Data Visualization for US Accident Analysis

Revna Altınöz
Artificial Intelligence and Data Engineering Department
Istanbul Technical University
150220756
altinoz20@itu.edu.tr

Eda Selin Küçükkara
Artificial Intelligence and Data Engineering Department
Istanbul Technical University
150210325
kucukkara21@itu.edu.tr

***Abstract*—This project focuses on analyzing the US Traffic Accident Dataset to uncover patterns and insights for traffic accidents in the United States. Through data preprocessing, we ensure clean and standardized data for further analysis. The visualizations and analyses help identify high risk periods, accident prone locations, and environmental factors contributing to accidents. The findings aim to assist policymakers, urban planners, and traffic safety organizations in reducing accident frequencies and improving road safety.**

I. INTRODUCTION

This project aims to analyze traffic accidents in the United States using the US Traffic Accident Dataset which is from 2016 to 2023. The dataset provides detailed information on accident severity, location, time, and environmental factors, enabling a multi-dimensional exploration of accident patterns. Using data preprocessing techniques, the data set is converted into a format with only the columns we need for effective analysis. The project includes temporal analyses, such as yearly, monthly, and hourly trends, as well as geospatial studies to identify accident hotspots for cities and states. Additionally, the influence of weather conditions on accident frequency and severity is examined. Interactive dashboards are designed to make it easier to explore accident data dynamically, giving insights about locations that are prone to accidents, high-risk times, and contributing variables.

II. DATASET

The dataset used for this project is the **US Traffic Accident Dataset (2016–2023)**, sourced from Kaggle. It contains detailed information on over 7.7 million traffic accidents across the United States, with 46 columns with attributes such as accident severity, timestamps, geographical locations, weather conditions, and road features. For this project, we focused on a subset of columns: `ID`, `Severity`, `Start_Time`, `End_Time`, `Start_Lat`, `Start_Lng`, `Street`, `City`, `County`, `State`, `Wind_Speed(mph)`, and `Weather_Condition`, as they are critical for understanding accident trends, severity, and contributing factors.

The preprocessing step focuses on preparing the dataset for analysis. First, we filtered the dataset to include only the relevant columns necessary for our study. We addressed inconsistencies in the `Start_Time` column by identifying problematic rows with incorrect timestamp formats, standardizing the format, and converting it to a proper datetime type. Additionally, we extracted date features, such as year, month, and day, from the `Start_Time` column to enable temporal analysis. The improved dataset provides useful data on traffic safety through complete evaluation and visualizations of accidents across time, locations, and weather conditions.

III. ANALYSES AND DISCUSSION

The following visualizations are used to analyze the dataset:

- **Line charts** for yearly, monthly, and daily accident trends.
- **Heatmaps** for hourly vs. day-of-the-week accident frequency.
- **Bar plots and pivot tables** to show the number of accidents per state.
- **Boxplots** to visualize how severity varies under different weather conditions.
- **Seasonal decomposition charts** to highlight recurring patterns.
- **Geospatial heatmaps** to identify accident hotspots across states and cities.
- **Bubble maps** highlighting the most dangerous cities.
- **Interactive Dashboard**

A. Line charts for yearly, monthly, and daily accident trends

These plots analyze accident trends for different time scales. The yearly trend plot combines a bar chart and line chart to show accident counts and their progression over the years, shows a rise from 2016 to 2022, followed by a drop in 2023. The monthly trend plot displays seasonal variations, with an increase during winter months such as December, may be related to weather conditions or holiday traffic. Lastly, the daily trend shows smooth trends and anomalies while highlighting the daily variations in accident numbers. Together, these visualizations offer a general view of temporal accident trends, long-term, seasonal, and short-term accident dynamics.

