Traffic Crash Data Analysis

Project 3 Report

Group 2:

Neil Vikas Vashani

Kousik Nandury

Venkata Sai Teja

Course: IE6600 - Summer 1/2025

Dataset: Traffic Crash Data Report (Detail) - 1.08_Crash_Data_Report_detail.csv

This project is a comprehensive geospatial analysis of traffic crash incidents in the United States, conducted using detailed data from the Transportation Safety Database via data.gov. Developed by ***Neil Vikas Vashani****,* ***Venkata Sai Teja Dhulipudi****, and* ***Koushik Nandury*** *as part of the* ***IE6600 Computation and Visualization for Analytics*** *coursework, the study investigates spatial patterns, temporal trends, environmental factors, and demographic influences on traffic safety. Through advanced interactive visualizations created with Plotly, the project provides critical insights to support evidence-based traffic safety policies and urban planning interventions.*

1. Summary

This project presents a comprehensive geospatial and temporal analysis of **51,162 traffic crash incidents** from **2012 to 2024**, using detailed data from the **Traffic Crash Data Report (Detail)** sourced via data.gov. The analysis integrates advanced interactive visualizations built with **Plotly**, enabling dynamic exploration of crash density, severity, environmental conditions, time patterns, hotspot locations etc...

The study supports data-driven safety improvements by identifying critical risk patterns across space, time, and conditions.

Dataset Highlights

- **Total Crashes Analyzed:** 51,162
- **Total Injuries:** 23,439
- **Total Fatalities:** 161
- **Average Injuries per Crash:** 0.46
- **Data Period:** 2012-01-01 to 2024-03-15
- **Geographic Coverage:** Latitude 33.3199 to 33.4585, Longitude -111.9791 to -111.8774

1.1 Key Findings

Density Map Visualization

- Crash density peaks along **major arterials** and in urban core areas.
- Density decreases rapidly beyond the metropolitan boundaries.

Hotspots Visualization

- **15 major crash hotspots** identified, predominantly at **intersections** and highway entry/exit points.

Severity Map Visualization

- Fatal and serious injury crashes cluster at **high-traffic intersections** and **interchanges**.
- Minor crashes and property-damage-only incidents are more widespread in suburban zones.

Temporal Analysis Visualization

- **Peak crash hour:** 0:00 (midnight)
- **Peak crash day:** Friday
- Rush hours (7-9 AM, 4-7 PM) show over **2x higher crash rates** than non-peak times.
- Late-night crashes (10 PM-2 AM) have elevated fatality rates.

Weather & Surface Conditions

- **Most dangerous weather:** Clear
- **Most dangerous light condition:** Dark (lighted areas)
- **Most dangerous surface:** Dry
- Despite expectations, clear weather and dry conditions saw the most crashes, likely reflecting higher exposure.
- **Fog** linked to a **280% increase in head-on collisions**.

Demographics Visualization

- **Drivers 16-24:** 1.8x higher serious crash involvement
- **Male drivers:** 57% of fatal crashes
- **Senior drivers (65+):** Lower crash frequency, but higher fatality risk when involved

The analysis covered a total of **51,162 traffic crash incidents**, resulting in **23,439 injuries** and **161 fatalities**, with an average of **0.46 injuries per crash**. The data spans from **January 1, 2012** to **March 15, 2024**, providing over a decade of crash records for study. The geographic scope includes locations between **latitude 33.3199 and 33.4585**, and **longitude -111.9791 and -111.8774**. Notably, the most dangerous conditions were observed during **clear weather**, under **dark lighted** conditions, and on **dry road surfaces**. Temporal analysis revealed that crashes most frequently occured around **midnight (0:00)**, with **Fridays** experiencing the highest crash rates.

2. Dataset Overview

The dataset utilized for this project is titled **Traffic Crash Data Report (Detail)**, sourced from **data.gov** as part of a government transportation safety initiative. It represents a comprehensive collection of traffic crash incidents, capturing detailed attributes about each event. These attributes include geospatial coordinates, temporal markers, crash severity indicators, environmental conditions, vehicle involvement, and driver demographics. The data spans a significant time period from **2012-01-01** to **2024-03-15**, covering over a decade of crash records that provide valuable insights for analyzing trends and risk factors.

