Analyzing Fatal Traffic Crashes in the USA (2015–2022)

Abhishek KUMAR
Master 1 – MLDM
Université Jean Monnet de Saint-Étienne

March 2025

1 Problem Understanding

Fatal road crashes present one of the most persistent public safety threats in the United States. Despite improvements in vehicle safety, human error and environmental factors continue to contribute to significant loss of life. This project aims to analyze a large dataset of traffic fatalities from 2015 to 2022 to extract hidden patterns, temporal and spatial trends, and key correlates of fatal crash severity.

Objectives:

- Explore crash trends by time, geography, and severity
- Cluster states based on crash statistics to identify risk profiles
- Study correlations between crash elements
- Build and interpret predictive models for fatalities
- Provide actionable, data-driven policy insights

2 Data Overview and Cleaning

Dataset: Kaggle – *Fatal Crash Data Across U.S. States (2015–2022)*
Rows: 286,311 crashes **Columns:** 9 variables **Size:** 70 MB

The dataset captures fatal crash events in the United States over an 8-year span, including key attributes like crash date, state, county, total fatalities, persons involved, and vehicles. It was imported using the `readr` package, and column types were validated and parsed accordingly.

Data Cleaning: Missing values were detected in `County` (742 rows) and `CrashDateTime` (2,146 rows). These were dropped prior to modeling as they represented a small portion and their removal didn't affect the analysis outcomes.

Outliers such as crashes involving over 50 people or 10+ vehicles were noted during summary review. These records were kept for integrity, but handled with caution during modeling to reduce skew impact.

Summary Statistics:

Variable	Mean	Min	Max
Fatalities	1.09	1	20
Persons	2.23	0	93
Vehicles	1.57	1	10
Pedestrians	0.23	0	73

Table 1: Summary of key variables

This exploratory phase revealed an important insight: most fatal crashes involve 1–2 people and a single vehicle, making these the dominant crash scenario for prediction modeling.

3 Exploratory Data Analysis

Crash Count by State

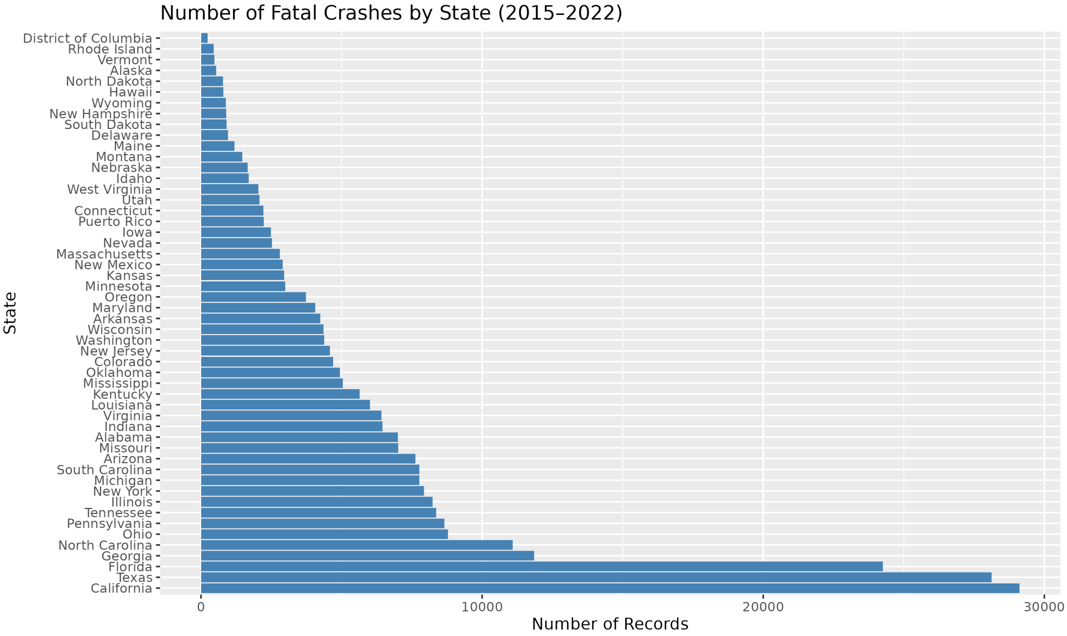


Figure 1: California, Texas, and Florida have the highest number of fatal crashes, reflecting large populations and heavy traffic.

