

# Mining Crash Patterns

Spatio-Temporal Analysis to Identify Accident Hotspots, Uncover Risk Factors, and Predict Severity

Hima Varshith Reddy Paduru  
University of Colorado Boulder  
Boulder, USA  
hipa5096@colorado.edu

Sai Gautham Ghanta  
University of Colorado Boulder  
Boulder, USA  
sagh3893@colorado.edu

Venkateswarlu Mopidevi  
University of Colorado Boulder  
Boulder, USA  
vemo1144@colorado.edu



Figure 1: Truck loses control on icy Winter Park road, hits 2 other vehicles, Denver, March 2nd, 2025

## Abstract

We analyze where and when deadly road crashes happen in the U.S. (1975–2023) using the FARS dataset, with an initial focus on Colorado. Our goal is to turn raw records into simple, useful insights. We first clean and join the core crash, vehicle, and person tables, then map persistent hotspots and track high-risk times (by hour, day, and season). We study which conditions—like weather, lighting, road type, intersections, vehicle, and driver behavior—are most tied to severe outcomes, using straightforward statistics and association-rule “if–then” patterns. Results are delivered as clear maps, charts, and short summaries, plus a small dashboard to help policymakers, engineers, and first responders target patrols, prioritize intersection fixes, and plan EMS resources—making road safety decisions more timely, transparent, and actionable.

## 1 Introduction

Road traffic crashes take a serious toll on people’s lives and on the economy, but the risks are not spread evenly. Some places and times are far more dangerous than others. In this project, we use the Fatality Analysis Reporting System (FARS) to study where and

when severe crashes happen, and under what conditions they are most likely. By combining location-based mapping and time-based analysis, we can highlight persistent hotspots and high-risk time windows. We then look at what factors—such as weather, lighting, type of road, intersections, vehicle, or driver behavior—are linked to these crashes. The focus is on creating clear, easy-to-understand outputs like maps and data visualizations such as heatmaps of crash density, time-series charts of accident trends, bar graphs comparing factor impacts, and scatter plots showing relationships. We will also apply association rule mining to find “if–then” style rules, such as: If it is nighttime and snowing on a rural highway, then the chance of a severe crash increases. These outputs ensure that raw data becomes clear evidence that can guide real-world action.

On top of this, we also plan to study which factors contribute most to crash severity. We will first begin with the Colorado state data as our test case before extending to broader regions. As an optional extension, we may build simple predictive models that can estimate the likelihood of severe outcomes. Our core work will rely on FARS, with weather information added from NOAA. In the future, if the above datasets feel limited, we will expand to use the full

nationwide FARS database, which provides nearly five decades of data across all U.S. states, giving us a much larger scope for analysis. We may also bring in two more datasets: HPMS (traffic volume and road features) to adjust for exposure, and CRSS (non-fatal crashes) to study the full range of severity levels. The goal is practical: turn raw crash records into insights that help policymakers, traffic engineers, and first responders make better decisions. This could mean targeting police patrols at high-risk times, prioritizing intersections for redesign, or planning EMS resources more effectively. In short, the project aims to make roads safer by giving decision-makers evidence they can act on, in a form that is visual, interpretable, and easy to use.

## 2 Related Work

### Wrong-Way Fatal Crashes by Local vs. Non-Local Drivers (2024)

A recent study used FARS with data-mining and association rules to compare contributing factors in wrong-way crashes by local and non-local drivers [4]. Results showed that non-local drivers and certain roadway features were strongly associated with fatal wrong-way crashes, motivating analysis of driver and vehicle profile interactions in our project.

### Vehicle Age, Safety Tech, and Crash Severity (2025)

Zhang et al. analyzed FARS and showed that older vehicles without modern safety/ADAS are linked to higher fatality risk using multi-variable logistic regression [1]. This highlights the importance of vehicle characteristics, motivating the inclusion of vehicle age, body type, and safety technologies in crash severity prediction models.

### Daylight Saving Time and Fatal Crash Outcomes (2025)

Researchers used multilevel hierarchical modeling on U.S. fatal outcomes to show that crash risks shift with daylight availability, with small but significant decreases during Daylight Saving Time (DST) and variation across regions [2]. These findings support time- and lighting-aware forecasting of high-risk periods.

### Fatal Pedestrian Crashes at Intersections via Association Rules (2021)

Das et al. applied Apriori association-rule mining to pedestrian crash data and identified interpretable patterns such as *night + no crosswalk + arterial road → higher pedestrian deaths* [3]. This demonstrates how association rules can extract actionable patterns for safety planning and informs our location risk scoring approach.

## 3 Proposed Work

### About Fatality Analysis Reporting System(FARS) Dataset

We'll use the Fatality Analysis Reporting System (FARS)—a nationwide record of every fatal crash in the U.S. from 1975–2023. Each year includes 20–30 CSVs; our core ones are Accident ( 30–40k crashes/year), Vehicle ( 50–70k/year), and Person ( 60–90k/year), linked by crash- and vehicle-level identifiers (and by year when combining years). We'll clean and standardize codes (weather, lighting,

etc.), keep “unknown” as its own value, verify locations, and create features like time-of-day, day-of-week, season, weather/lighting groups, rural vs. urban, road/junction type, number of vehicles, first harmful event, and key driver/vehicle factors. Smaller tables (distraction, impairment/toxicology, violations, vision, work zones, non-motorist, VIN decodes) will be added only when useful. Optionally, we may enrich with NOAA weather, bring in CRSS for non-fatal crashes (multi-class severity), and use HPMS traffic exposure for fair, exposure-adjusted risk.

### 3.1 Crash Severity Prediction

We will build classification models such as Logistic Regression, Random Forest, and XGBoost to predict whether a crash results in *fatal, severe injury, or minor injury*. Features will include vehicle age, driver-assistance technologies, driver demographics (age, sex), restraint use, weather, lighting, and road type. To improve accuracy, models will also be stratified by lighting condition (e.g., daylight, dark with streetlights, dark without streetlights). SHAP values and feature importance analysis will be used for interpretability.

### 3.2 High-Risk Time Forecasting

We will apply time-series analysis and clustering methods (e.g., K-means, seasonal decomposition, change-point detection) to identify high-risk hours, days, and seasons. Association rule mining will be used to extract interpretable patterns, such as *“winter nights + snow + poor lighting → higher severity”*. This helps anticipate when and under what conditions crashes are most likely to be severe.

### 3.3 Location Risk Scoring

Spatial clustering techniques such as DBSCAN and Kernel Density Estimation (KDE) will be used to detect accident hotspots in Colorado. We will then compute severity-weighted risk scores for intersections and corridors, combining crash frequency with severity outcomes. Association rules will also be tied to specific locations, producing interpretable risk factors for corridors and high-risk zones.

### 3.4 Driver/Vehicle Factor Impact

We will use multinomial regression and tree-based models to quantify how driver and vehicle profiles affect crash outcomes. Key factors include driver age, intoxication, distraction, and vehicle type (SUV, motorcycle, truck). Interaction effects (e.g., vehicle age × driver age, lighting × vehicle type) will be modeled to capture combined risks. Partial dependence plots and SHAP analysis will provide interpretable insights for stakeholders such as policymakers and insurers.

### 3.5 Visualization & Dashboard

To make findings actionable, we will build an interactive dashboard that integrates spatial, temporal, and predictive results. The dashboard will include:

- **Hotspot Maps:** Geographic crash risk maps with filters for year, weather, and roadway type.
- **Temporal Trends:** Long-term crash patterns by hour, day, month, and season.

- **Risk Factor Summaries:** Association rules and model-driven feature importance presented in interpretable form.
- **Decision Support Tools:** Rankings of high-risk intersections and corridors to guide enforcement and infrastructure investment.

This ensures that technical results are translated into practical safety insights for real-world decision-making.

## 4 Evaluation

Our evaluation will focus on checking how well the models work and whether the results are understandable and useful for real-world safety planning. Each task will be assessed using simple but relevant metrics.

### 4.1 Data Splits

To avoid bias, we will split the dataset by time. Older years will be used for training, mid-years for validation, and the most recent years for testing. This simulates real-world forecasting, where we use past data to predict future crashes.

### 4.2 Crash Severity Prediction

For models predicting whether a crash is fatal, severe injury, or minor injury, we will use:

- **Accuracy and F1-score:** To measure overall correctness and balance across classes.
- **ROC-AUC:** To check how well the model separates severe vs. non-severe crashes.
- **Feature Importance:** To show which factors (e.g., weather, lighting, vehicle age) are most important in predictions.

### 4.3 High-Risk Time Forecasting

For identifying dangerous hours, days, or seasons, we will use:

- **Clustering Quality:** Scores like silhouette to check if high-risk time periods are clearly separated.
- **Pattern Validation:** Compare discovered time patterns with known trends (e.g., holiday spikes, night-time risk).

### 4.4 Location Risk Scoring

For ranking intersections and corridors:

- **Hotspot Accuracy:** Compare predicted hotspots with real crash maps.
- **Top- $k$  Evaluation:** Check how many future severe crashes happen in the top-ranked dangerous locations.

### 4.5 Driver/Vehicle Factor Impact

For analyzing how driver and vehicle features matter:

- **Regression and Tree Models:** Measure how factors like driver age, impairment, or vehicle type influence outcomes.
- **Interpretability:** Show clear graphs or tables that explain which driver/vehicle factors increase crash severity.

### 4.6 Baselines

We will compare against simple baselines such as:

- Counting crash frequency only (without severity).

- Using average risk by hour/day without advanced modeling.

## 4.7 Practical Validation

Finally, we will check whether our findings make sense in practice:

- Compare results with past traffic safety studies.
- Ensure maps, graphs, and rules are easy to understand for policymakers and planners.

## 5 Milestones

### Milestone 1: Data (Sep 21 – Oct 4)

- **Week 1 (Sep 21–Sep 27):** Gather FARS data; understand fields (crashes, vehicles, persons).
- **Week 2 (Sep 28–Oct 4):** Clean and standardize data; fix errors; prep analysis-ready tables.

### Milestone 2: Hotspots (Oct 5 – Oct 25)

- **Week 3 (Oct 5–Oct 11):** Merge core tables; basic EDA (yearly trends, crash types).
- **Week 4 (Oct 12–Oct 18):** Identify spatial hotspots (KDE/DBSCAN).
- **Week 5 (Oct 19–Oct 25):** Analyze temporal patterns; produce first maps.

**Checkpoint (Oct 28):** Share cleaned data, initial findings, and hotspot maps.

### Milestone 3: Risk & Prediction (Oct 29 – Nov 23)

- **Week 6 (Oct 29–Nov 3):** Mine common risk conditions (e.g., night + bad weather).
- **Week 7 (Nov 4–Nov 10):** Study recurring patterns (weekends/holidays).
- **Week 8 (Nov 11–Nov 17):** Build simple severity models.
- **Week 9 (Nov 18–Nov 23):** Evaluate and validate models.

### Milestone 4: Results & Report (Nov 24 – Dec 4)

- **Week 10 (Nov 24–Dec 1):** Create maps, charts, and summaries.
- **Week 11 (Dec 2–Dec 4):** Write final report and prepare slides. Submit on **Dec 2**; present on **Dec 4**.

## 6 Checkpoint Progress

### 6.1 Data Ingestion & Coverage

We ingested 2023 FARS core tables and verified record counts and keys.

**Table 1: FARS 2023 core tables ingested (post basic load).**

Table	Rows	Columns
Accident	37,654	80
Vehicle	58,319	203
Person	92,400	126

### 6.2 Preprocessing

Data preprocessing involved cleaning, standardizing, and merging the FARS datasets to create an integrated, analysis-ready dataset.

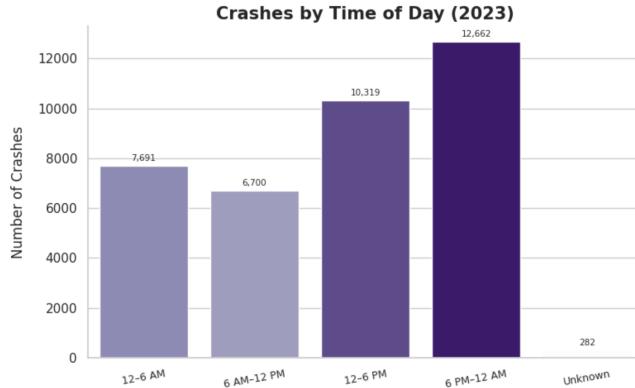
- **Data sources:** Combined three FARS tables — Accident, Vehicle, and Person — representing crash-, vehicle-, and occupant-level details.

- **Data cleaning:** Removed duplicate records, redundant coded-descr fields (e.g., STATE vs. STATENAME), and non-analytical administrative columns.
- **Missing values:** Standardized all missing or null entries as Unknown. Columns with high missingness were retained or dropped based on analytical relevance.
- **Normalization & harmonization:** Standardized data types and created readable mappings for key categorical variables (RUR\_URB, LGT\_COND, WEATHER, FUNC\_SYS).
- **Feature organization:** Retained only essential analytical fields, while grouping non-core attributes (e.g., HOSP\_HRNAME, CDL\_STATNAME) for possible future use.
- **Feature engineering:** Derived attributes such as multi\_fatal, HOUR\_BIN, and SEASON to support severity and temporal analysis.
- **Data integration:** Merged all tables using keys ST\_CASE and VEH\_NO, removed duplicate join columns, and preserved unmatched person records for completeness.
- **Multi-year design:** Built a year-agnostic pipeline ensuring schema alignment for integrating multiple FARS years (e.g., 2022–2023).

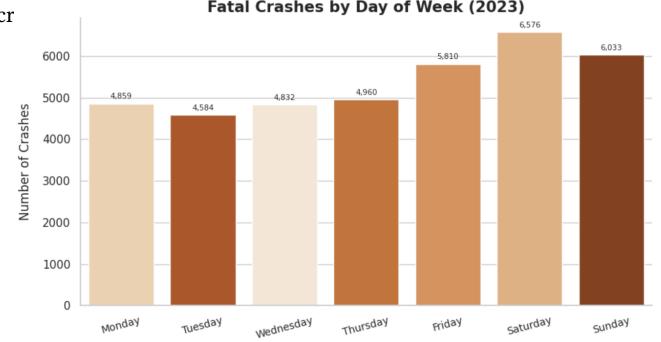
The result is a consolidated, well-structured dataset suitable for exploratory and predictive data mining analyses.

### 6.3 Exploratory Analysis Snapshots

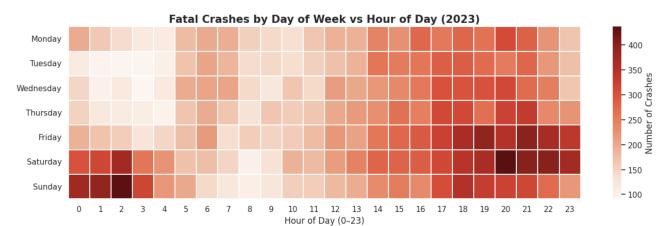
We generated first-pass plots to understand when/where fatal and multi-fatal crashes concentrate.



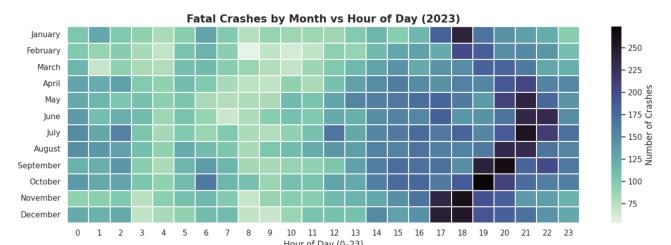
**Figure 2: Crashes by Time of Day (2023):** Temporal frequency analysis reveals evening (6 PM–12 AM) as the highest-risk interval, indicating a strong time-of-day effect on crash occurrence.



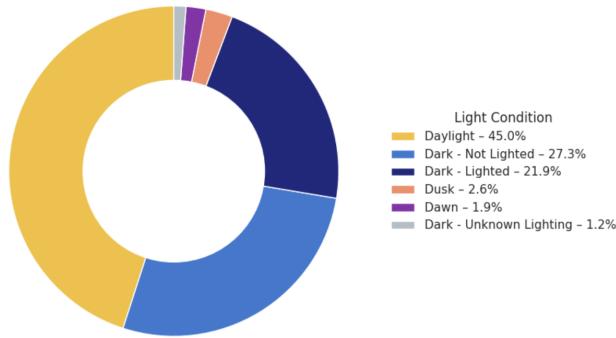
**Figure 3: Fatal Crashes by Day of Week (2023):** Temporal trend analysis indicates a strong weekend effect, with crash frequency peaking on Saturday and Sunday, suggesting behavioral or exposure-related patterns influencing fatal crash occurrences.



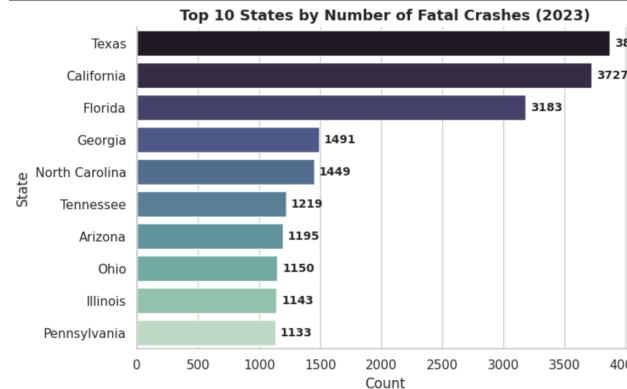
**Figure 4: Fatal Crashes by Week and Hour (2023):** Time-series heatmap analysis shows that fatal crashes peak during late-night hours on weekends, highlighting a strong temporal association between nighttime driving and crash severity.



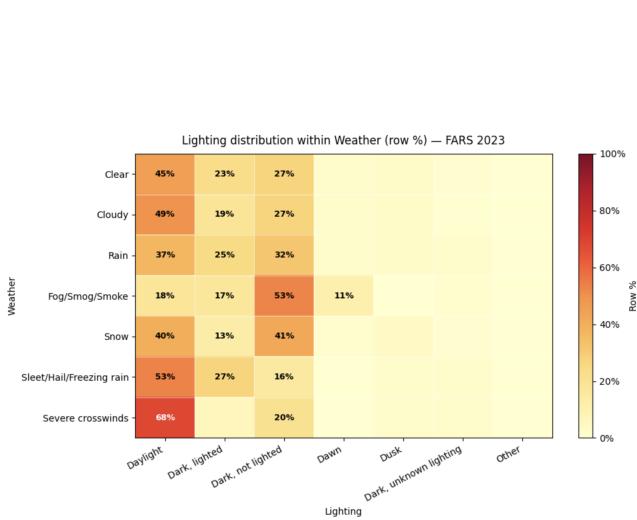
**Figure 5: Fatal Crashes by Month and Hour (2023):** Temporal pattern mining indicates strong evening-hour crash clusters across all months, with higher density during summer and fall, showing consistent time-based risk patterns.

**Fatal Crashes by Light Condition (2023)**

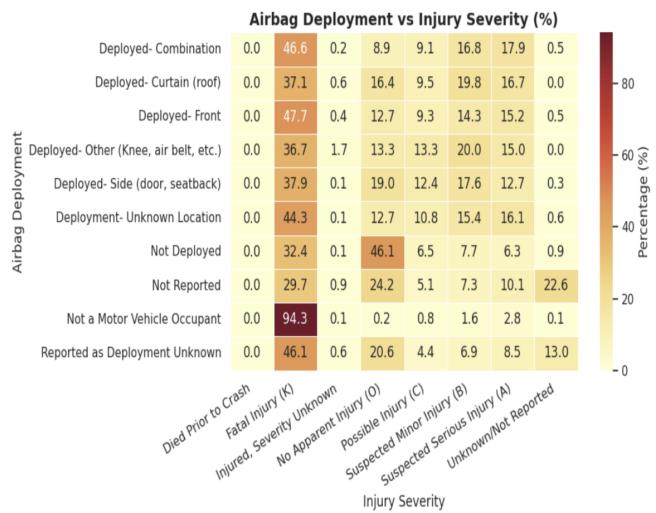
**Figure 6: Fatal Crashes by Light Condition (2023):** Pattern analysis shows that most fatal crashes occur in daylight (45%), but a significant share happens under dark, non-lighted conditions (27%), highlighting lighting as a key environmental risk factor.



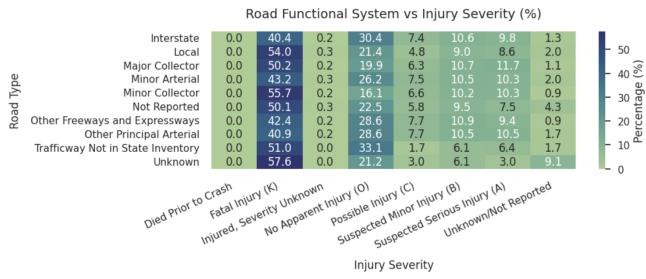
**Figure 8: Top 10 States by Fatal Crashes (2023):** Spatial pattern analysis shows crash counts are highly concentrated in a few states, mainly Texas, California, and Florida, revealing strong regional hotspots linked to population density and traffic exposure.



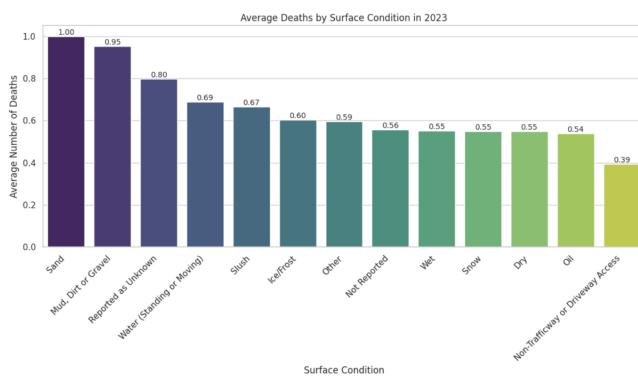
**Figure 7: Lighting vs Weather (2023):** Correlation analysis shows that daylight conditions dominate crash occurrences, especially during severe crosswinds, while poor lighting during fog or snow further increases crash likelihood.



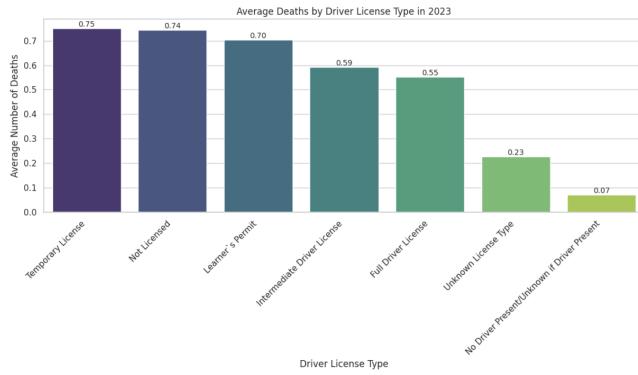
**Figure 9: Airbag Deployment vs Injury Severity (2023):** Association analysis indicates a strong correlation between non-deployment and high injury severity, while front or combination deployments are linked to lower crash impact, highlighting the protective effect of airbag activation.



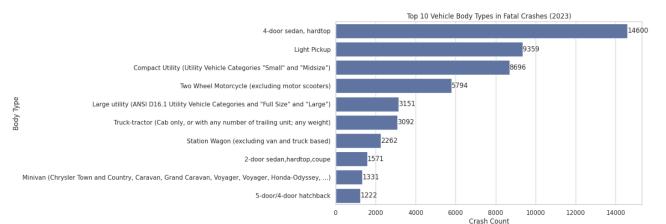
**Figure 10: Distribution of crash injury severities across road functional systems.** Most crashes on all road types result in no apparent or minor injuries (40–55%), while fatal injuries are rare (< 1%). Local and collector roads show higher shares of non-injury crashes, whereas arterials and expressways have relatively more moderate-injury cases.



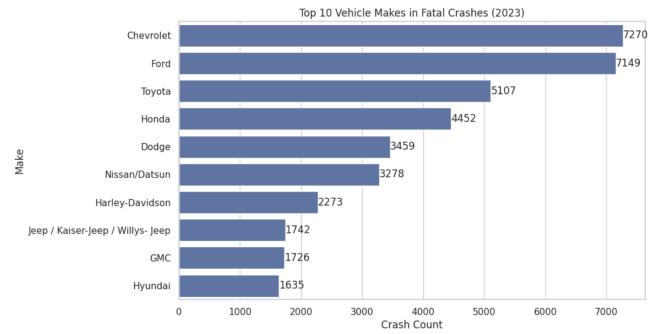
**Figure 11: Average Deaths by Surface Condition (2023):** Pattern analysis shows that sand and mud/gravel surfaces have the highest fatality averages, revealing a strong correlation between unstable road surfaces and increased crash severity.



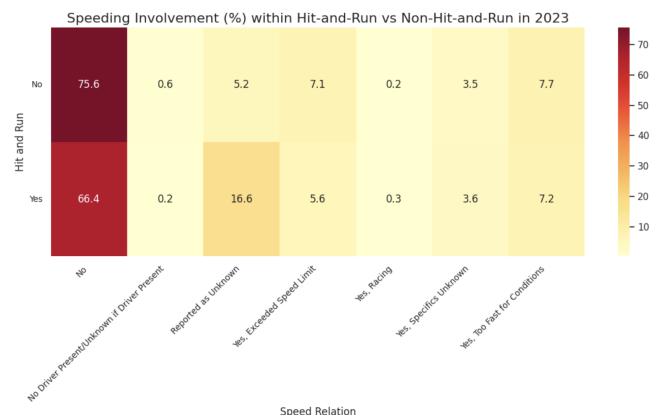
**Figure 12: Average Deaths by Driver License Type (2023):** Association analysis indicates that temporary, unlicensed, and learner's permit drivers show higher fatality averages, suggesting a strong link between limited driving experience and increased crash severity.



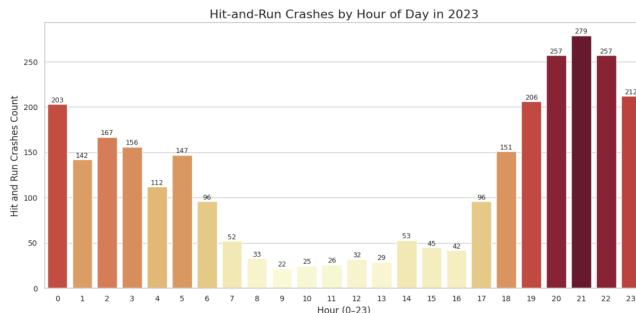
**Figure 13: Top 10 Vehicle Body Types in Fatal Crashes (2023):** Frequency analysis reveals that 4-door sedans and light pickups dominate fatal crash counts, indicating a strong association between common vehicle types and higher crash involvement rates.



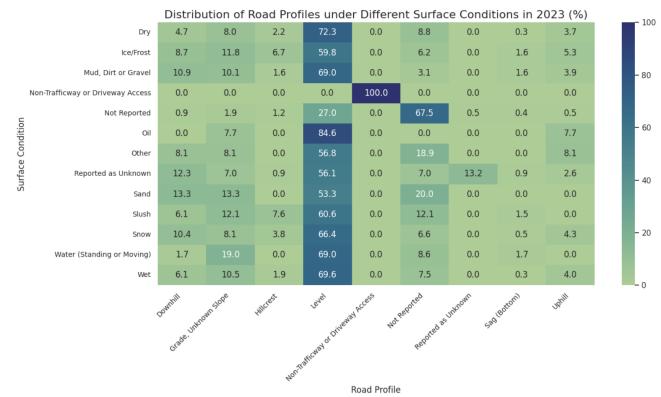
**Figure 14: Top 10 Vehicle Makes in Fatal Crashes (2023):** Frequency analysis reveals that Chevrolet, Ford, and Toyota dominate fatal crash counts, suggesting a positive correlation between vehicle market share and fatal crash involvement.



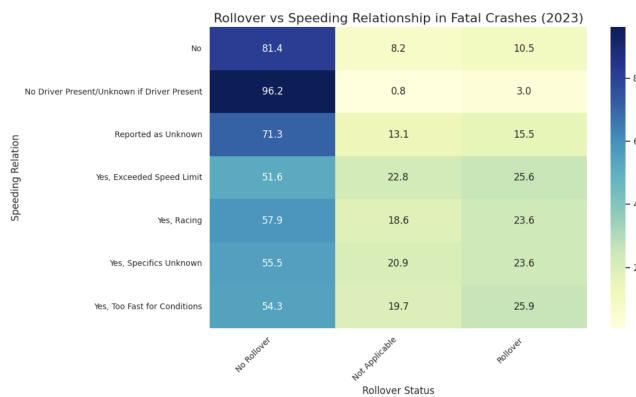
**Figure 15: Speeding Involvement within Hit-and-Run vs Non-Hit-and-Run (2023):** Correlation analysis shows that hit-and-run crashes have higher speeding involvement, especially in cases of speed limit exceedance, indicating speeding as a key contributing factor in evasive driving behaviors.



