CSE564: Visualization

Final project proposal

VISUAL ANALYTICS ON TRAFFIC ACCIDENT TRENDS IN NEW YORK

Submitted By:

Group 27

Goutham Kalla (116385130)

Shreyas Reddy Gaddampally (116439954)

BACKGROUND AND INSPIRATION

The inspiration for us behind selecting this specific data is the current ongoing traffic safety crisis which has been constantly in discussion. Factors such as distracted driving, autonomous vehicle integration, post-pandemic travel patterns, and increasing extreme weather events leading to hazardous road conditions and ultimately traffic accidents. What intrigued us was whether these accident patterns existed even before the pandemic period (2018 – 2021) or similar patterns have been observed before. Excellent Data Visualizations can help us infer and find valuable patterns and relationships.

As of 2025, New York emerged as the state with the fourth highest traffic fatality rate (1.12 per 100 million vehicle miles traveled) followed by Florida (1.41), South Carolina (1.73) and Mississippi (1.90). Similar patterns have emerged in the last 5 years with northeastern states consistently ranking among the most accident-prone regions. There has been extensive research on the underlying factors and patterns behind this alarming rate of traffic incidents in the state. Data collected by the NHTSA and various traffic monitoring systems has been widely used in drawing comparisons of scale.

The data for 2024 and 2025 has not been completely gathered as 2025 is the current ongoing year, the current dataset which ranges from 2016 to 2023 (7 years) provides a comprehensive overview for all regions of New York state. This is one of the most reliable available datasets from traffic monitoring systems providing authentic and detailed trends of how traffic accidents prevail over the state with 49 different feature parameters encapsulating all relevant scenarios (weather conditions, road features, time of day, visibility, traffic conditions and more).

This dataset has been widely used by transportation departments, insurance companies, urban planners, researchers and the Government to scrutinize, identify potential factors for the continuing high accident rates in the state and the ways to reduce/mitigate them through underlying trends and continuous improvement. Especially in traffic safety circles, this dataset is widely used for creating visualizations and inferring valuable results such as accident rates by county, by weather condition, by time of day, and by intersection type.

PROBLEM STATEMENT

Being one of the most densely populated states with significant traffic volume, New York is constantly under attention and scrutiny from all concerned parties about the underlying traffic safety conditions and their repercussions. The accident trends in New York and their visualization can bring in important information and can be beneficial for transportation planners, safety researchers, and policy makers in reaching valid, concrete conclusions and designing policies to improve road safety conditions. **Traffic analysts can identify which road features contribute most to accident frequency**

and severity. Safety researchers can identify underlying cumulative trends of New York accident patterns over a period of 7 years. Urban planners can make important decisions about infrastructure improvements based on accident hotspots, environmental conditions, and temporal patterns. **Weather impacts and visibility conditions can be monitored using attributes such as Temperature, Humidity, and Weather Condition for extensive safety research and emergency response planning. Proper planning, analysis and visualization of the data are essential to achieve the following objectives:**

- How has the **accident frequency and severity changed over time** in New York from 2016 to 2023?
- Which counties and cities in the state of New York consistently had the **highest and lowest accident rates** throughout the years?
- On which days of the week and times of day do accidents most frequently occur, and is there a pattern to this distribution?
- Is there any **underlying pattern or correlation** between weather conditions (temperature bins, humidity bins) and accident frequency or severity?
- How do accident patterns differ between daytime and nighttime (Sunrise_Sunset parameter) and what factors might explain these differences?
- Do we get any insights from the **road feature analysis**, particularly related to junctions, crossings, and traffic signals in determining accident likelihood?
- Which year has the **highest and lowest accident severity ratings** for the state of New York?
- Which streets and zip codes have the highest concentration of accidents, and what common features might they share?
- How did the COVID-19 pandemic period impact traffic accident patterns in New York compared to pre-pandemic levels?
- What **weather conditions** are most strongly associated with higher accident rates and greater severity?
- Is there a **correlation between visibility conditions** and accident severity that could inform weather-related travel advisories?
- How do different road features such as **junctions, traffic signals, and railway crossings** influence accident frequency and severity?