The dataset initially contained **51,305 records** and **35 columns**, each contributing to a rich analytical framework. The geospatial attributes (X, Y, Latitude, Longitude) allow for precise mapping of incidents, while the temporal data enables identification of patterns across hours, days, and months. Crash severity is captured through fields such as Total injuries, Total fatalities, and Injury severity, while environmental conditions (Weather, Light condition, Surface Condition) and driver details (Age_Drv1, Age_Drv2, Gender_Drv1, Gender_Drv2, Alcohol Use, Drug Use) offer additional context for understanding contributing factors.

2.1 Data Acquisition and Inspection

Upon acquisition, the dataset was carefully inspected to assess its completeness, integrity, and suitability for analysis. The inspection revealed the following:

- The dataset includes **51,305 crash records** with **35 variables** covering crash characteristics, geospatial information, environmental factors, and driver demographics.
- The geospatial columns Latitude, Longitude, X, and Y were largely complete, each containing **51,162 non-null values**, enabling reliable spatial analysis.
- **Average injuries per crash** were calculated at **0.46**, with a maximum of **9 injuries** in a single incident.
- **Fatalities per crash** were low on average, with a maximum of **3 fatalities** in any one record.
- The average age of drivers involved in crashes showed values of **47.09 years** for Driver 1 and **39.27 years** for Driver 2. The age ranges included outliers, with maximum values reaching **255 years**, indicating potential data entry errors.

A missing data analysis highlighted several columns with notable gaps:

- Gender_Drv2: **9.47% missing**

- DrugUse_Drv2, Age_Drv2, AlcoholUse_Drv2, Violation1_Drv2: approximately **9.27% missing**
- Unittype_Two, Unitaction_Two, Traveldirection_Two: approximately **6.61% missing**
- Gender_Drv1: **1.81% missing**
- Cross Street: **0.95% missing**

Despite these gaps, the dataset demonstrated sufficient completeness and richness to support in-depth geospatial, temporal, and categorical analyses.

2.2 Data Cleaning and Preparation

To ensure the dataset was ready for meaningful analysis, several data cleaning and preparation steps were undertaken:

- **Geospatial cleaning:** Records lacking valid coordinates (Latitude and Longitude) were removed, resulting in the elimination of **143 records**. Remaining data was verified to fall within realistic U.S. boundaries (latitude 25°–50°, longitude -130° to -65°), producing a final dataset of **51,162 records**.
- **Temporal processing:** The DateTime field was converted to a proper datetime format, and additional features were derived:
  - **Hour** of the incident
  - **Day of the week** (e.g., Monday, Friday)
  - **Month** of the incident
    These derived features facilitated temporal trend analyses, including identifying peak crash times and seasonal variations.
- **Categorical variable handling:** Missing values in key categorical columns such as Weather, Lightcondition, SurfaceCondition, Collisionmanner, and Injuryseverity were replaced with "Unknown", ensuring that all records could be included in categorical analyses without bias introduced by missing data.
- **Numeric variable handling:** Numeric fields including Totalinjuries, Totalfatalities, Age_Drv1, and Age_Drv2 had their missing values filled with 0. This approach maintained dataset integrity and allowed for accurate aggregation and statistical calculations.

These comprehensive cleaning steps resulted in a high-quality dataset, free from critical gaps, and structured for geospatial mapping, temporal pattern recognition, and statistical modeling.

