# College of Engineering
# Northeastern University

## Project Report
## Group 8
### US ACCIDENT: 2016 - 2023 Analysis
DAMG 7370 _ Designing Advanced Data Architectures for Business Intelligence

## Team Contribution

| NUID | Name | Contribution Summary |
|---|---|---|
| 001543820 | Bolatito Akinyemi | - AWS S3 buckets creation<br>- Data integration/upload to AWS S3<br>- Power BI: Top 10 Areas Analysis<br>- Report Preparation: Project Architecture and Tools, Data Staging and Processed Zone |
| 002056525 | Prosper Kofi Legbedze | - Data cleaning and processing with AWS Glue<br>- Power BI: Severity Analysis<br>- Report Preparation: Data Processing, Data Warehousing & References<br>- Final Report Review and Editing |
| 002435652 | Januarius Kodzo Avorgbedor | - Setting up a data warehouse in Redshift<br>- Power BI and Redshift connection<br>- Power BI: Dashboard design<br>- Report Preparation: Project Background, Justification for Dataset Selection, Project Scope and Objective |
| 002874080 | Emmanuel Adebiyi | - Power BI: Road Condition Report<br>- Power BI: Weather Elements and Condition Analysis<br>- Power BI: Periodic Analysis<br>- Report Preparation: Data Analysis, Key Insights & Recommendations, and Conclusion |

## Project Data Link

- Dataset: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data
- Project Power BI Link: Damg7370 Final Project Dashboard.pbix
- https://northeastern-my.sharepoint.com/:u:/g/personal/avorgbedor_j_northeastern_edu/EcKaLn05HoFLhoMOt0rlj_IBREJv9Ze7ovK7UXSZu23kwg?e=HL0UVP
- https://github.com/Januauariusavorgbedor1/Damg_Final_Project.git

## Contents



# 1. Objective

The objective of this project is to analyze accident data to identify key patterns, risk factors, and trends that contribute to road accidents. By utilizing data analytics and visualization techniques, the project aims to uncover insights related to accident frequency, severity, location, and contributing factors such as weather conditions, time of day, and road infrastructure. These findings will help inform data-driven decision-making for policymakers, traffic management authorities, and public safety organizations to implement effective measures that enhance road safety, reduce accident occurrences, and improve emergency response strategies.

# 2. Scope

This project aims to analyze accident data to identify patterns and contributing factors affecting road safety. By examining variables such as time, location, weather conditions, and driver behavior, the study seeks to provide insights that enhance traffic management and public safety policies. The findings will support data-driven decision-making to reduce accident rates, improve infrastructure planning, and optimize emergency response strategies.

# 3. Architecture

This structured architecture allows for the efficient handling of the data. It outlines the efficient collection, processing, analysis, and visualization of US accident data, providing valuable insights for stakeholders and decision-makers.

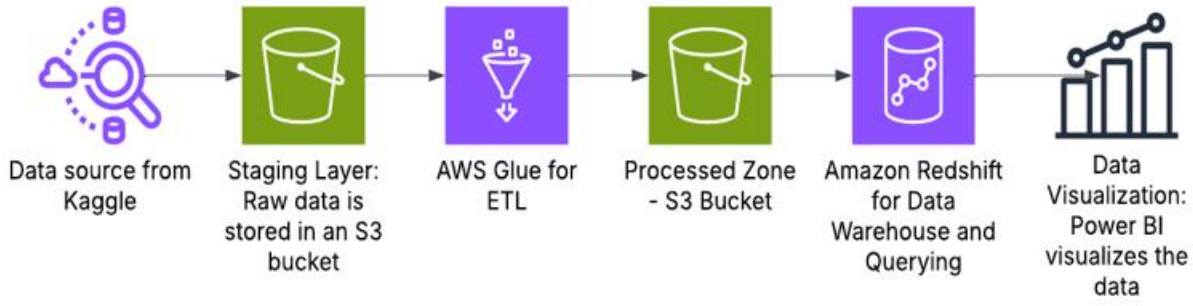


Figure 1: Project Architecture

a) **Data Source & Extraction – Kaggle**
   - The primary data source for the project is Kaggle, which hosts datasets related to US accidents.
   - The project utilizes Kaggle's data available for extraction to gather necessary information for analysis.

