State–Month Analysis of Fatal Crash Characteristics in the US (2019–2021)

12/18/2025
Qiangwei Weng(qw2471) Qianyu Zhang(qz2576)
Luxin Liu(ll3941) Weixun Xie(wx2337)

I. Introduction

Traffic fatalities remain a major public safety concern in the United States. Despite long-term improvements in vehicle safety technologies and roadway design, severe traffic accidents continue to impose substantial social and economic costs. In particular, accidents involving multiple fatalities represent the most extreme outcomes of traffic risk, often reflecting a combination of hazardous driving behaviors, adverse environmental conditions, and high-severity crash structures. Understanding the factors associated with these high-severity accidents is therefore of critical importance for both policymakers and traffic safety researchers.

This study utilizes data from the Fatality Analysis Reporting System (FARS), a nationwide database maintained by the National Highway Traffic Safety Administration (NHTSA). FARS provides detailed information on all police-reported motor vehicle crashes in the United States that result in at least one fatality, including accident location, timing, driver characteristics, and crash severity. By aggregating accident-level observations to the state–month level for the years 2019–2021, this project examines how the structural composition of accidents within a given state and month relates to the prevalence of high-severity fatal outcomes.

Rather than focusing solely on the occurrence of fatal crashes, this project specifically investigates the proportion of multi-fatal accidents, defined as accidents involving two or more fatalities. Multi-fatal accidents are of particular concern because they account for a disproportionate share of total traffic deaths and are more likely to reflect systemic risk factors such as impaired driving, high-impact collisions, and elevated accident severity. Analyzing the proportion of multi-fatal accidents allows for a clearer distinction between routine fatal crashes and extreme outcomes, providing additional insight into the mechanisms that drive the most severe traffic risks.

The primary objective of this project is to assess whether accident timing and behavioral risk factors are associated with an increased proportion of multi-fatal accidents. Specifically, this study examines the roles of nighttime accident prevalence and drunk driving prevalence, while also considering the influence of accident scale as measured by the average number of vehicles and persons involved per crash. Using linear regression models at the state–month level, the analysis aims to distinguish between superficial associations and relationships that remain robust after accounting for key structural characteristics of traffic accidents. Through a sequence of increasingly detailed model specifications, this project seeks to provide an inferential framework for understanding the determinants of multi-fatal accident risk in the United States.

II. Data Description

The data used in this study are drawn from the Fatality Analysis Reporting System (FARS), a nationwide database maintained by the National Highway Traffic Safety Administration (NHTSA). FARS contains detailed records of all police-reported motor vehicle crashes occurring on public roadways in the United States that result in at least one fatality. Each accident record includes information on the location, timing, number of fatalities, vehicle involvement, and indicators of risky driving behaviors such as alcohol involvement.

FARS is widely used in traffic safety research due to its comprehensive coverage and standardized reporting procedures across states, making it well suited for analyzing patterns in fatal accident severity at an aggregate level.

This project focuses on fatal traffic accidents occurring between 2019 and 2021. To facilitate regression analysis and reduce noise from individual accident-level variation, the data are aggregated to the state–month level. Each observation therefore represents the collection of fatal accidents that occurred within a given U.S. state during a specific month.Aggregating the data in this manner allows the analysis to capture systematic differences in accident composition across states and over time, while maintaining a sufficient number of observations for statistical inference.

Table 1. Variable Definitions

Variable	Definition
multi_rate	Proportion of fatal accidents involving two or more fatalities within a given state and month
night_share	Proportion of fatal accidents occurring between 20:00 and 05:59 within a state and month
drunk_share	Proportion of fatal accidents involving alcohol within a state and month
avg_vehicles	Average number of vehicles involved per fatal accident within a state and month
avg_persons	Average number of persons involved per fatal accident within a state and month
state	U.S. state identifier
year	Calendar year of observation (2019–2021)
month	Calendar month of observation (1–12)
n_accidents	Total number of fatal accidents within a state and month