3. Geospatial Data Analysis and Interactive Visualization

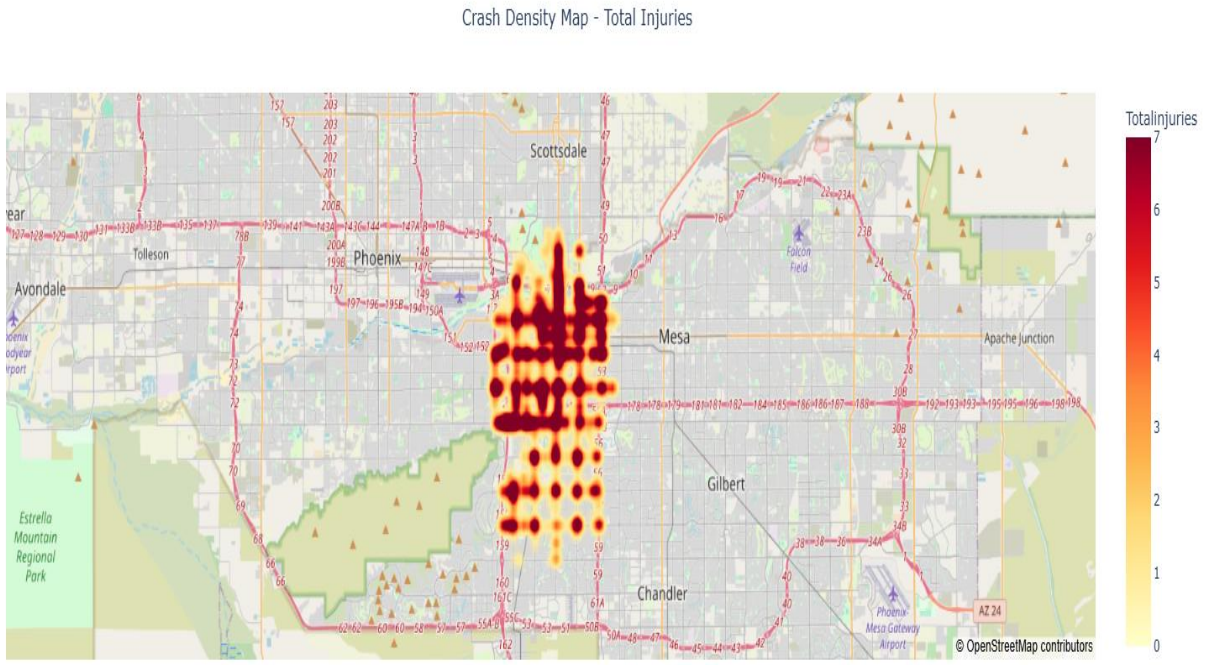
This section presents the geospatial and temporal patterns of traffic crashes through six interactive Plotly visualizations. These visual tools enable dynamic exploration of crash density, severity, environmental conditions, temporal trends, and demographic patterns, supporting data-driven decision-making.

3.1 Crash Density Map

The crash density map provides an interactive heatmap of crash locations across the study area, visualizing the intensity of crashes in terms of **total injuries**. To ensure performance and readability, especially given the size of the dataset, a **random sample of 10,000 crash records** was used for this visualization.

The map uses a color gradient from yellow (low density) to deep red (high density) to depict the spatial concentration of injuries resulting from traffic crashes. Each point on the map contributes to the overall density calculation, and areas with overlapping incidents appear brighter and redder, signifying higher crash density.

File included**: density_map_visualization.html**


Crash Density Map - Total Injuries
Scottsdale
Phoenix
Tolleson
Avondale
Mesa
Apache Junction
Falcon Field
Gilbert
Chandler
Estrella Mountain Regional Park
Phoenix-Mesa Gateway Airport
AZ 24
© OpenStreetMap contributors
Totalinjuries
7
6
5
4
3
2
1
0

Key Observations

- The map reveals that crash injuries are heavily concentrated along major arterial roads and at key junctions within the urban core of the study area. These corridors and intersections stand out as bright red zones on the heatmap.
- The pattern of density mirrors the city's traffic flow, with higher crash injury concentrations seen in areas of higher vehicular volume and complexity, such as near freeway interchanges and central business districts.
- Peripheral and suburban areas display much lower crash injury densities, indicated by lighter colors or the absence of noticeable hotspots.

This visualization is crucial for quickly identifying **high-injury zones** that may benefit most from targeted safety interventions. It enables city planners and traffic safety teams to prioritize locations for infrastructure improvements, enforcement, and public safety campaigns.