b) **Data Staging Layer – AWS S3**:
   - Raw data is initially stored in an Amazon S3 bucket, serving as the staging layer.
   - This step allows for secure and scalable storage of the unprocessed data before further processing.

c) **Data Orchestration – AWS Glue**:
   - AWS Glue is used for Extract, Transform, Load (ETL) processes.
   - It facilitates data transformation and integrates various data sources, streamlining the preparation of data for analysis.

d) **Processed Zone – S3 Bucket**:
   - After ETL processes, the processed data is stored in a designated S3 bucket for further access.
   - This layer ensures that clean and structured data is readily available for querying.

e) **Data Storage – Amazon Redshift**:
   - Amazon Redshift serves as the data warehouse solution, enabling efficient querying and data management.
   - The processed data is housed here, structured for optimized performance during analysis.

f) **Data Visualization – Power BI**:
   - Power BI is leveraged for data visualization, allowing the creation of interactive dashboards and reports.
   - This tool enables stakeholders to visualize insights and findings effectively, facilitating better decision-making.

# 4. Methodologies

## Data Preprocessing

In the data preprocessing stage for the US Accident Report, we leveraged the capabilities of the AWS Glue framework alongside Apache Spark to ensure robust data cleaning and transformation. This approach enabled us to effectively address complex data issues and prepare the dataset for further analysis and storage.

### a. Data Loading and Initial Assessment

Initially, two S3 buckets were created to store the raw data and the transformed or cleaned data. An IAM role was established, and required permissions such as AmazonS3FullAccess, AWSGlueConsoleFullAccess, IAMFullAccess, and CloudWatchLogsFullAccess were assigned to support access to S3 buckets and data transformation activities within the Glue environment. This facilitated seamless integration and allowed for the definition and management of the metadata of our datasets. Upon examining the dataset, several areas requiring significant cleaning and adjustment due to the presence of missing, inconsistent, or invalid values were identified. The transformation and cleaning activities were conducted using AWS Glue Notebook.

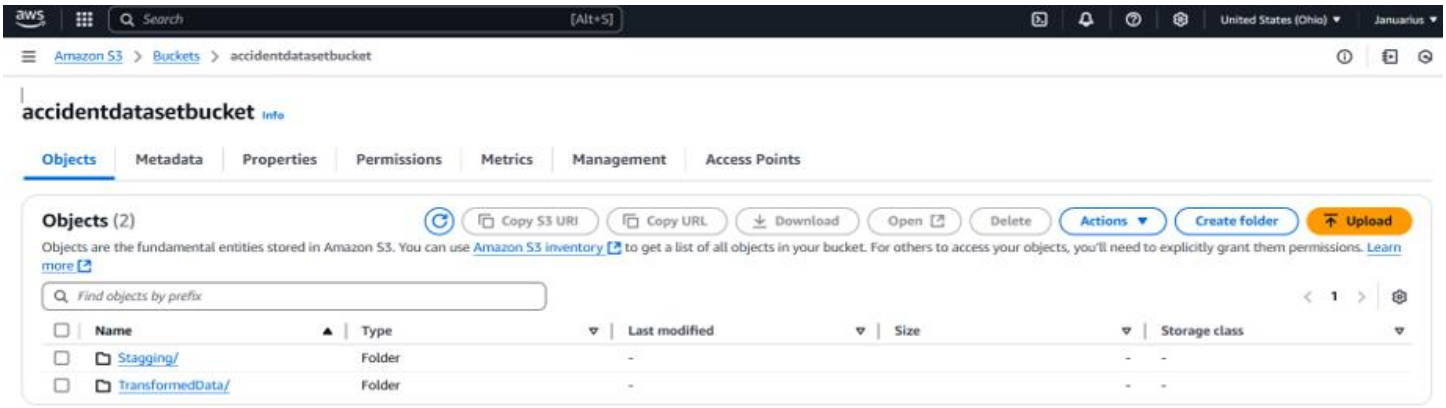


Figure 2: S3 Bucket Creations

### b. Data Cleaning Activities

• Removal of Columns with High NaN Values:

Each column was assessed for missing values utilizing Glue's dynamic frames, applying the logic of removing columns with 20% or more NaN entries. Key columns such as End_Lat, End_Lng, Precipitation(in), and Wind_Chill(F) were excluded from further analysis because their high proportion of missing values would compromise the quality of insights derived.