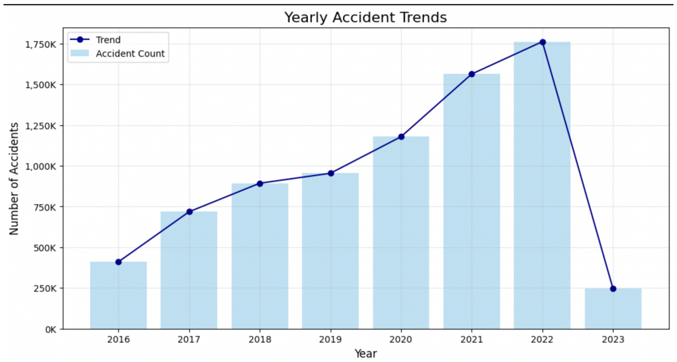

Fig. 1. Yearly Accident Trends

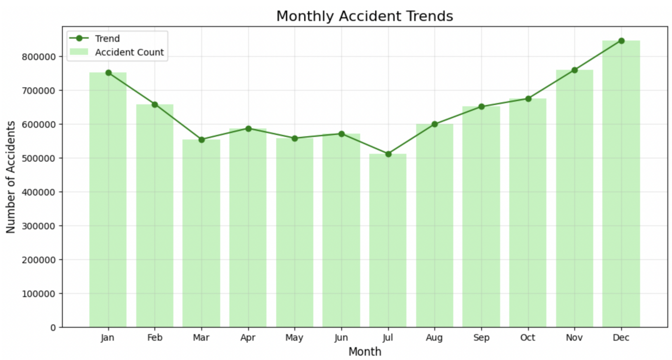

Fig. 2. Monthly Accident Trends

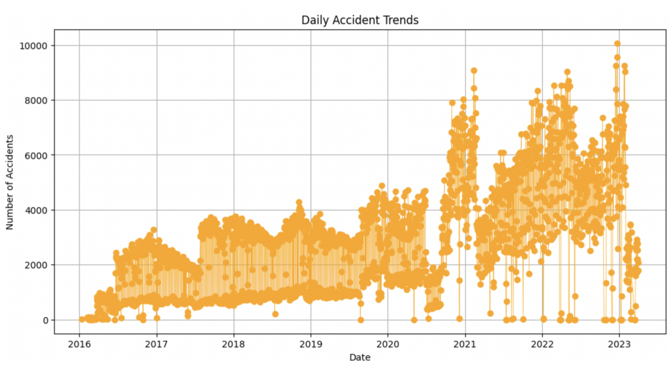

Fig. 3. Daily Accident Trends

B. Heatmaps for hourly vs. day-of-the-week accident frequency

The heatmap visualizes accident frequency as a percentage, distributed for different hours of the day and days of the week. It shows that accidents are most frequent during weekday rush hours, particularly between 7 AM and 9 AM and again between 4 PM and 6 PM, corresponds to peak commuting times. Tuesday and Wednesday mornings, as well as Friday evenings, stand out with higher accident frequencies. In contrast, weekends show relatively lower accident activity during rush hours but higher activity in the late morning and early afternoon, showing a shift in driving behavior.

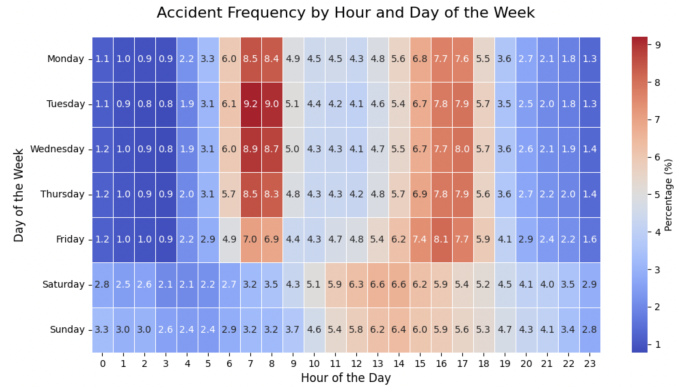

Fig. 4. Accident Frequency by Hour and Day of the Week

C. Bar plots and pivot tables to show the number of accidents per state

The visualizations shows the top 10 states with the highest number of accidents and the bottom 10 states with the fewest accidents. In the first plot, California (CA) leads, contributing 22.5% of the total accidents, followed by Florida (FL) and Texas (TX). This shows us that highly populated states with heavy traffic are more prone to accidents.

In contrast, the second plot shows the least accident states, with South Dakota (SD), Vermont (VT), and Maine (ME) at the bottom. These states have lower populations and reduced traffic congestion, which likely contributes to their fewer accidents. In order to develop specific policy measures and allocate resources to enhance road safety, the visualizations highlight the regional variations in accident patterns.