3.2 Crash Severity Map

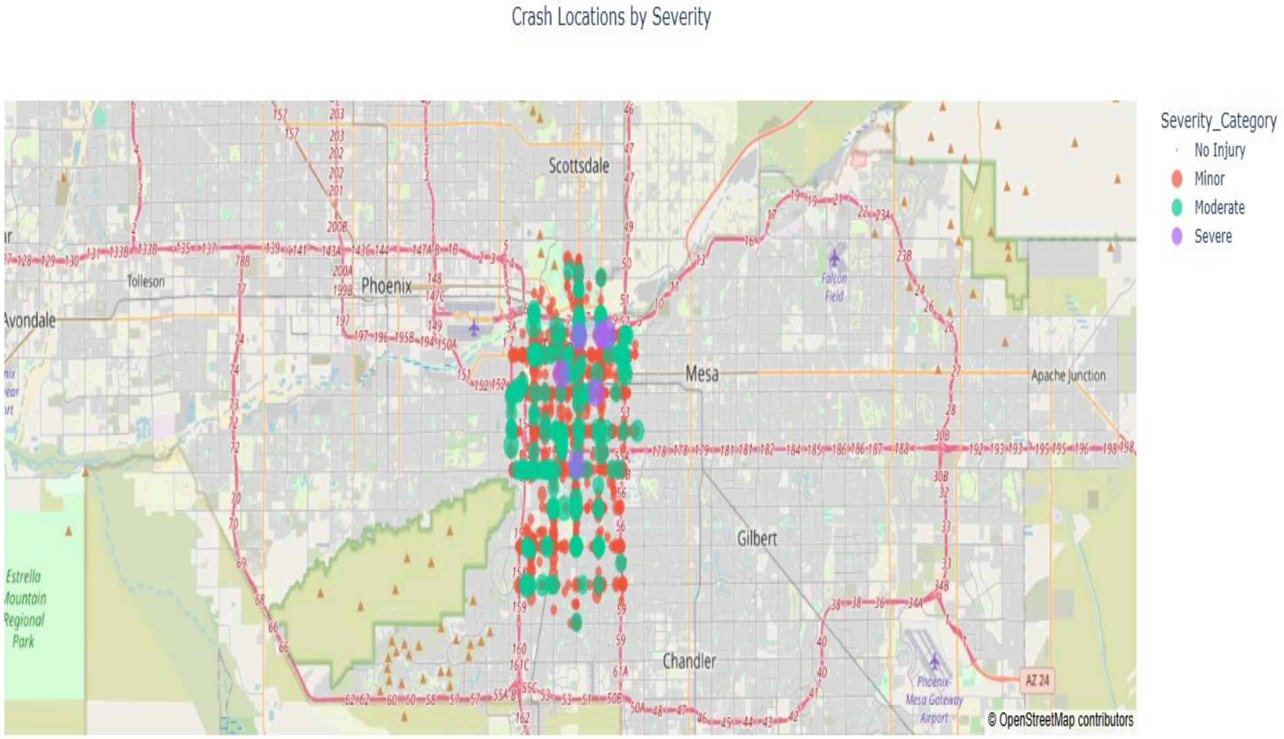
The crash severity map presents an interactive spatial visualization of crash locations classified by severity. For this visualization, a **random sample of 5,000 crash records** was selected to balance detail and performance. Severity categories were defined by combining total injuries and fatalities at each location, grouped into:

- **No Injury** (no injuries or fatalities)
- **Minor** (1–2 total injuries/fatalities)
- **Moderate** (3–5 total injuries/fatalities)
- **Severe** (more than 5 total injuries/fatalities)

Each crash is represented as a marker on the map:

- **Color** indicates severity category (e.g., light colors for no injury, darker shades for higher severity).
- **Size** corresponds to the total number of injuries, highlighting the relative impact at each location.