PROBLEM APPROACH AND ANALYSIS:

For this project, we have devised a comprehensive Visual Analytics Dashboard with the following data visualizations clearly explained as below:

- **Choropleth map:** Accident frequency, severity levels, and accident density rates can be mapped based on counties and zip codes in New York to visualize geographical hotspots and safety disparities.
- **Bar chart:** Accident counts by severity level, road features (junctions, traffic signals, crossings), and time of day can be compared for specific years or averaged over the entire period. Distribution of accidents across different weather conditions can also be visualized using bar charts.
- **Line chart:** Trend visualization of accident frequency, severity ratings, and weather-related variables over the years for specific counties or aggregated for the entire state. Line charts can

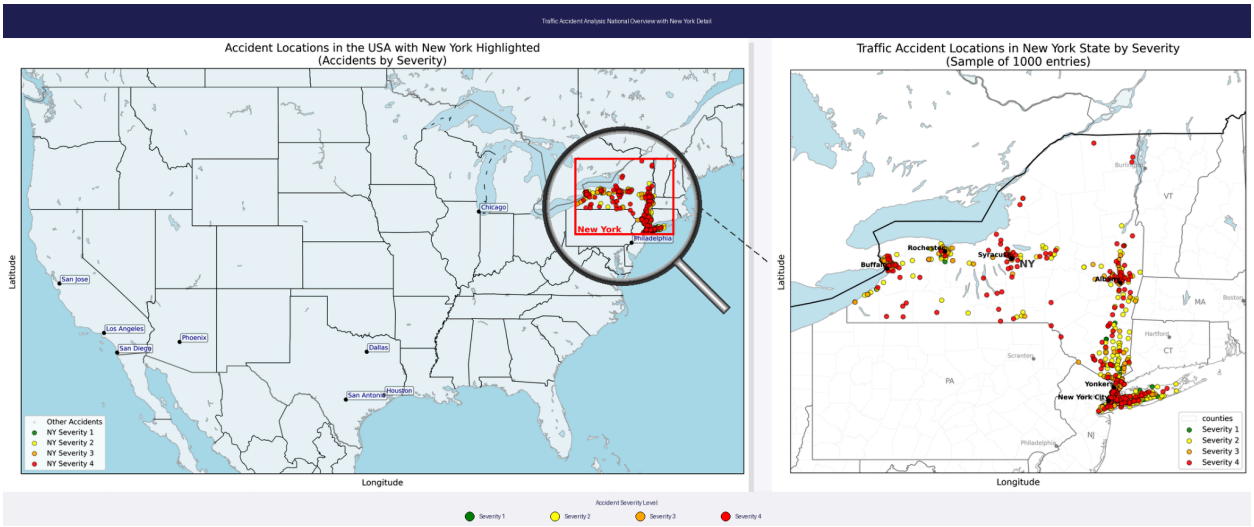
be plotted for attributes which clearly depict the temporal patterns through the seven-year period or can be aggregated by county.