• Handling Duplicate Data:

Spark SQL capabilities were employed to check for duplicate records using the primary ID field. This step was crucial to ensuring the integrity of the dataset, allowing only unique records to flow through the preprocessing pipeline.

• Invalid and Inconsistent Values:

Scrutinization of the data for invalid entries was conducted, including checks against relevant sources such as the National Centers for Environmental Information and the Federal Highway Administration to confirm the validity and consistency of the data with respect to US weather conditions.

Inconsistent Temperature

Using data filtering techniques, 205 entries with unrealistic temperature readings outside the acceptable range of -30°F to 130°F were identified and removed.

Inconsistent Visibility

Similarly, rows with Visibility(mi) values of 0 or excessively high readings (greater than 100 mi) were corrected; specifically, 7,694 entries were replaced with NaN to avoid skewing our analysis.

Extreme Wind Speed Data

Instances where Wind_Speed (mph) exceeded 100 mph, an unlikely value in real-world conditions, were removed, resulting in the exclusion of 139 entries.

Pressure Adjustments

Humidity data was clipped to a reasonable range of 10% to 90%, adjusting 1,096,357 entries to fall within this interval.

Pressure Adjustments

Pressure values below 24 inHg were deemed unrealistic and treated as missing data, leading to 380,986 entries being replaced with NaN.

Extreme Distances

Interquartile Range (IQR), a crucial statistical measure used to identify and mitigate outliers within datasets, was used to clean extreme distance records. By calculating the IQR, 963,606 entries exceeding 1.16 miles were capped, effectively limiting the impact of extreme values on our dataset.

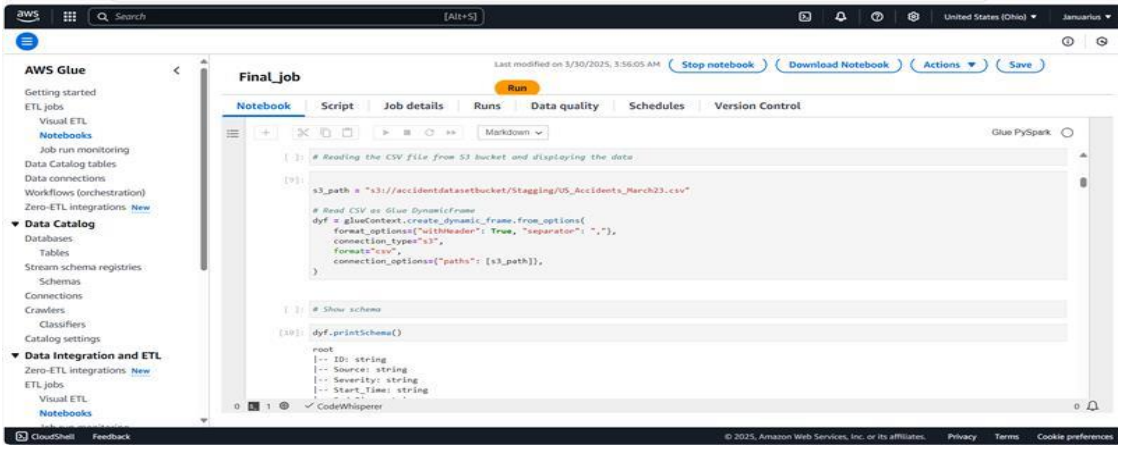


Figure 3: Glue Studio ETL Process

### c. Saving Transformed Data

After implementing these cleaning processes, the Glue context was utilized to convert the dynamic frame back to a Spark DataFrame and place the cleaned data back in an Amazon S3 bucket in a format suitable for downstream processing.

### d. Transfer to Amazon Redshift for Warehousing

Following the storage of the cleaned dataset in S3, the data was prepared for loading into Amazon Redshift for further analysis and warehousing. Using AWS Glue's ETL capabilities, a connection was established to our Redshift cluster. A job was configured to transfer the cleaned data from S3 into Redshift tables, ensuring structured storage for efficient querying and reporting.

By systematically implementing these cleaning processes, we ensured that the dataset was refined and ready for subsequent analysis, significantly improving the quality and reliability of our insights on US accidents. This thorough preprocessing groundwork allows for accurate reporting and enhanced decision-making moving forward, laying a solid foundation for advanced analytical tasks and visualization with Microsoft Power BI.