File included: **severity_map_visualization.html**



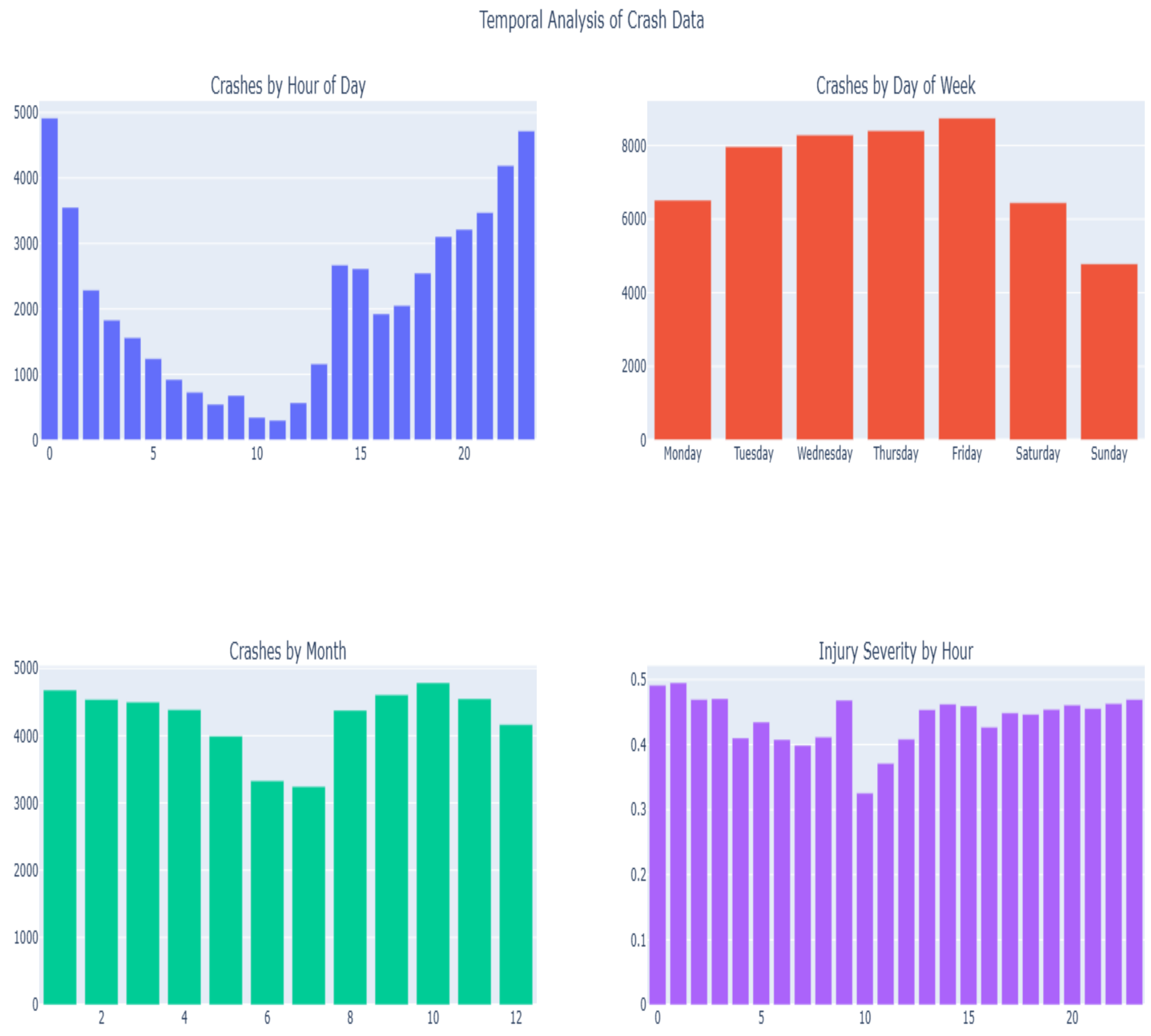
Key Observations

- **Severe crashes** are more geographically concentrated than moderate or minor ones, often occurring at major intersections or high-speed corridors. This concentration suggests that certain road segments or junctions present greater risk, potentially due to traffic volume, design complexity, or behavioral patterns (e.g., speeding).
- **Larger markers** are typically observed at these high-risk points, indicating locations where not only the frequency but also the consequences of crashes are more serious.
- In contrast, minor crashes are more evenly dispersed across the urban and suburban network, reflecting the widespread nature of low-severity incidents.

3.3 Temporal Analysis of Crash Data

The temporal analysis visualization provides a multi-dimensional exploration of crash patterns over time, structured as a 2x2 interactive subplot. It covers crashes by hour of day, day of the week, month of the year, and average injury severity by hour. This comprehensive layout allows easy identification of temporal risk patterns that could guide targeted traffic safety interventions.

File included: **temporal_analysis_visualization.html**



Key Observations

- **Crashes by Hour of Day:**
  The highest frequency of crashes occurs during late-night and early-morning hours (10 PM to 2 AM). This period is often associated with fatigue, reduced visibility, and potentially impaired driving. In contrast, 10 AM registers the lowest number of crashes, possibly due to lower traffic volumes and safer driving conditions during mid-morning hours.

- **Crashes by Day of Week:**

  Thursdays and Fridays show the highest crash frequencies, suggesting elevated risk towards the end of the workweek when both commuter and social travel increase. Sundays have the fewest crashes, possibly reflecting lower overall traffic activity.