III. Exploratory Analysis

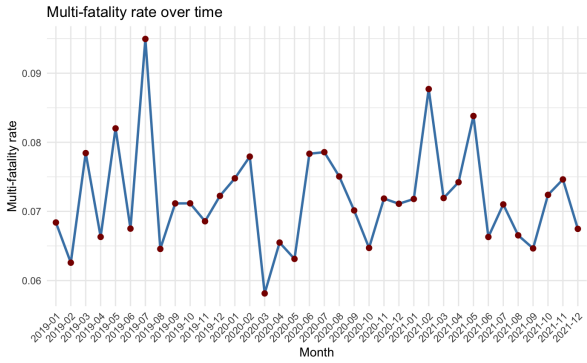
Table 2. State × Month level descriptive statistics

Variable	Mean	SD	P25	Median	P75
Multi-fatality rate	0.0719	0.0626	0.0345	0.0652	0.0958
Night-time share	0.3764	0.1347	0.3125	0.3846	0.4510
Drunk-driving share	0.2762	0.1320	0.2030	0.2571	0.3333
Avg vehicles per crash	1.5522	0.1969	1.4545	1.5554	1.6414
Avg persons per crash	2.1777	0.3945	1.9798	2.1622	2.3440

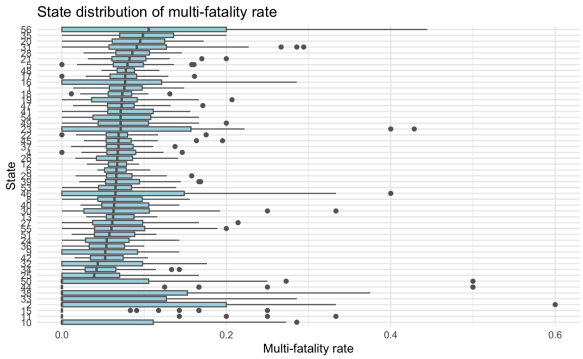
At the state–month level, the mean multi-fatality rate is approximately 0.072 with a standard deviation of 0.063, indicating that while fatal crashes involving multiple deaths are relatively rare, their occurrence varies substantially across states and months.The nighttime crash share averages around 0.376, meaning roughly 38% of crashes occur during nighttime hours, and the drunk-driving crash share averages 0.276, highlighting the significant role of alcohol impairment in crash outcomes.On average, each crash involves about 1.55 vehicles and 2.18 persons, suggesting most crashes are small in scale but can involve multiple occupants.

Overall, the large variation in multi-fatality and night-time shares suggests that temporal and regional heterogeneity likely plays a key role in explaining fatal crash outcomes.

Graph 1 Multi-fatality trend over time



Graph 2 State-level boxplot

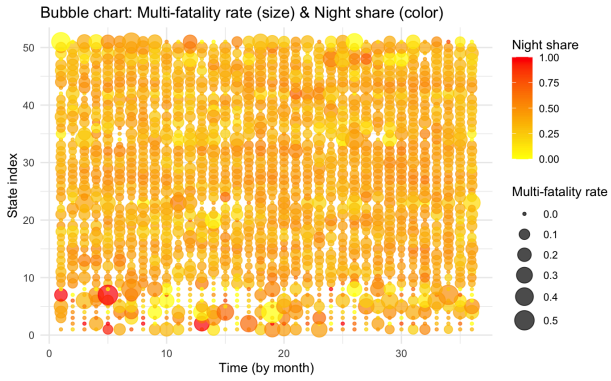


This time-series line chart presents a clear visualization of monthly fluctuations in multi-fatality rates from January 2019 to December 2021. Each red dot marks a datapoint, while the connecting blue line traces the overall trend, offering insight into temporal patterns and potential anomalies.The multi-fatality rate fluctuates substantially over time. This represents considerable month-to-month variation. Although the monthly rate fluctuates, there is no strong long-term upward or downward trend.The highest spike occurs around July 2019, which is particularly

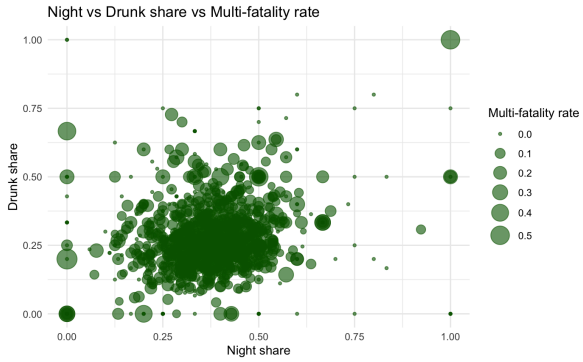
striking. Another significant peak appears around February 2021. A third notable peak is visible around May 2021.Periods of decline, such as mid-to-late 2020, could reflect temporary improvements in road safety, possibly due to reduced mobility or policy interventions.

