Fatal US School Shootings Are More Likely in Targeted and Multi-Victim Events: Evidence from Logistic Regression on Contextual Factors

Amanda Ng

2025-10-25

Abstract

1. Introduction

Gun violence in schools remains one of the most distressing forms of firearm-related harm in the United States. Each incident causes immediate loss and trauma while also increasing public concern over the safety of educational environments. Despite ongoing research, relatively little is known about why some school shootings result in fatalities while others do not. Understanding these factors is important for informing prevention strategies and school safety policies. Previous studies have examined individual-level factors such as shooter motivation [Schibler et al., 2024], access to firearms, and mental health [Peterson et al., 2022]. Fewer studies have examined contextual-level characteristics—features of the incident and school environment that may affect whether a shooting results in a fatality. This study examines these contextual factors to identify conditions that are associated with higher likelihoods of fatal outcomes.

The analysis draws on The Washington Post School Shootings Database, which documents every act of gunfire on U.S. primary and secondary school campuses during school hours since 1999. The estimand of interest is the probability that a school shooting results in at least one killing. Using logistic regression, we modeled this probability as a function of several contextual characteristics, including whether the incident involved any injuries, the shooting type, the shooter's demographics and relationship to the school community, and the broader racial and socioeconomic composition of the school.

Results show that that shootings with no non-shooter injuries, those classified as targeted or indiscriminate, and incidents involving shooters with a connection to the school were more likely to result in fatalities. These patterns suggest a higher degree of planning, precision, and familiarity with the school environment, reflecting more deliberate and focused attacks. Such findings underscore the importance of preventive strategies such as internal monitoring of students and staff, including robust threat assessment protocols, enhanced communication with law enforcement, and comprehensive emergency preparedness training. Understanding these factors can guide interventions aimed at reducing fatalities and improving overall school safety.

Section 2: Data describes the data source, alongside the cleaning and preparation of the dataset for analysis. A summary of key variables and exploratory data visualizations is also presented. Section 3: Modelling explains the logistic regression approach used to analyze the relationship between fatality likelihood and contextual characteristics, along with justifications for the model design. Section 4: Results interprets the effects of predictors and practical applicability of the model in terms of performance metrics. Section 5: Discussion summarizes the main findings, limitations of current study and outlines directions for future studies. Supporting information, including details of full dataset and additional plots are provided in Section 6: Appendix.

2. Data

2.1 Data source

The study draws on The Washington Post School Shootings Database (citation), a systematically compiled record of firearm incidents occurring at U.S. primary and secondary schools during school hours since the Columbine High School shooting on April 20, 1999. The database was developed by integrating multiple information channels, including Nexis archives, news reports, open-source databases, law enforcement records, school websites, and direct communications with school and police officials. Incidents were included only if gunfire occurred on campus immediately before, during, or after classes, thereby excluding after-hours events, accidental discharges injuring only the shooter, and suicides that posed no threat to others. Since no federal agency tracks school shootings, The Washington Post database represents one of the most comprehensive and systematically verified datasets available for studying these events. It further incorporates school-level demographic and enrollment information from the U.S. Department of Education's Common Core of Data and Private School Universe Survey, allowing for contextual analyses of how school characteristics relate to the severity of shooting incidents.

2.2 Data overview

The raw data consists of 428 recorded school shooting incidents in the United States, spanning the period from April 20, 1999 onward. Each observation represents a distinct shooting incident, and the dataset contains 50 variables. These variables include temporal and geographical information about the shooting incident, shooter(s) demographic characteristics, shooter(s) relationship to school community, shooting type, shooting outcomes (number of killings and injuries), weapon used and source, school type, number of students at school(total enrollment and by racial groups), number of students at school eligible to receive a free or reduced-price lunch, presence of a school resource office, and grade-levels offered by school. The dataset records a maximum of two shooters per incident. Every incident has at least one shooter, and details for a second shooter are included only if applicable. The full list of variables can be found in the Appendix.

2.3 Data cleaning

To address inconsistencies and typos in categorical variables, we cleaned and standardized their string representations. For "shooting type", the original categories included "unclear", "public suicide", "hostage suicide", "accidental" and "targeted." To simplify analysis and ensure sufficient sample sizes within each group, we combined "public suicide" and "hostage suicide" into a single "suicide" category, as these types were rare (8 and 2 observations, respectively). One observation labeled "public suicide (attempted)" was also included under "suicide". Six observations labeled "targeted and indiscriminate" were grouped under "targeted," and two labeled "accidental or targeted" were grouped under "accidental". Missing values were assigned as "unknown". We also re-coded "shooter's relationship to the school community" into broader, interpretable categories. Specifically, we grouped the textual entries into six main categories: "Current Student", "Former Student", "Family/Guardian of Student", "Non-Security Staff", "Family of Staff", and "Police/Security". For example, "student in summer school" and "suspended student" were re-coded as "Current Student" and "Former Student," respectively. All remaining or missing entries were categorized as "Other, Unknown or No Connection". This standardization resolves inconsistent textual entries, and facilitates simpler comparisons across levels embedded in the categorical variables in subsequent analyses.

One observation had shooter 1's gender recorded as "h," which was presumed to be a data entry error. Since this case did not involve a second shooter and the intended gender could not be inferred from other available information, the observation was excluded from the dataset. Three observations had number of students eligible for free or reduced-price lunch greater than total enrollment, they were removed.

Below is a summary of (non-zero) count and proportion of missing values in our variables of interested after basic cleaning.

Table 1: Summary of Missing Values in Interested Variables

Variable	Missing Count	Missing Proportion (%)
Age of shooter 1	127	29.95
Gender of shooter 1	94	22.17
Age of shooter 2	412	97.17
Gender of shooter 2	412	97.17
Number of white students	10	2.36
Number of students eligible for free/reduced-price lunch	56	13.21
Resource officer present	187	44.10
Weapon used	170	40.09
Weapon source	321	75.71
Shooter 2 relationship with school community	418	98.58

According to Table 1, more than 97% of incidents did not involve a second shooter, indicating that the vast majority of incidents are perpetrated by a single individual. Variables "presence of a resource officer", "weapon used", and "source of the weapon" contains missing values for over 40% of the observations. The high proportion of missing values, combined with the specificity uncoded in weapon-related variables, makes reliable imputation impractical. Therefore, in the following analysis, we exclude all information related to a second shooter, as well as presence of a resource officer and weapons.