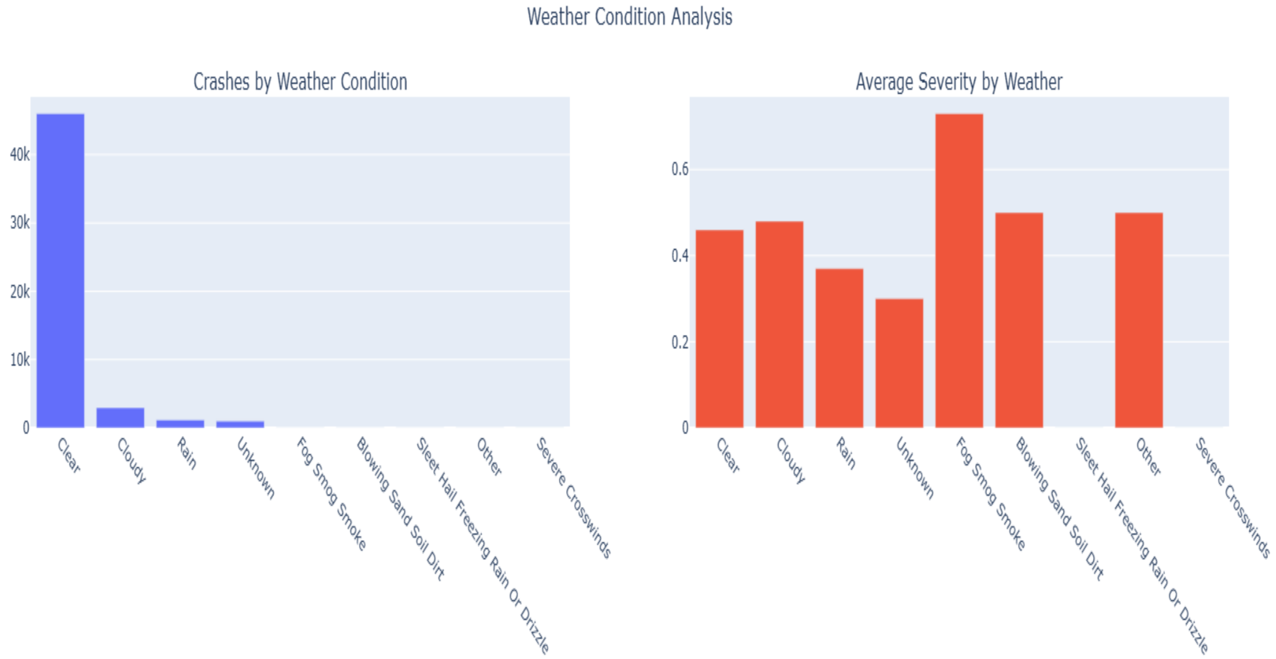
- **Crashes by Month**:
  Crashes are most frequent in October, potentially reflecting seasonal factors such as changes in daylight, weather conditions, or increased travel during holiday periods. Mid-year months like June and July show relatively lower crash frequencies.

- **Injury Severity by Hour**:
  Average injuries per crash are highest during the same late-night and early-morning hours when crash frequencies peak. This correlation indicates that crashes occurring at these times tend to be not only more frequent but also more harmful.

3.4 Weather Condition Analysis

The weather condition analysis provides an interactive view of how different weather types correlate with both crash frequency and injury severity. This visualization comprises two bar charts: one illustrating the **total number of crashes under each weather condition**, and the other depicting the **average number of injuries per crash** for each condition. Together, these charts offer a nuanced understanding of how environmental factors contribute to both the likelihood and severity of traffic incidents.