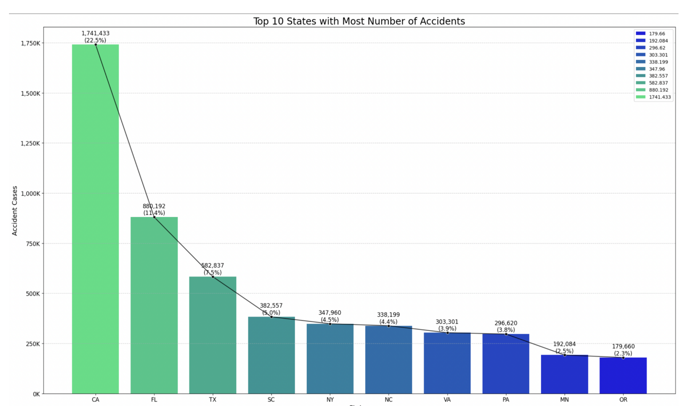

Fig. 5. Top 10 States with Most Number of Accidents

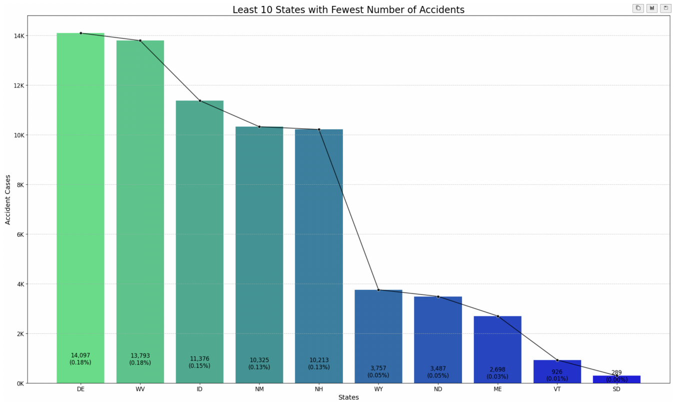

Fig. 6. Least 10 States with Fewest Number of Accidents

D. Boxplots to visualize how severity varies under different weather conditions

The boxplot visualizes the variation in accident severity under different weather conditions. The severity is measured on a scale of 1 to 4, with 4 being the most severe. The plot indicates that most weather conditions, including ”Light Rain,” ”Mostly Cloudy,” and ”Overcast,” tend to cluster around a severity level of 2 or 3. However, there is variability, as seen by the presence of outliers at higher severity levels for certain weather types.

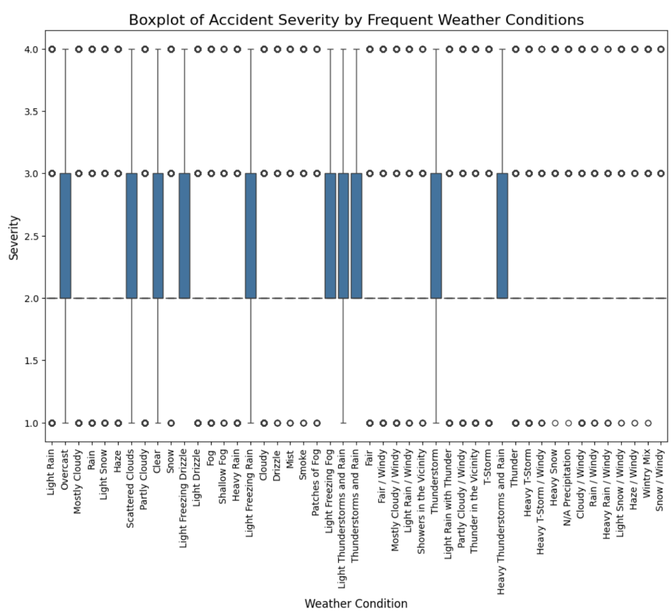

Fig. 7. Boxplot of Accident Severity by Frequent Weather Conditions

The bar chart below shows the top 10 weather conditions associated with the highest number of accident cases. ”Fair” weather conditions account for the largest proportion, representing 33.90% of the total cases, followed by ”Mostly Cloudy” (13.45%), ”Cloudy” (10.82%). The chart demonstrates that even under favorable weather conditions, such as ”Fair” or ”Clear,” accidents can occur frequently. This highlights the importance of factors beyond weather, such as human behavior and road conditions, in contributing to accidents. Adverse weather conditions like ”Light Rain,” ”Overcast,” and ”Fog” are also contributors, though they occur less frequently.

