EDA Summary -

Shervan Shahparnia and Nathan Cohn - Group 1

Our analysis focuses on traffic accidents across the U.S., with Severity (scale 1–4) and Location (latitude, longitude, state) as the target variables. The dataset originally consisted of 7.7 million records but for our analysis we reduced it to 5.39 million records. The data includes accident details, weather conditions, and geographic information.

1. Dataset Summary

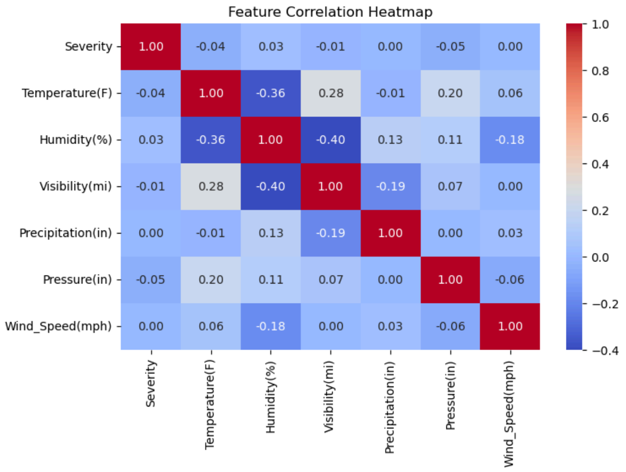
- Target Variables: Severity (int), Location (Start_Lat, Start_Lng, State)
- Features: Start_Time (datetime), Temperature(F) (float), Humidity(%) (float), Visibility(mi) (float), Precipitation(in) (float), Pressure(in) (float), Wind_Speed(mph) (float), Weather_Condition (str)

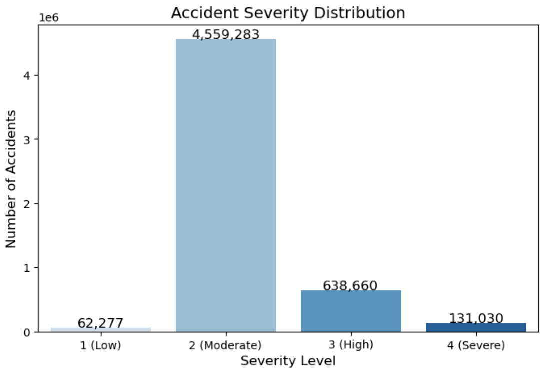
2. Data Cleaning & Handling Missing Values

- Missing Data: Identified in Precipitation, Wind Speed, and some weather-related fields.
- Handling Strategy: Dropped missing values in columns not used for analysis and ensured no loss of important data.
- Duplicates & Inconsistencies: Checked for and removed where necessary to ensure data integrity (however, the dataset was very clean so barely any of this was required).
- Final Dataset Size: 5,391,250 records after data cleaning.

3. Descriptive Statistics & Correlation Analysis

- Mean Severity: 2.16 (Most accidents are moderate in severity)
- Temperature: Ranges from -45°F to 196°F, with a mean of 61.42°F
- Humidity: Ranges from 1% to 100%, with a mean of 65.55%
- Visibility: Mostly 10 miles, but as low as 0 miles in extreme cases
- Weather Impact: Weak correlation (-0.03 to 0.04) between weather factors and severity
- Strongest correlation: Humidity vs. Visibility (-0.41) (High humidity reduces visibility)



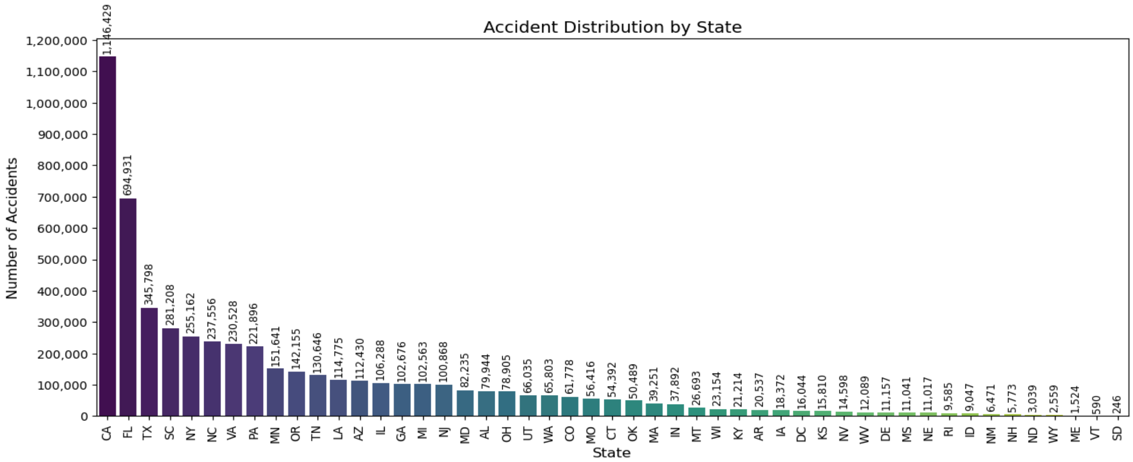


4. Severity Distribution Analysis

- Most accidents fall under Severity Level 2 (~3.9M).
- Severe accidents (Level 4) are significantly less frequent.
- Accident severity does not appear to be heavily influenced by weather conditions.

5. Geographic Accident Distribution

- California, Florida, and Texas have the most accidents.
- Urban areas experience more accidents due to higher traffic density.
- Severity distribution varies by state, suggesting potential infrastructure or policy influences.



Key Insights from EDA & Key Questions for Future Analysis::

- Severity is mostly moderate (Level 2), with fewer severe accidents (Level 4).
- Weather factors (temperature, humidity, precipitation) have weak correlations with severity.
- Humidity significantly affects visibility (-0.41 correlation), which may contribute to accidents in foggy or humid conditions.
- Urban areas experience more accidents, but severity distribution varies by location.

1. Does accident severity differ between urban and rural locations?
   - ○ Hypothesis: Rural accidents tend to be more severe due to higher speed limits and longer emergency response times.
2. How do road network characteristics (speed limits, lane count, intersections) impact severity?
   - ○ Hypothesis: High-speed roads (highways) and areas with fewer traffic controls may have more severe accidents.
3. Are certain states more prone to severe accidents, and what factors contribute to this?
   - ○ Hypothesis: States with extreme weather conditions, high-speed highways, or poor road infrastructure may experience more severe accidents.