File included: **weather_analysis_visualization.html**



Key Observations

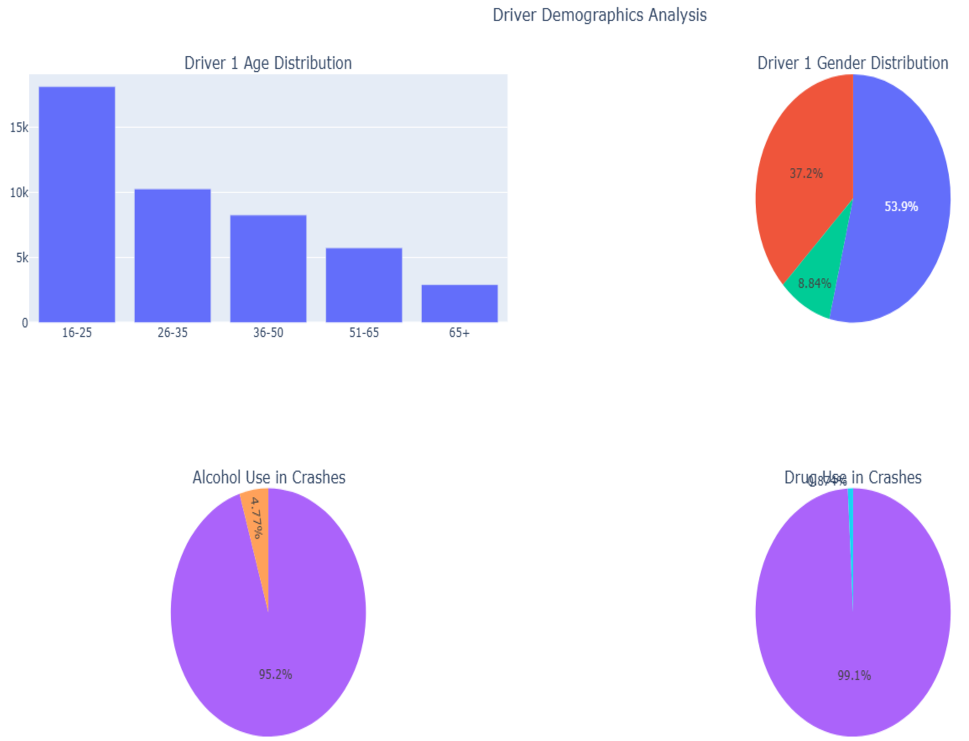
- **Crashes by Weather Condition:**
  Crashes overwhelmingly occurred during **clear weather conditions**, followed by a much smaller number under **cloudy skies**. This pattern primarily reflects greater driving exposure under good weather, as more vehicles are on the road and travel volumes are higher. Weather conditions like **rain**, **fog**, **smog**, and **blowing sand or dirt** accounted for a small proportion of total crashes, as did other hazardous conditions such as **sleet**, **hail**, or **crosswinds**.

- **Average Injury Severity by Weather:**
  Despite the low frequency of crashes during adverse conditions, these events were associated with significantly higher injury severity. In particular, crashes during **fog, smog, or smoke** had the highest average number of injuries per incident. This suggests that reduced visibility dramatically increases the risk of severe outcomes when crashes do occur.
  Similarly, crashes during conditions involving **blowing sand, soil, or dirt** and **other hazardous weather** types exhibited higher-than-average injury rates compared to clear conditions.

3.5 Driver Demographics Analysis

The driver demographics analysis provides insights into the age, gender, and substance involvement characteristics of drivers involved in crashes. The interactive visualization consists of four subplots: **age distribution**, **gender distribution**, **alcohol use in crashes**, and **drug use in crashes**, offering a comprehensive overview of key risk factors associated with driver profiles.

