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 14, 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 originates from the Washington Post (Post 2021) and spans 1999 to 2024, documenting school shootings in the United States. The raw dataset includes 416 observations and 50 variables, detailing information such as the time, location, racial demographics of schools, and shooter characteristics. This study extends the dataset by simulating additional observations to balance the binary outcome variable (shooting occurrence), resulting in a dataset with 830 observations and 31 variables. The dataset provides a robust basis for examining the predictors of school shootings, including temporal trends, demographic composition, and urbanicity.

2.2 Measurement

The dataset is a product of rigorous investigation by the Washington Post, combining diverse and credible sources to create a comprehensive record of school shootings from 1999 to the present (Post 2021). Real-world incidents—where a gun was fired and at least one person was injured or killed on school property—were systematically transformed into structured data entries through a transparent and detailed process.

2.2.1 Data Sources and Process

The measurement process involved triangulating information from:

- **News Reports**: Verified through multiple outlets for accuracy.
- **Law Enforcement Records**: Provided timelines, locations, and key details.
- **School Data**: Publicly available records enriched the dataset with demographic and institutional contexts.

Each incident was identified, verified, and contextualized to ensure accuracy. Variables such as school type, urban locale description, and racial composition were drawn from NCES data and school enrollment records, with additional details cross-referenced for completeness. Where gaps existed, synthetic data were carefully generated to maintain analytical consistency.

2.2.2 Strengths of the Dataset

The dataset's strengths lie in its:

- **Comprehensiveness**: Captures a wide range of school shootings with detailed contexts.
- **Transparency**: Relies on well-documented and publicly accessible sources.
- **Relevance**: Includes key variables such as urbanicity and demographics critical to understanding school shootings.

These qualities ensure that the dataset provides a reliable foundation for examining the complex factors contributing to school shootings.

2.3 Data Cleaning

We used the`R`programming language (R Core Team 2023), the `here` package (Müller 2020),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 datasets by labeling 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 Table 1, only specific columns shows in the table.

Table 1: The selected columns in the dataset

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 Data visualization

The binary outcome variable indicates whether a school shooting occurred (1) or not (0). As shown in Figure 1, public schools dominate the dataset, reflecting their prominence in the U.S. educational landscape. Private schools are significantly underrepresented, necessitating careful modeling to account for this imbalance.