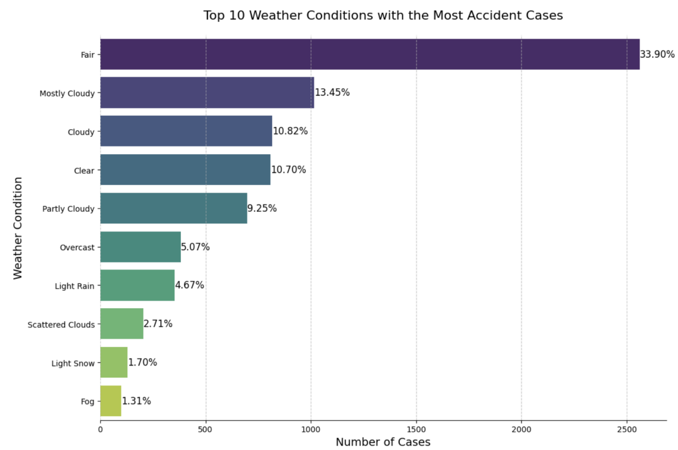

Fig. 8. Top 10 Weather Conditions with the Most Accident Cases

E. Seasonal decomposition charts to highlight recurring patterns

The original series shows the monthly fluctuations in accident numbers over time, while the trend component highlights the overall upward trajectory of accidents, possibly indicating increasing traffic volumes. The seasonal component uncovers patterns in accidents, with periodic up and down moves indicating seasonal variations, like increased accidents in specific months due to weather or holiday traffic. The residual component captures irregular variations, showing anomalies or unexpected fluctuations not explained by the trend or seasonality.

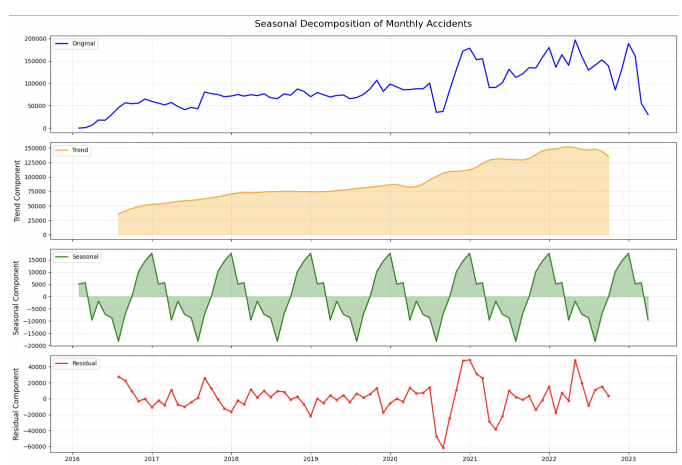

Fig. 9. Seasonal Decomposition of Monthly Accidents

F. Geospatial heatmaps to identify accident hotspots across states and cities

The geospatial heatmap provides a visual representation of accident density across the United States using a hexagonal binning approach. Areas with higher accident frequencies, such as California, Texas, and Florida, are highlighted in darker shades, stand out as points where accidents occur frequently.

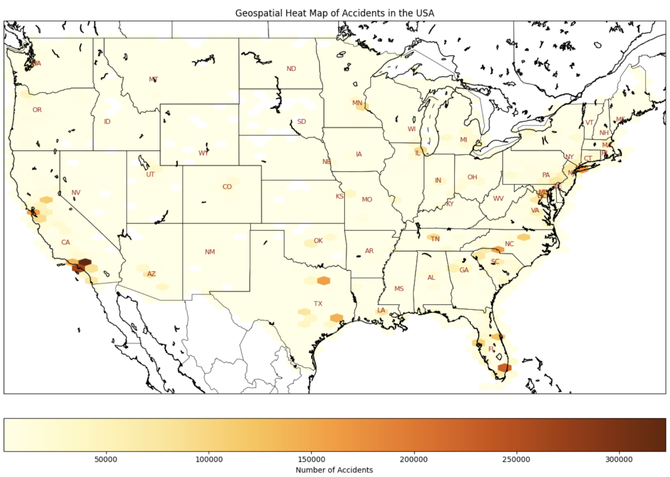

Fig. 10. Geospatial Heat Map of Accidents in the USA

G. Bubble maps highlighting the 50 most dangerous cities

The bubble plot shows the most dangerous cities in the United States based on accident frequency. Larger and darker bubbles represent cities with a higher number of accidents. Cities like Los Angeles, Miami, Houston, and Dallas stand out with higher accident counts. Color gradients further emphasize the accident volume, shades of red indicate a higher number of accidents.