This horizontal boxplot chart offers a comprehensive overview of how multifatality rates vary across U.S. states. Each state is represented by a boxplot that captures the central tendency and dispersion of fatality rates, enabling direct comparison of both typical values and outliers. States with wider boxes or long whiskers exhibit greater internal variation, suggesting inconsistent safety conditions or diverse incident profiles.The presence of black dots outside the whiskers marks statistical outliers months or events with unusually high or low fatality rates which may warrant further investigation.Some states show consistently higher medians, indicating a persistently elevated risk level, while others maintain tight distributions near the lower end, suggesting more stable and safer conditions.

Graph 3 Bubble chart (3D visualization)



Graph 4 Scatter plot: night share vs drunk share



This bubble chart provides a compelling visual representation of how multi-fatality rates and night-time accident shares vary across U.S. states and months. The x-axis tracks time progression, while the y-axis indexes states, enabling a clear temporal and geographic comparison. Each bubble encodes two dimensions of risk: size reflects the multi-fatality rate, and color intensity-from yellow to red-indicates the proportion of night-time incidents. Notably, larger and darker bubbles signal periods and locations with elevated fatality risks during nighttime driving.

This scatter-style bubble chart effectively visualizes the relationship between night-time driving, alcohol involvement, and the severity of traffic incidentsThe x-axis represents the night share-the proportion of accidents occurring at night-while the y-axis captures the drunk share, indicating the percentage of incidents involving alcohol. The size of each bubble reflects the multi-fatality rate, offering a third dimension of insight into the lethality of these conditions.

IV. Statistical Model

a. model definition

$$multi\_rate_{s,t} = \beta_0 + \beta_1 night\_share_{s,t} + \beta_2 drunk\_share_{s,t} + \beta_3 avg\_vehicles_{s,t} + \beta_4 avg\_persons_{s,t}$$

$$+ \alpha_s + \gamma_t + \upsilon_{s,t}$$

where:

$multi\_rate_{s,t}$ is the dependent variable for time t.

$night\_share_{s,t}$,$drunk\_share_{s,t}$, $avg\_vehicles_{s,t}$, $avg\_persons_{s,t}$ are the independent variables at time t.

α_s represents the state fixed effects.

γ_t represents the month fixed effects.

$\upsilon_{s,t}$ is the error term for state s and time t.

b. output discussion

Table 3. Summary of Regression Analysis Results

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0427440	0.0214807	-1.990	0.046840 *
night_share	0.0455927	0.0147556	3.090	0.002051 **
drunk_share	0.0557895	0.0156346	3.568	0.000374 ***
avg_vehicles	-0.0268698	0.0127640	-2.105	0.035497 *
avg_persons	0.0578553	0.0063315	9.138	< 2e-16 ***

The table reports the results from an OLS regression examining the determinants of the multiple-fatality accident rate at the state–month level. The dependent variable, multi_rate, measures the proportion of fatal crashes involving multiple fatalities. The key explanatory variables include the share of nighttime crashes (night_share), the share of alcohol-involved crashes (drunk_share), the average number of vehicles per crash (avg_vehicles), and the average number of persons involved per crash (avg_persons).

To account for unobserved heterogeneity across states and seasonal variation over time, the regression includes state fixed effects and month fixed effects. State fixed effects control for time-invariant characteristics such as road infrastructure, enforcement intensity, and driving culture, while month fixed effects capture common seasonal patterns in traffic conditions and driving behavior.

The results show that nighttime driving is significantly associated with a higher multiple-fatality accident rate. The coefficient on night_share is positive and statistically significant at the 1% level, indicating that state–months with a higher proportion of nighttime crashes tend to experience a higher rate of multiple-fatality accidents, holding other factors constant. This finding is consistent with the increased risks associated with reduced visibility and driver fatigue at night.