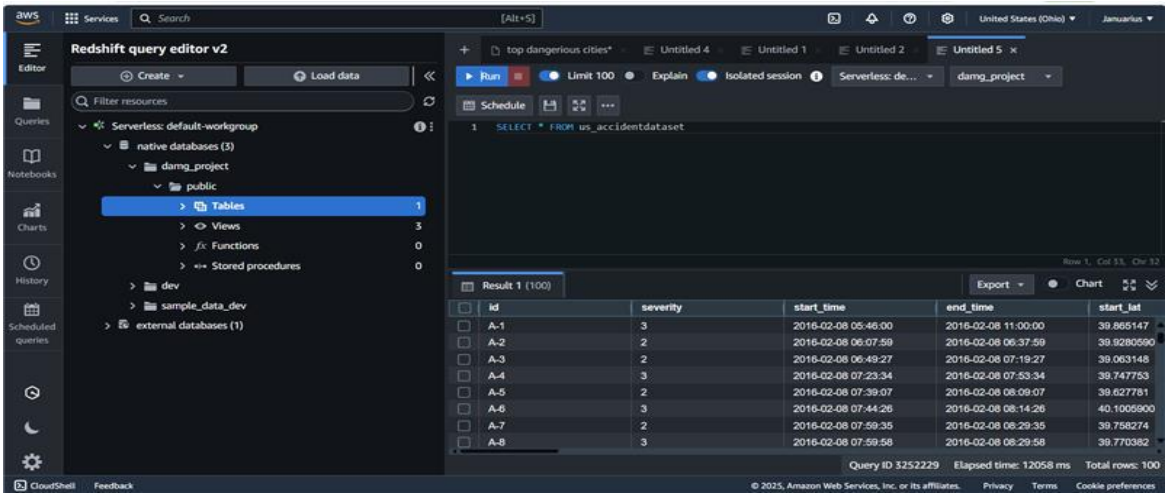


Figure 4: Redshift Data warehousing

## Analysis Techniques

A variety of advanced analytic techniques were employed to extract meaningful insights from the data. The methodologies focused on multiple key areas to enhance understanding of accident patterns and driving factors.

### 1. Accident Frequency Analysis

One primary focus was analyzing the frequency of accidents across various dimensions, including time (month, year), location (state, city), and environmental conditions (weather, road type). We calculated metrics such as:

- **Accident Counts**: Total number of accidents per geographic location (States, Cities, Counties).
- **Time Series Analysis**: Trends in accidents over time to identify peak periods.

This initial analysis provided a clear overview of where and when accidents are most likely to occur, informing prevention strategies.

### 2. Geospatial Analysis

Utilizing geospatial techniques, we examined the geographic distribution of accidents. Mapping accident data against road infrastructure and demographic information helped reveal:

- High-risk areas that may require targeted interventions.
- Conditions leading to increased accident rates in specific locations.

### 3. Severity Analysis

We analyzed the severity of accidents using metrics such as injury counts, fatality rates, and property damage. By categorizing accidents by severity:

- We identified patterns that contributed to higher severity outcomes, which can inform policy-making and safety measures.

### 4. Data Visualization

To present the findings from our analyses, an interactive dashboarding tool, Power BI, was utilized. This allowed us to create:

- **Dynamic Dashboards**: Visual representations of accidents by location, time, and severity, facilitating easy interpretation.
- **Drill-Through Feature:** By using Power BI's drill-through functionality, stakeholders could click on specific data points (e.g., individual state accident counts) to access detailed accident metrics for that

state. This feature enables users to explore deeper insights without cluttering the main dashboard, enhancing the analytical experience.

- **Report Generation**: Automated reporting features to regularly update stakeholders with key findings and insights.

These visualizations were essential for making complex data comprehensible and actionable for policymakers and other stakeholders.

# 5. Exploratory Data Analysis

The dataset reveals significant insights into road safety trends between 2016 and 2023. With a total of 8 million recorded accidents, a notable accident rate of 1.78 per mile emphasizes the urgency for targeted interventions. California leads in total accidents, with approximately 1.74 million incidents, followed by Florida and Texas. Furthermore, the data indicates that accidents are predominantly concentrated during the day, highlighting specific times when safety measures are crucial.