For shooter 1's age, missing values were imputed by the mean age within the same shooting type and state, with state- or overall means used as fallbacks. Shooter 1's gender was imputed probabilistically based on the observed state-level gender proportions, preserving underlying shooter gender distribution in the data. No fallback is needed here since male shooter proportions are available in all states. For the number of students eligible for free or reduced-price lunch, we first set all private schools to zero, as such programs typically apply only to public schools. For public schools, missing values were imputed using the average lunch proportion within each state, with the overall proportion used as a fallback. The imputed value was then scaled by the school's total enrollment. Finally, missing counts of white students were imputed using the mean proportion of white students within each combination of shooting type, state, and school type and multiplied by school's total enrollment. Mean at state–school type level, and school type mean were used as fallbacks.

We collapsed the number of killings and injuries variables into binary indicators. For observations with number of killed or injured greater than zero, the corresponding indicator was coded as "1"; otherwise, it was coded as "0". This is appropriate because our study focuses on the presence or absence of harm in each incident, rather than the exact number. This approach hence simplifies the analysis, and facilitates clear interpretation of the relationships between incident characteristics and the occurrence of harm. Additionally, we created two standardized variables to capture key aspects of students' socioeconomic and racial composition in each school. The racial composition was measured as the proportion of non-white students calculated as $1-\frac{\text{Number of white students}}{\text{Total enrollment}}$, and the economic situation was measured as the proportion of students receiving subsidized meals calculated as $\frac{\text{Number of students eligible for free or reduced lunch}}{\text{Total enrollment}}$. Standardizing both measures by total enrollment allows for meaningful comparisons across schools of different sizes and ensures that these variables reflect relative composition rather than absolute counts.

R version and packages (citation)

2.4 Data description

The cleaned data consists of 424 observations and 9 variables. Below is a summary description of all variables in the cleaned dataset.

Table 2: Summary Description of Variables

Variable	Type	Categories/Range
Killed	Binary	1 = At least one killing (excluding shooter) occured
Injured	Binary	1 = At least one injured (excluding shooter) occured
School Type	Categorical	Public, Private
Shooting Type	Categorical	Indiscriminate, Targeted, Accidental, Suicide, Unclear
Shooter Age	Numerical	6 to 74
Shooter Gender	Categorical	Male, Female
Shooter Relationship	Categorical	Current Student, Former Student, Family/Guardian of Student, Non-Security Staff, Police/Security, Other/Unknown/No Connection
Proportion of Non-white Studens	Numerical	0 to 1
Proportion of Students Eligible for Subsizied Lunch	Numerical	0 to 1

2.5 Data summary and visualizations

In total, the dataset includes 424 shooting incidents. Among them, 328 cases involved no killings, while 99 cases resulted in at least one killing. This distribution indicates that the majority of school shootings (approximately 77%) did not involve any death excluding the shooter.

Table 3: Shootings With Killing Occured

Measure	Shooter Age	Proportion of Non-white Students	Proportion of Students with Subsidized Lunch
Mean	20.41	0.57	0.48
Standard Deviation	9.63	0.35	0.30
Median	17.00	0.60	0.48
Min	6.00	0.00	0.00
Max	53.00	1.00	1.00

Table 4: Shootings Without Killing Occured

Measure	Shooter Age	Proportion of Non-white Students	Proportion of Students with Subsidized Lunch
Mean	19.48	0.69	0.58
Standard Deviation	8.74	0.33	0.29
Median	17.00	0.88	0.60
Min	6.00	0.00	0.00
Max	74.00	1.00	1.00

Tables 3 and 4 show that shootings resulting in at least one killing were associated with a slightly older average shooter age, while non-fatal shootings occurred more often in schools with higher proportions of non-white students and students receiving subsidized lunch. These patterns suggest that these variables may influence the likelihood of a shooting being fatal. However, the substantial standard deviations in each numerical variable indicate considerable heterogeneity within each group, highlighting that themselves alone does not fully explain shooting outcomes. This heterogeneity should be kept in mind when interpreting the model results. Additional visualizations are provided in the Appendix.

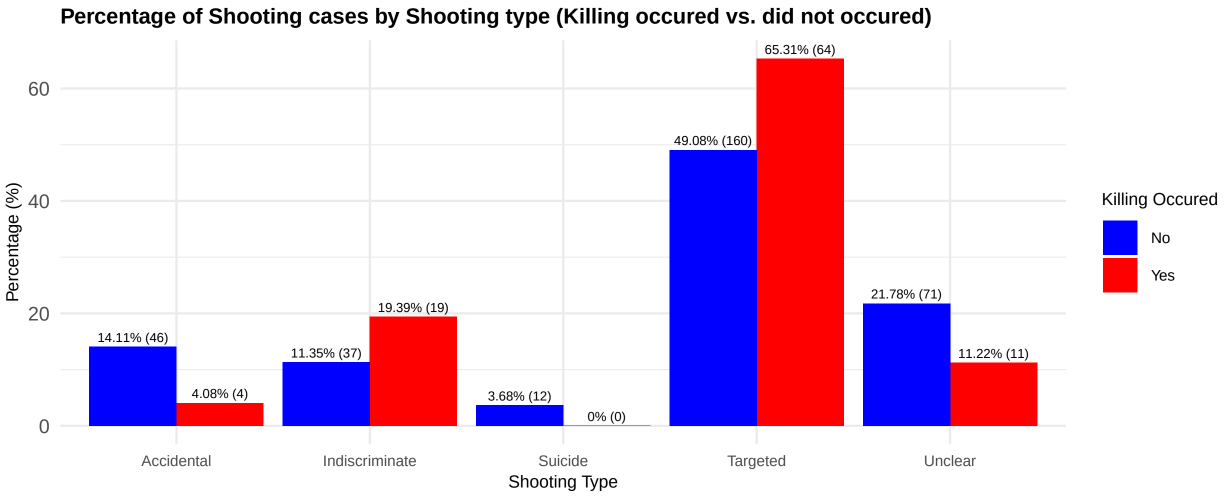


Figure 1: Percentage of Shooting cases by Shooting type (Killing occured vs. did not occured). This bar plot illustrates the distribution of all shooting cases based on their shooting type, categorized into five groups: "Accidental", "Indiscriminate", "Suicide", "Targeted" and "Unclear". Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category. Percentages belonging to the same group sum up to 100%.

Figure 1 suggests that targeted shootings were the most common shooting type and were substantially more prevalent when killings occurred (65.3% in killing occurred shootings vs. 49.1% in no killing shootings), suggesting a strong association between targeted intent shootings and fatal outcomes. In contrast, accidental and unclear shootings were more frequent in non-fatal incidents, while indiscriminate shootings showed a moderate increase when killings occurred (19.4% vs. 11.4%). These patterns indicate that the nature of the shooting event may play a critical role in determining whether killing would occur. Hence, shooting type should be deemed as a key predictor in the model to examine its contribution to the likelihood of a killing.

