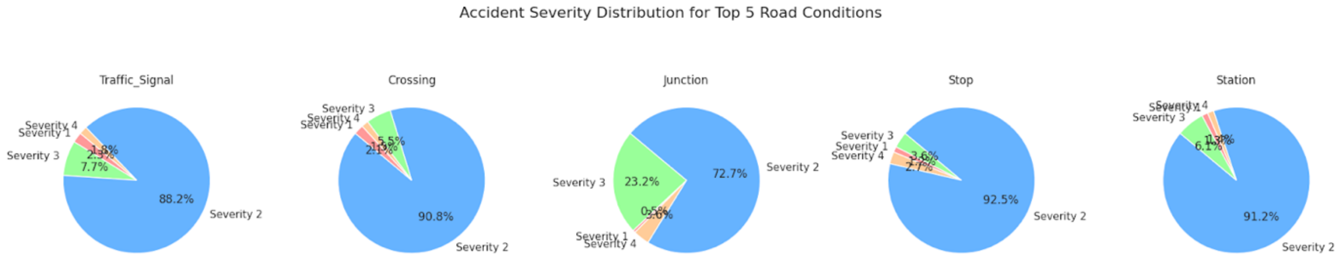
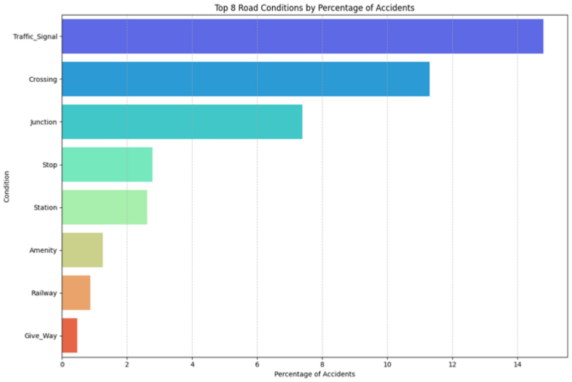


Figure 6. Bar Chart Showing Number of Accidents by Day of the Week

This bar chart shows the distribution of accidents across days of the week. We can see that accidents occur more frequently during weekdays than on weekends, which fits into our intuition. People travel to school or work during weekdays, therefore, there is higher traffic density and hence higher chance of road accidents. The high density of accidents in metropolitan areas also supports this intuition.

VI. Road Condition Analysis





Traffic signals and crossings account for the highest percentage of accidents, with traffic signals leading the chart. Among different road conditions, junctions are the most hazardous, with 23.2% of Severity 3 (serious) accidents, significantly higher than traffic signals (7.7%) and crossings (5.5%). This underscores the complexity and risks associated with vehicle convergence, turning, and potential misjudgments at junctions. In contrast, stop signs emerge as the safest, with 92.5% of accidents at stop signs classified as Severity 2 (minor accidents), and only 3.6% and 0.1% classified as Severity 3 and Severity 4, respectively. This indicates that stop signs are effective in mitigating severe accidents, likely due to reduced vehicle speeds and increased driver caution.

VII. Weather Analysis

Serious accidents (Severity 3) are notably more common in overcast and scattered cloud conditions, with overcast weather accounting for 31.3% and scattered clouds for 32.2% of such accidents. These weather conditions, while not as extreme as rain or snow, still lead to a higher proportion of severe accidents. This trend could be attributed to factors such as reduced visibility or driver misjudgment, which may increase the likelihood of serious incidents. Clear weather, surprisingly, also shows a high proportion of Severity 3 accidents, accounting for 30.3%. While clear weather is typically considered ideal for driving, this proportion is higher than in conditions like light rain (20.1%) or light snow (17.2%). This finding suggests that drivers may become overconfident in clear weather, leading to risky behaviors such as speeding or distracted driving, which contribute to severe accidents. These patterns highlight the complexity of accident causation, where both environmental and human factors play a role. While adverse weather conditions like rain or snow naturally pose challenges for drivers, seemingly benign conditions such as overcast skies or clear weather can also lead to a significant number of serious accidents due to factors like visibility issues or overconfidence behind the wheel.

VIII. Conclusion

This analysis of the US Accidents Dataset (2016–2023) highlights key trends and factors contributing to road accidents in the United States. Accidents are concentrated in metropolitan areas, with Miami, Los Angeles, and Dallas being major hotspots. Time-based analysis reveals that accidents are more frequent during weekdays, winter months, and peak traffic hours, indicating the influence of traffic density and seasonal patterns. Road condition analysis identifies junctions as the most hazardous, while stop signs are effective in mitigating severe accidents. Weather analysis underscores the complexity of accident causation, with overcast and clear weather conditions contributing to a high proportion of serious accidents due to reduced visibility or risky driving behaviors.

These insights emphasize the need for targeted interventions, such as improved traffic management in urban areas, enhanced road infrastructure at junctions, and driver education to mitigate risks in diverse weather and road conditions. By addressing these factors, policymakers and stakeholders can work toward improving road safety and reducing accident frequency nationwide.

Reference

[1] S. Moosavi, “US accidents (2016 - 2023),” Kaggle, https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents (accessed Nov. 29, 2024).

[2] “Quick start,” Quick Start - Spark 3.5.3 Documentation, https://spark.apache.org/docs/latest/quick-start.html (accessed Nov. 29, 2024).

[3] “Advanced pyspark for Exploratory Data Analysis,” Kaggle, https://www.kaggle.com/code/tientd95/advanced-pyspark-for-exploratory-data-analysis (accessed Nov. 29, 2024).