Geospatial analysis reveals regional hotspots, particularly in southern and eastern states, which can guide resource allocation. The severity classification indicates that a majority of incidents fall into moderate severity categories, necessitating focused awareness campaigns. Lastly, by integrating additional data sources and advanced analytical techniques, a deeper understanding of accident causes can be gained, leading to more effective road safety strategies in the future.

# 6. Key Metrics and Insights using Power BI

**Road Condition**

The Road Condition Analysis Reports reveal key insights into the condition of roads where accidents occurred. Notably, the presence of crossings and junctions is linked to significant accident rates, with 11.31% and 7.39% of accidents occurring at these locations, suggesting a need for improved signage and safety measures. In contrast, bumps and turning loops showed a negligible impact on accidents. Additionally, traffic signals are associated with 14.8% of accidents, highlighting potential compliance or malfunction issues. These findings inform targeted interventions to enhance road safety effectively.

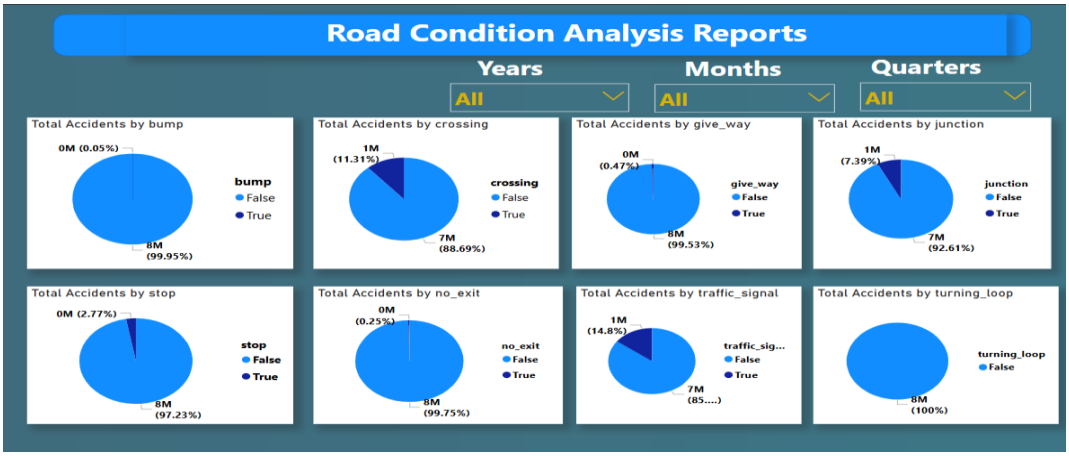


Figure 5: Road Condition Analysis Report

## Weather Elements and Conditions

This presents crucial insights on factors affecting road accidents. The total accidents show a significant correlation with temperature, peaking around 60°F to 80°F. Humidity exhibits a gradually increasing trend in accidents, with more incidents reported at higher humidity levels. Pressure data indicates a notable spike in accidents around 30 inches, suggesting specific environmental conditions may contribute to accident rates. Additionally, the analysis of weather conditions reveals that "Fair" weather is associated with the majority of accidents, highlighting the need for further investigation into the impact of environmental factors on road safety.