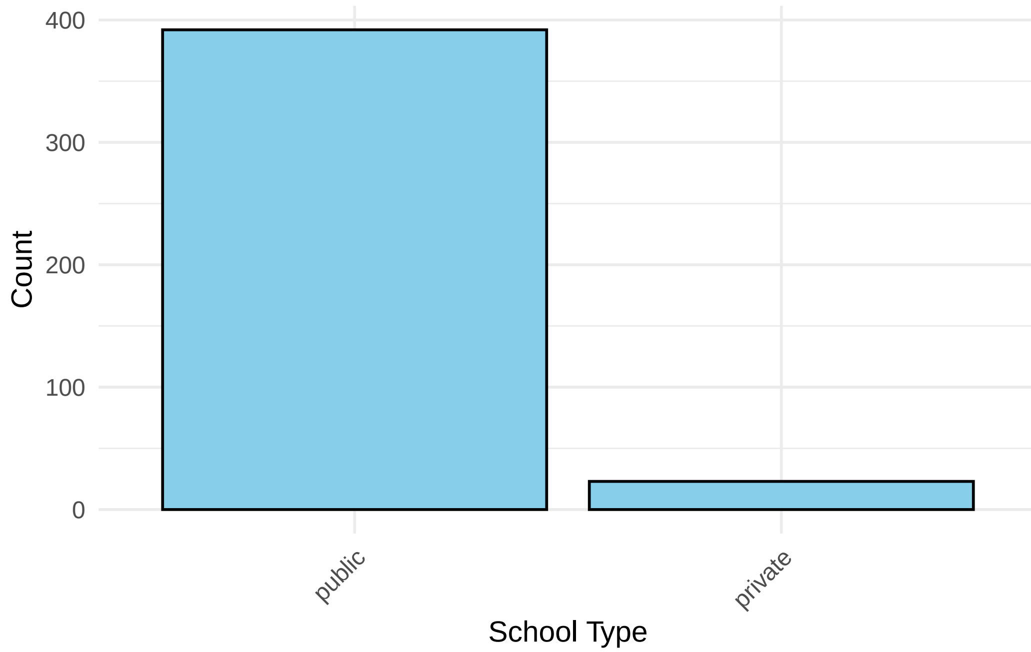

Figure 1: Distribution of School Types

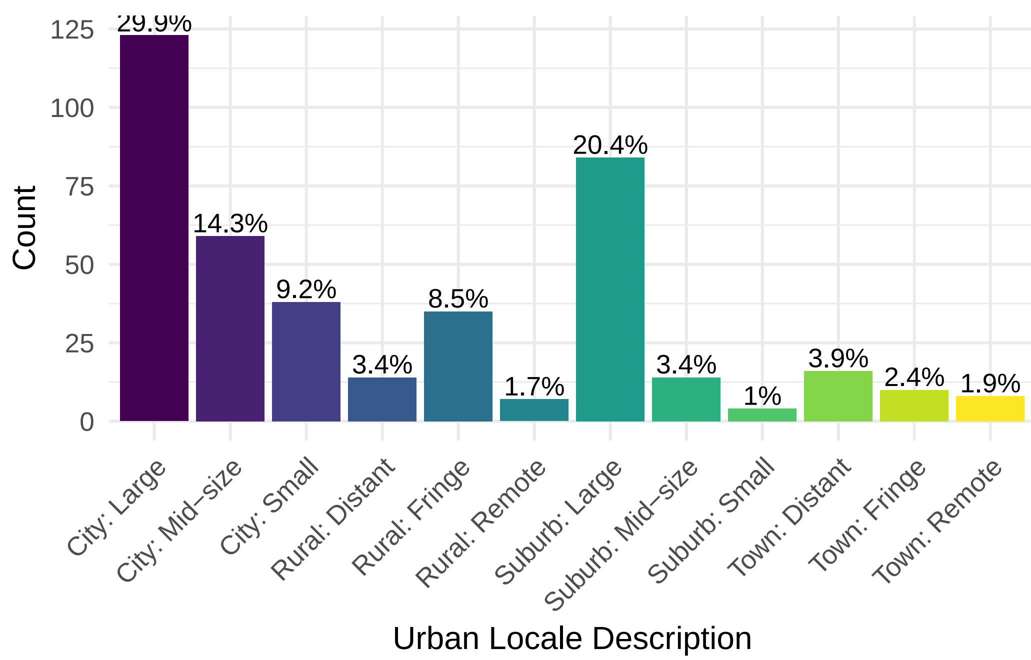

Figure 2: Distribution of Urban Locale Descriptions

Figure 2 illustrates the distribution of urban locales for shootings, with **City: Large** and **Suburb: Large locales** representing the majority of incidents. Smaller locales, such as **Town: Fringe** and **Rural: Remote**, are underrepresented. This trend reflects the impact of population density and urbanicity on school shooting occurrences.