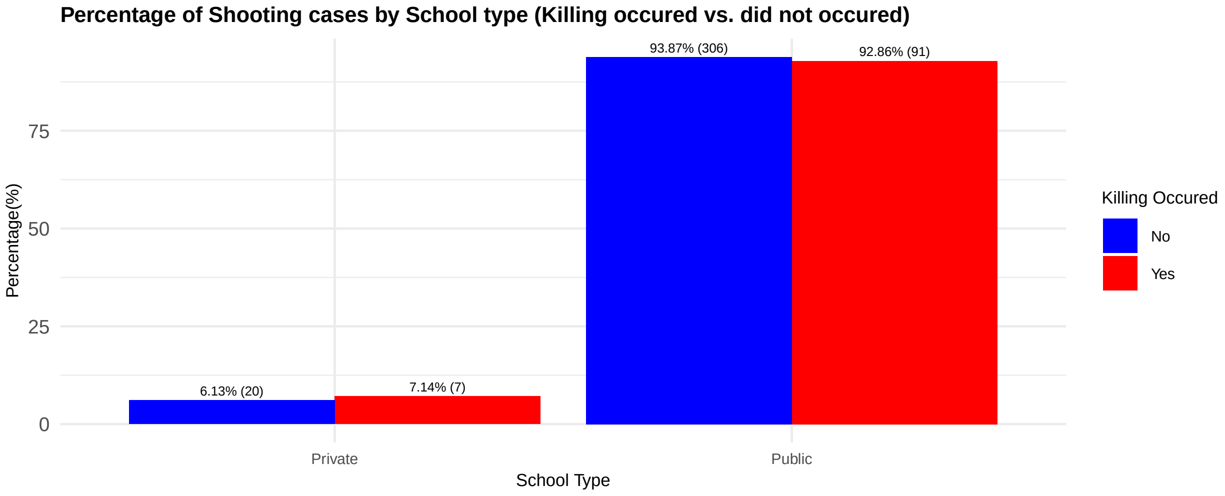


Figure 2: Percentage of Shooting cases by School type (Killing occured vs. did not occured). This bar plot illustrates the distribution of all shooting cases based on their school type, categorized into two groups: "Private" and "Public". Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category. Percentages belonging to the same group sum up to 100%.

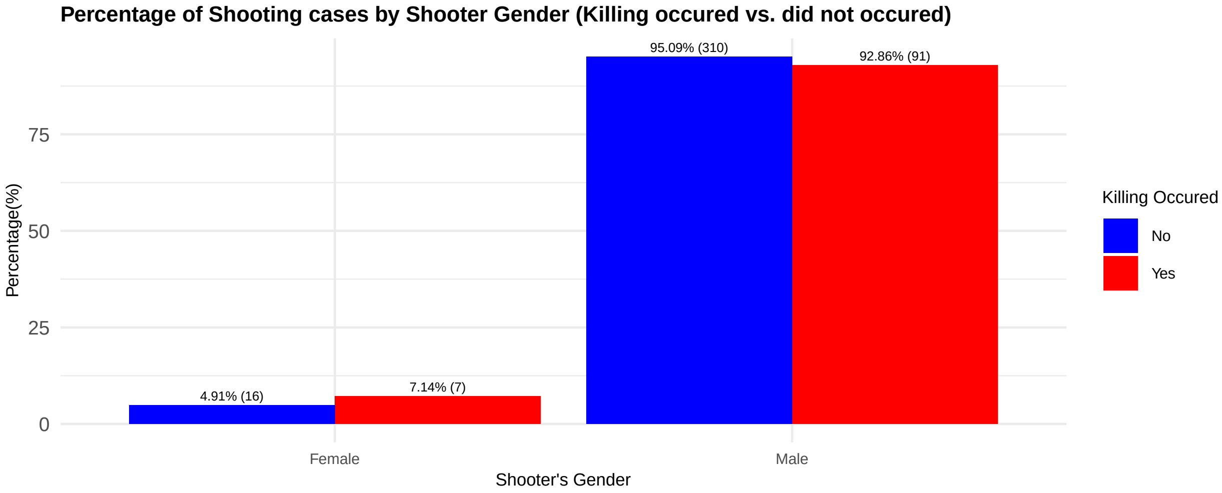


Figure 3: Percentage of Shooting cases by Shooter Gender (Killing occured vs. did not occured). This bar plot illustrates the distribution of all shooting cases based on the shooter's gender, categorized into two groups: "Female" and "Male". Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category. Percentages belonging to the same group sum up to 100%.

Figure 2 and 3 suggest that majority of both fatal and nonfatal school shooting incidents occurred in public schools and shooters are predominately male, accounting for more than 92% of all cases. The near-identical percentages between fatal and nonfatal incidents across the underlying categories suggest that whether a shooting results in a fatality may not strongly associated with whether the school is public or private or the shooter's gender. This pattern indicates that institutional and shooter's gender classification alone may offer limited explanatory power for modeling fatal outcomes. Hence, additional contextual factors, such as the nature of the shooting, the presence of injuries, or the relationship between the shooter and the school, are needed in explaining fatal outcomes.