Fatalities Distribution

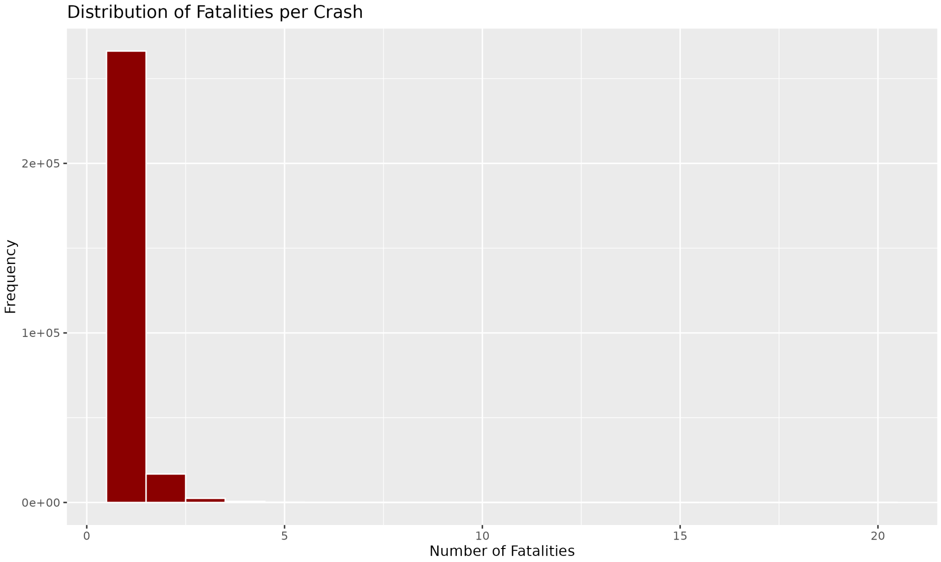
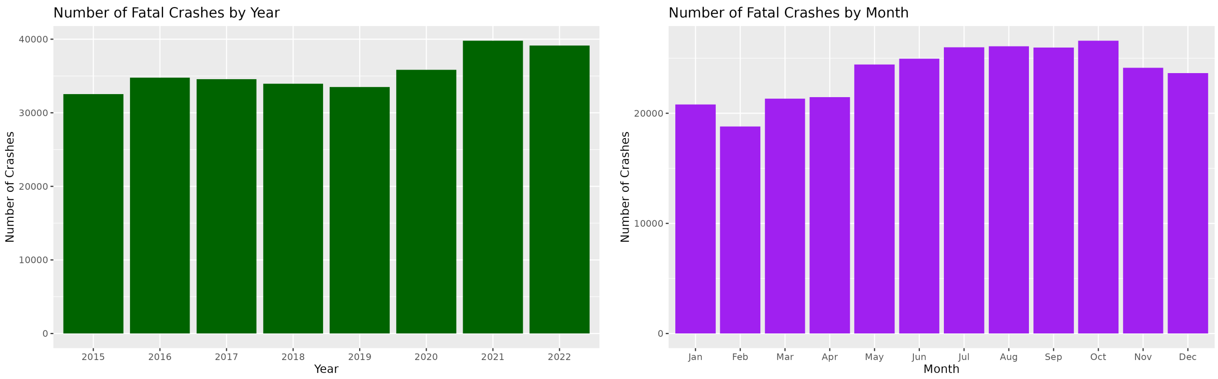


Figure 2: The vast majority of crashes involve 1 fatality; extreme fatal events are rare outliers.

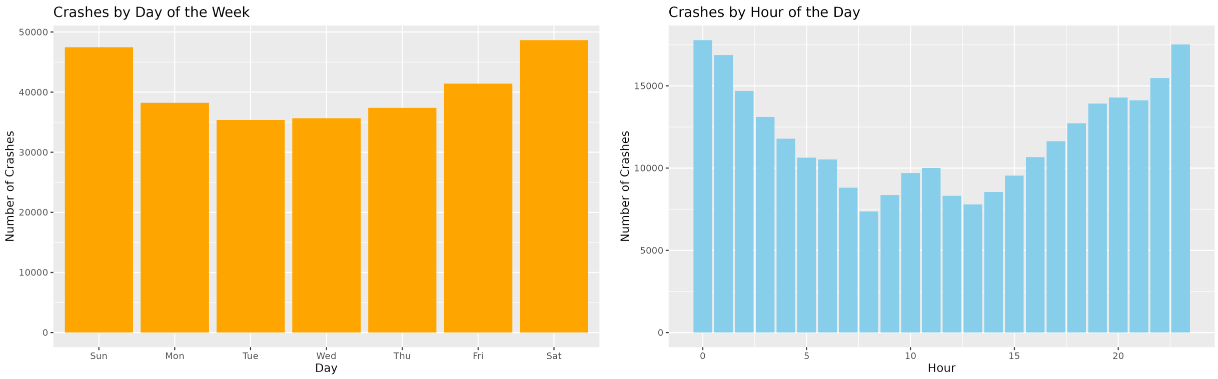
Temporal Trends



(a) Annual trend

(b) Monthly variation

Figure 3: Crashes peaked in 2021. Summer to early fall months are more dangerous.