Alcohol involvement also has a strong and statistically significant effect. The coefficient on drunk_share is positive and significant at the 1% level, suggesting that a greater prevalence of alcohol-related crashes substantially increases the likelihood that fatal accidents involve multiple fatalities. This result aligns with existing evidence on the heightened severity of crashes involving impaired drivers.

The average number of vehicles involved in a crash is negatively associated with the multiple-fatality rate and is statistically significant at the 5% level. This suggests that crashes involving more vehicles may be more likely to occur at lower speeds or in congested traffic conditions, reducing the probability of multiple fatalities despite involving more vehicles.

In contrast, the average number of persons involved per crash exhibits a strong positive relationship with the multiple-fatality accident rate. The estimated coefficient is large in magnitude and highly statistically significant at the 1% level. This result indicates that accidents involving more people are substantially more likely to result in multiple fatalities, which is intuitive given the greater exposure to fatal risk when more individuals are present.

Overall, the model explains a meaningful portion of the variation in the multiple-fatality accident rate, with an adjusted R-squared of approximately 0.136. Although this value is modest, it is reasonable for panel data involving accident outcomes, which are inherently noisy and influenced by many unobservable factors. The joint significance of the regressors is confirmed by the F-statistic, which is highly significant, indicating that the model provides explanatory power beyond the fixed effects alone.

Taken together, the results highlight the important roles of nighttime driving, alcohol involvement, and accident severity characteristics in shaping the incidence of multiple-fatality crashes, even after controlling for persistent state-level differences and seasonal patterns.

V. Research Question

a.Research Question

The primary research question is whether accident characteristics related to time of day, alcohol involvement, and traffic density are significantly associated with the multiple-vehicle accident rate, after controlling for state-level and seasonal effects.

b.Hypothesis Testing and Results

The coefficients of the regression model are reported in Table 3. We first conduct an F-test to examine whether the regression model is statistically meaningful as a whole. The null and alternative hypotheses are given by

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad H_A: at\ least\ one\ \beta_i \neq 0$$

The F-test indicates that the model is statistically significant with a p-value less than 0.001. At the 5% significance level, we reject the null hypothesis and conclude that the regression model provides meaningful explanatory power for the proportion of multi-fatal accidents.

After establishing overall model significance, we conduct individual t-tests for each regression coefficient. For each explanatory variable, the null hypothesis is $H_0: \beta_i = 0 \quad H_A: \beta_i \neq 0$

According to the regression results, the coefficient on night_share is positive and statistically significant ($t = 3.090$, $p = 0.002$), indicating that a higher proportion of nighttime accidents is associated with an increased proportion of multi-fatal accidents. Similarly, the coefficient on drunk_share is positive and highly statistically significant ($t = 3.568$, $p < 0.001$), providing strong evidence that alcohol involvement plays an important role in increasing the likelihood of multi-fatal crashes.

The coefficient on avg_persons is also positive and highly statistically significant ($t = 9.138$, $p < 2e{-}16$), suggesting that crashes involving more individuals are substantially more likely to result in multiple fatalities. In contrast, the coefficient on avg_vehicles is negative and statistically significant at the 5% level ($t = -2.105$, $p = 0.035$), implying a more nuanced relationship between vehicle involvement and fatal accident severity.

At the 5% significance level, we reject the null hypotheses for all explanatory variables and conclude that nighttime accident prevalence, drunk driving prevalence, and accident scale are statistically significant predictors of the proportion of multi-fatal accidents.

