Understanding School Shootings: Analyzing Trends, Demographics, and Geographic Factors Using Bayesian Methods*

A Data-Driven Investigation into the Predictors of School Shootings in the United States (1999–2024)

Ruizi Liu

December 3, 2024

School shootings remain a critical issue in the United States, with far-reaching social and psychological consequences. This study investigates the factors influencing the occurrence of school shootings, focusing on temporal trends, school type, demographic composition, and geographic context. Using a Bayesian logistic regression model, this analysis incorporates both observed and simulated data to explore the relationship between these predictors and the likelihood of a shooting. The results reveal significant associations between the predictors and school shootings. Temporal trends indicate an alarming increase in incidents over time, while public schools and urban locales exhibit higher risks compared to private schools and rural areas. Demographic composition also plays a role, with schools characterized by specific racial majorities showing distinct patterns in the likelihood of shootings. This report emphasizes the need for targeted, data-driven interventions that consider institutional and geographic variations in risk. By leveraging probabilistic modeling, this study provides actionable insights for policymakers aiming to mitigate the occurrence of school shootings and improve the safety of educational environments across the United States.

1 Introduction

School shootings are among the most tragic and impactful events in the United States, leaving lasting scars on communities and prompting urgent calls for effective prevention measures. While public discourse often focuses on the immediate aftermath of such events, it is equally

*Code and data are available at: https://github.com/RIRI0527/school_shooting_analysis.git.

important to examine the underlying factors contributing to their occurrence. This report aims to analyze the temporal, institutional, demographic, and geographic predictors of school shootings, drawing insights from historical data to inform future policies.

This analysis utilizes a Bayesian logistic regression model to explore whether factors such as the year, school type, demographic composition, and urbanicity influence the likelihood of school shootings. Bayesian methods are particularly suitable for this investigation, as they allow for the incorporation of prior knowledge while providing probabilistic interpretations of the results. By using both observed and simulated data, the model offers a comprehensive perspective on the risk factors associated with these incidents.

The findings presented in this report highlight critical trends and disparities that warrant further attention. For example, the temporal trend of increasing school shootings underscores the urgency of targeted interventions. Similarly, the variations in risk between public and private schools, urban and rural areas, and schools with differing demographic compositions suggest that a one-size-fits-all approach may not be sufficient. By identifying and analyzing these patterns, this study seeks to contribute to the broader conversation on preventing school shootings in the United States.

2 Data

2.1 Overview

The dataset is from the wasington post (Post 2021).The data provides a data record of school shootings in the United States since 1999 to 2024, with the original data set consisting of 50 variables and 416 observations, which recorded when, where, and where they occurred in the data set. The racial proportion of the schools where the shootings took place and the gender of the shooters, among other variables. These variables provide detailed information about each school shooting. We will conduct statistical analysis on this data set and explore whether these factors affect the occurrence of school shooting.

2.2 Measurement

Measurement data is the process by which we turn events that happen in reality into visual data by collecting information, taking measurements, and so on. In this data, the Washington Post spent a year counting how many children were affected by the events of school shootings (not just the number of casualties). Reporters compiled this data through Nexis, news articles, open-source databases, law enforcement reports, information from school websites, and calls to schools and police departments (Post 2021) to compile this dataset.

2.3 Data Cleaning

We used theRprogramming language (R Core Team 2023), the `here` package [], the `dplyr` package (Wickham, François, et al. 2023), the `tidyr` package (Wickham and Henry 2023), the `maps` package (Brownrigg et al. 2023), the `ggplot2` package (Wickham, Chang, et al. 2023), and the`scale` package (Wickham and Seidel 2023) to clean the data. We used the `tidyverse` package (Wickham, Hester, et al. 2023), the `maps` package (Becker et al. 2018), the `ggplot2` package (Wickham, Chang, et al. 2023), the `knitr` package (Xie 2015), the `kableExtra` package (Zhu 2021) to plot the dataset and the visualization.

In doing so, R code was adapted from (Alexander 2023).

First we did some general cleaning of this dataset, including replacing NAs in the data with means and changing the form of dates in the data.In addition, because each school had different enrolments, we calculated the percentage of different ethnicities in each school and selected the highest percentage of ethnicities in each school.

In addition, because the original data only recorded the shootings that occurred and not the ones that did not, but we wanted to simulate the logistic regression model, we generated synthetic data based on the cleaned dataset to reflect the data that did not occur. The main components that were simulated were latitude, longitude, school ethnicity, lunch, and staffing, and to ensure that the data were reasonable, the synthetic data were generated within the allowable range of variability (a reasonable float between the maximum and minimum of the variable), and to ensure that the model was implementable, the synthetic data were generated within the allowable range of variation.Synthetic data and raw data both have 415 observations.at the same time, we made a distinction between synthetic data and raw data inside the dataset by labelling their data source in data_source and creating a new binary variable to indicate whether a shooting occurred or not.

Finally, we cleaned the dataset of some unnecessary variables, such as the name and id of the school, the specific city neighbourhoods where the shootings took place (we focused our analysis on those states specifically), the shooter's access to the gun, and so on.

In the end, we cleaned up the data to include 830 observations and 31 variables, the dataset can be find in **?@tbl-data-overview**, only specific columns shows in the table.