(a) By day of week

(b) By hour of day

Figure 4: Saturdays and Sundays are riskier. Crashes peak at night and early morning hours.

4 Clustering and State-Level Risk

K-means clustering was used to classify states into three groups based on total crashes, fatalities, people involved, and vehicles.

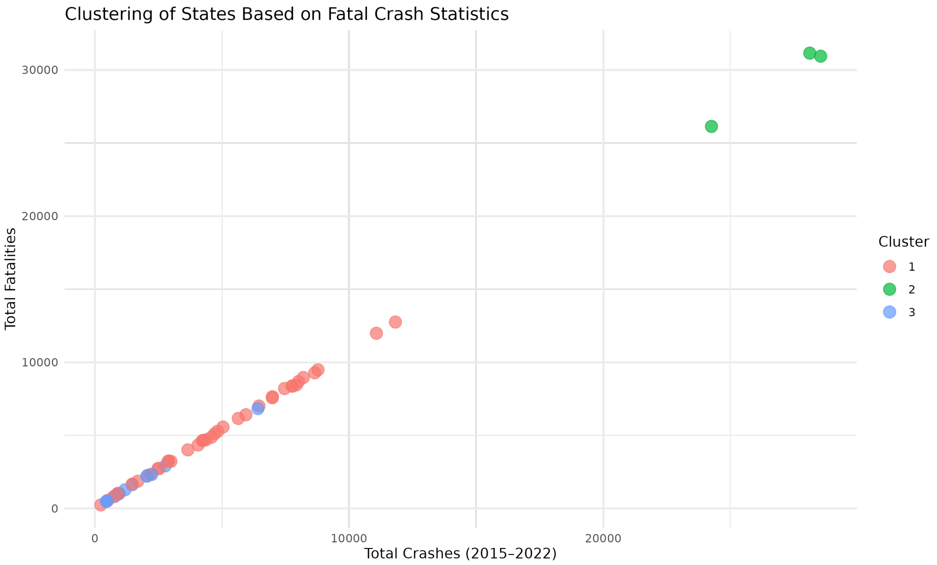


Figure 5: Cluster 3 includes states with highest crash volumes and fatality rates.

Insights:

- Cluster 1: Low-crash states — typically smaller or less urbanized (e.g., North Dakota, Vermont)
- Cluster 2: Medium-risk — moderate populations or mixed urban-rural profiles
- Cluster 3: High-risk states — California, Texas, Florida, where population and vehicle density are high

This classification supports tiered strategies — awareness in Cluster 1, behavioral interventions in Cluster 2, and aggressive enforcement and infrastructure investment in Cluster 3.

5 Correlation Analysis

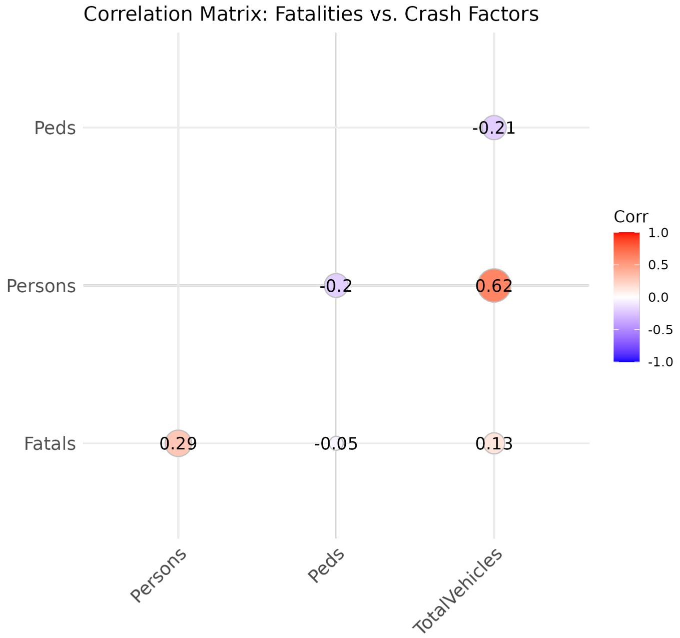


Figure 6: Fatalities correlate most strongly with number of persons involved (0.62), then vehicles. Pedestrians show weaker links.