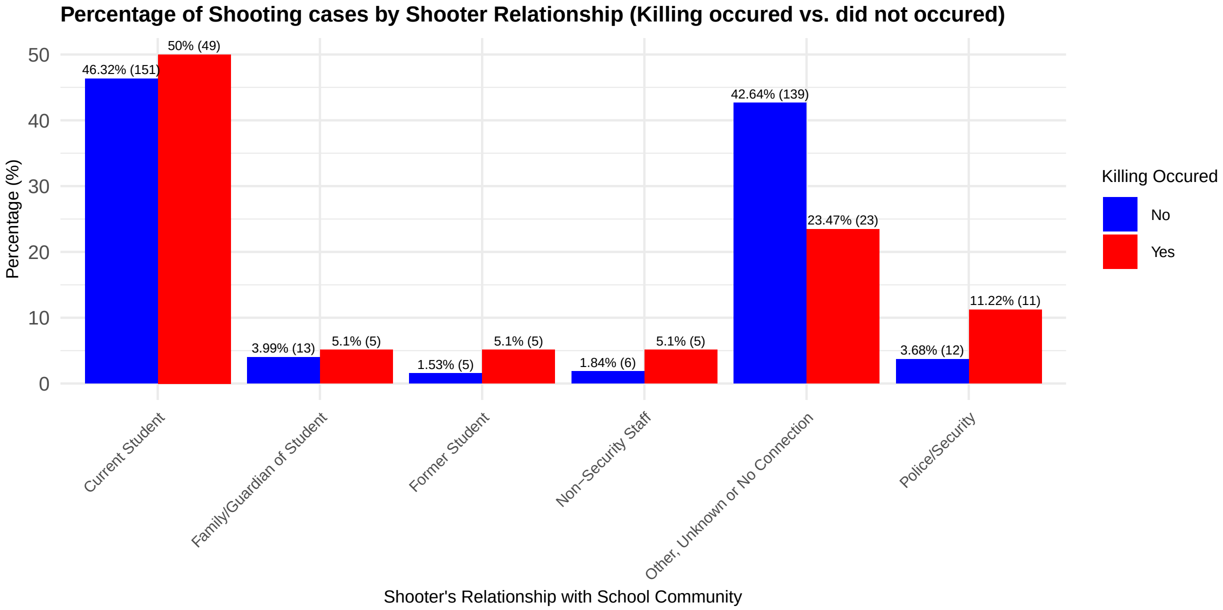


Figure 4: Percentage of Shooting cases by Shooter Relationship (Killing occured vs. did not occured). This bar plot illustrates the distribution of all shooting cases based on the shooter's relationship with the school community, categorized into six groups: "Current Student", "Family/Guardian of Student", "Former Student", "Non-security Staff", "Police/Security", and "Other/Unknown/No Connection". Each bar is labeled with the corresponding percentage and the total counts in brackets within each category. Percentages belonging to the same group sum up to 100%.

Figure 4 suggests that shooters who were current students account for nearly half of both fatal and nonfatal incidents, indicating that student-perpetrated shootings are the most frequent overall. Notably, percentage of shootings carried out by individuals who are police or security personnel in fatal incidents is more than 3 times to that in nonfatal incidents. This reflects that their access to weapons, gun training, and the high-risk situations in which they are involved may put others at greater risk of death. Conversely, shootings committed by individuals with no known connection to the school are more common in nonfatal incidents (over 40%). These patterns suggest that the shooter's relationship to the school strongly influences the likelihood of a fatal outcome, and we expect this variable to be a key predictor in the regression model.

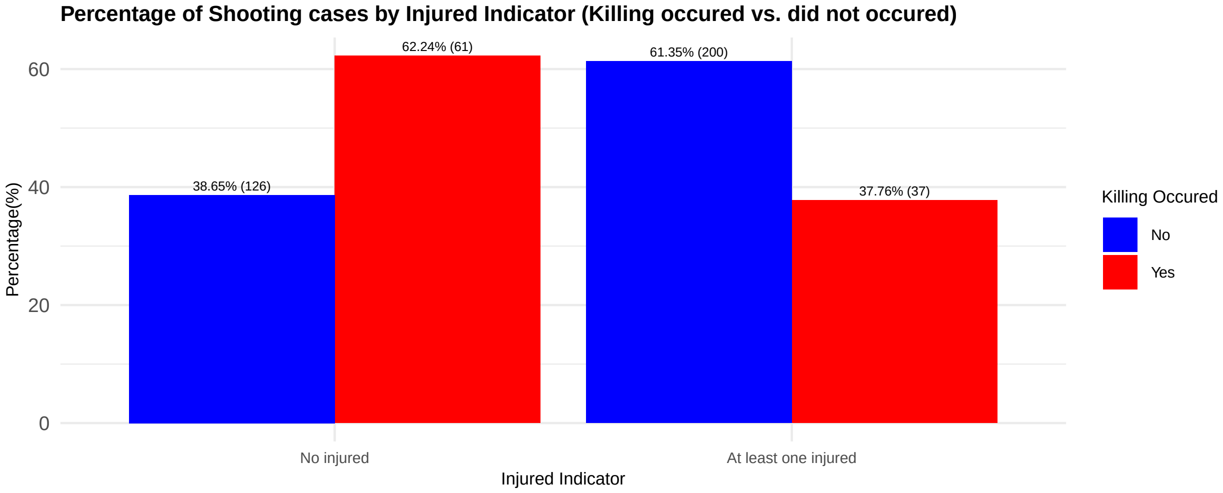


Figure 5: Percentage of Shooting cases by Injured Indicator (Killing occured vs. did not occured). This bar plot illustrates the distribution of all shooting cases based on whether there were injuries excluding the shooter, categorized into six groups: "No injured", and "At least one injured". Each bar is labeled with the corresponding percentage and the total counts in brackets within each category. Percentages belonging to the same group sum up to 100%.

Figure 5 illustrates that over 60% of fatal cases did not involve additional non-shooter injuries, while majority of non-fatal shootings involved at least one non-shooter injury. This contrast highlights that fatal shootings tend to be more targeted or severe, while non-fatal ones may reflect less severe killing intent or lower accuracy. These descriptive patterns underscore the relevance of injury occurrence as a meaningful variable for modeling fatality risk in school shooting incidents.

3. Model

3.1 Model description

In this study, we employed a logistic regression model to predict the likelihood of a student dropping out after their first year. Logistic regression is a statistical method that models the relationship between a binary outcome variable and a set of predictor variables. The outcome variable in our models is the logit of the fatal probability. In our study, the logit refers to the natural logarithm of "ratio of the probability of at least one killing occurred to the probability of no killing occurred". The predictors include whether injured occurred, school type, shooting type, shooter's age, shooter's gender, shooter's relationship to school community, proportion of non-white students, and proportion of students eligible for subsidized lunch.

We set the reference level of:

- School type as "Private"
- Shooting type as "Unclear"
- Shooter's gender as "Female"
- Shooter's relationship with school community as "Other/Unknown/No relationship"

Below is the model equation:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 \cdot I_{\text{Injured occured(i)}} + \beta_2 \cdot I_{\text{Public(i)}} \\ &+ \beta_3 \cdot I_{\text{Shooting type = Accidental (i)}} + \beta_4 \cdot I_{\text{Shooting type = Indiscriminate (i)}} + \beta_5 \cdot I_{\text{Shooting type = Targeted (i)}} + \beta_6 \cdot I_{\text{Sh}} \\ &+ \beta_7 \cdot \text{Shooter Age}_{(i)} + \beta_8 \cdot I_{\text{Shooter Gender = Male(i)}} \\ &+ \beta_9 \cdot I_{\text{Shooter Relationship = Current Student(i)}} + \beta_{10} \cdot I_{\text{Shooter Relationship = Former Student(i)}} + \beta_{11} \cdot I_{\text{Shooter Relationsh}} \\ &+ \beta_{14} \cdot \text{Proportion of non-white students}_{(i)} \\ &+ \beta_{15} \cdot \text{Proportion of students eligible for subsidized lunch}_{(i)}\end{aligned}$$

where

- i represent the ith incident
- p_i represents the probability of the ith incident involving at least one killing excluding the shooter

3.2 Model assumptions

The logistic regression model relies on 5 key assumptions:

1. **Binary Dependent Variable**: The dependent variable in logistic regression must be binary, meaning it takes on only two possible outcomes. In our case, the dependent variable is the killing indicator, which indicates whether a killing occurred (1) or not (0). This assumption is satisfied as the response variable is binary by design.

2. **Independence of Observations**: Logistic regression assumes that the observations are independent of one another. This assumption is satisfied since each observation in our dataset corresponds to an independent shooting incident.

3. **No Multicollinearity**: Logistic regression assumes that there is little or no multicollinearity among the independent variables. This means that the predictors should not be highly correlated with one another, as high correlation can make it difficult to separate out the individual effects of each predictor on the dependent variable. To verify this assumption, we calculated VIF values for each predictor in each model. A VIF value below 5 indicates no significant multicollinearity in our model. Should any predictor exhibit a VIF value exceeding 5, we address multicollinearity by removing the predictor with the highest VIF and constructing a reduced model. Fortunately, multicollinearity check showed that all VIF values are below 5, indicating no severe issue is observed in any of the predictors in our model.

4. **Linearity of Independent Variables and Log Odds**: Logistic regression assumes a linear relationship between the independent variables and the log odds of the dependent variable. To check for this assumption, we identify if there are systematic non-linear patterns using the "deviance residual vs fitted" plot. Despite the slight curvatures in the residuals, we observe no significant non-linear pattern observed in any of the model residual plots. Hence, we may assume that the linear requirement is satisfied in the model.

5. **No Overdispersion**: Overdispersion occurs when the variance of the outcome variable is greater than what the logistic regression model assumes. This implies that the model underestimates the true variance of the model parameter estimates, leading to incorrect standard errors and inflated Type 1 error. We check for overdispersion by examining the Pearson chi-squared statistic. A statistic which is significantly greater than 1 indicates overdispersion. If overdispersion is present, we might consider using alternative methods, such as quasi-binomial regression, which accounts for overdispersion. Fortunately, Pearson chi-squared statistic is close to 1, indicating no severe issue is observed in any of the model.

3.3 Justification

Given that our response variable, the killing indicator, is binary, logistic regression is an appropriate modeling choice. Logistic regression models the outcome as the logit of the probability that a killing occurs, allowing us to predict fatality risk while accounting for multiple predictors. A key strength of logistic regression is interpretability: the beta coefficients represent the log odds of a killing occurring for a one-unit change in a predictor, holding all other variables constant. The odds reflect the ratio of the probability that a killing occurs to the probability that it does not, and the log odds are the natural logarithm of this ratio. Positive log odds indicate an increased likelihood of a killing, while negative log odds indicate a decreased likelihood, providing a straightforward way to assess each predictor's impact on fatality risk in school shootings.

Predictors were chosen to capture key contextual and demographic factors associated with the severity of school shooting outcomes (Section 2.5). Shooting type was included as a key predictor because it reflects the shooter's intention. As shown in Figure 1, shootings resulting in fatalities were more likely to be targeted rather than accidental or unclear. Including shooting type allows the model to examine whether planned and deliberate targeting is associated with fatal outcomes. Shooter's relationship to the school community further captures intention and social proximity. Figure 4 suggests that shooters with close ties to the school, such as current students or police and campus security, may commit targeted attacks arising from interpersonal conflicts. Including this variable helps assess whether familiarity with the school affects the likelihood of fatalities.

Shooter characteristics, including age and gender, were included to account for personal factors that may influence fatality risk. Age may capture maturity or experience-related differences, while gender accounts for potential psychological differences between males and females (citation).

Presence of non-shooter injuries was included as a binary indicator to capture whether any non-shooter was injured. This approach was chosen over modeling the number of injuries due to substantial skewness, as most incidents involved either no additional injuries or a single injury. Figure 5 suggests that fatal shootings are less likely to involve non-shooter injuries, indicating that these incidents may be more deliberate and targeted. Including this variable allows us to formally test whether the presence of non-shooter injuries differentiates less severe shootings from those resulting in fatalities.

School type was retained in the model despite Figure 2 showing little difference between groups, since it represents structural characteristic of the educational environment and allows examination of whether institutional context, resource availability, or security infrastructure indirectly affect fatality in shootings.

Finally, proportion of non-white students and proportion of students eligible for subsidized lunch were included as measures of school racial and socioeconomic composition. These variables may reflect differences in students' past exposure to community violence, which could influence their responses during a shooting and, indirectly, affect fatality risk.

3.4 Model Evaluation

For model evaluation, the data were randomly divided into a training set (85%) and a testing set (15%). The training set was used to fit the logistic regression model, while the testing set was held out for performance assessment. Random sampling ensured that both subsets were representative of the overall data distribution, allowing for an unbiased evaluation of the model's predictive ability on unseen observations. Model performance was assessed across a range of probability thresholds using sensitivity (true positive rate), specificity (true negative rate), and overall misclassification rate. Comparing these metrics under different thresholds provides insight into the trade-off between correctly identifying fatal shootings and minimizing false classifications.

4. Results

4.1 Model Estimates and Hypothesis tests

Below is a summary table of the odd ratios based on coefficient estimates from the fitted model:

Table 5: Summary table of Model Coefficient Estimates

Term	Coefficient Estimate	Standard Error	95% CI	p-value
Intercept	-0.7177	0.9794	[-2.6945, 1.1715]	0.4637
Injured	-1.2622	0.2907	[-1.8449, -0.7019]	< 0.0001
Public	-0.1616	0.6846	[-1.4875, 1.2211]	0.8133
Accidental	-0.4631	0.6815	[-1.9034, 0.8243]	0.4968
Indiscriminate	0.7509	0.5475	[-0.3154, 1.8464]	0.1702
Suicide	-15.9813	684.2636	[-282.6329, 5.7403]	0.9814
Targeted	0.8206	0.4269	[0.0183, 1.7073]	0.0546
Shooter Age	0.0136	0.0202	[-0.0277, 0.0525]	0.5024

Male	-0.5199	0.5683	[-1.617, 0.6372]	0.3603
Current Student	0.9088	0.3743	[0.1885, 1.6616]	0.0152
Family/Guardian of Student	0.0437	0.7510	[-1.5342, 1.4485]	0.9536
Former Student	0.8807	0.8679	[-0.9278, 2.5885]	0.3103
Non-Security Staff	0.9444	0.9864	[-1.0399, 2.9194]	0.3384
Police/Security	1.1309	0.5974	[-0.0634, 2.2985]	0.0583
Proportion of Non-White Students	-0.7242	0.5698	[-1.8494, 0.3929]	0.2037
Proportion of Students Eligible for Subsidized Lunch	-0.3045	0.7268	[-1.7367, 1.1235]	0.6752

Interpretation:

The fitted logistic regression model estimates the log-odds of a school shooting resulting in at least one fatality based on the incident context, shooter characteristics, and school-level demographics. The baseline category represents a shooting occurring in private schools which did not involve injuries, the shooter was female, were not affiliated with the school community and had unclear shooting intention.

For all categorical variables, the exponentiated coefficients represents the odds ratio of killing occurrence in a school shooting between the indicator level as compared to the reference level of the predictor. Holding all other predictors constant, shootings in which at least one person was injured were significantly less likely ($p < 0.0001$) to involve fatalities, suggesting that incidents resulting in injuries alone tend to be less severe or potentially interrupted before turning deadly. Compared to private schools, shootings in public schools were also less likely to involve fatalities. Relative to unclear shooting intention, targeted and indiscriminate shootings were more likely to result in fatalities with odd ratios greater than 2, whereas accidental and suicide-related shootings were less likely to result in fatalities. Note that suicide-related shootings yielded an extremely large negative estimate with high uncertainty, possible due to very few such cases in the data. However, the difference between any categorical level as compared to the reference level of school type and shooting type are statistically insignificant. Male shooters were not found to differ significantly from female shooters in fatality likelihood. Shooters who had some affiliation to the school community were more likely to be fatal compared to those committed by unaffiliated individuals, with police/security staff posing the highest risk, followed by non-security staff and current students. Such difference is statistically significant for current students ($p = 0.0152$) and marginal significant for police/security staff ($p = 0.0583$).

For numerical variables, the exponentiated coefficients represents the odds ratio of killing occurrence in a school shooting associated with an "one-unit increase" in the variable. Assuming all other predictors are fixed, a shooter who is a year older is associated with a statistically insignificant small increase (1.3%) in killing occurrence probability. This implies that age differences among shooters do not substantially affect whether a shooting results in fatalities. Since both school-level demographic variables (proportion of non-white students and proportion of students eligible for subsidized lunch) are defined as proportions, it does not make sense to increase a proportion (which is between 0 and 1) by 1. So, we conduct interpretation in terms of a factor of the coefficient. The odds of killing occurrence if a school has 10% of students being non-white is $exp(-0.7242 * 0.1) = 0.93$ times compared to a school with no non-white students, assuming fixed levels for all other predictors. The odds of killing occurrence if a school has 10% of students eligible for subsidized lunch is $exp(-0.3045 * 0.1) = 0.97$ times compared to a school with no students eligible for subsidized lunch. This suggests that schools with higher proportions of non-white and economically-disadvantaged student were, if anything, slightly less likely to result in fatalities, though these associations were not statistically significant.

Overall, these results indicate that contextual factors of the shooting incident, particularly injuries occurrence and whether the shooter was affiliated with the school community play an important role in determining whether a school shooting becomes fatal.

We conducted a Likelihood Ratio Test to assess the overall significance of the model, rejecting the null hypothesis that none of the covariates are linearly related to the log-odds of killing occurred (p = 5.078e-07 < 0.05). This suggests that at least one predictor in the model is significantly associated with killing occurrance.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