Table 1: Summary of School Shooting Data

Year	State	School Type	Top 1 Race	Ulocale	Data Source	School Shooting
1999	CO	public	white	Suburb: Large	raw	1
1999	LA	public	black	City: Mid-size	raw	1
1999	GA	public	white	Suburb: Large	raw	1
1999	PA	public	black	City: Large	raw	1
1999	MA	public	black	City: Large	raw	1

1999	NM	public	hispanic	Town: Remote	raw	1
1999	OK	public	white	Town: Distant	raw	1
2000	FL	public	white	Suburb: Large	raw	1
2000	CA	public	hispanic	City: Small	raw	1
2000	IL	public	black	City: Large	raw	1

2.4 Outcome variables

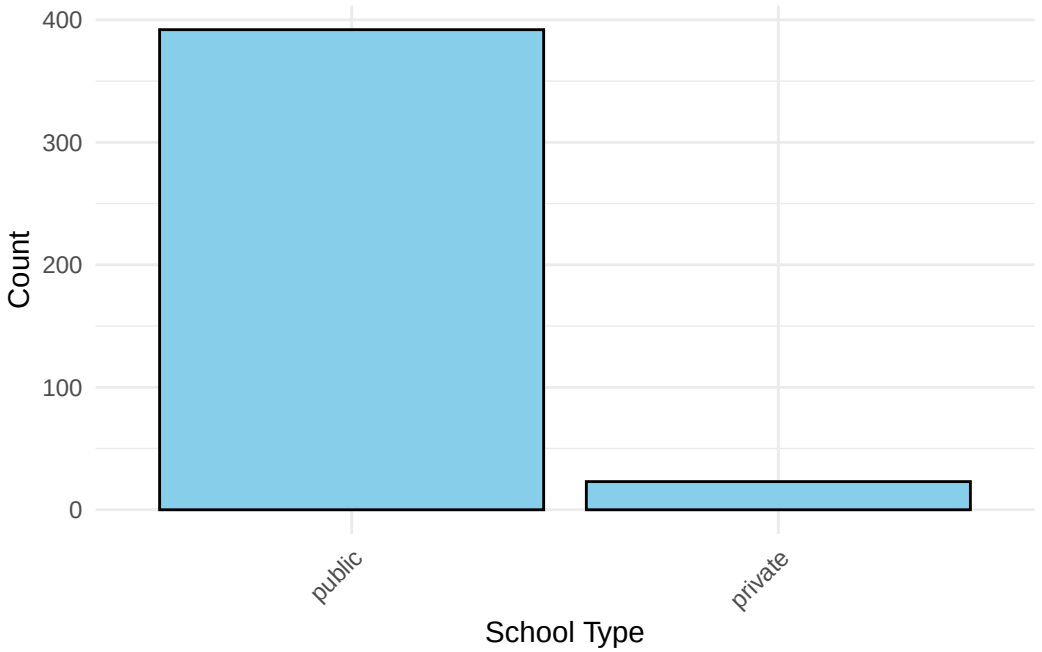

Figure 1: Distribution of School Types

The dataset reveals a notable disparity in the distribution of school types. As shown in Figure 1, public schools account for the vast majority of cases, while private schools are significantly underrepresented. This imbalance reflects the broader landscape of education in the United States, where public schools serve a larger student population. Moreover, this distribution is critical for our analysis, as it highlights the need to account for institutional differences when examining factors contributing to school shootings.

Figure 2 descriptions for schools with raw data highlights key geographic trends. As visualized in the bar chart, Suburb: Large and City: Large locales account for the majority of school shootings, reflecting the higher density of schools and populations in these areas. Smaller categories, such as Rural: Remote and Town: Fringe, are underrepresented, suggesting that sparsely populated regions experience fewer incidents. This pattern underscores the role of urbanicity in shaping exposure to risk factors like population density, socioeconomic conditions,