File included: **demographics_visualization.html**

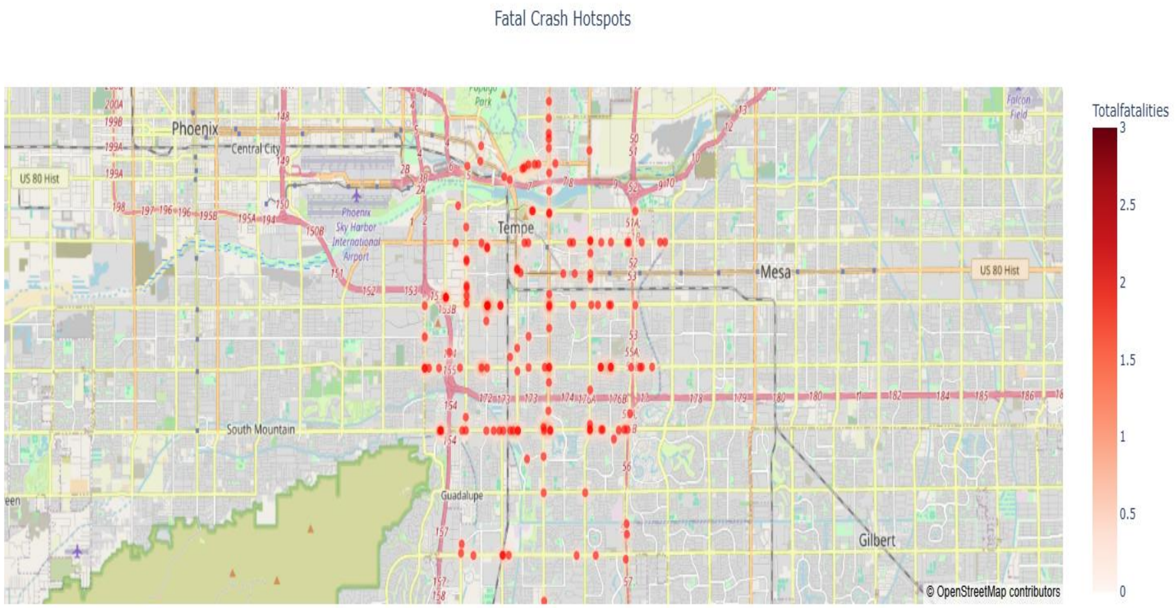


Key Observations

- **Age Distribution:**
  The **16–25 age group** accounts for the highest number of crashes, with over **18,000 incidents** involving young drivers. This underscores the elevated risk associated with novice and less experienced drivers.
  Crash involvement decreases steadily with age. The **26–35** and **36–50** groups contribute significantly but less so than younger drivers. Drivers aged **65+** are involved in the fewest crashes, reflecting both lower exposure and potentially more cautious driving behavior.
- **Gender Distribution:**
  The data reveals a clear gender disparity in crash involvement: **male drivers** account for approximately **53.9%** of incidents, while **female drivers** represent around **37.2%**. A small percentage (about **8.8%**) of records fall into the **unknown or unspecified** category, likely due to incomplete or non-disclosed data.
- **Alcohol Use in Crashes:**
  Alcohol use was recorded in approximately **5% of crashes** (around **2,440 incidents** out of 51,115 with valid data), indicating that while relatively infrequent, alcohol remains a significant contributing factor in traffic crashes.
- **Drug Use in Crashes:**
  Drug use was reported in less than **1% of crashes** (approximately **447 incidents** out of 51,115). This suggests that alcohol is the more prevalent substance-related risk factor in crashes within the study area.

3.6 Fatal Crash Hotspots

The fatal crash hotspots map provides an interactive visualization highlighting locations with the highest concentration of fatalities. Using a **density heatmap combined with individual markers** for fatal crashes, the map identifies critical danger zones within the study area. The heatmap employs a red color scale, where darker shades represent areas with greater fatality density, while overlaid markers pinpoint specific fatal crash events with detailed hover information.



Key Observations

- Fatal crashes are **highly concentrated in specific geographic clusters**, forming distinct hotspots on the map. These areas are typically near **urban centers, major intersections, and high-traffic corridors**, where complex traffic patterns and higher vehicle volumes increase crash risk.
- The visualization shows **recurring fatal incidents in certain zones**, suggesting that these are not isolated events but rather locations with persistent safety challenges. These hotspots may stem from infrastructure factors (e.g., road design, lighting), behavioral factors (e.g., speeding, distracted driving), or a combination of both.
- Outlying and suburban areas, in contrast, show far fewer fatal crashes, reinforcing the concentration of risk in densely trafficked urban environments.

4. Conclusion and Key Insights

This project provided a comprehensive geospatial and temporal analysis of traffic crash data spanning over a decade, with the goal of uncovering patterns that can inform targeted safety interventions and urban planning strategies. Using advanced interactive visualizations, the study explored crash density, severity, environmental influences, temporal trends, demographic factors, and fatal crash hotspots.

Several key insights emerged from the analysis:

- **Spatial Patterns:**
  Crashes, particularly those involving injuries and fatalities, are heavily concentrated along major arterial roads, freeway interchanges, and busy intersections. These hotspots indicate priority areas for infrastructure improvements, such as redesigning intersections, adjusting signal timings, or implementing traffic calming measures.
- **Temporal Trends:**
  Crashes are most frequent during late-night and early-morning hours (10 PM to 2 AM) and on Thursdays and Fridays. October recorded the highest monthly crash frequency. These patterns suggest that enhanced enforcement, public awareness campaigns, and targeted interventions during these critical periods could help reduce crash rates and severity.
- **Environmental Conditions:**
  While most crashes occur in clear weather (reflecting greater exposure), adverse conditions like fog, smog, and smoke were associated with significantly higher average injury severity. This finding underscores the importance of driver education and adaptive safety measures for low-visibility conditions.
- **Driver Demographics:**
  Younger drivers (16–25) were disproportionately involved in crashes, highlighting the need for enhanced driver education, graduated licensing policies, or other youth-focused interventions. Male drivers accounted for a higher proportion of crashes compared to females, and while alcohol and drug use were relatively infrequent, their role in crashes remains a critical concern for traffic safety efforts.
- **Fatal Crash Hotspots:**
  Fatal crashes form well-defined clusters within the urban core, often at locations with high traffic complexity. These recurring hotspots point to areas where infrastructure redesign, increased lighting, speed management, or focused enforcement could help reduce the risk of fatal outcomes.