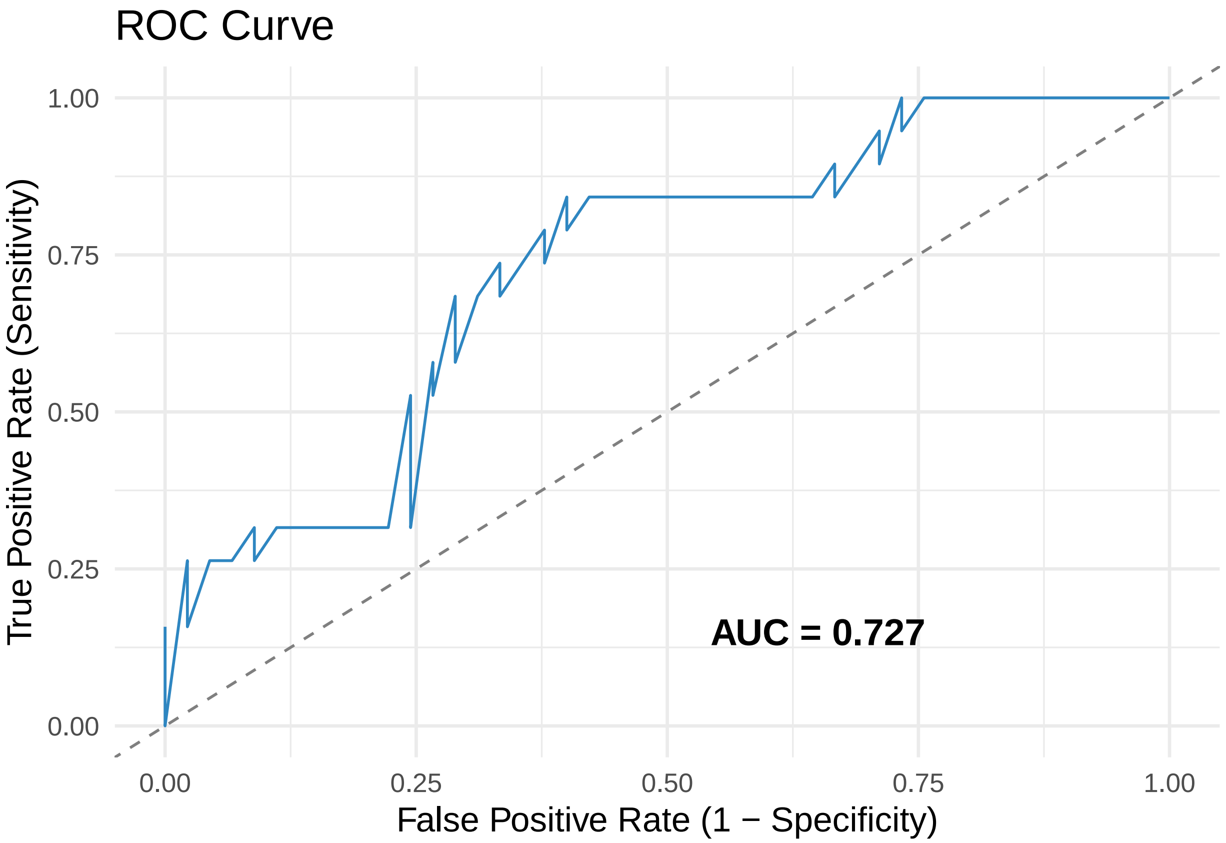


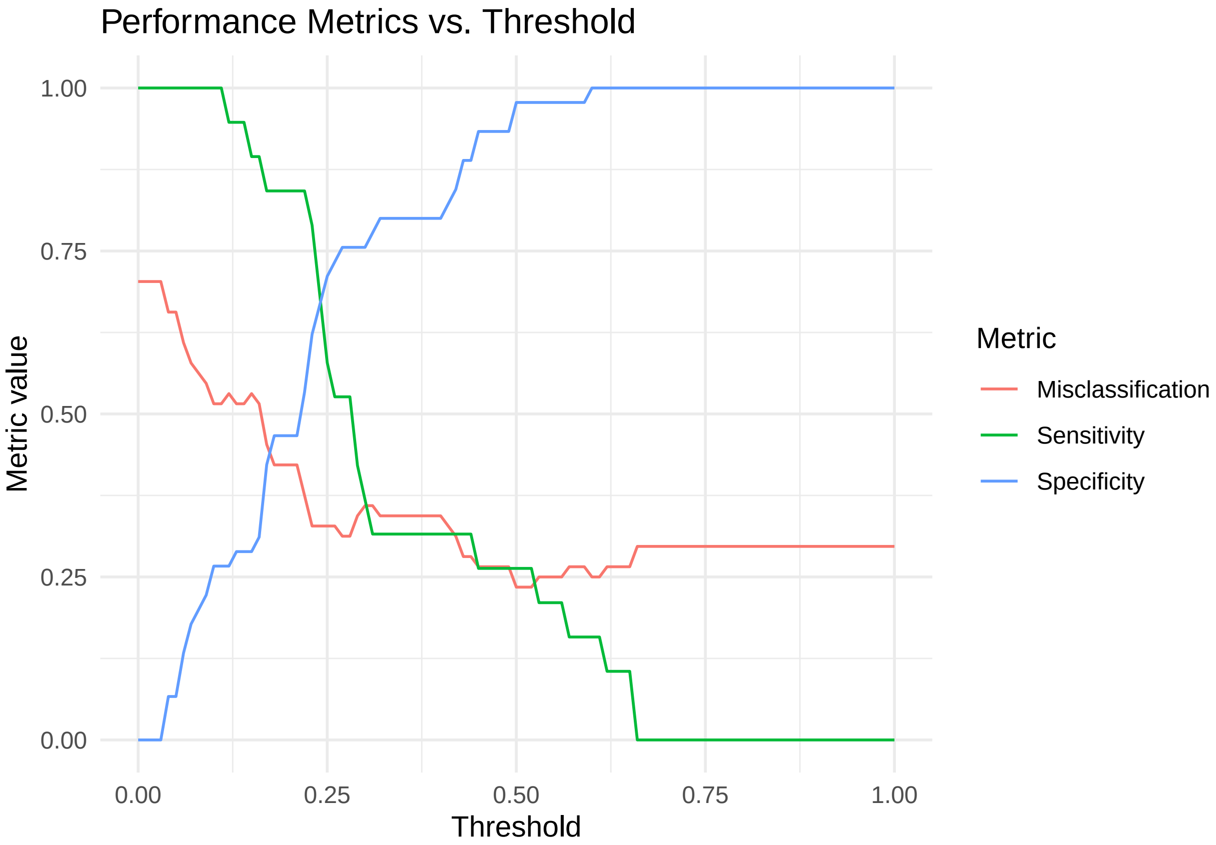
Figure 6: Receiver operating characteristic (ROC) Curve for the logistic regression model. The ROC curve displays the diagnostic performance of the logistic regression model using shootings context-related predictors, with true positive rate plotted against false postive rate. The area under the curve (AUC) is 0.727.

The area under the ROC curve is 0.727, indicating good discriminative ability of the model to distinguish between shooting incidents involving killings and those which did not.

4.2 Sensitivity, Specificity, and Misclassification rates

Sensitivity, specificity, and misclassification rate are key performance metrics used to evaluate the effectiveness of the classification models. Sensitivity, also known as the true positive rate, measures the model's ability to correctly identify shootings which actually resulted in killings. Specificity, on the other hand, is the true negative rate and represents the model's capacity to correctly identify shootings which did not result in killings. The misclassification rate is the proportion of all predictions that are incorrect, combining both false positives and false negatives. A lower misclassification rate indicates better overall model accuracy, but it must be interpreted alongside sensitivity and specificity to understand the model's strengths and weaknesses in identifying killing occurrence.

Below is a figure summarizing sensitivity, specificity, and misclassification rates on the unseen testing data using different threshold values (i.e. the cut-off probability to classify a shooting incident into the killing occurred group).



The performance evaluation of the model highlights the fundamental trade-off between sensitivity and specificity when selecting a classification threshold for identifying whether a shooting involved a killing. Lower threshold values, such as 0.1, yield higher sensitivity (100%), meaning the model successfully identifies all fatal shootings. However, this comes with lower specificity (26.67%), indicating that some non-fatal shootings are also classified as fatal. While both fatal and non-fatal shootings are serious incidents, a higher sensitivity ensures that cases involving killings are not missed although over half of the shootings are misclassified at this level.

Conversely, higher threshold values, such as 0.3, substantially increase specificity (97.78%) but sharply reduce sensitivity (26.32%). This means that while most non-fatal shootings are correctly identified, a considerable number of fatal cases go undetected. Given the severe consequences of missing a fatal shooting, this represents an important limitation. Therefore, selecting an appropriate threshold requires balancing sensitivity and specificity to ensure the model reliably identifies shootings that result in killings while maintaining reasonable overall accuracy.

Ultimately, the optimal threshold depends on the acceptable trade-off between false positives and false negatives. Since the goal is to support timely intervention and situational awareness, it may be preferable to flag more shootings that could result in killings rather than miss those that are truly at fatal risk. In this study, a threshold of 0.23 offers a reasonable balance, achieving a sensitivity of 78.94%, specificity of 62.22%, and an overall misclassification rate of 32.81%. These values suggest that the model performs adequately at this cut-off, capturing the majority of fatal shootings while maintaining moderate accuracy in distinguishing between fatal and non-fatal cases.

Overall, the results indicate that the model can differentiate between fatal and non-fatal shootings with fair accuracy when an appropriate threshold is selected. Although its performance leaves room for improvement, particularly in increasing specificity without compromising sensitivity, the model shows practical potential for identifying shootings most at risk of resulting in fatalities.