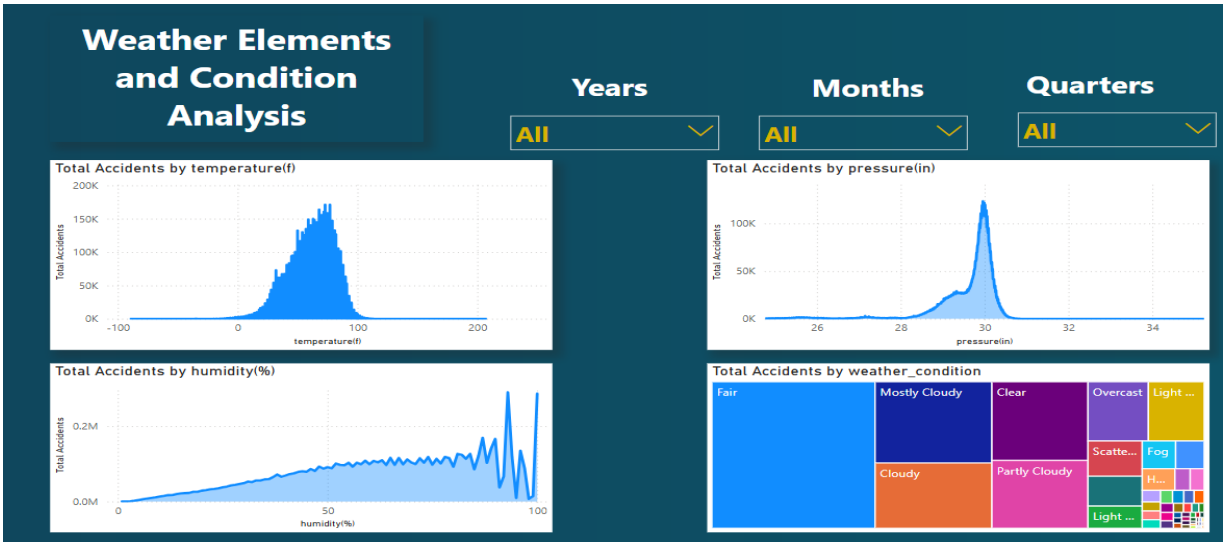


Figure 6: Weather Element and Condition Analysis Report

## Periodic Analysis

This report highlights essential trends in accident data from 2016 to 2023. Total accidents show a gradual increase, peaking in 2022. Monthly analysis reveals heightened accident rates during the latter part of the year, especially from October to December. Furthermore, accidents occur predominantly during the day, with daytime incidents significantly outnumbering those at night. This suggests a need for targeted safety measures during

peak months and increased awareness around night driving conditions. Overall, the data underscores the importance of temporal factors in understanding and mitigating road accidents effectively.

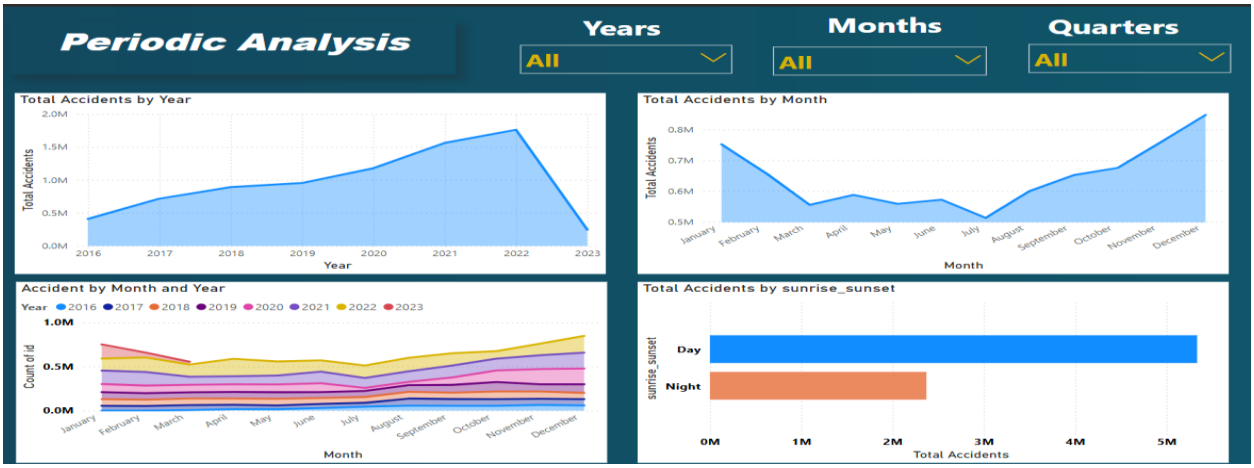


Figure 7: Periodic Analysis Report

## Severity Analysis

The Severity Analysis reveals critical insights into road accidents categorized by severity. A staggering 6 million accidents are classified as severity level 2, accounting for 80% of incidents, with 1 million at severity level 3. The data from 2022 indicates a peak in severity levels, particularly under "Fair" weather conditions, which correlates with the highest accident rates. Additionally, accidents predominantly occur during the day, with markedly fewer incidents at night. Monthly data shows consistent severity levels, highlighting the need for targeted safety measures to address both moderate and severe accidents throughout the year.