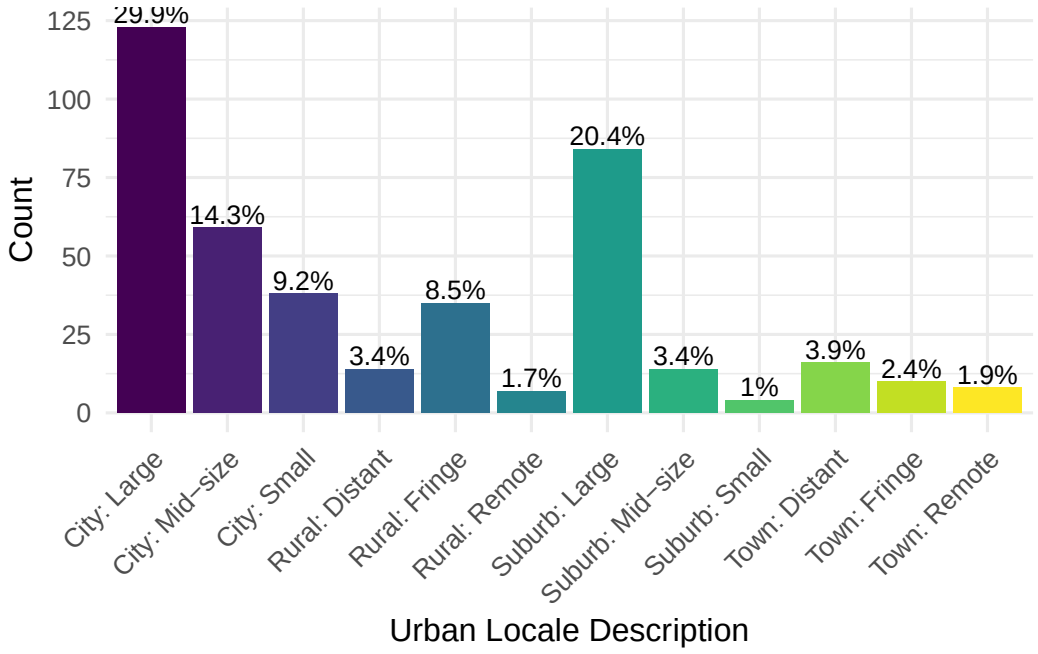

Figure 2: Distribution of Urban Locale Descriptions

and resource allocation. The findings emphasize the need for geographically tailored interventions, with a focus on urban and suburban areas that bear a disproportionate burden of school shootings. These insights can guide policymakers in targeting resources and preventive measures effectively.

The geographic distribution of school shootings across states is visualized in Figure 3. States such as California, Texas, and Florida exhibit higher frequencies of school shootings, reflecting their larger populations and more extensive school systems. Conversely, smaller states and those with lower populations, such as Wyoming and Vermont, report fewer incidents. This map underscores the importance of contextualizing school shootings within regional characteristics, including population size, school density, and state-specific policies.

The temporal trend in school shootings is depicted in Figure 4, showing a concerning increase in incidents over the years. This upward trend highlights the growing urgency of addressing the factors contributing to school shootings. Spikes in certain years may correspond to specific events or policy changes, suggesting that temporal patterns warrant further investigation to understand their underlying causes. These findings emphasize the importance of considering temporal dynamics in any analysis of school shooting data.

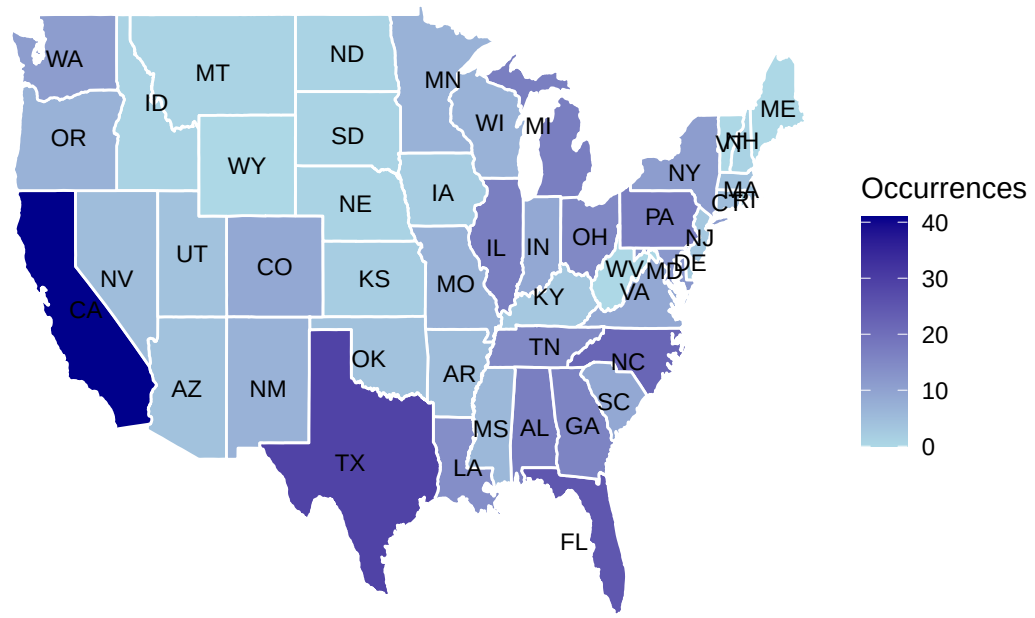

Figure 3: Shooting Occurrences by State

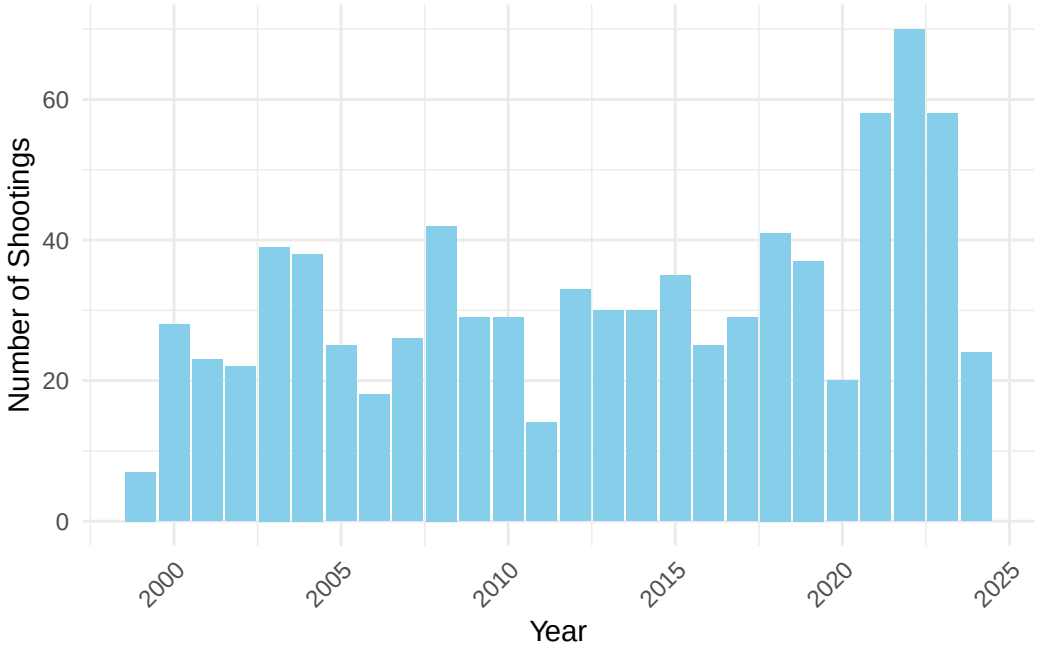

Figure 4: Number of Shootings by Year

2.5 Predictor variables

3 Model

We use the `brms` package (Bürkner 2017) and the `bayesplot` package (Gabry and Mahr 2021) to fit the model.

This analysis aimed to examine the relationship between school characteristics, demographic composition, and geographical location on the occurrence of school shootings.To achieve this, a Bayesian logistic regression model was employed. Logistic regression was deemed appropriate as the outcome variable (i.e., whether or not a school shooting occurred) is binary in nature. Bayesian methods were chosen to incorporate prior knowledge and provide probabilistic interpretations of the model parameters.

3.1 Model set-up

The Bayesian logistic regression model can be expressed as follows:

$$y_i \mid \pi_i \sim \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \cdot \text{year}_i + \beta_2 \cdot \text{school\_type}_i$$

where:

- y_i: Occurrence of a school shooting (1 if a shooting occurred, 0 otherwise).
- π_i: Probability of a school shooting occurring for observation i.
- β_0: Intercept, representing the baseline log-odds of a school shooting for the reference categories of the predictors.
- β_1, β_2, β_3, β_4: Effects of the predictors on the log-odds of a school shooting.

Priors:

- $\beta_0 \sim \text{Cauchy}(0, 2)$ (Intercept prior)
- $\beta_k \sim \text{Normal}(0, 10)$ for $k = 1, 2, 3, 4$ (Predictor priors)

3.2 Model justification

The inclusion of predictors was guided by theoretical considerations:

- **Year** (β_1): To assess temporal trends in school shootings, capturing whether incidents have increased or decreased over time.
- **School Type** (β_2): To examine if public versus private schools are differently associated with the likelihood of school shootings.
- **Dominant Race in the School Population** (β_3): To investigate the potential role of demographic composition in school shootings.
- **Urbanicity** (β_4): To explore how geographical location and urban-rural classifications affect school shooting occurrences.

Bayesian methods were used for their ability to incorporate prior knowledge and provide credible intervals for parameter estimates. Logistic regression was preferred as it ensures probabilities remain between 0 and 1, appropriate for binary outcomes.

3.3 Diagnostics and Visualizations

To evaluate the model, several diagnostic tools and visualizations were utilized:

1. **Posterior Predictive Check** Figure 5 was performed to compare the observed data with predictions from the model.
2. **Comparison of Posterior and Prior** Figure 8 illustrates how the posterior distribution was updated from the priors.
3. **Trace Plots** Figure 6 were used to check for MCMC convergence.
4. **Rhat Diagnostic Plot** Figure 7 was used to ensure all parameters had converged ($\hat{R} < 1.1$).
5. **90% Credibility Intervals** Figure 10 display the uncertainty around the parameter estimates.

These diagnostics confirmed good convergence and model fit, providing confidence in the estimated effects.

4 Results

This section reports the findings from the Bayesian logistic regression model examining the likelihood of school shootings. Parameter estimates, convergence diagnostics, and the implications of the results are presented below.

4.1 Parameter Estimates

The parameter estimates from the Bayesian logistic regression model are shown in Figure 10. Positive coefficients indicate an increased likelihood of school shootings, while negative coefficients indicate a decreased likelihood.

The results highlight several significant effects:

- **Temporal Trends**: The coefficient for `year` was positive ($\beta = 0.065$, 95% CI: [0.042, 0.088]), suggesting an increasing trend in school shootings over time.
- **School Type**: Public schools were associated with higher odds of shootings compared to private schools ($\beta = 0.788$, 95% CI: [0.069, 1.523]).
- **Demographic Composition**: Schools with a majority of Asian ($\beta = -6.007$, 95% CI: [-13.346, -0.230]) or Black students ($\beta = -5.194$, 95% CI: [-12.128, -0.255]) showed significantly reduced odds of school shootings.
- **Urbanicity**: Schools in rural and suburban areas had reduced odds of shootings compared to urban schools:
  - Remote rural areas ($\beta = -3.162$, 95% CI: [-4.168, -2.229])
  - Mid-sized suburban areas ($\beta = -2.528$, 95% CI: [-3.314, -1.792])

These findings suggest temporal, institutional, demographic, and geographic factors all contribute to the likelihood of school shootings.

4.2 Diagnostics

Model diagnostics confirmed the robustness of the Bayesian logistic regression.

Key diagnostic plots include:

1. **Posterior Predictive Check** Figure 5: The posterior predictive distribution aligns closely with the observed data, indicating a good model fit.
2. **Convergence**: Figure 6 displays trace plots for key parameters, showing well-mixed chains that reached stationarity. All parameters had ($\hat{R} < 1.1$), as shown in Figure 7.
3. **Credibility Intervals** Figure 10: The 90 credibility intervals for each parameter are plotted, highlighting significant predictors (intervals that exclude 0).

These diagnostics provide strong evidence of model convergence and fit, supporting the reliability of the results.

4.3 Implications

The results have several implications for understanding and addressing school shootings:

- The increasing trend over time underscores the urgency of targeted interventions to reduce school shootings.
- The higher risk in public schools suggests that policies and resources should prioritize these institutions.
- Geographic patterns indicate that rural and suburban schools have lower risks compared to urban schools, which may reflect differences in community or school characteristics.
- Demographic findings highlight the need for equitable and inclusive approaches to address structural inequalities potentially influencing school shootings.

Overall, these findings provide a robust framework for designing policies to mitigate the risk of school shootings.

5 Discussion

This study examined the temporal, institutional, demographic, and geographic factors influencing the likelihood of school shootings using a Bayesian logistic regression model. The model provided robust insights, with credible intervals indicating the degree of uncertainty around the estimated effects Table 2.

5.1 Key Findings

The results revealed several significant predictors of school shootings:

- **Temporal Trends**: The positive association between `year` and the likelihood of school shootings highlights a concerning trend over time. This finding underscores the urgency of implementing effective preventative measures to reverse this trajectory.
- **Institutional Context**: Public schools showed a significantly higher risk compared to private schools. This result suggests the need for targeted interventions in public school settings, potentially focusing on resource allocation, security measures, and community engagement.
- **Demographic Composition**: The negative coefficients for schools with predominantly Asian or Black populations indicate a reduced likelihood of shootings in these contexts.

  These results call for further investigation into how demographic, cultural, or social factors may create protective environments.

- **Geographical Variation**: Rural and suburban schools were consistently associated with lower odds of school shootings compared to urban schools. This pattern emphasizes the importance of location-specific prevention strategies that account for urban challenges, such as population density and resource distribution.

5.2 Strengths and Limitations

5.2.1 Strengths:

1. **Bayesian Framework**: The use of Bayesian methods allowed for the incorporation of prior information and the quantification of uncertainty, resulting in more nuanced interpretations of the model coefficients.
2. **Model Diagnostics**: Comprehensive diagnostic checks (Figures 1-5) confirmed model convergence and goodness-of-fit, lending credibility to the findings.

5.2.2 Limitations:

1. **Generalizability**: While the model captures key predictors, external factors such as policy changes or economic conditions were not included. These factors may mediate or moderate the observed relationships.
2. **Data Availability**: The quality and completeness of the data used to fit the model may have introduced bias or omitted variable effects. Future studies should aim to include additional covariates to improve explanatory power.

5.3 Implications and Future Directions

The findings have several implications for policy and practice:

- **Policy Interventions**: Addressing the rising trend in school shootings will require concerted efforts at the institutional level, including targeted investments in public schools.
- **Equity and Inclusion**: The protective effects associated with certain demographic groups suggest that inclusive, community-driven strategies may play a role in prevention.
- **Location-Specific Strategies**: Policymakers should tailor interventions to the unique needs of urban, suburban, and rural schools, recognizing the contextual factors that influence risk.

Future research should explore additional variables, such as socioeconomic factors, legislative impacts, and mental health resources, to provide a more comprehensive understanding of the dynamics driving school shootings.

Table 2: Summary of the Bayesian Logistic Regression Model

	Estimate	Lower CI	Upper CI
Intercept	-125.423	-173.459	-78.394
year	0.065	0.042	0.088
school_typepublic	0.788	0.069	1.523
top_1_racesasian	-6.007	-13.346	-0.230
top_1_racesblack	-5.194	-12.128	-0.255
top_1_raceshispanic	-5.164	-12.066	-0.228
top_1_raceswhite	-4.595	-11.516	0.279
ulocale_descCity:MidMsize	-1.137	-1.704	-0.593
ulocale_descCity:Small	-0.491	-1.203	0.224
ulocale_descRural:Distant	-2.094	-2.952	-1.259
ulocale_descRural:Fringe	-1.599	-2.239	-0.969
ulocale_descRural:Remote	-3.162	-4.168	-2.229
ulocale_descSuburb:Large	-0.325	-0.922	0.294
ulocale_descSuburb:MidMsize	-2.528	-3.314	-1.792
ulocale_descSuburb:Small	-3.589	-4.902	-2.521
ulocale_descTown:Distant	-2.423	-3.215	-1.658
ulocale_descTown:Fringe	-2.104	-3.027	-1.241
ulocale_descTown:Remote	-3.259	-4.151	-2.389

6 Appendix

6.1 A Additional Data Details

This analysis relied on data from the Washington Post dataset, which provides detailed records of school shootings in the United States from 1999 to 2024. To enhance the robustness of the model, synthetic data were generated for schools where shootings did not occur, following the methodology described in Section 2.3. The inclusion of synthetic data allowed for a balanced dataset of 830 observations and 31 variables.

The data cleaning process included: - Replacing missing values with means. - Converting date formats for consistency. - Calculating demographic percentages for each school and retaining only the most prominent demographic group per school. - Excluding variables that were not relevant to the analysis, such as school names and shooter-specific details.

Table 2 provides a summary of the cleaned dataset and key variables.

6.2 B Model Details

6.2.1 B.1 Posterior Predictive Check

The posterior predictive check was performed to evaluate how well the model captures the observed data distribution. Figure 5 demonstrates that the model predictions align closely with the observed data, indicating a good model fit.

6.2.2 B.2 Diagnostics

To ensure the reliability of the Bayesian logistic regression model, the following diagnostics were conducted: 1. **Trace Plots:** Figure 6 shows well-mixed chains for all parameters, confirming convergence. 2. **Rhat Plot:** Figure 7 illustrates that all parameters have $\hat{R} \leq 1.1$, indicating no convergence issues.

6.2.3 B.3 Comparison of Priors and Posteriors

Figure 8 compares the posterior distributions with the priors, highlighting how the data influenced the parameter estimates.

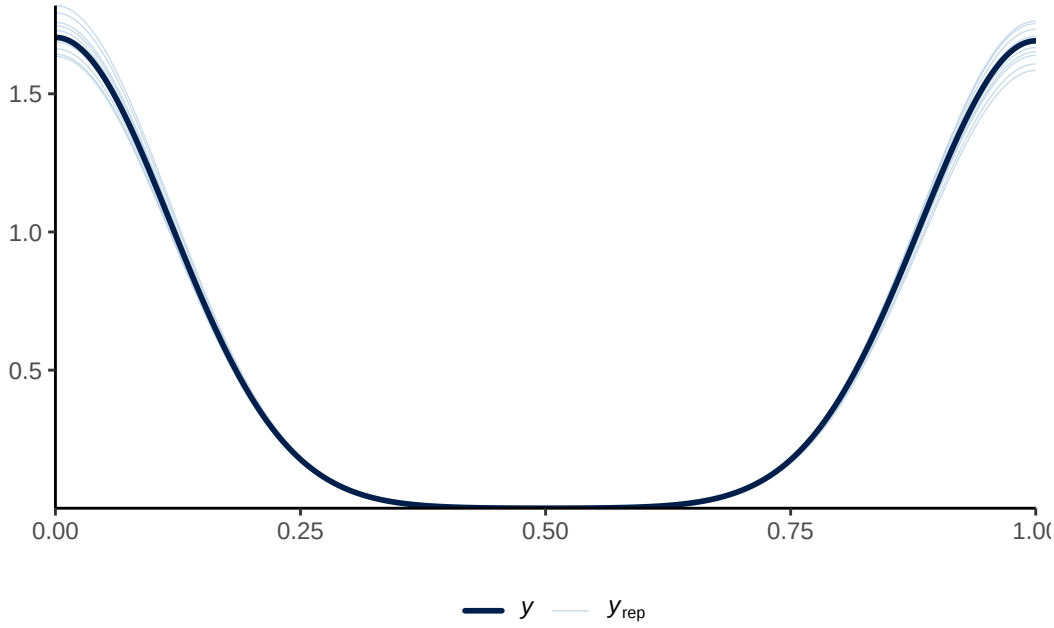

Figure 5: Posterior predictive check

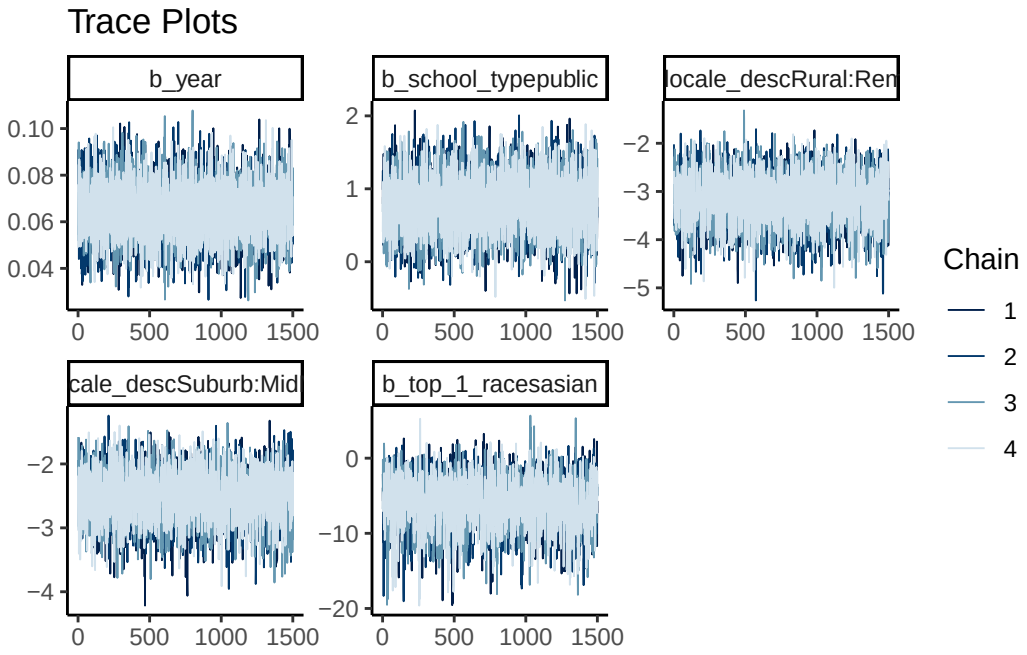

Figure 6: Trace plots for selected parameters

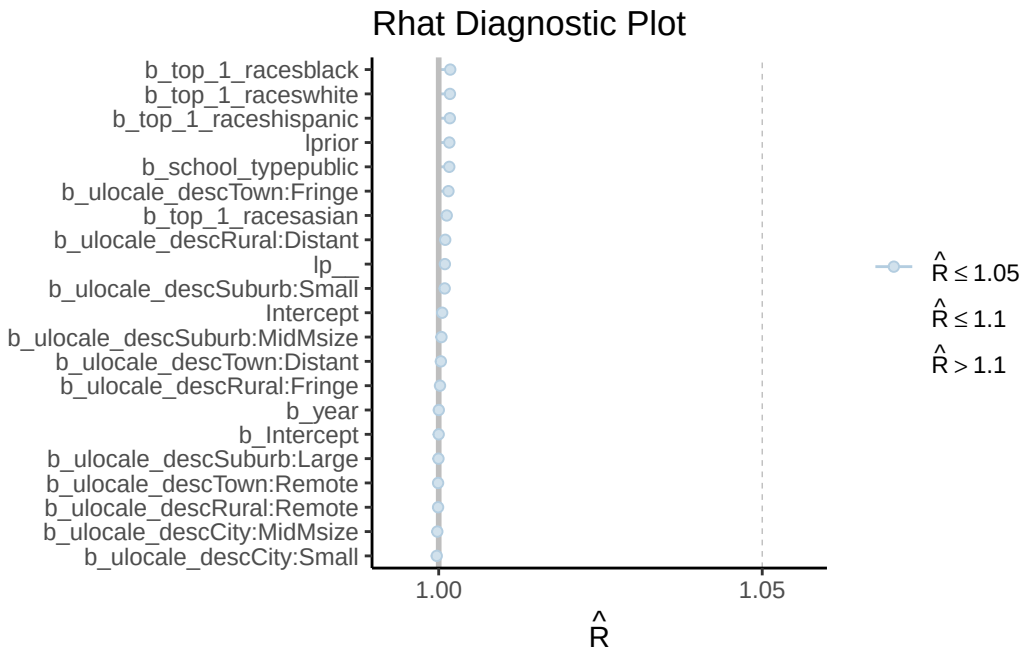

Figure 7: Rhat diagnostic plot for model convergence

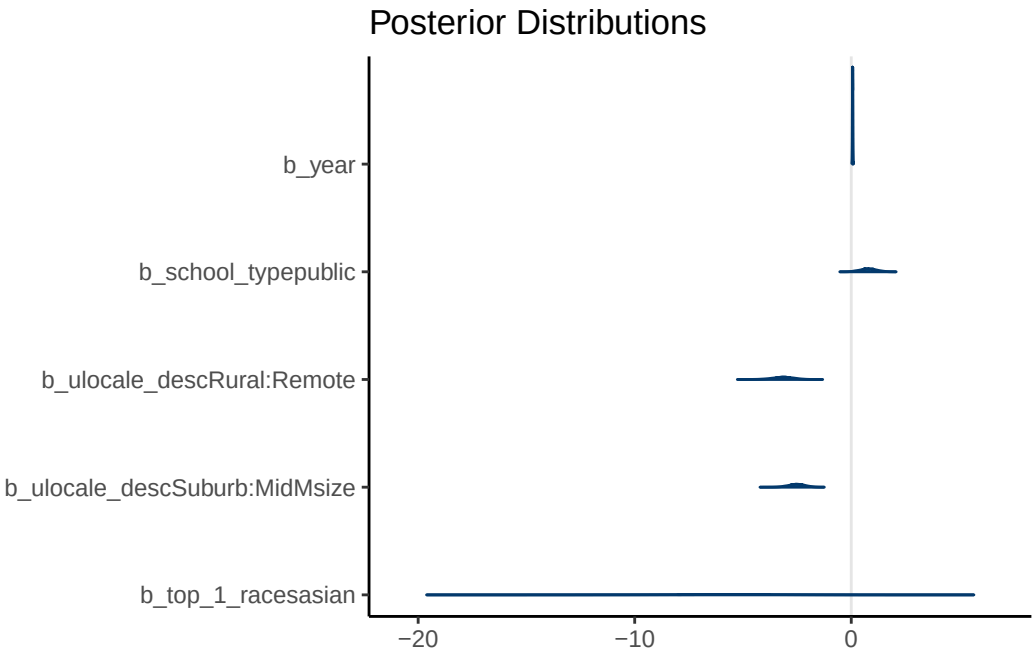

Figure 8: Comparison of posterior and prior distributions

6.3 C Survey and Sampling Methodology

6.3.1 C.1 Population, Frame, and Sample

- **Population:** All K-12 schools in the United States.
- **Frame:** Schools from the Washington Post dataset, combined with synthetic data for schools without shootings.
- **Sample:** Stratified by school type, urban locale description, and dominant demographic group.

6.3.2 C.2 Sampling and Recruitment

- Synthetic data were generated within observed variable ranges to represent schools where shootings did not occur.
- Observed data were supplemented by demographic, geographic, and enrollment metrics.

6.3.3 C.3 Questionnaire and Validation

- **Survey Focus:** School security measures, demographic composition, geographic details, and historical incidents.
- **Validation:** Cross-referenced data with administrative records to ensure accuracy.

6.4 D Simulation Details

To simulate the occurrence of school shootings, synthetic data were generated with random variability: - Geographic coordinates were adjusted within observed ranges. - Demographics and enrollment were assigned proportionally based on observed distributions.