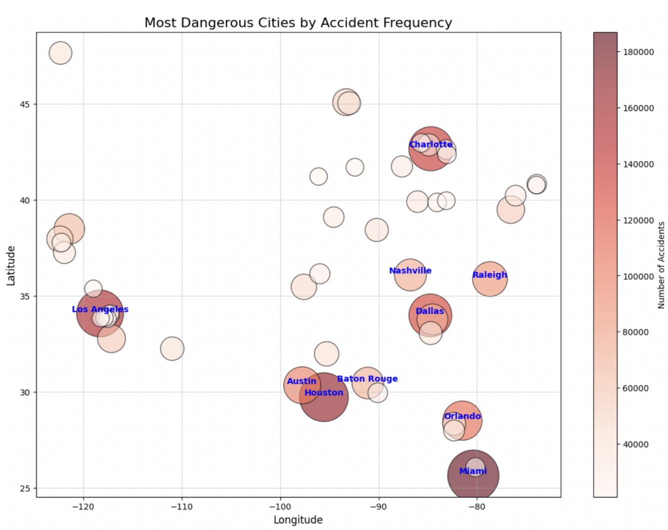

Fig. 11. Most Dangerous Cities by Accident Frequency

H. Interactive Dashboard

The interactive dashboard is divided into four main tabs: Overview, By State, By Time, and Severity Analysis.

The "Overview" tab provides a temporal analysis of accidents for selected years.

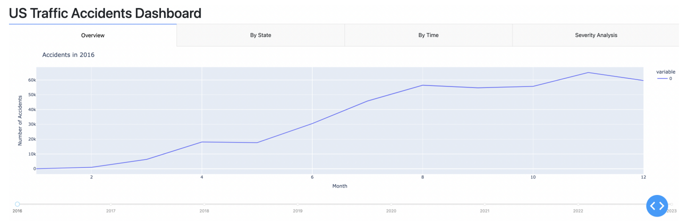

Fig. 12. Overview

The "By State" tab enables filtering of accidents by state, presents monthly accident distributions for selected locations.

Fig. 13. By State

The "By Time" tab includes a heatmap to explore the relationship between the day of the week and hour of the day.

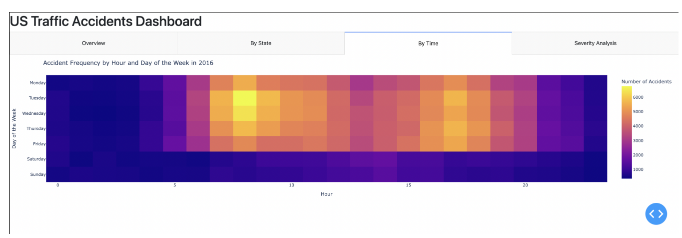

Fig. 14. By Time

Lastly, the "Severity Analysis" tab presents a boxplot showing the distribution of accident severities across different weather conditions.

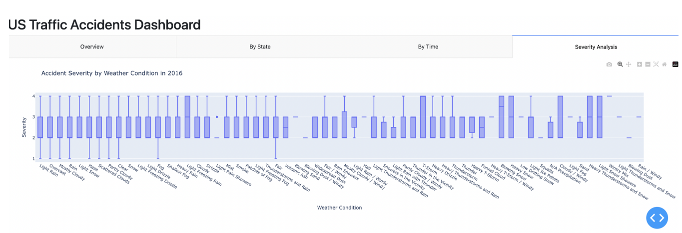

Fig. 15. Severity Analysis

IV. CONCLUSION

In conclusion, this project provides an analysis of traffic accident data in the United States, using visualizations and an interactive dashboard to obtain useful information. The project displays important trends using a variety of charts and maps, such as accident frequency by state, time, and weather conditions, as well as trends over years, months, and hours of the day. The interactive dashboard provides a simple interface for dynamic data exploration and the identification of high-risk states, cities, and conditions contributing to accidents. The findings from this analysis can be utilized by policymakers, urban planners, and transportation authorities to design targeted interventions and improve road safety. This project shows how methods based on data may be used to address complex social issues like traffic accidents by combining data preprocessing, advanced visualizations, and an interactive interface.

REFERENCES

[1] https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents