Course: **Data Mining** (CSCI 5502)

Mining Crash Patterns

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

Professor: **Qin Lv**
TAs: **Mohsena Ashraf, Julia Romero**

THE TEAM

Hima Varshith Reddy Paduru CSCI 5502

Sai Gautham Ghanta CSCI 5502

Venkateswarlu Mopidevi CSCI 5502

INTRODUCTION

Context:
Road traffic accidents remain a major public safety issue. The **Fatality Analysis Reporting System** (FARS, 1975–2023) offers nationwide fatal crash data to support data-driven safety research and urban planning.

Why Spatio-Temporal Analysis?
Crash risk follows spatio-temporal patterns influenced by **weather, lighting, road type, and driver behavior**. Traditional methods reveal correlations but lack localized, predictive, and actionable insight.

Approach:
This project applies data mining techniques such as **clustering, association rule mining, and predictive modeling** to uncover hidden crash patterns, identify high-risk factors, and move toward severity prediction.

Scope and Purpose:
This project analyzes five decades of Colorado crash data from the Fatality Analysis Reporting System (FARS) to **identify hotspots and key risk factors**. The state-level focus provides a balanced, comprehensive view of long-term crash trends while keeping the analysis manageable.
If findings are limited or resources allow, the study may expand to other states or integrate complementary datasets to enhance predictive insights and support data-driven safety decisions for policymakers and planners.

Potential datasets for future integration::

- CRSS: Includes non-fatal crashes for multi-class severity prediction.
- NOAA: Provides weather data for climate-related crash analysis.
- HPMS: Adds traffic exposure data for more accurate hotspot detection.

MINING CRASH PATTERNS

RELATED WORK

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

→ A recent FARS-based study used data mining and association rules to compare factors in wrong-way crashes. It found that non-local drivers and certain road features were strongly linked to fatal incidents, motivating our analysis of driver - vehicle interactions.

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

→ Zhang et al. used FARS data and logistic regression to show that older vehicles lacking modern safety or ADAS features face higher fatality risks. This underscores the need to include vehicle age, body type, and safety technologies in crash severity models.

Daylight Saving Time and Fatal Crash Outcomes (2025)

→ Researchers found that fatal crash risk varies with daylight, showing slight decreases during Daylight Saving Time and regional differences - supporting time- and lighting-aware risk forecasting.

Fatal Pedestrian Crashes at Intersections via Association Rules (2021)

→ Das et al. used Apriori association-rule mining on pedestrian crash data, revealing patterns like night + no crosswalk + arterial road → higher fatalities. This shows how such rules can uncover actionable insights, guiding our location risk scoring.

PROPOSED WORK (PART 1)

- → Use five decades of nationwide FARS crash data.
- → Focus on core tables: Accident, Vehicle, & Person.
- → Clean and standardize categorical codes (weather, lighting, roadway).
- → Enhance with derived features: time, season, location type, and driver/vehicle factors.
- → Optionally integrate NOAA (weather), CRSS (non-fatal), and HPMS (traffic exposure) data.

- → Train Logistic Regression, Random Forest, and XGBoost models.
- → Predict crash severity: fatal, severe, minor injury.
- → Use key features: vehicle age, safety tech, driver demographics, restraints, weather, lighting, road type.
- → Stratify results by lighting (daylight, dark with/without streetlights).
- → Apply SHAP values and feature importance for model interpretability.

- → Use K-means and seasonal decomposition for spatio-temporal clustering.
- → Identify high-risk hours, days, and seasons.
- → Apply association rules to reveal interpretable risk patterns (e.g., winter nights + snow → higher severity).
- → Enable forecasting of high-risk time windows for proactive safety measures.

- → Detect Colorado crash hotspots using DBSCAN and KDE.
- → Compute severity-weighted risk scores for intersections and corridors.
- → Link association rules to locations for interpretable risk insights.

PROPOSED WORK (PART 2)

05 Driver and Vehicle Factor Impact

- Analyze driver factors: age, impairment, distraction.
- Analyze vehicle factors: type, age, safety features.
- Model key interactions (e.g., driver age × vehicle age, lighting × vehicle type).
- Use regression and tree-based models for analysis.
- Explain results with SHAP values and partial dependence plots.

06 Visualization and Dashboard

- Build an interactive dashboard to integrate all results.
- Include hotspot maps with filters (year, weather,roadway).
- Add temporal trend charts by hour, day, and season.
- Summarize risk factors from association rules and machine learning models.
- Provide decision tools to rank high-risk intersections and corridors for policymakers and planners.

EVALUATIONS

Data Splits

- Train on older years, validate on mid-years, and test on recent years.
- Simulates real-world forecasting with past → future prediction.

Crash Severity Prediction

- Metrics: Accuracy, F1-score, ROC-AUC.
- Check feature importance (weather, lighting, vehicle age, etc.).

High-Risk Time Forecasting

- Use clustering quality scores (e.g., silhouette) to validate patterns.
- Compare discovered time patterns with known trends (holidays, night-time risk).

Location Risk Scoring

- Compare predicted hotspots with real crash maps.
- Top-k evaluation: check how many future crashes happen at top-ranked locations.

Driver and Vehicle Factor Impact

- Use regression and tree-based models to measure effects of driver/vehicle factors.
- Provide clear visual explanations (graphs, tables).

Baselines & Practical Validation

- Crash frequency only (ignoring severity).
- Average risk by hour/day without modeling.
- Compare with past traffic safety studies.
- Ensure outputs (maps, graphs, rules) are easy for policymakers and planners to interpret.

PROJECT PLAN AND MILESTONES

Milestone 1

Data Collection and Cleaning

Sep 21 - Oct 4

Week 1:
Gather the FARS data and learn what information is available (crashes, vehicles, people).

Week 2:
Clean the data so it is consistent, remove errors, and prepare it for analysis.

Milestone 2

First Analysis and Hotspot Detection

Oct 5 - Oct 25

Week 3: Combine the data into one set and look at simple statistics (trends by year, crashes by type).

Week 4: Start finding accident "hotspots" - places where crashes happen more often.

Week 5: Look at how crashes change over time (seasons, years) and make first maps to show results.

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

Milestone 3

Risk Factors and Predictions

Oct 29 - Nov 23

Week 6: Look for common conditions that lead to crashes (for example, night + bad weather).

Week 7: Study how patterns repeat at certain times (like weekends or holidays).

Week 8: Build simple models to see if we can predict how severe a crash might be.

Week 9: Test the models and see how well they work.

Milestone 4

Results, Dashboard, and Final Report

Nov 24 - Dec 4

Week 10:
Create visuals (maps, charts, summaries) to clearly show the results.

Week 11:
Write the final report and prepare the presentation. Submit the report on Dec 2 and present on Dec 4.

Checkpoint Presentation Update

Updates and progress since the initial project proposal

Data Preprocessing and Exploratory Pattern Analysis

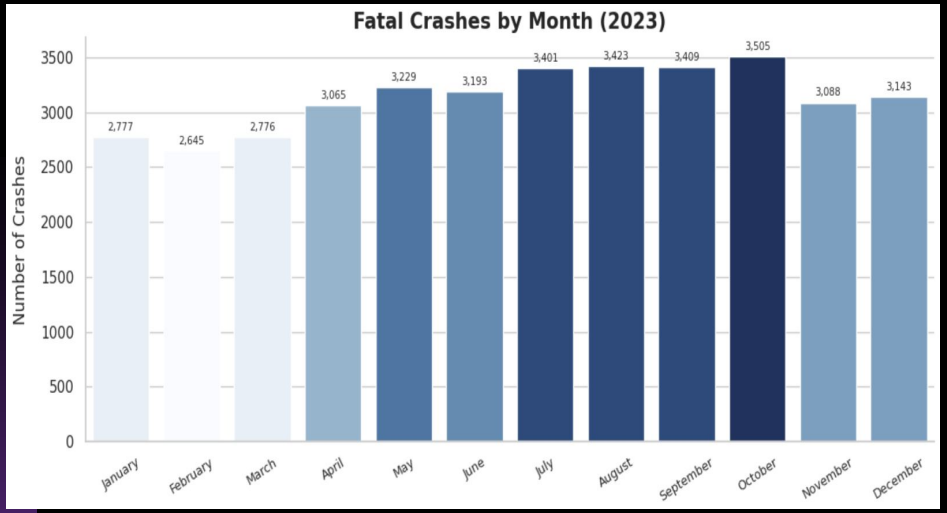
Data Sources:
We worked with three main FARS datasets (2023) —
Accident Table (37654, 80): crash-level info — location, time, weather, lighting, road characteristics (surface conditions, type, junctions, rural vs urban).
Vehicle Table (58319, 203): vehicle-level info — driver characteristics, hit and run, speed limits, type, maneuver, and damage.
Person Table (92400, 126): person-level info — alcohol, use, seatbelt use, airbag deployment, injury severity, age, gender, and role.
Together, they provide a complete multi-layered view of each crash (environment → vehicle → person).

Data Preprocessing:
- **Data Cleaning:** checked for duplicates, removed redundant or administrative columns, standardized missing values as "Unknown".
- **Data Reduction:** kept only relevant attributes while dropping less useful ones. (Columns reduced to total of **150 columns** from the initial 400 columns)
- **Data Integration:** merged all three datasets on common **IDs (ST_CASE, VEH_NO)**, removed duplicate columns, and retained unmatched person records to preserve fatality counts.
- **Attribute Organization:** grouped columns into crash-level, vehicle-level, and person-level for modular analysis. Kept core features like weather, lighting, seatbelt use, injury severity, driver chars and then archived fields such as hospital time, heavy vehicle type or commercial license for possible future studies.

The final result was a clean, unified dataset, consistent across attributes and ready for pattern discovery.

Analysis/Work done Till Now:
So far, our focus has been on understanding data and discovering key patterns using data-mining concepts:
- **Exploratory statistics:** overall crash counts, **injury distribution, and statewise trends**.
- **Temporal analysis:** variations by month, day, and hour to identify repeating time patterns.
- **Condition-based correlation:** studied combinations such as night + bad weather or **weekend + speeding**, showing strong co-occurrence with fatalities.
- **Environmental influence:** explored weather, lighting, and road conditions to see how external factors affect crash severity.
- **Human & behavioral factors:** analyzed **seatbelt use, airbag deployment**, alcohol involvement, & licensing to understand human impact on fatal outcomes.
- **Spatial clustering:** identified states and counties contributing to most fatalities — revealing concentration in mobility-dense regions.

We applied **association rule mining, correlation checks, and pattern discovery concepts** learned in class to uncover these recurring relationships.

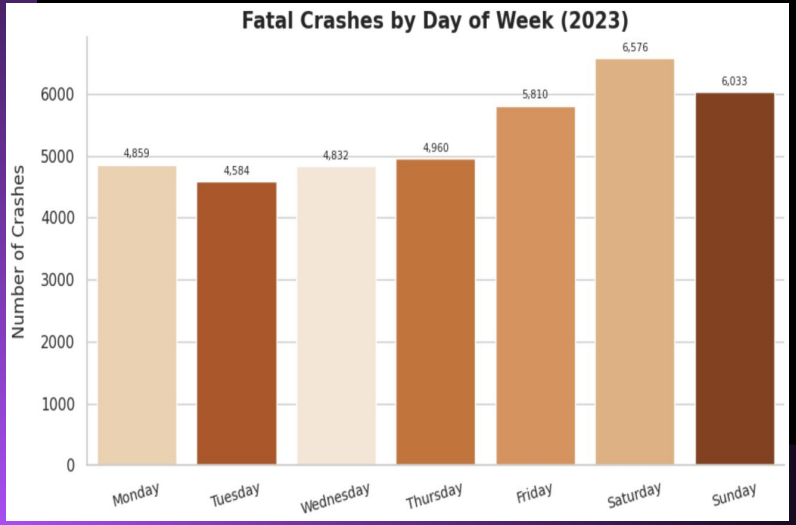


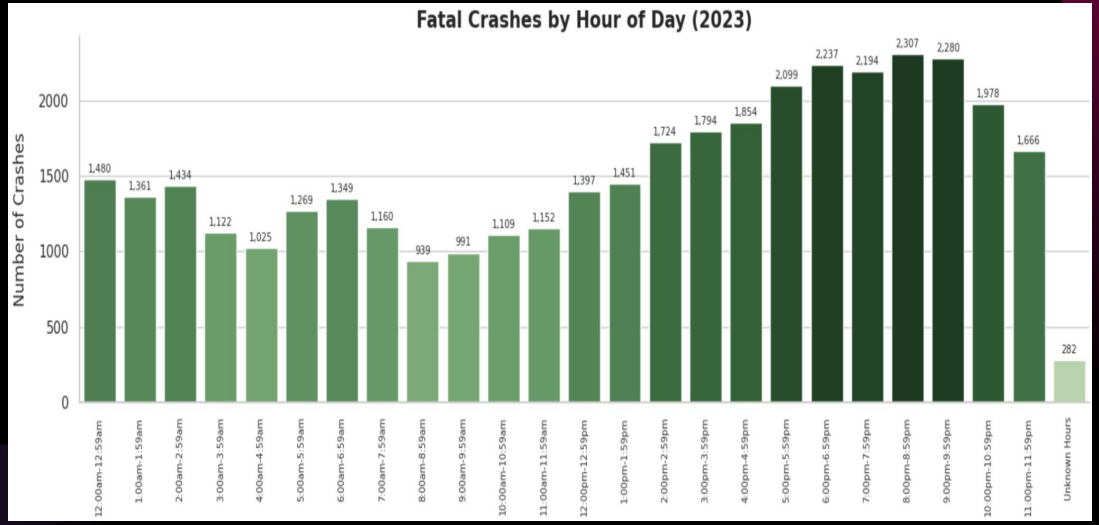
Temporal Crash Dynamics: Month, Day, and Hour Patterns (2023)
Crash frequencies show strong temporal clustering, aligning with high mobility, social, & visibility variance periods, especially weekends & evening commute hours

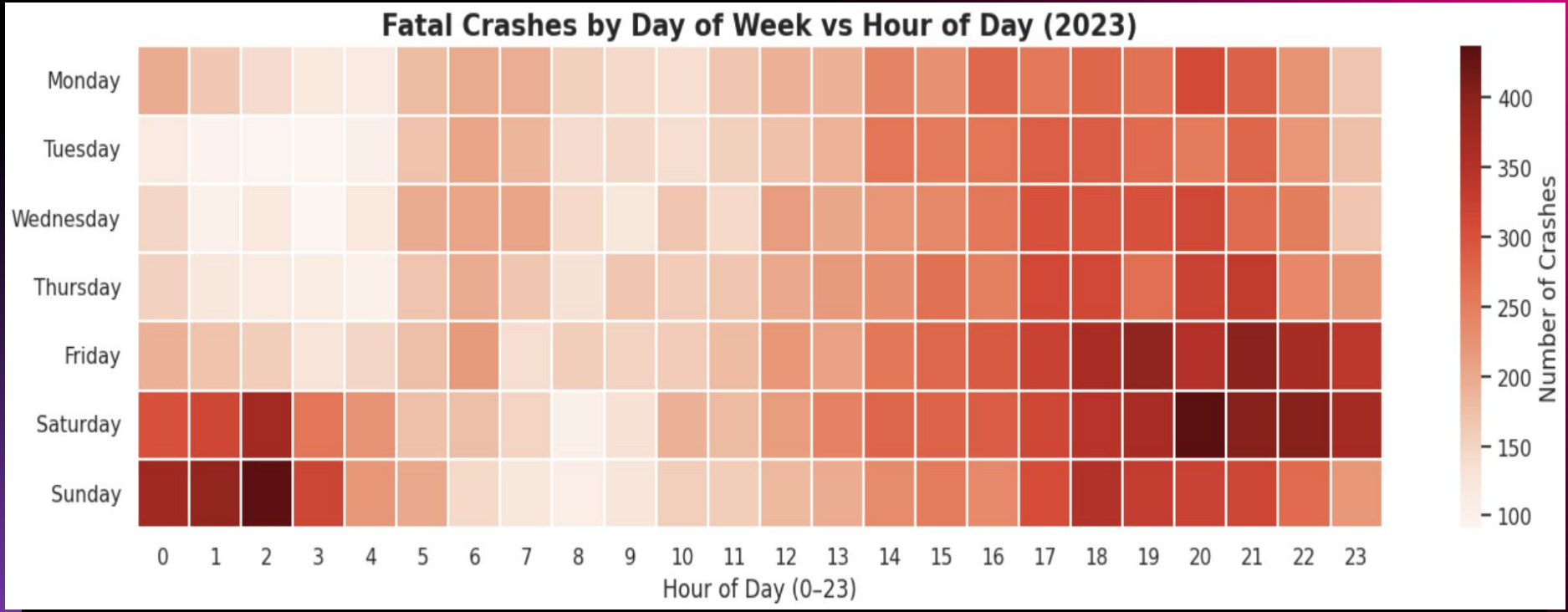
Monthly pattern: Fatal crashes peak during **summer (Jul–Oct)** with ~3.4K–3.5K incidents per month — **≈25%** higher than early-year months — likely linked to vacation traffic and **extended daylight activity**.

Weekly cycle: Saturday alone accounts for **17.5%** of weekly crashes (6,576 cases), confirming a **weekend mobility-risk pattern**, often reinforced by leisure and impaired driving behavior.

Hourly trend: Clear bimodal peaks — **early evening (6–9 PM)** and **post-midnight (12–3 AM)** — align with low visibility and fatigue effects, collectively representing **~30%** of total daily fatalities.





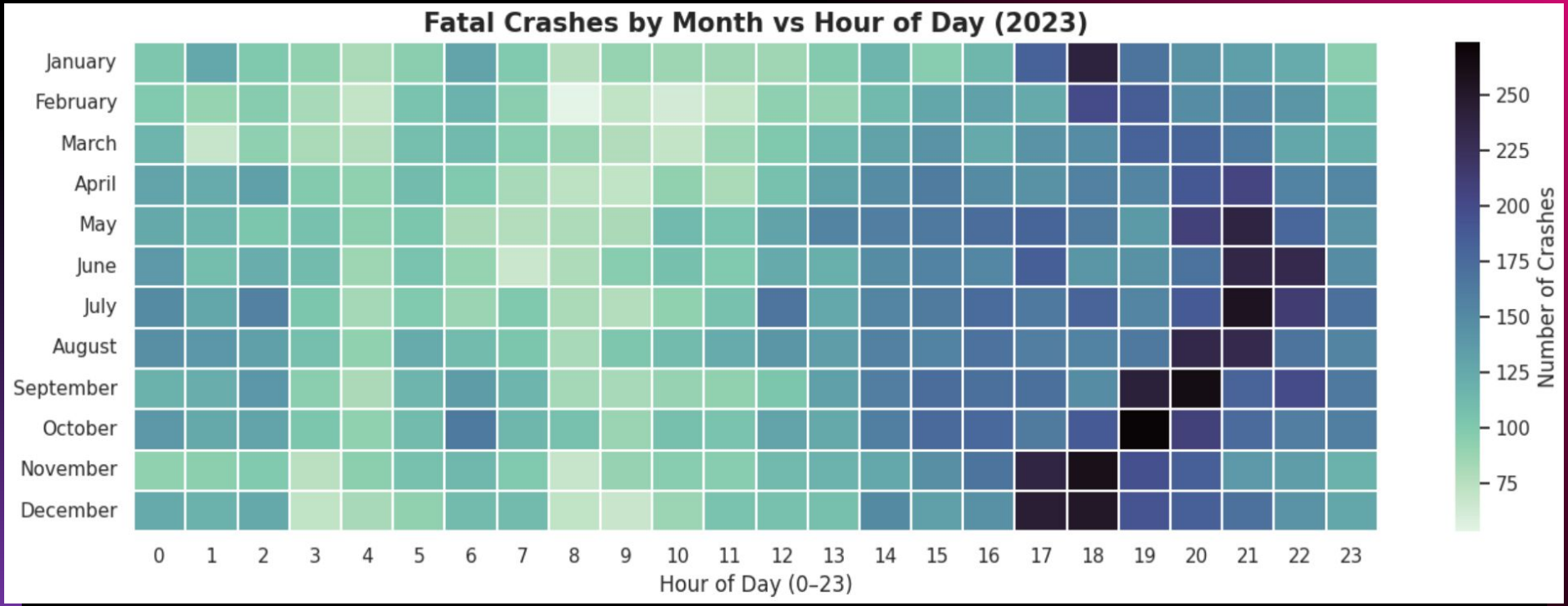


Weekend nights show the highest fatal crash density, especially **between 6 PM – 3 AM**.

Strong temporal association — crashes peak when weekend and nighttime driving coincide.
High-intensity clusters observed:

- **Sunday (12 AM – 3 AM) → late-night peak**
- **Saturday (6 PM – 10 PM) → evening surge**

Friday evenings also show a sharp rise, marking the start of the weekend surge.

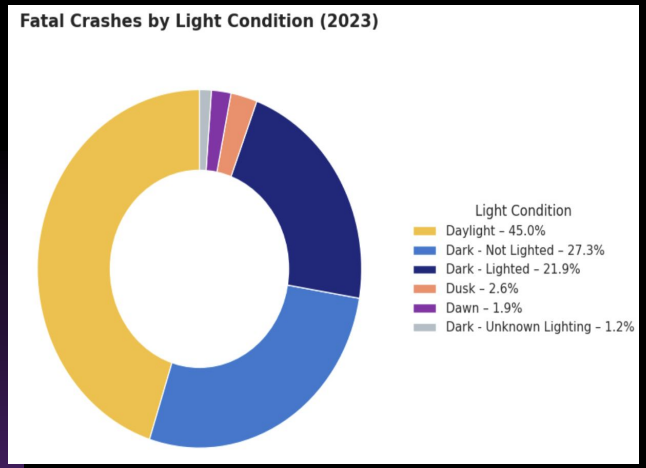


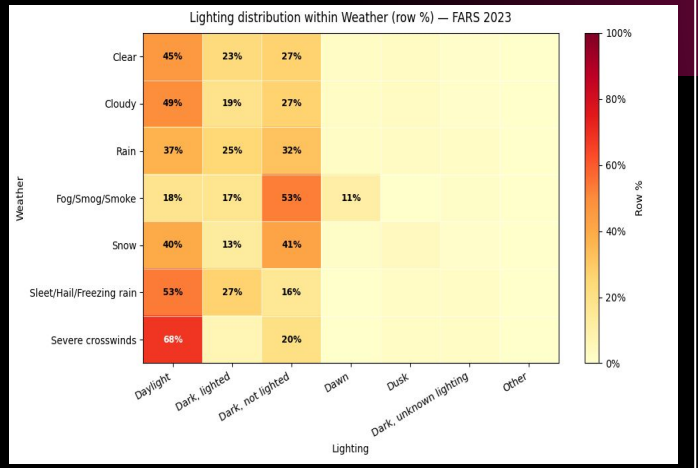
Fatal crashes concentrate sharply during **evening hours (6 PM – 9 PM), peaking in late summer and early fall (August – October).**

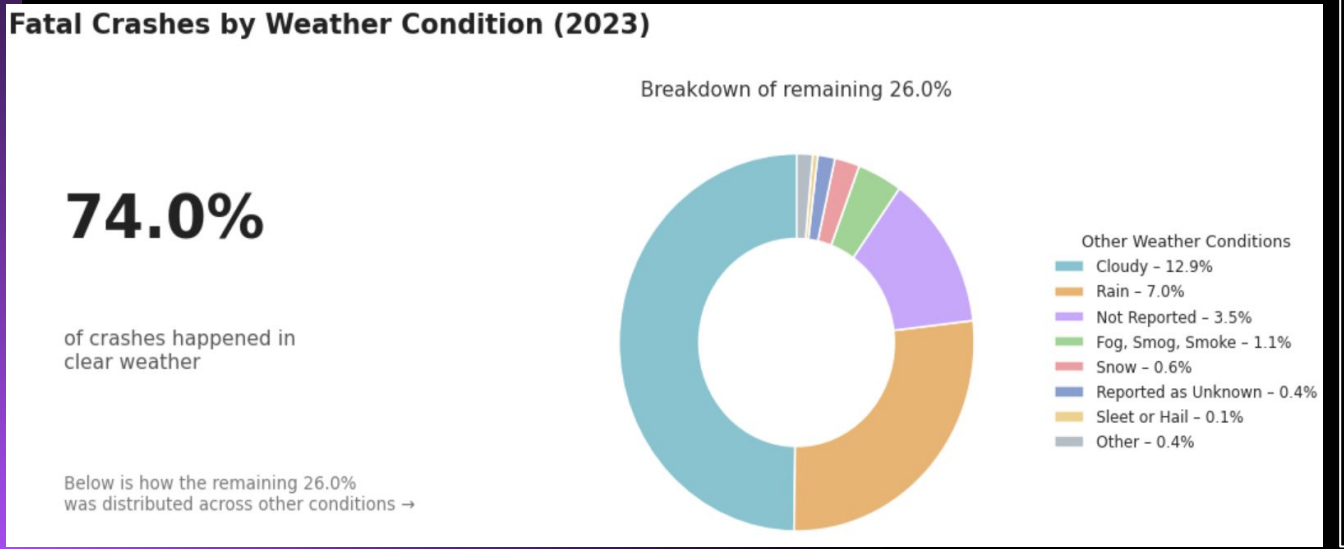
Strong season–time interaction — crash frequency rises when daylight decreases and traffic activity remains high. High-intensity clusters observed:

- **August – October (6 PM – 9 PM) → strongest crash concentration**
- **May – July (5 PM – 8 PM) → moderate evening surge**
- Minimal activity during early-morning hours (12 AM – 6 AM) across all months.

(Key Finding) The **post-summer daylight reduction** coinciding with rush-hour traffic leads to a spike in fatal crashes — highlighting a possible daylight-saving transition effect where reduced visibility and fatigue amplify evening-time risks.







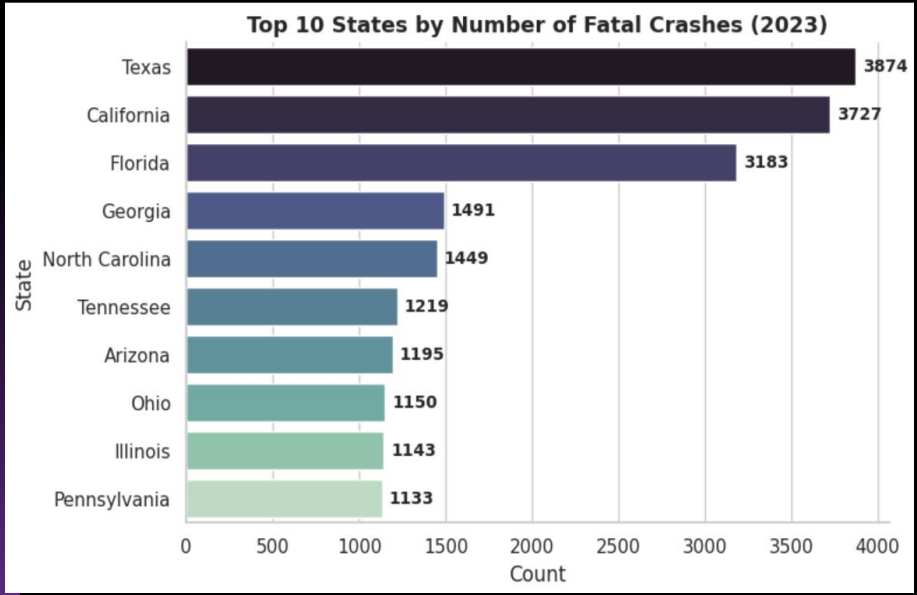
Crashes are not confined to adverse conditions instead, **lighting quality and visibility interplay** with weather to amplify risk in specific contexts.

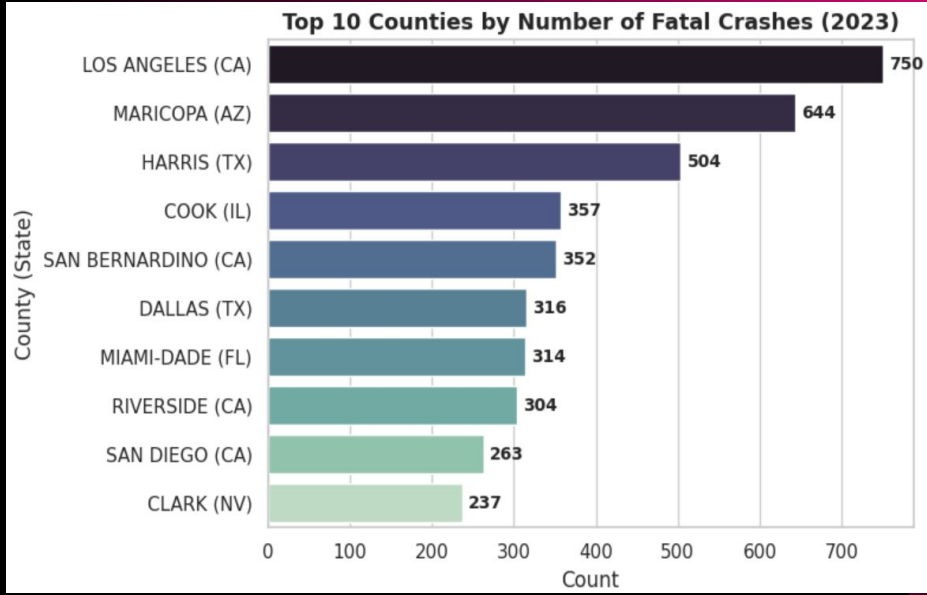
Counter-intuitive pattern: Even with clear weather (74%), fatal crashes remain high implying that normal conditions + high mobility pose greater risk than storms or snow.

Lighting–weather correlation:

- Under fog, smog, or snow, >50% of crashes occur in **dark, unlighted zones, indicating visibility-driven vulnerability**.
- Cloudy and rain conditions show balanced distributions across daylight and dark hours - suggesting **illumination rather than precipitation** is the dominant factor.
- **Anomaly pattern:** Severe crosswinds (68% daylight) mark rare but distinct daytime risk clusters, likely linked to high-speed open-road driving.

(Key Finding) A multi-attribute dependency exists — crash severity spikes when **moderate weather overlaps with low or inconsistent lighting**, especially near daylight-saving transitions, where sudden evening darkness disrupts driver adaptation.



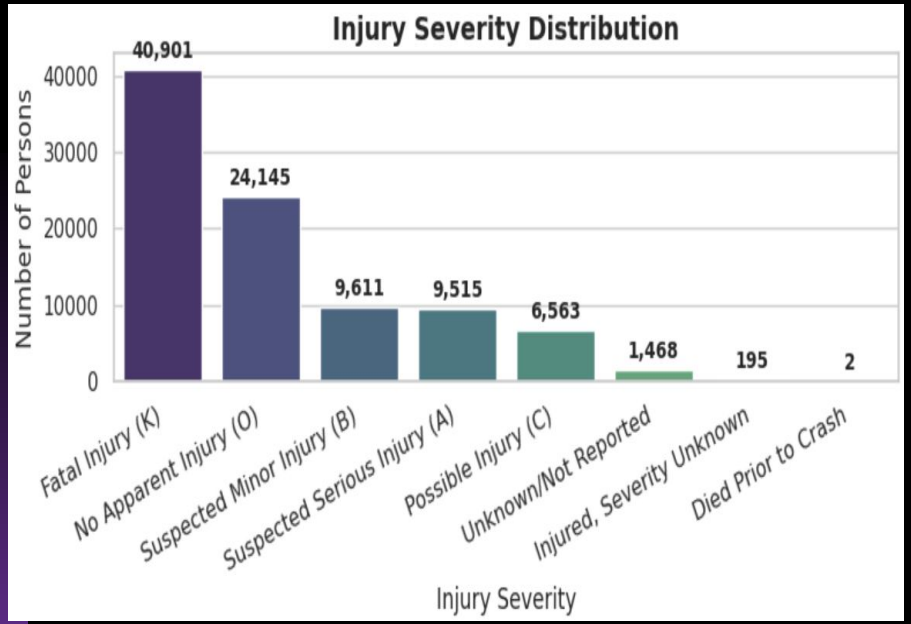


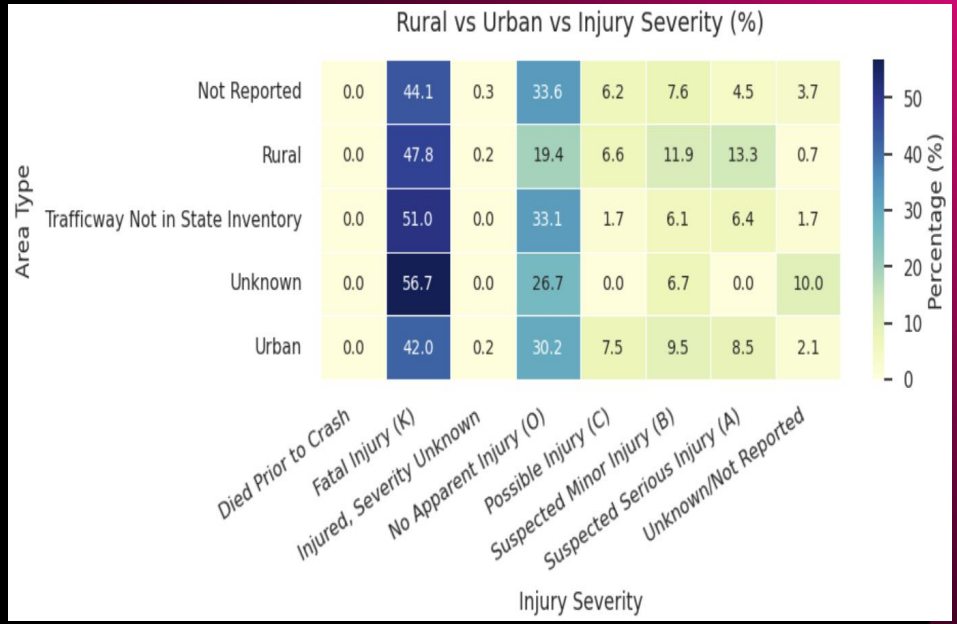
A small cluster of states contributes disproportionately to national fatalities. It reveals heavy right-skewed distribution - **nearly half of national fatalities** originate from just 10 states.

- The top 3 states - **Texas, California, and Florida** - account for **~29%** of all fatal crashes (≈10,784 out of 37,654).
- Top 10 states combined contribute nearly **45%** of total U.S. crashes, underscoring macro-level spatial concentration.
- These states share high mobility, urban density, and extensive interstate corridors, magnifying exposure risk.

Urban mobility hubs form micro-clusters driving state-level crash totals.

- The top 10 counties collectively contribute **~4,000 fatalities**, i.e., about **12%** of total national crashes (37,654).
- **Los Angeles County alone (750 crashes) accounts for ~2% of the U.S.** total, the highest single-county share.
- High-volume metro areas like **Maricopa (AZ), Harris (TX), and Miami-Dade (FL)** show similar spikes, confirming urban traffic density as a key local driver.



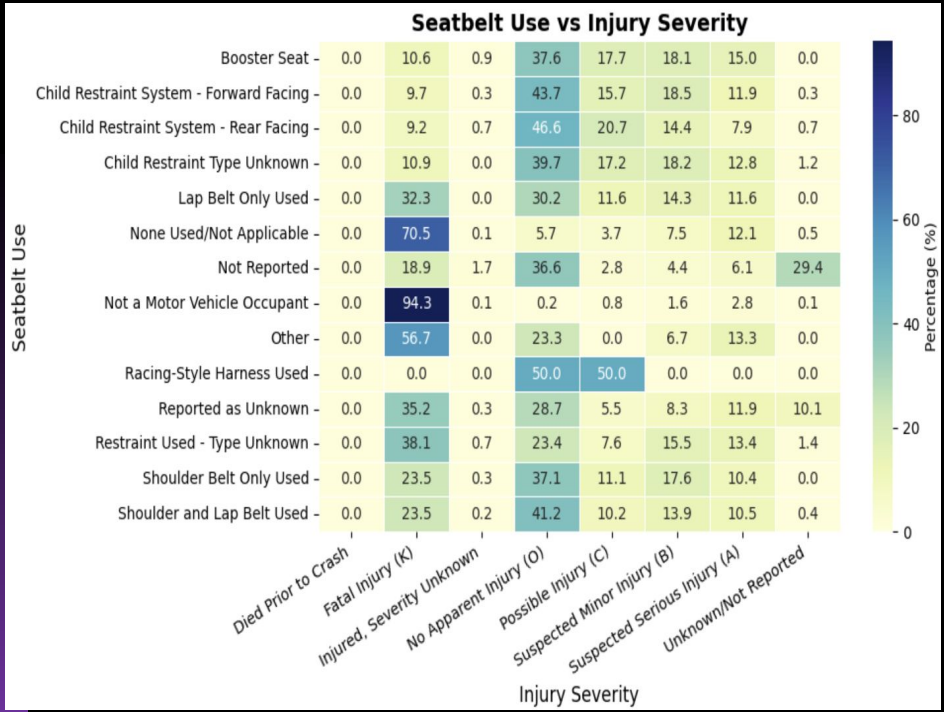


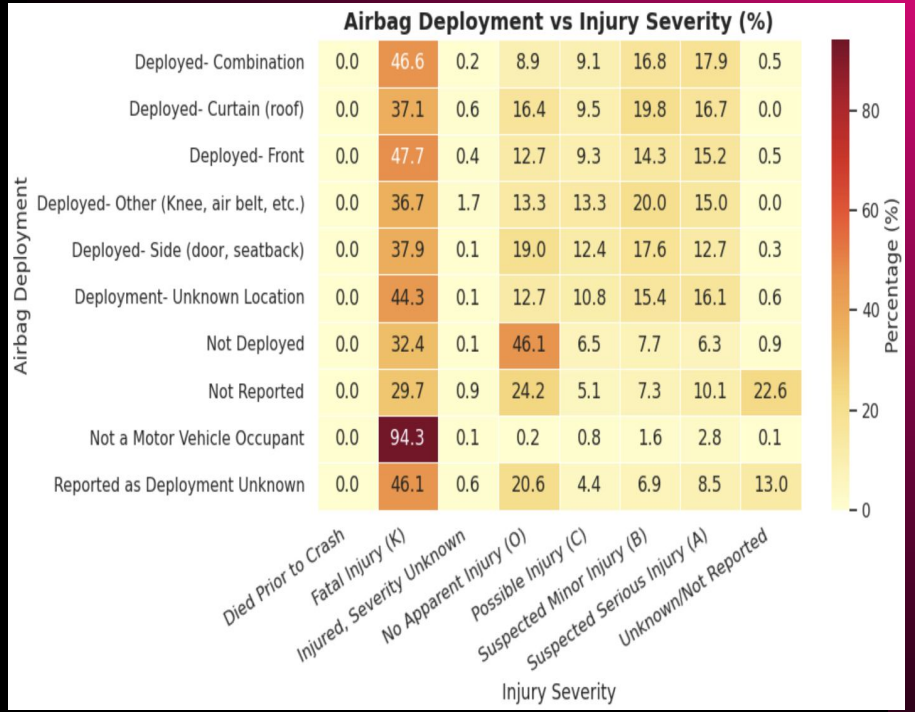
Fatal outcomes dominate crash data, indicating a high severity ratio despite non-fatal incidents being frequent.

- Out of **~92K** persons involved, fatal injuries form **44%** (40,901 cases) – a strikingly high share compared to 26% with no apparent injury.
- Minor and serious injuries (A-B) together account for **~21%**, showing a long-tail severity pattern.
- Distribution is right-skewed toward fatal outcomes, reflecting the dataset's fatal-crash bias (only fatal crash events included).

Rural environments show a sharper fatality gradient than urban areas, indicating speed- and response-related vulnerability.

- Fatal injury share: **Rural – 47.8%, Urban – 42%** → confirming higher lethality outside cities.
- Non-fatal injuries: urban regions show greater diversity (A–C injuries totaling ~25%), likely due to **faster medical access & lower impact speeds**
- Anomaly: "Trafficway not in state inventory" areas **show >50% fatality**, suggesting unmonitored or **infrastructure-poor zones are risk hotspots**.



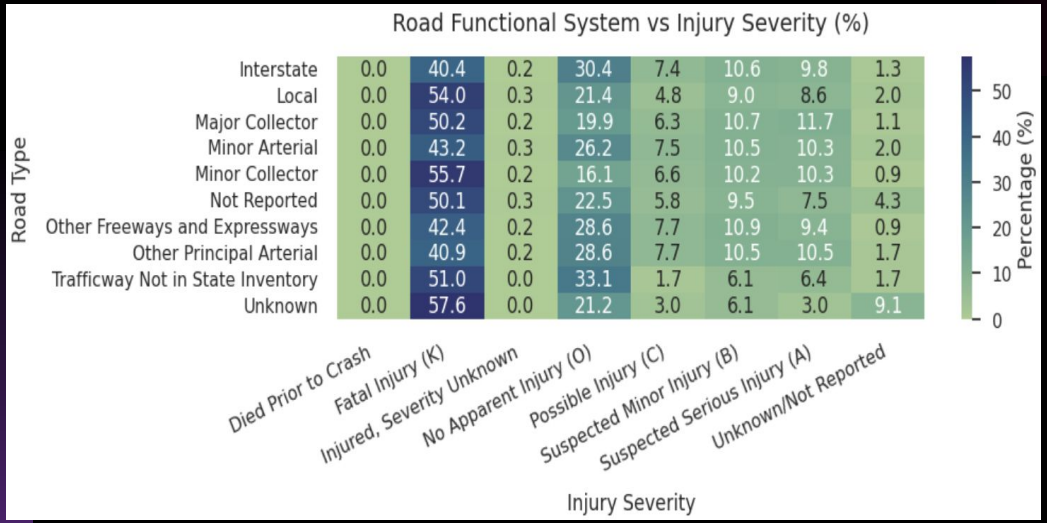


Seatbelt usage shows a strong inverse correlation with fatal injury rates - confirming restraint effectiveness.

- **No seatbelt used:** Fatal injury rate peaks at **~70.5%**, indicating drastically higher lethality.
- **Lap or shoulder-lap belts:** Fatal share drops to **~23–32%**, showing a >50% risk reduction compared to unrestrained occupants.
- **Child restraints (rear/forward facing):** Majority of cases are non-fatal (**40–47%** no injury), confirming correct restraint systems protect effectively

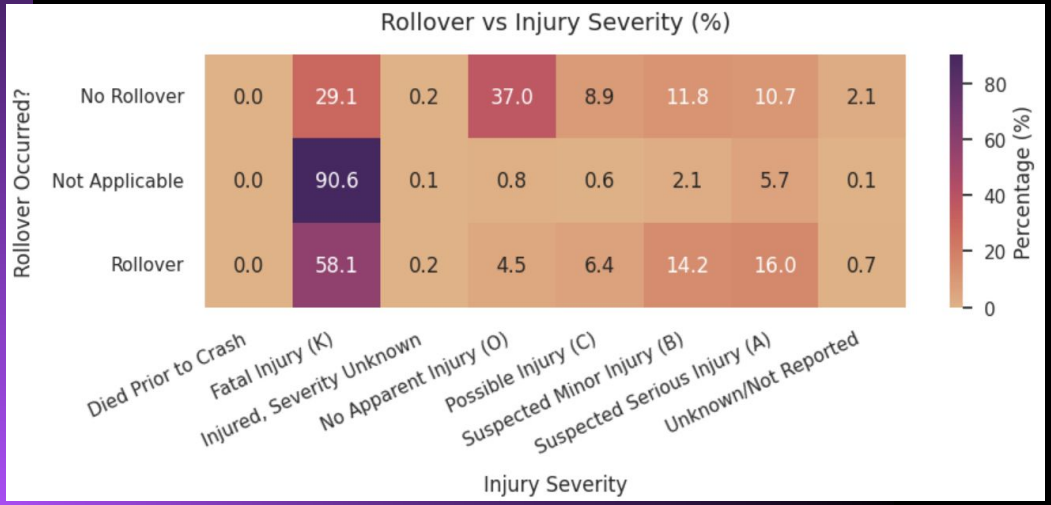
Airbag deployment correlates with moderate injury reduction, but not all types yield equal protection.

- **Front or combination airbags:** Fatal injury share ≈ **46–48%**, lower than non-deployed (32%), but still significant — implying partial mitigation in high-impact crashes.
- **Curtain and side airbags:** Higher proportions of minor/possible injuries **(15–20%)**, showing effective absorption in lateral collisions.
- **No deployment:** Has highest no-injury rate **(46%)**, often due to low-speed or non-triggering impacts — not necessarily safer crashes.



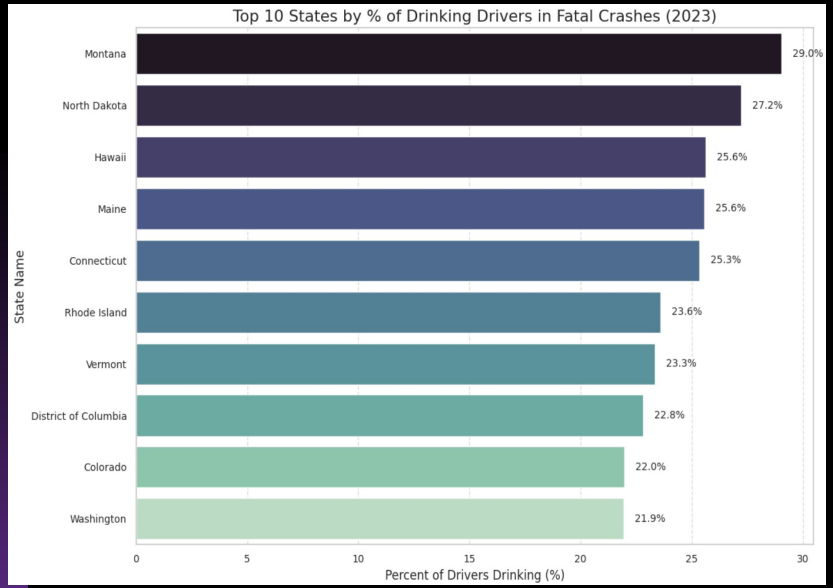
Non-highway and unmonitored road segments carry a disproportionately high share of fatal outcomes — showing that **local infrastructure and data coverage strongly influence crash lethality**.

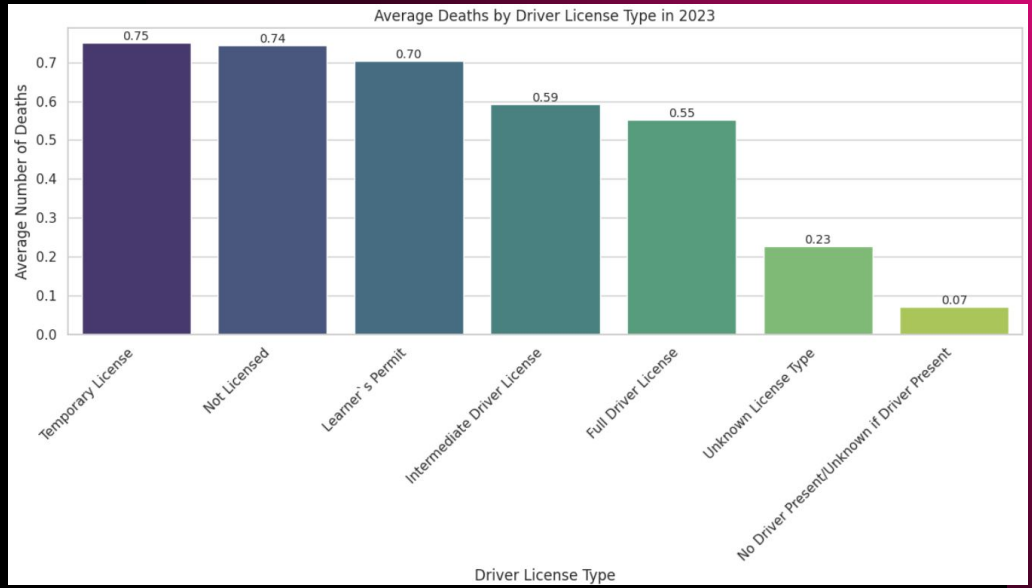
- **Local & minor collector roads:** Highest fatal share **(≈54–56%)** → **weak safety design & limited lighting**.
- **Interstates & arterials:** Lower fatality ratios (**~40–43%**) → benefit from controlled traffic flow.
- Unknown and Not in State Inventory categories also exceed 50%, revealing **data blind spots aligned with risk zones**.

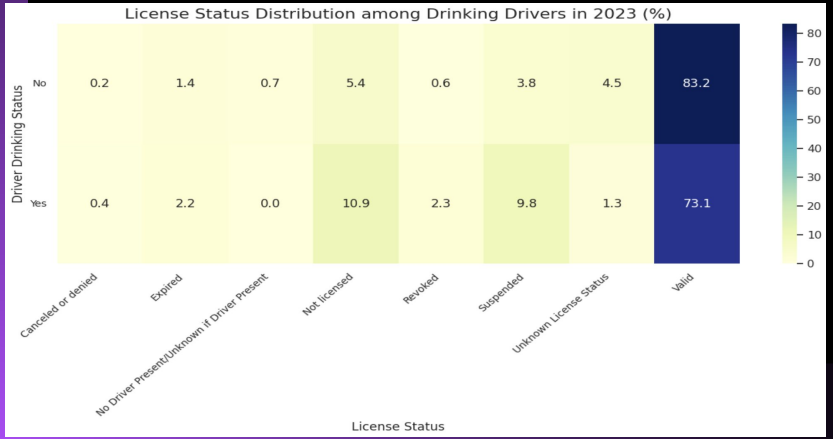


Rollover events are extreme outliers — if a rollover occurs, the **probability of fatal or serious injury more than doubles,** confirming their status as high-lethality crash types.

- **Rollover crashes:**
  Fatality rate ≈ 58%, nearly 2× non-rollover **(29%).**
- **Severe injuries (A/B):** Jump to ~30%, showing massive impact energy transfer.
- **Non-rollover crashes:**
  Broader injury spread → less lethal but more frequent.







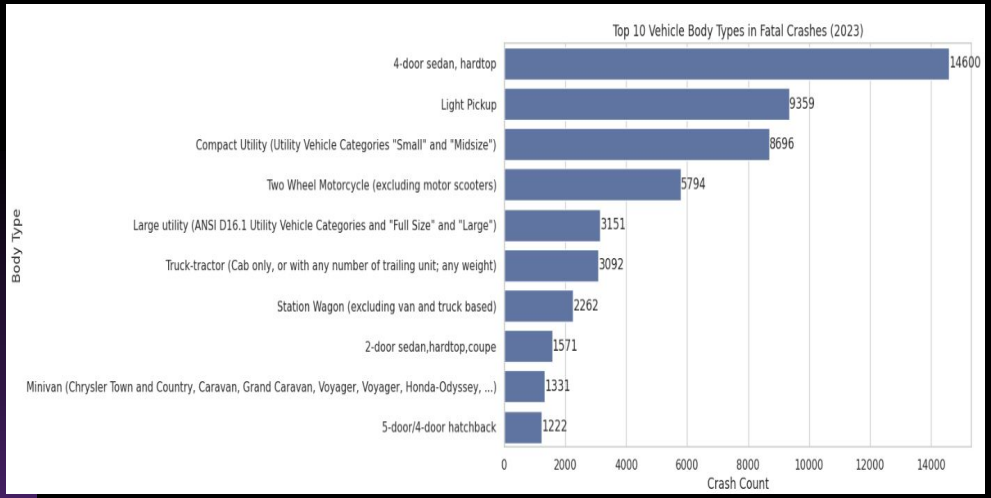
Alcohol Involvement & Licensing Patterns in Fatal Crashes (2023)

Alcohol-related fatalities are not just a behavioral issue — they are strongly tied to licensing irregularities and state-level enforcement gaps, showing that weak control over unlicensed or suspended drivers amplifies fatal crash risk.

Geographical pattern: Northern states like **Montana (29%), North Dakota (27%), and Hawaii/Maine (~25%)** show the highest share of drinking drivers, revealing localized enforcement or cultural tolerance issues.

Licensing correlation: Drivers with temporary or no licenses show **~0.7–0.75** average deaths per case, nearly **40%** higher than fully licensed drivers (~0.55).

Regulatory gap: Over **25%** of drinking drivers were not validly licensed (expired, suspended, or unlicensed), highlighting a policy enforcement failure—license status acts as a hidden risk multiplier for alcohol-related crashes.

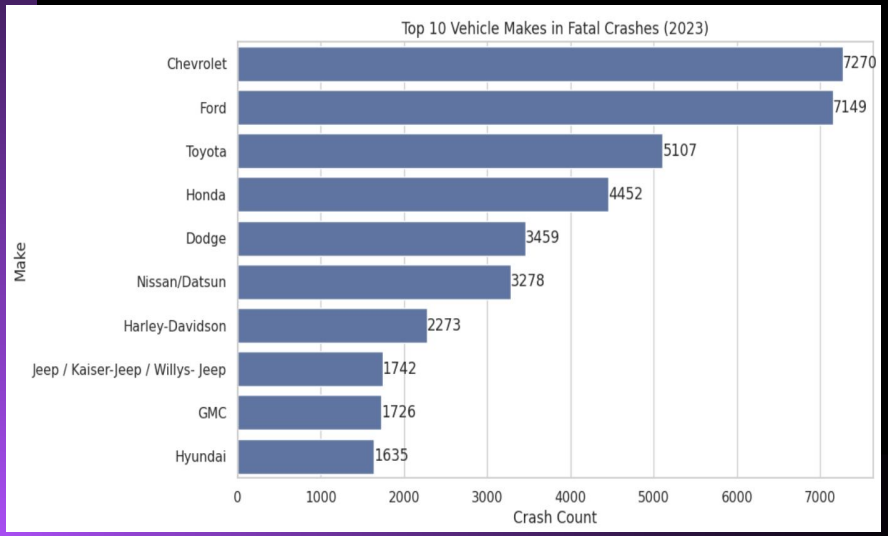


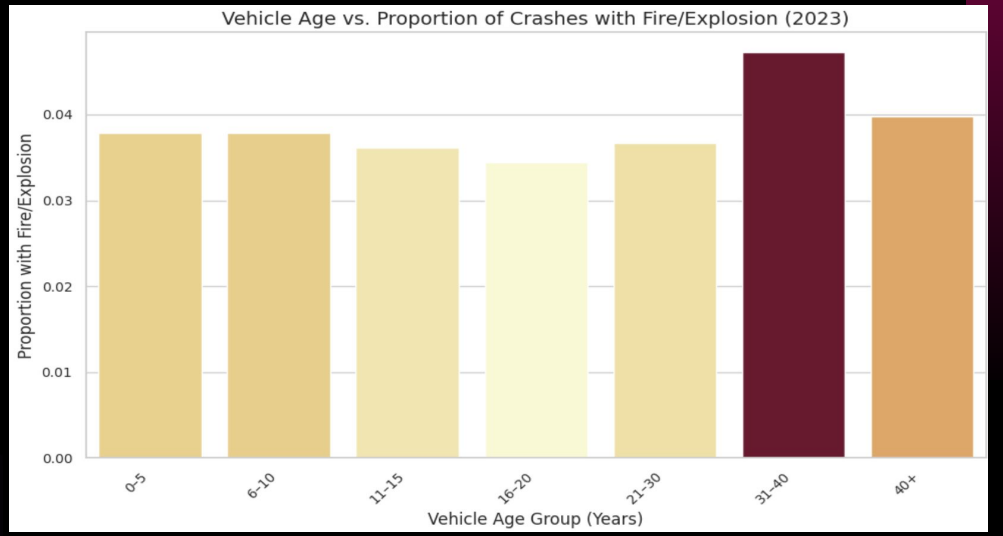
Vehicle Characteristics, Make, and Age Influence on Fatal Crashes (2023)
Fatal crash risk is strongly concentrated in light personal and utility vehicles, with aging fleets showing rising mechanical vulnerability — especially toward fire or explosion incidents in older vehicle groups.

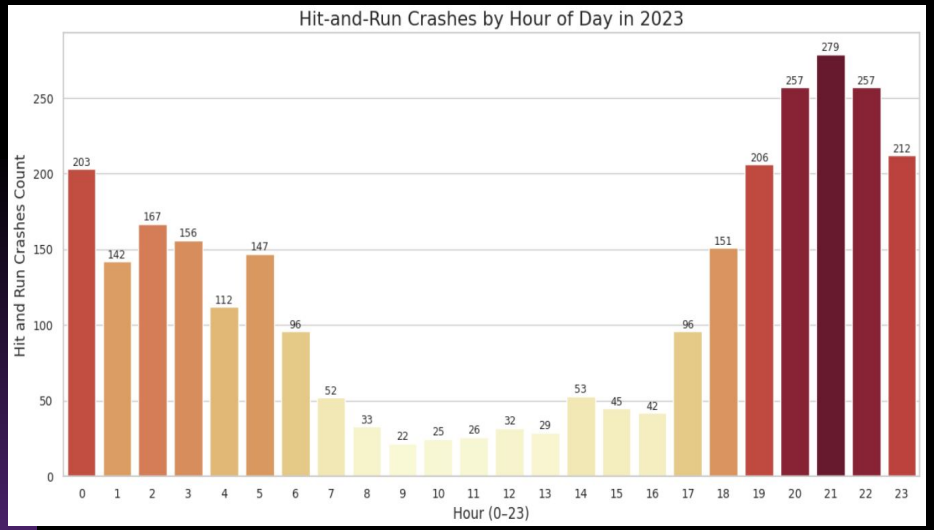
Vehicle body trends: **4-door sedans** and pickups dominate fatal crashes (≈14.6K and 9.3K cases), together accounting for >**55%** of total fatal incidents, reflecting exposure bias from high daily use.

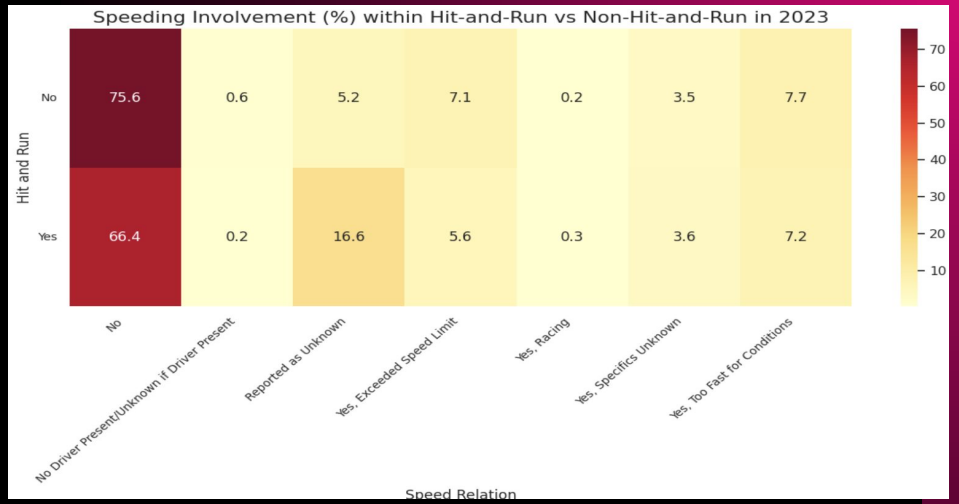
Vehicle make pattern: **Chevrolet, Ford, and Toyota** alone contribute nearly **40%** of all fatal crashes, suggesting both market dominance and potential safety design disparities among high-volume brands.

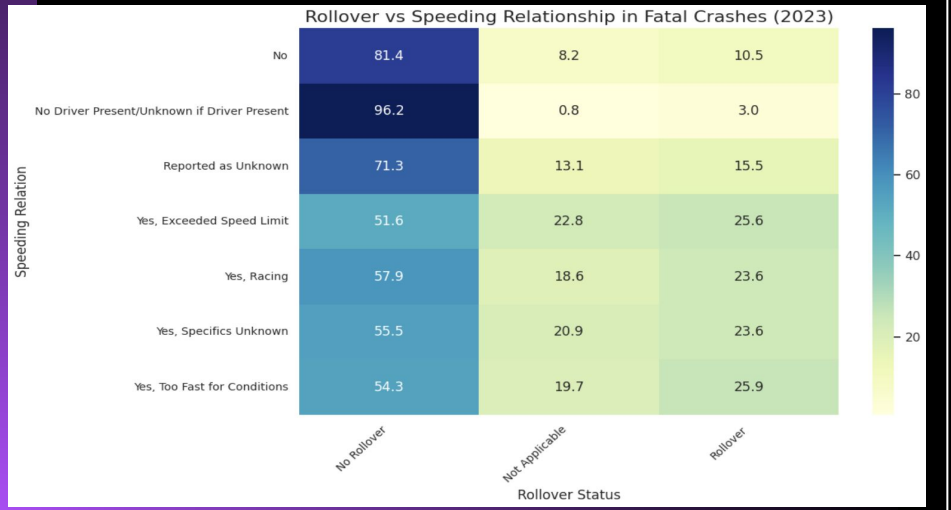
Vehicle age correlation: Fire/explosion probability spikes for 31–40-year-old vehicles (~0.045) — **almost 50% higher than the fleet average** — highlighting risk accumulation from aging components and fuel system degradation.











Temporal, Behavioral, and Dynamic Factors in Hit-and-Run and Speed Linked Fatal Crashes (2023): Hit-and-run and rollover crashes cluster sharply around late-night & early-evening hours, showing a behavioral triad of speeding, impaired driving, & evasion, where speed escalation amplifies rollover probability.

Temporal pattern: Hit-and-run incidents surge at midnight (200+) and again at **8-10 PM** (≈270–280), mirroring peak alcohol and fatigue exposure windows.

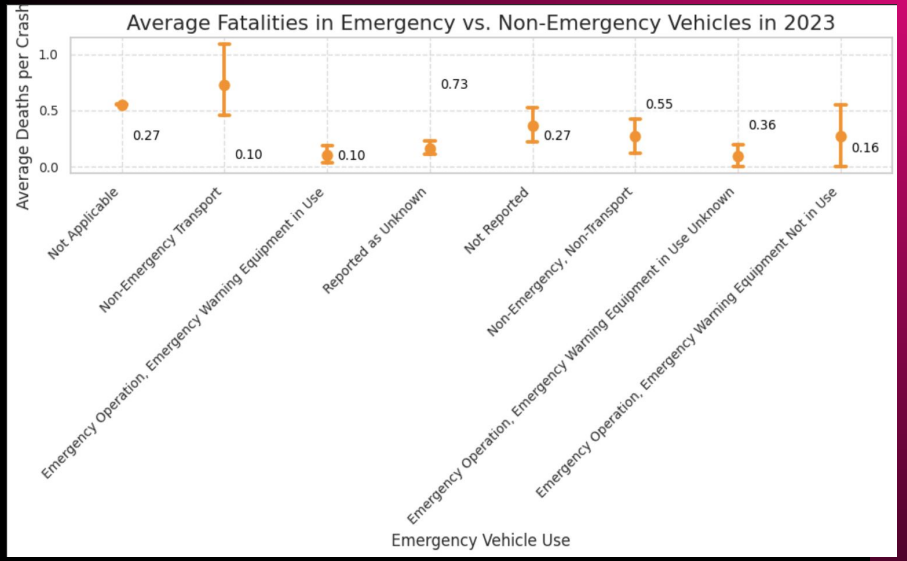
Speeding correlation: Among hit-and-runs, speeding accounts for **~17 %** of reported cases vs **~5 %** in non-hit-and-runs — pointing to riskier, deliberate flight behavior rather than situational loss of control.

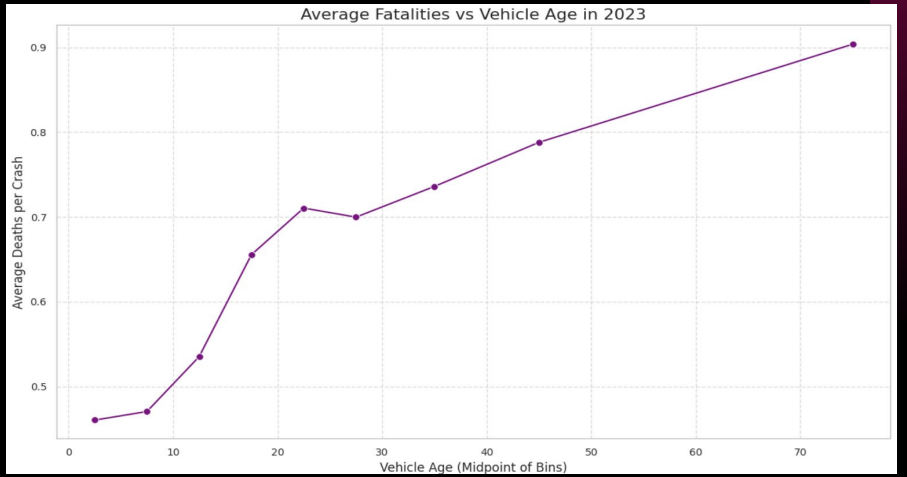
Rollover dynamics: When drivers exceed speed limits, rollover likelihood rises from **~10 % to >25 %**, revealing velocity is a dominant mechanical trigger for fatal instability.

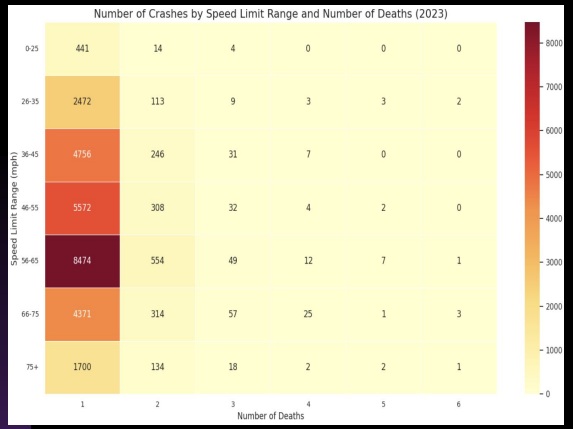
Nighttime hit-and-runs form a distinct **high-risk cluster** where speed, evasion, and rollover dynamics converge, demanding targeted enforcement between **8 PM–2 AM.**

Emergency Operations, Vehicle Integrity, and Aging Risk Factors in Fatal Crashes (2023): Fatality risk is lowest during active emergency operations but rises sharply with vehicle aging & structural deformation, showing that crash survival hinges more on vehicle resilience than situational urgency.

- **Emergency context:** Non-emergency vehicles show the highest fatality average (≈**0.73 deaths/crash**), while vehicles on active emergency duty record just ~0.10–0.16, indicating effective procedural mitigation (sirens, clear lanes).
- **Deformation severity:** Over **65–75%** of 4-door sedans, and motorcycles experience disabling or extensive deformation, confirming that smaller body structures absorb impact less efficiently.
- **Aging factor:** Average fatalities steadily increase with vehicle age, rising from ~0.45 for <10 yrs to ~0.9 beyond 70 yrs, suggesting cumulative wear on safety systems (airbags, crumple zones) as a **silent fatality amplifier**.







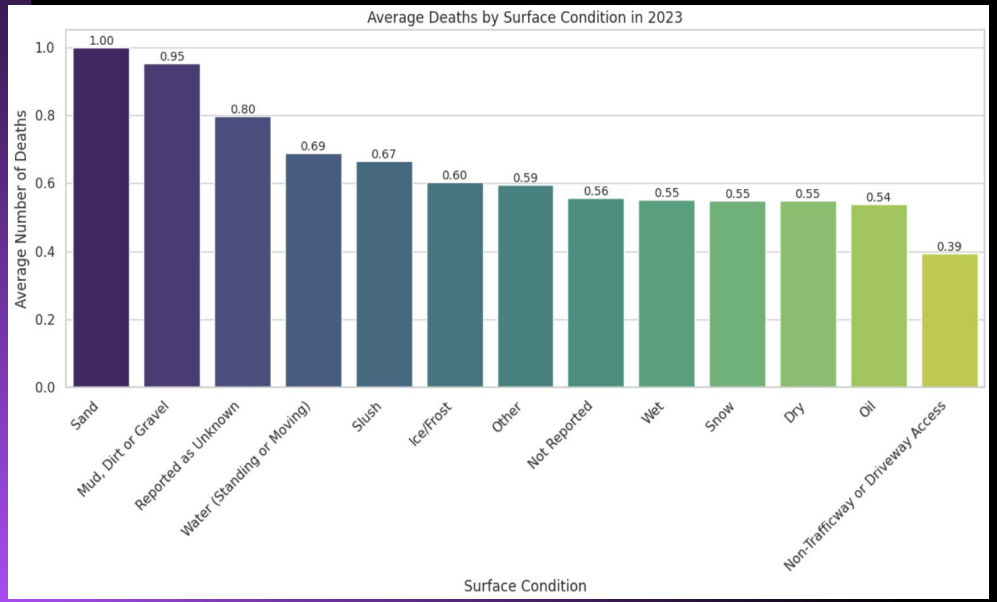
Speed Limits, Surface Conditions, and Road Profile Correlations in Fatal Crashes (2023):

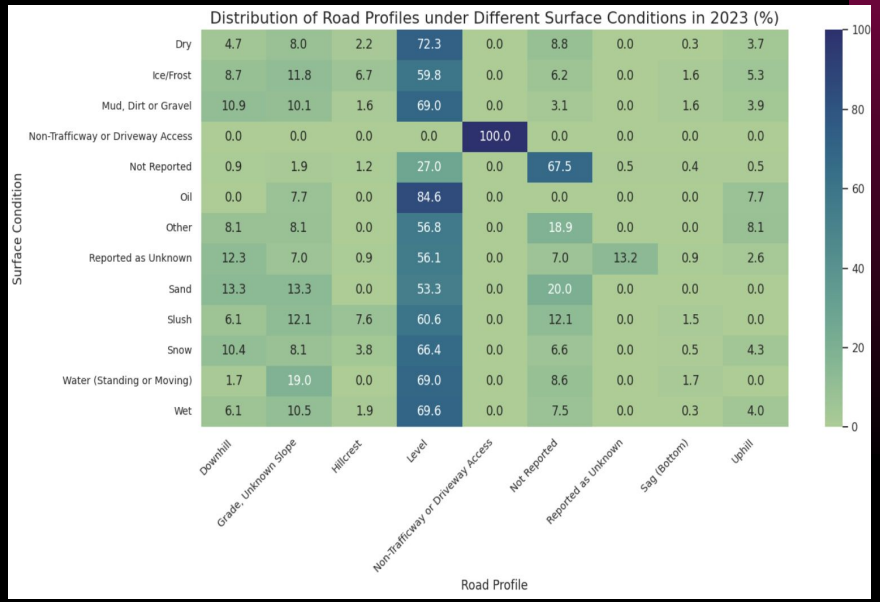
Higher speed limits and unstable surface conditions amplify fatality rates — with sandy, gravel, and uneven terrains showing the strongest interaction between loss of control and environmental instability.

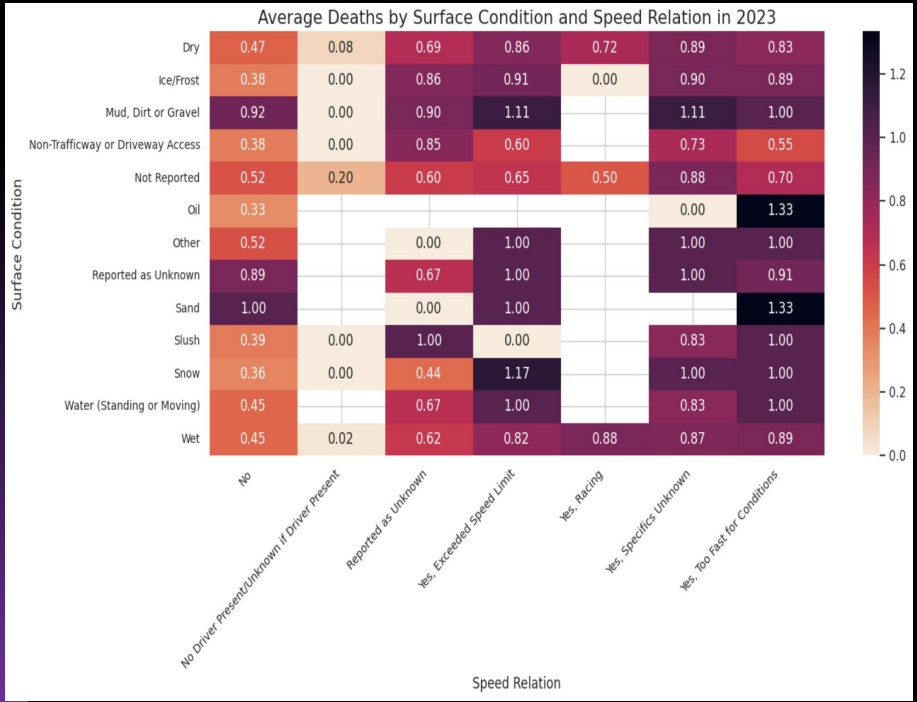
Speed–fatality gradient: Fatalities rise steeply beyond 55 mph, peaking at **~8,400 crashes** in the 65–70 mph band, where even **small visibility or traction losses** become lethal.

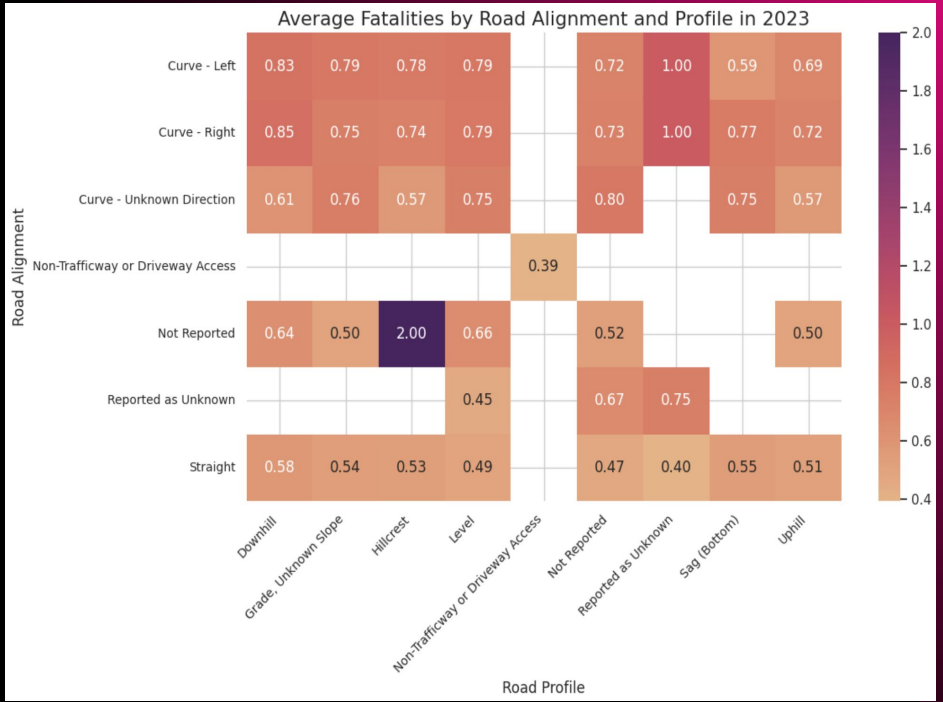
Surface risk pattern: Crashes on **sand and gravel** show the highest average deaths (~1.0), nearly 2× greater than dry or paved surfaces (~0.5), revealing **friction loss** as a critical multiplier.

Profile link: Over **70%** of fatalities on unstable surfaces occur on **level or hillcrest roads**, suggesting reduced reaction opportunity rather than curve-induced errors.









Surface Condition & Speed Relation vs Fatalities (2023): Crash lethality peaks when speeding interacts with poor surface conditions — confirming a compound risk effect between **traction loss and driver behavior**.

- **Mud, gravel, or water-covered roads:** Fatality index rises to ~1.1–1.3× average when paired with speeding or "too fast for conditions."
- **Dry surfaces:** Maintain moderate values (~0.7–0.9) even under speeding, indicating driver misjudgment dominates over surface friction.
- **Ice/frost and snow:** High risk when speed exceeds safe limits, revealing low-friction amplification of crash severity.

Road Alignment & Profile vs Fatalities (2023): Curved and sloped road geometries magnify fatality risk, especially when grade and alignment interact — implying combined geometric instability effects.

- **Curves (left/right):** Maintain consistently higher fatality ratios (~0.8–1.0) than straight roads (~0.5–0.6).
- **Hillcrest and downhill slopes:** Intensify fatal outcomes (values reaching ~1.0–2.0), reflecting visibility loss and momentum buildup.
- **Straight-level segments:** Show lower fatality averages but contribute the most volume-wise — risk dilution due to higher exposure, not safety.

Project Checkpoint Summary and Next Steps

Progress Till Checkpoint

- Built a clean, unified dataset by merging Accident, Vehicle, and Person tables from FARS 2023
- Completed full data preprocessing — cleaning, reduction, integration, and modular organization into crash-, vehicle-, and person-level features.
- Developed a reusable preprocessing framework that keeps schema consistent so the historical FARS years (past 25 to 50) can be easily added for long-term trend analysis.
- Applied key data-mining techniques — association rules, correlation checks, and pattern discovery — to uncover meaningful insights about Temporal, Environmental, Behavioral, Spatial patterns and rules
- So far, we've put major effort into understanding the data and exploring diverse patterns, creating a strong analytical base and frameworl for all future model development and historical data ingestion.

Next Steps (Final Phase)

- **Predictive Modeling:**
  Build models to predict crash severity using regression and tree-based algorithms (Decision Tree, Random Forest)
- **Hotspot Detection:**
  Use Colorado crash data to locate high-risk zones through threshold-based filtering and clustering (DBSCAN/K-Means). (for past 25 to 50 years)
- Validate cluster quality using silhouette scores and interpret key features.
- **Temporal Verification:**
  Compare discovered time patterns (weekends, night hours, holidays) with known traffic risk trends.
- **Model Testing & Visualization:** Evaluate models for accuracy and reliability.
- Present findings through clear visuals — maps, graphs, and tables that highlight severity predictions and hotspot zones.

Goal Moving Forward

Shift from descriptive pattern analysis → predictive and prescriptive modeling, to forecast **crash severity, identify hotspots, and support data-driven road-safety planning**.

THANK YOU