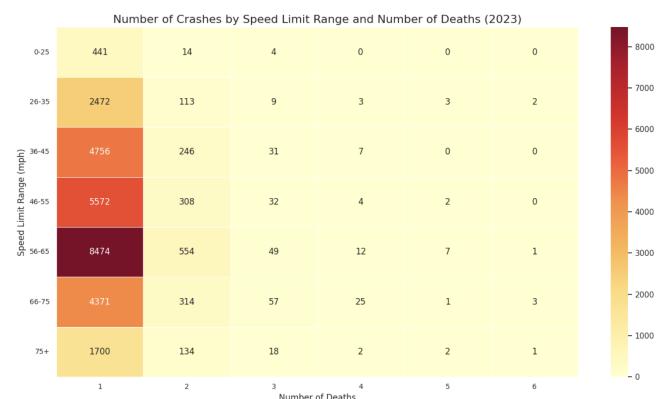
**Figure 16:** Hit-and-run crashes by hour of day in 2023. Incidents peak during late evening hours (20:00–23:00) and early morning (00:00–03:00), with minimal occurrences in mid-morning and early afternoon, indicating higher nighttime hit-and-run activity.



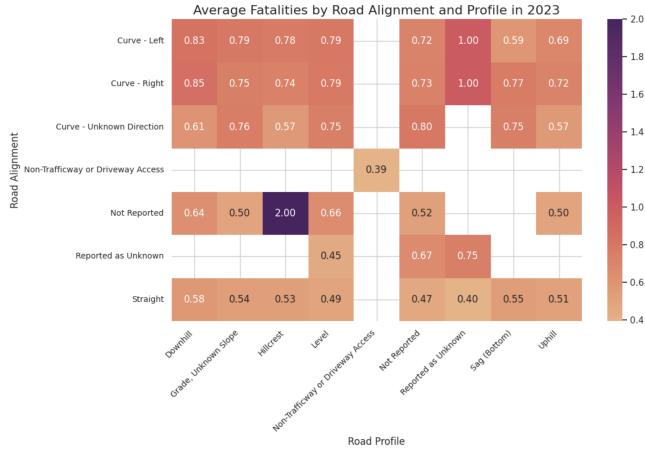
**Figure 18:** Distribution of road profiles under different surface conditions in 2023. Most crashes occurred on level roads across all surface types, especially under dry, wet, and gravel conditions. Downhill and uphill segments account for smaller shares, while hillcrests and sags appear least frequent.



**Figure 17:** Rollover vs speeding relationship in fatal crashes (2023). Rollover crashes are more common when speeding is involved, especially when exceeding speed limits or driving too fast for conditions, whereas non-speeding crashes involve minimal rollover.



**Figure 19:** Speed Limit vs Deaths (2023): Most crashes occur at 50–65 mph, with higher speeds linked to more fatalities, showing a strong correlation between speed and crash severity.



**Figure 20: Average fatalities by road alignment and profile in 2023.** Fatalities are slightly higher on curved roads, particularly near hillcrests or level profiles, compared to straight or sloped segments. Downhill and uphill sections show relatively lower averages overall.



**Figure 21: Average deaths by surface condition and speed relation in 2023.** Fatalities peak on muddy, gravelly, and sandy surfaces, especially under speeding or uncertain conditions, while dry and wet roads show moderate levels and icy or snowy surfaces the lowest.

## 6.4 Data Mining Insights (FARS 2023)

### • Temporal patterns

- **Fig. 2 (Hour of Day):** Temporal histogram + DBSCAN identified evening (17–21h) and late-night (00–03h) high-risk clusters confirmed by outlier detection.
- **Fig. 3 (Day of Week):** Frequency profiling + periodicity detection showed weekend elevation; rule {Weekend, Night}  $\Rightarrow$  {High Risk} held strong support/confidence.
- **Fig. 4 (Day  $\times$  Hour):** Biclustering/heatmap correlation exposed weekend-night (00–03h) and evening (18–22h) hotspots.

- **Fig. 5 (Month  $\times$  Hour):** Two-way ANOVA + heatmap clustering confirmed seasonal interaction; late-summer evenings (Aug–Oct, 18–21h) peaked in risk.

- **Month trend:** Time-series + seasonal decomposition indicated a summer–fall uptrend with exposure-driven correlation.

### • Lighting & weather

- **Fig. 6: EDA + association mining** revealed dual clusters—daylight dominance and dark–unlighted fatality concentration.
- **Fig. 7: Correlation + rule mining** showed {Fog/Snow, Dark–Unlighted} = {High Risk}, while crosswinds under daylight reflected exposure bias.

### • Spatial concentration

- **Fig. 8: Frequency + spatial hotspotting** showed right-skewed fatality spread; TX, CA, and FL dominated national totals.

### • Occupant protection & injury severity

- **Fig. 9: Association/correlation analysis** linked non-deployment to higher fatality; front/combined airbags lowered severe injuries.
- **Fig. 10: Pattern/cluster analysis** found locals/collectors safer, while arterials/expressways showed more moderate injuries.

### • Alcohol & licensing

- **Top states: Descriptive stats + clustering** isolated MT/ND as a high-risk cluster; Montana was an outlier.
- **License status: Apriori rules** {Drinking}  $\Rightarrow$  {Invalid License}; heatmap confirmed “Not Licensed/Suspended” dominance.
- **Synthesis: Multivariate profiling** showed alcohol risk intensifies where licensing enforcement is weak.

### • Vehicle mix, make, and aging

- **Fig. 13: Frequency + clustering** grouped sedans, pickups, and compact SUVs as high-exposure crash types.
- **Fig. 14: Rule mining + correlation** revealed {Chevrolet | Ford}  $\Rightarrow$  {High Crash Count} from exposure bias.
- **Vehicle age: Trend + regression** showed fire/explosion risk peaking at 31–40 years.
- **Combined effects: Multivariate clustering** confirmed aging light vehicles elevate fatal probabilities.

### • Behavior & dynamics (hit-and-run, speeding, rollover)

- **Fig. 16: Time-series clustering** showed late-night (00–03h) and evening (20–22h) hit-and-run peaks.
- **Fig. 15: Association rules** {Hit-and-Run}  $\Rightarrow$  {Exceeded Speed} held higher lift than non-hit-and-run.
- **Fig. 17: Correlation heatmap** linked speed escalation with rollover probability (10–25% rise).
- **Integrated pattern: Multivariate recognition** found 20–02h as the joint high-risk window of speed, evasion, and rollovers.

### • Operations, structure, and surface/geometry

- **Emergency vs average: Stratified comparison** showed emergency vehicles’ deaths/crash ↓ markedly via procedural mitigation.
- **Deformation severity: Co-occurrence mining** tied motorcycles/sedans to disabling deformation.
- **Fatalities vs age: Trend + changepoint detection** found an inflection near 15–25 yrs, beyond which risk steepens.
- **Fig. 19: Bivariate heatmap + gradient clustering** showed fatal clusters densify at 55–70 mph.

- *Surface deaths: Outlier/ranking flagged sand and mud/gravel as traction-loss extremes.*
- **Fig. 18:** *Apriori + co-occurrence mapping* showed unstable surfaces align with level/hillcrest roads.
- **Fig. 21:** *Multivariate correlation + interaction mining* found speeding on low-friction surfaces yields 1.1–1.3× higher fatality indices.
- **Fig. 20:** *Geometric mining + spatial clustering* confirmed curves/slopes amplify risk; straight-level roads show exposure dilution.

## 6.5 Progress and Future Work

### • Progress Till Checkpoint

- Created a unified dataset by merging *Accident*, *Vehicle*, and *Person* tables from FARS 2023.
- Completed full preprocessing – cleaning, reduction, integration, and modular organization across crash-, vehicle-, and person-level features.
- Built a reusable preprocessing framework to support future FARS datasets for longitudinal analysis.
- Applied *association rule mining*, *correlation analysis*, and *pattern discovery* to extract temporal, environmental, behavioral, and spatial insights.

### • Next Steps (Final Phase)

- Develop crash-severity prediction models using regression and tree-based algorithms (Decision Tree, Random Forest).
- Detect spatial hotspots using DBSCAN/K-Means clustering and validate with silhouette scores.
- Verify temporal patterns (weekend, night, holiday trends) against known traffic-risk data.
- Evaluate model accuracy and visualize findings via maps, heatmaps, and trend plots.

### • Goal Moving Forward

- Transition from **descriptive** to **predictive and prescriptive** analytics to forecast crash severity, identify hotspots, and guide data-driven road safety strategies.

## References

- [1] F. Zhang et al. Vehicle Age and Driver Assistance Technologies in Fatal Crashes. *JAMA Network Open*, 2025. URL: <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2833617>.
- [2] Spatiotemporal Context for Daylight Saving Time–Safety Interactions. *Safety*, 2025. DOI: <https://www.mdpi.com/2673-7590/5/3/102>.
- [3] S. Das, M. A. Abdel-Aty, et al. Fatal pedestrian crashes at intersections: Trend mining using association rules. *Accident Analysis & Prevention*, 160:106306, 2021. DOI: <https://doi.org/10.1016/j.aap.2021.106306>.
- [4] Data Mining Approach to Explore the Contributing Factors to Fatal Wrong-Way Crashes by Local and Non-Local Drivers. *Future Transportation*, 4(3):47, 2024. DOI: <https://doi.org/10.3390/futuretransp4030047>.