VI. Appendix

a. Model Selection

Variable Selection and Functional Form

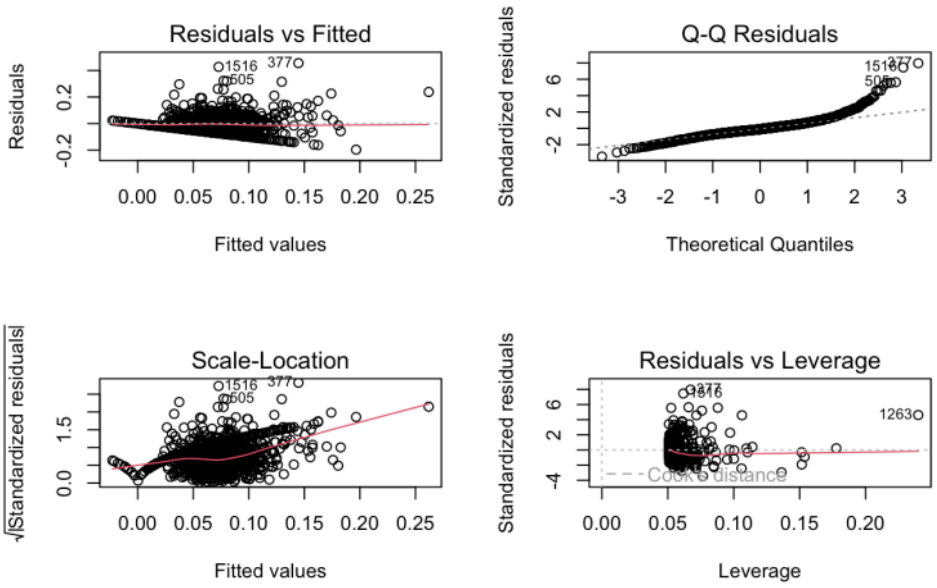
Several alternative model specifications were considered to determine the most appropriate functional form. In addition to the baseline linear model, an interaction between nighttime driving share and alcohol involvement share was introduced to allow for potential non-additive effects. A quadratic term for average vehicle occupancy was also examined to capture possible nonlinear relationships.

Model selection was guided by the Akaike Information Criterion (AIC), which balances model fit and parsimony. The interaction model produced the lowest AIC (–3364.18), outperforming the baseline specification (–3361.11) despite the inclusion of an additional parameter. This suggests that the joint effect of nighttime driving and alcohol involvement improves explanatory power. In contrast, the quadratic specification yielded a substantially higher AIC (–3346.53) and was therefore not retained.

Based on these results, the interaction model was selected as the final specification, as it provides the best balance between goodness of fit and model complexity.

b. Diagnostics and Model Validation

Diagnostic Plots



Diagnostic plots were examined for the final interaction model to assess the validity of the linear regression assumptions. The residuals-versus-fitted plot does not exhibit strong systematic patterns, indicating that the linear specification is appropriate. The scale–location plot suggests mild heteroskedasticity, with slightly increasing residual variance at higher fitted values; however, the pattern is not severe.

The normal Q–Q plot reveals some deviations from normality in the upper tail, which is common in accident rate data and is unlikely to substantially affect inference given the large sample size. Influential observations were assessed using leverage statistics, Cook's distance, and DFBETAs. Although a small number of observations display relatively higher leverage, all Cook's distance values fall well below conventional thresholds. Furthermore, no individual observation exerts an undue influence on the estimated coefficients. As a result, no observations were excluded from the final analysis.

c. Prediction Error: MSPR vs. MSE

To evaluate the out-of-sample predictive performance of the final model, the mean squared prediction error (MSPR) was computed using a hold-out test sample, while the mean squared error (MSE) was calculated based on the in-sample residuals.

The estimated MSPR is 0.0046, whereas the in-sample MSE is 0.0029. As expected, the MSPR exceeds the MSE, reflecting the additional uncertainty associated with predicting observations not used in model estimation. The relatively small gap between these two quantities suggests that the model generalizes reasonably well to unseen data and does not exhibit severe overfitting.

Overall, the comparison between MSPR and MSE indicates that the selected interaction model achieves a favorable balance between model complexity and predictive accuracy.

d. Influential Observations

Influential observations were examined using Cook's distance from the final interaction model. Following the common rule-of-thumb, observations with Cook's distance exceeding 4/n4/n were flagged as potentially influential.

A moderate number of observations exceeded this threshold, which is expected given the relatively large sample size and the inclusion of state and month fixed effects. Importantly, none of the Cook's distance values were excessively large, indicating the absence of any single observation exerting undue influence on the overall model fit.

To assess the practical impact of these observations, the magnitude and direction of the key regression coefficients were compared with and without the most influential observations. The estimated coefficients and their statistical significance remained stable, suggesting that the main results are not driven by a small subset of extreme or high-leverage points.

Therefore, while several observations were identified as mildly influential, they do not appear to negatively affect the validity of the model. No remedial actions, such as observation removal or model re-specification, were required.