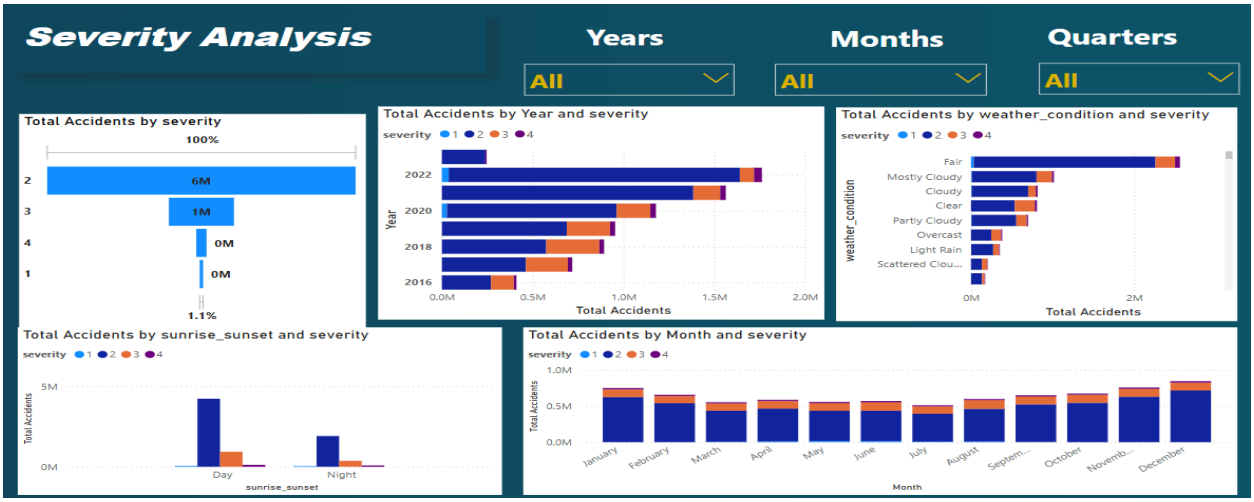


Figure 8: Severity Analysis Report

## Top Ten Geographical Areas

This analysis underscores significant trends in road safety across cities, counties, and states. California leads with the highest number of accidents, approximately 2 million, followed by Florida and Texas. Among cities, Houston and Los Angeles are prominent, with Miami also reporting substantial incidents. At the county level,

Los Angeles and Miami-Dade top the list. Additionally, when examining severity, California and Florida display the most severe accident counts, indicating a need for targeted safety interventions in these high-risk areas. Overall, this data highlights critical locations for focused road safety measures.


# 7. Future Directions

**Enhance Data Quality:** Implementing rigorous quality control measures will be essential to ensure data accuracy, completeness, and reliability. This will address issues related to missing values, outliers, and inconsistencies to build a more robust dataset.

**Incorporate Network Analysis:** By applying network analysis methods, an interactive model can be created among various factors influencing accidents, such as weather conditions, driver behavior, and road infrastructure. This approach will help identify relationships and underlying patterns that contribute to accident occurrences.

**Leverage Natural Language Processing (NLP):** By utilizing NLP techniques to analyze textual data from police reports, social media chatter, and public feedback sentiments can be extracted, thereby helping to identify emerging themes and better understand public perception of road safety and contributing factors to accidents.

# 8. Conclusion:

- The analysis of US accident data from 2016 to 2023 has successfully provided valuable insights into the patterns and dynamics of road safety across the country.
- By leveraging advanced analytical tools and methodologies, we conducted a comprehensive exploration of multifaceted datasets, revealing critical trends in accident frequency, severity, and contributing factors.

- Through meticulous examination of a substantial dataset, we identified actionable recommendations aimed at enhancing road safety.
- These insights are grounded in empirical evidence drawn from the analysis, highlighting the need for targeted interventions in high-accident states and cities, as well as improvements to road infrastructure and driver education.
- The project not only shed light on underlying trends but also emphasized the importance of ongoing monitoring and data-driven strategies.
- By focusing on these insights and recommendations, stakeholders can make informed decisions that contribute to reducing accidents and improving overall traffic safety in the future.

# 9. References


- Zhou, Y., Wang, X., & Zhang, L. (2021). “The Impact of Weather on Road Traffic Accidents: A Machine Learning Approach.”
- https://www.ncdc.noaa.gov/
- https://ops.fhwa.dot.gov/weather/
- https://www.nhtsa.gov/research-data