5. Discussion

5.1 Key findings

This study uses logistic regression model to examine the factors associated with fatal outcomes in school shooting incidents, focusing on both shooter characteristics and school-level contextual variables. Model results show that school shooting incidents that did not result in injuries, as well as those classified as targeted or indiscriminate, were associated with significantly higher odds of resulting in fatalities. Additionally, shooters affiliated with the school community, particularly current students and security personnel, posed a greater risk of fatal outcomes compared to unaffiliated individuals. In contrast, no statistically significant differences in fatality likelihood were observed based on shooter gender or age, school type (public vs. private), or the school's socioeconomic and racial composition.

These findings suggest that fatal school shootings may be more predictable when they originate within the school community's existing social or disciplinary networks. These shooters often exhibit early-warning signs, such as behavioral issues, disciplinary history, or peer concerns. Hence, school-based prevention systems focusing on students and staff may be effective in identifying or deterring them beforehand. Moreover, affiliated shooters may enter with greater intent or preparation, potentially bringing more fatal weapons or executing more deliberate plans. Therefore, in addition to internal behavioral monitoring, there is a need for enhanced external security coordination, including improved access control, visitor screening, and robust information-sharing between schools, law enforcement, and community organizations. The findings underscores the importance of rapid emergency response protocols and preparedness training, as these incidents tend to escalate quickly and with consequences beyond injuries.

5.2 Limitation

As shown in Section 2.5, shooter's age, proportion of non-white students, and proportion of students eligible for subsidized lunch exhibited high standard deviations within both fatal and non-fatal incident groups. This suggests that these variables alone may not fully capture the underlying characteristics that differentiate fatal from non-fatal shootings. There may be unobserved variables, such as local law enforcement response times or shooter mental health history, influencing the likelihood of fatality that are not captured by the current model.

Additionally, the model estimate for suicide-related shootings exhibited an extremely large standard error. This likely stems from the very limited number of cases in the dataset with only 12 in the fatal group and none in the non-fatal group, which severely constrains the reliability and generalizability of findings for this category. As a result, any conclusions drawn regarding suicide-related shootings should be interpreted with caution, and future analyses would benefit from larger, more representative samples to improve statistical stability.

Lastly, data imbalance is an issue in our dataset. Specifically, we observe a disproportionate number of non fatal incidents (more than 77%). This imbalance creates challenges in training and evaluating our model. In particular, when applying a naive "predict-all-negative" approach (i.e. predicting all shootings to be non fatal), the misclassification rate is about 30% on the testing dataset. In fact, our model with the chosen "optimal" threshold demonstrates slightly worse misclassification rate, this suggest that it is not offering any improvement over the naive model and might not provide the expected level of overall predictive power.

5.3 Next Steps

An improvement to address the wide dispersion observed in numerical variables within fatal and non-fatal incident groups is to adopt more flexible modeling strategies that better account for within-group variability. Approaches such as random effects models, hierarchical modeling, or non-linear transformations on numerical variables may help capture latent heterogeneity that simple linear terms overlook, ultimately yielding more stable and interpretable estimates.

To further enhance model reliability, reducing the number of levels in categorical variables should be considered. Currently, variables such as shooting intention and shooter affiliation contain more than five levels, some with very few observations, leading to inflated standard errors and unstable estimates. Future analysis should collapse similar categories into broader, conceptually coherent groups, such as combining accidental and suicide-related shootings into a "non-aggressive intent" category, to improve statistical power and clarity.

The Wald test results indicate that most school-context predictors, including school type, racial composition, and subsidized lunch eligibility, are statistically insignificant. Building on this insight, future improvement of the model should incorporate feature selection techniques such as LASSO or StepAIC to systematically eliminate non-contributory predictors. This refinement would reduce overfitting, enhance predictive accuracy, and improve interpretability.

Lastly, further validation is necessary to assess the model's generalizability. The current evaluation relies on training and testing subsets drawn from the same dataset. While this provides initial insight into predictive performance, testing the model on an independent dataset, such as school shootings from subsequent years, would offer a more robust assessment of its reliability across varying social, demographic, and situational contexts.

6. Appendix

6.1 Raw data variables

- Unique identifier for each shooting incident
- National Center for Education Statistics (NCES) unique school ID
- Name of the school
- NCES unique district ID
- Name of the school district
- Date of the shooting
- School year of shooting
- Year of the shooting
- Approximate time of the shooting
- Day of the week when the shooting occurred
- City where the school is located
- State where the school is located
- Type of school (Public or Private)
- Total enrollment at the school at the time of the shooting
- Number of people killed (excluding the shooter)
- Number of people injured (excluding the shooter)
- Total number killed and injured (excluding the shooter)
- Type of shooting: accidental, hostage suicide, indiscriminate, public suicide, targeted, or unclear
- Age of the first shooter
- Gender of the first shooter (Male or Female)
- Race or ethnicity of the first shooter (American Indian, Black, Hispanic, White, Asian)
- First shooter's relationship to the school
- Indicates if the first shooter died in the incident (yes or no/unknown)

- How the first shooter died (if deceased)
- Age of the second shooter (if any)
- Gender of the second shooter
- Race or ethnicity of the second shooter
- Second shooter's relationship to the school
- Indicates if the second shooter died (yes or no/unknown)
- How the second shooter died (if deceased)
- Number of White students enrolled at the time of shooting
- Number of Black students enrolled
- Number of Hispanic students enrolled
- Number of Asian students enrolled
- Number of American Indian or Alaskan Native students enrolled
- Number of Hawaiian Native or Pacific Islander students enrolled (unavailable before 2009)
- Number of students of two or more races (unavailable before 2009)
- Presence of a school resource officer or security guard (present or absent/unknown)
- Weapon(s) used in the shooting
- Source from which the shooter acquired the weapon(s)
- Latitude of the school
- Longitude of the school
- Full-time equivalent teachers at the school at the time of shooting
- Lowest grade level offered by the school
- Highest grade level offered
- Number of students eligible for free or reduced-price lunch
- County where the school is located
- Two-digit state Federal Information Processing Standards (FIPS) code
- Five-digit county FIPS code
- NCES urban-centric locale code

6.2 Additional summary plots

6.3 Model assumption diagnostic plots

Reference

Jillian K Peterson, James A Densley, Kyle Knapp, Stasia Higgins, and Amanda Jensen. Psychosis and mass shootings: A systematic examination using publicly available data. *Psychol. Public Policy Law*, 28(2): 280–291, May 2022.

Katherine Schibler, Lindsay Hahn, and Adam Lankford. Investigating 55 years of mass shooter statements in the united states: A study of perpetrators' stated motivations and their association with attack severity. *Commun. Monogr.*, pages 1–24, May 2024.