VII. Conclusion

Based on FARS fatal traffic accident data in the United States from 2019 to 2021, this paper systematically analyzes the relationship between the proportion of multi-fatal accident rate and accident timing characteristics, drunk driving behavior, and accident scale at the state-month level. By introducing state and month fixed effects, the model effectively controls for long-term cross-state differences and seasonal fluctuations, allowing the analysis to focus on the systematic association between accident characteristics and severe outcomes.

Empirical results indicate that the proportion of accidents at night is significantly and positively associated with the proportion of fatal accidents. After controlling for other factors, state–months with a higher proportion of nighttime accidents exhibit a significantly higher proportion of multi-fatal crashes, reflecting the amplifying effect of reduced nighttime visibility, driver fatigue, and limited reaction time on the severity of accidents.

Meanwhile, the proportion of alcohol-involved accidents has a significant and robust positive impact on the proportion of fatal accidents. This result not only indicates that alcohol involvement increases the risk of accidents, but also further illustrates that drunk driving significantly increases the severity of fatal consequences of accidents. This finding is highly consistent with existing empirical research in the field of traffic safety.

With respect to accident scale, the average number of people involved is the strongest structural predictor of the proportion of multi-fatal accidents. The more people involved, the higher the probability that an accident will escalate into a fatal one, highlighting the central role of the number of people exposed in determining the scale of fatal consequences. In contrast, the average number of vehicles involved per accident is significantly negatively correlated with the proportion of multi-fatal accidents, suggesting that multi-vehicle accidents are more likely to occur under low-speed or congested conditions, where fatal severity is not necessarily higher than in high-speed single-vehicle accidents.

In terms of model specification and robustness testing, the model that includes the interaction term between the proportion of nighttime accidents and the proportion of drunk driving performed best in the comparison of AIC index and out-of-sample prediction error, indicating

that the combined effect of the two has additional informational value in explaining the risk of multiple fatal accidents. Although the adjusted R^2 level of the model is limited, its explanatory power is reasonable in the context of accident data that is highly random and affected by multiple unobservable factors.

Overall, the results demonstrate that multi-fatal traffic accidents are not random events, but are highly correlated with identifiable behavioral characteristics and temporal structure. Unlike traditional studies that focus on the frequency of accidents, this paper characterizes the "severity structure" of fatal risks from the perspective of the proportion of accident components. This perspective provides a complementary approach for understanding the outcomes of extreme traffic accidents and also provides empirical evidence to support nighttime traffic management and the governance of drunk driving.

VIII. Limitations

Although this paper strives for rigor in data processing and model specification, the findings should be interpreted within the following limitations.

First, we analyze aggregated data based on the state-month level. This setting helps identify systemic risk structures and reduce noise at the individual level, but it also limits the characterization of individual accident mechanisms and driver heterogeneity, and is therefore unsuitable for making causal inferences at the individual level.

Second, this paper identifies statistical correlations rather than strict causal effects. While the proportion of nighttime accidents and drunk-driving is significantly correlated with fatal accidents, these variables may also be affected by unobservable factors such as enforcement intensity, road conditions, or emergency response capabilities, making it difficult to completely eliminate potential endogenous problems.

Third, the model intentionally focuses on the combination of accident components and behavioral characteristics, excluding factors such as weather conditions, road type, speed limit level, vehicle safety features, and driver age structure. This trade-off improves interpretability and robustness of the model, but it also limits the systematic analysis of environmental and infrastructure factors.

Fourth, the proportional dependent variables may exhibit greater volatility in states or months with lower total accident volumes. Although model diagnostics and robustness tests show that the main results are not driven by a few high-leverage observations, caution is still warranted when interpreting extreme values.

Fifth, linear models have limited capacity to capture potentially nonlinear risk structures. Although this paper partially alleviates this problem by comparing interaction terms with the model, the generation of multi-fatal accidents may still involve threshold effects or more complex nonlinear dynamics, which warrant further investigation using finer-grained data in future research.

In conclusion, the results of this paper should be regarded as an empirical characterization of multi-fatal traffic accident risk rather than a direct causal evaluation of specific policy interventions. Future research could extend and validate the conclusions of this paper by incorporating accident-level data, longer time spans, and richer covariates.