Figure 9 visualizes the geographic distribution of shooting occurrences, highlighting both observed and synthetic data points. ## E Implications of Results

The analysis revealed the following key insights: 1. **Temporal Trends:** The increasing trend in shootings over time (**?@fig-trend**) underscores the need for immediate interventions.

2. **Institutional and Demographic Context:** Public schools and urban locales were associated with higher risks, as shown in Figure 10. These findings provide actionable insights for policymakers aiming to mitigate risks associated with school shootings.

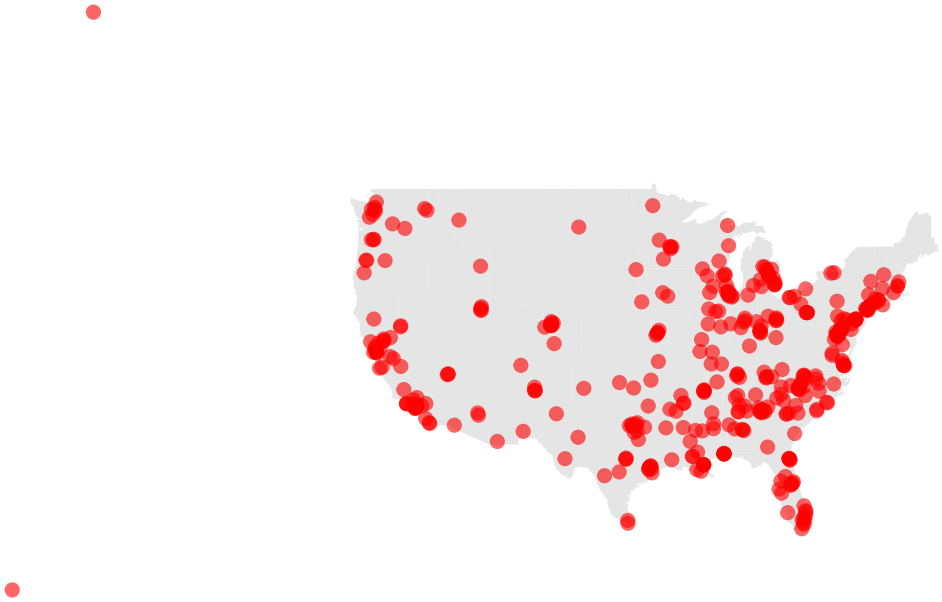

Figure 9: School Shooting Occurrence

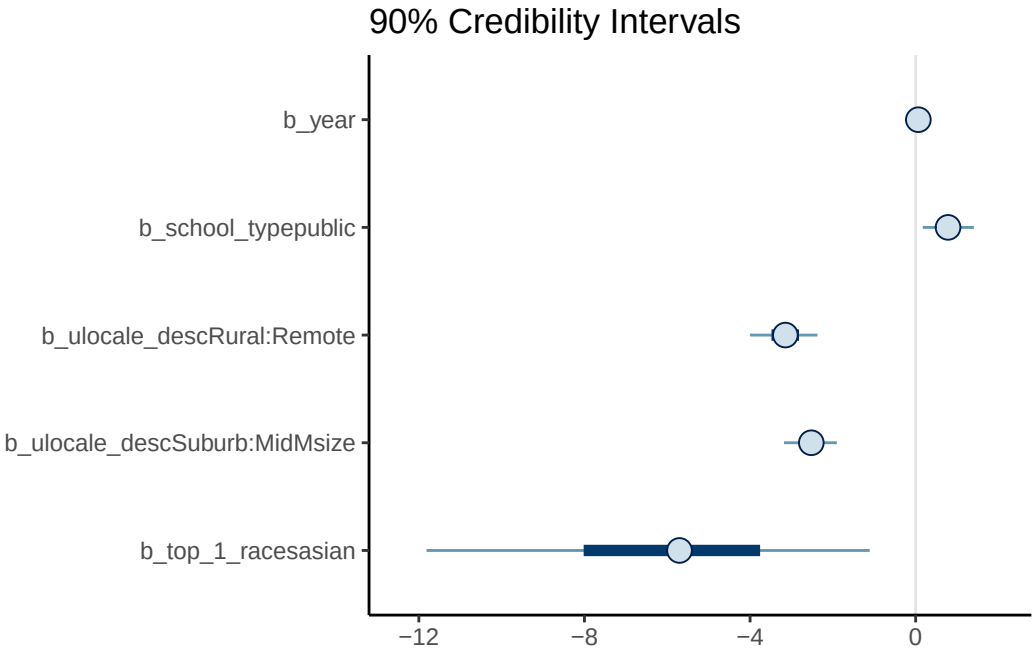

Figure 10: 90% credibility intervals for selected predictors

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Becker, Richard A., Allan R. Wilks, Ray Brownrigg, Thomas P. Minka, and Alex Deckmyn. 2018. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Brownrigg, Raymond J., Thomas P. Minka, Nicholas Lewin-Koh, and Roger Bivand. 2023. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Bürkner, Paul-Christian. 2017. *Brms: An r Package for Bayesian Multilevel Models Using Stan*. https://doi.org/10.18637/jss.v080.i01.

Gabry, Jonah, and TJ Mahr. 2021. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

Post, The Washington. 2021. "Data on School Shootings." https://github.com/washingtonpost/data-school-shootings/tree/master.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, Lionel Henry, and many others. 2023. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2015. *Dynamic Documents with r and Knitr*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.