- **Scatter Plot:** Investigating the relationship between variables such as Temperature and Accident Frequency, Humidity and Severity, Visibility and Accident Rate, etc. Scatter plot is the best visualization for this purpose as it helps in capturing valuable correlations between environmental factors and accident metrics.
- **Histogram:** Distribution of attributes like Severity, Temperature(F), Humidity(%), Visibility(mi), and Distance(mi) to understand their underlying distributions and patterns. Histograms are the most optimal visualizations to identify frequency distributions of these key variables.
- **Heatmap:** Time-based analysis showing accident frequency by hour of day and day of week to identify high-risk periods. Heatmaps can clearly display temporal patterns that might otherwise be difficult to discern in traditional charts.
- **Parallel Coordinates Plot:** PCP allows you to plot multiple variables along different vertical axes, with each axis representing one attribute (e.g., Temperature, Humidity, Visibility, Wind Speed, Precipitation, etc.). Each data point is then represented by a line connecting the values of the attributes across the axes. By examining the patterns in the plot, we can also identify clusters of accidents that exhibit similar environmental conditions, which can help us in understanding the complex interplay between different weather variables in determining accident likelihood.
- **Radar Chart:** Comparing accident profiles across different regions of New York by plotting multiple variables (severity, frequency, environmental factors) on different axes radiating from a center point. This allows for quick comparison of the overall accident characteristics between different counties or cities.
- **Pie Charts:** Breaking down accidents by categorical variables such as Sunrise_Sunset, Weather_Condition, or the presence of specific road features. Pie charts offer a clear visual representation of the proportional distribution of accidents across these categorical variables.
- **Box Plots:** Analyzing the distribution of numerical attributes like Temperature, Humidity, and Visibility across different severity levels. Box plots can reveal whether certain environmental conditions are associated with more severe accidents by showing the median, quartiles, and potential outliers.

Dataset:

Source: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

The **National Highway Traffic Safety Administration (NHTSA)** and the United States Department of Transportation provide comprehensive data on traffic accidents across the nation through various monitoring systems and reporting platforms. This particular dataset, compiled by researchers at Sobhan Moosavi and made available on **Kaggle**, provides a comprehensive collection of traffic accidents covering 49 states of the United States from February 2016 to March 2023. The full dataset contains approximately 7.7 million records, while our analysis focuses on a representative sample of 1,000 entries from **New York state**. The dataset includes over 45 different attributes covering all required statistics that explain traffic safety patterns, road conditions, weather impacts, and temporal distributions of accidents across the country. Given below is the detailed overview of the parameters:



Data Structure

1. **Accident Identifiers**: Unique identifiers for each accident record
2. **Temporal Features**: Start and end times, duration of the accident
3. **Geospatial Information**: Latitude, longitude, street addresses, cities, counties, states, and zip codes
4. **Environmental Conditions**: Weather, visibility, temperature, wind conditions, etc.
5. **Road Features**: Proximity to junctions, traffic signals, crossing, etc.
6. **Severity Indicators**: Rated on a scale of 1 (least severe) to 4 (most severe)

Key Attributes:

- **ID**: A unique identifier assigned to each accident record.
- **Severity**: A numerical value on a scale of 1 to 4 that indicates the impact level of the accident (1 being the least severe, 4 being the most severe).
- **Description**: A natural language description of the accident.
- **Street**: The street name where the accident occurred.
- **City**: The city where the accident occurred.
- **County**: The county where the accident occurred.
- **Zipcode**: The zipcode of the accident location.
- **Crossing**, **Junction**, **Traffic_Signal**: Binary indicators of specific road characteristics
- **Temperature(F)**: The temperature in Fahrenheit during the accident.
- **Humidity(%)**: The humidity percentage during the accident.
- **Pressure(in)**: The air pressure in inches during the accident.
- **Visibility(mi)**: The visibility distance in miles during the accident.
- **Wind_Direction**: The wind direction during the accident.
- **Wind_Speed(mph)**: The wind speed in miles per hour during the accident.
- **Precipitation(in)**: The precipitation amount in inches during the accident.
- **Weather_Condition**: The weather condition during the accident (rain, snow, etc.).

- **Sunrise_Sunset**: A categorical value indicating if the accident occurred during daytime or nighttime.
- **Day, Month, Year**: The day, month, year when the accident occurred (extracted from Start_Time).
- **Bump**: A categorical indicating presence of a speed bump near the accident.
- **Crossing**: A categorical indicating presence of a crossing near the accident.
- **Junction**: A categorical indicating presence of a junction near the accident.
- **Traffic_Signal**: A categorical indicating presence of a traffic signal near the accident.