Insights from Correlation Heatmap:

- Positive correlation (0.62) between `Persons` and `Fatals` suggests group crashes are more deadly.
- `TotalVehicles` shows moderate positive correlation with `Fatals` (0.11), indicating higher severity with more vehicles involved.
- Pedestrian presence has weak correlation with fatalities, likely due to rarity in fatal crashes.
- These observations guided the variable selection for our regression model.

6 Predictive Modeling

We applied multiple linear regression to predict fatalities based on `Persons`, `Peds`, and `TotalVehicles`.
Findings:

- `Persons` had the most significant influence on fatality prediction
- R^2 value of 0.0908 suggests limited explanatory power – other features likely required

- However, the general direction and trend of risk were captured

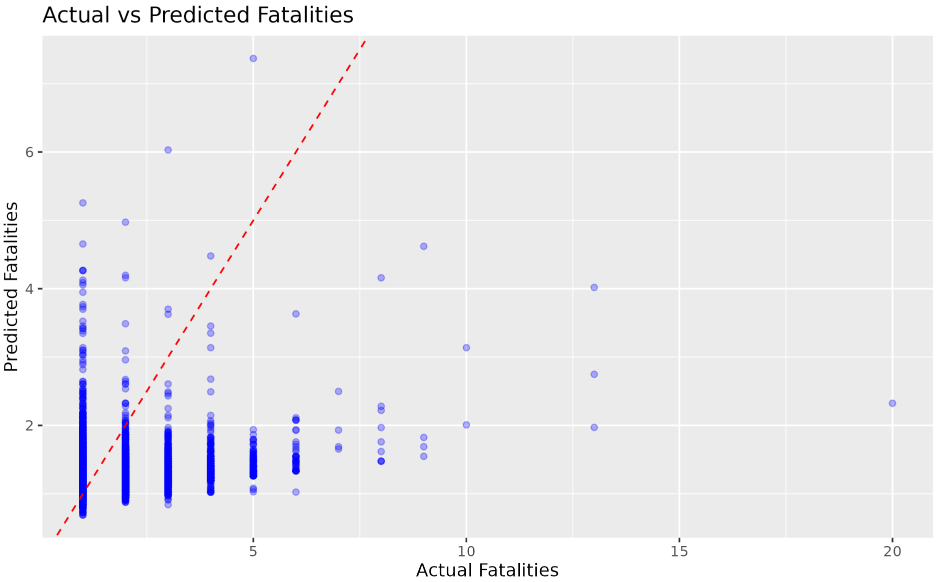


Figure 7: Actual vs. Predicted Fatalities. Points align loosely with the ideal line (red dashed), showing some trend learning but high residual variance.

7 Key Insights and Recommendations

Discovered Patterns

- Crashes rise in summer and fall seasons
- Night hours (especially 10PM–2AM) are high-risk for fatalities
- Weekends show 20–30% higher crash volume
- High-population states dominate crash statistics but also vary in risk per capita
- Person count is a better indicator of fatality likelihood than vehicle count

Recommendations

- Deploy real-time traffic alerts in high-risk hours
- Enhance enforcement and patrolling during weekends
- Prioritize policy reforms and investment in infrastructure in Cluster 3 states
- Extend dataset with weather, alcohol, seatbelt usage, and vehicle type for stronger prediction

8 Conclusion

This study has provided a comprehensive analysis of fatal road crashes in the United States from 2015 to 2022 using real-world crash-level data. Our findings highlight distinct temporal and spatial patterns, identify regional disparities in crash volumes, and pinpoint variables with the strongest influence on fatalities.

We successfully applied clustering to define risk-based state groupings and leveraged statistical modeling to assess variable impact on fatal outcomes. The strong correlation between the number of persons involved and fatalities emphasizes the significance of crash group dynamics.

While our regression model demonstrated limited predictive accuracy, it offers a foundation to build upon. Incorporating additional features like weather, alcohol influence, vehicle speed, and road conditions could substantially improve future model precision.

Overall, the project bridges exploratory insight with real-world application, reinforcing the value of data-driven road safety interventions. The methodologies implemented here serve as a replicable framework for future transport analytics and policy design.

References

- Fatal Crash Data Across U.S. States (2015–2022)
- Wickham, H. (2019). *Welcome to the Tidyverse.* Journal of Open Source Software.
- Wickham, H., et al. (2016). *ggplot2: Elegant Graphics for Data Analysis.*
- R Packages: `tidyverse, lubridate, ggcorrplot, scales, ggthemes`