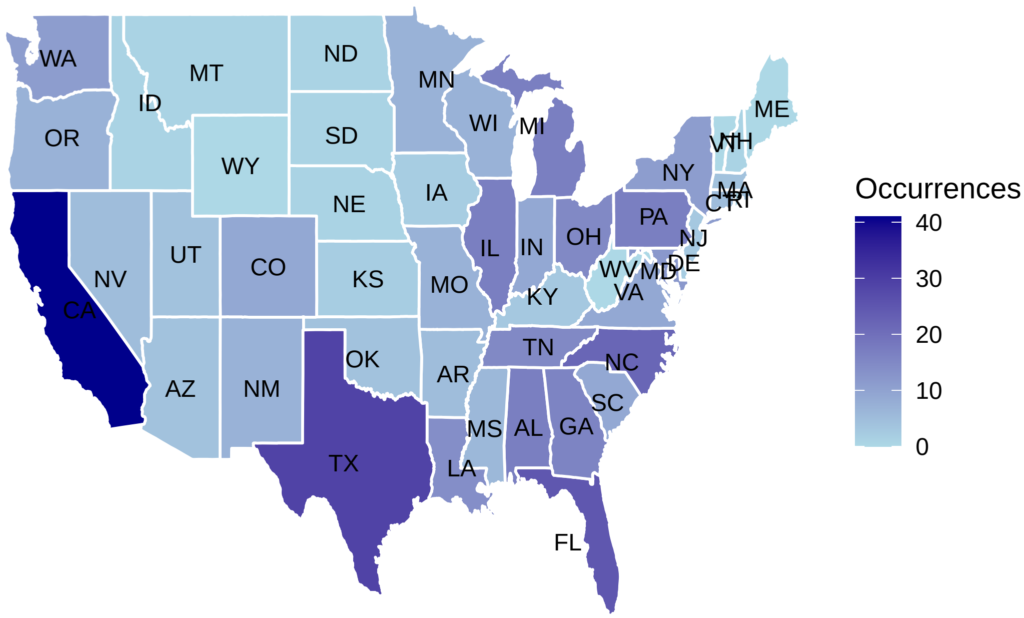

Figure 3: Shooting Occurrences by State

Figure 3 maps the distribution of shootings across states, highlighting regions like California, Texas, and Florida as hotspots due to their large populations and extensive school systems.

Figure 4 depicts an alarming increase in school shootings over time. This trend emphasizes the urgency of understanding and mitigating factors contributing to these incidents.

3 Model

3.1 Model set-up

To analyze the factors contributing to school shootings, we use a Bayesian logistic regression model. This approach is well-suited to modeling the binary outcome variable (whether a school shooting occurred) while incorporating prior knowledge about the predictors. The model is defined as:

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

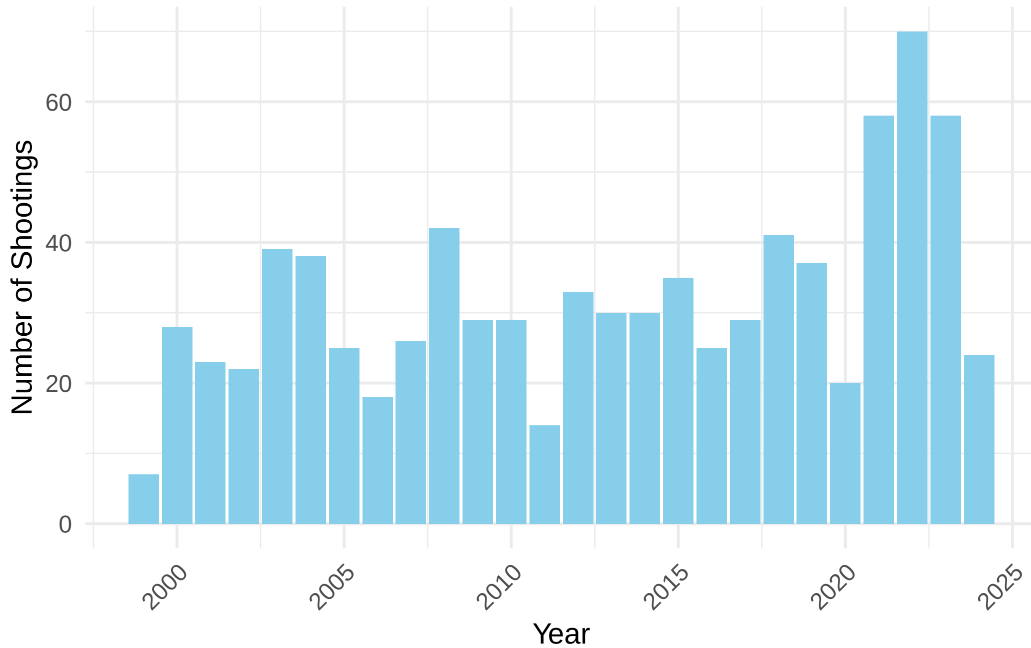

Figure 4: Number of Shootings by Year

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \cdot \text{year}_i + \beta_2 \cdot \text{school\_type}_i + \beta_3 \cdot \text{ulocale\_desc}_i + \beta_4 \cdot \text{enrollment}_i + \beta_5 \cdot \text{top\_1\_races}_i$$

where:

- y_i: Binary response variable indicating whether a school shooting occurred ((y_i = 1)) or not ((y_i = 0)).
- π_i: Probability of a school shooting for observation (i).
- β_k: Coefficients representing the influence of predictors on the log-odds of a school shooting.

3.1.1 3.2 Variables and Priors

3.1.1.1 Predictors

- **Year**: Captures temporal trends in shootings.
- **School Type**: Public vs. private schools.
- **Urban Locale**: Urbanicity levels from NCES data.

- **Enrollment**: Total number of students in the school.
- **Dominant Race**: Primary racial group in the school.

3.1.1.2 Priors

- Intercept β_0: Normal$(0, 5)$, reflecting baseline uncertainty.
- Coefficients β_k: Normal$(0, 2.5)$, allowing moderate variability.

These priors prevent overfitting while allowing data-driven insights.

3.2 Results

3.2.1 Model Estimates

Table Table 2 presents the Bayesian logistic regression coefficients, while Figure Figure 10 visualizes 90% credible intervals for the predictors.

Key findings include:

• Temporal Trends: Positive association with increasing years.

• School Type: Higher risk in public schools.

• Urban Locale: Elevated risk in urban and suburban schools.

• Enrollment: Modest increase in risk with larger enrollment.

• Dominant Race: Schools with predominantly White or Hispanic populations are at higher risk.

3.2.2 Posterior Predictive Checks

Figure Figure 5 shows the posterior predictive checks (PPCs), which validate the model's fit by comparing observed and simulated data.

Table 2: Bayesian Logistic Regression Coefficients

Table 3: Bayesian Logistic Regression Coefficients

	(1)
b_Intercept	−125.368
b_year	0.065
b_school_typepublic	0.782
b_top_1_racesasian	−5.705
b_top_1_racesblack	−4.869
b_top_1_raceshispanic	−4.829
b_top_1_raceswhite	−4.269
b_ulocale_descCity × MidMsize	−1.134
b_ulocale_descCity × Small	−0.492
b_ulocale_descRural × Distant	−2.092
b_ulocale_descRural × Fringe	−1.596
b_ulocale_descRural × Remote	−3.147
b_ulocale_descSuburb × Large	−0.325
b_ulocale_descSuburb × MidMsize	−2.521
b_ulocale_descSuburb × Small	−3.546
b_ulocale_descTown × Distant	−2.420
b_ulocale_descTown × Fringe	−2.095
b_ulocale_descTown × Remote	−3.251
Num.Obs.	827
R2	0.277
ELPD	−468.5
ELPD s.e.	14.6
LOOIC	937.0
LOOIC s.e.	29.3
WAIC	935.6
RMSE	0.43

3.3 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.

3.4 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.

3.5 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.

4 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 4.

4.1 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.

4.2 Strengths and Limitations

4.2.1 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.

4.2.2 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.

4.3 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 4: 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

5 Appendix

5.1 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 4 provides a summary of the cleaned dataset and key variables.

5.2 B Model Details

5.2.1 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.

5.2.2 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.

5.2.3 B.3 Comparison of Priors and Posteriors

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

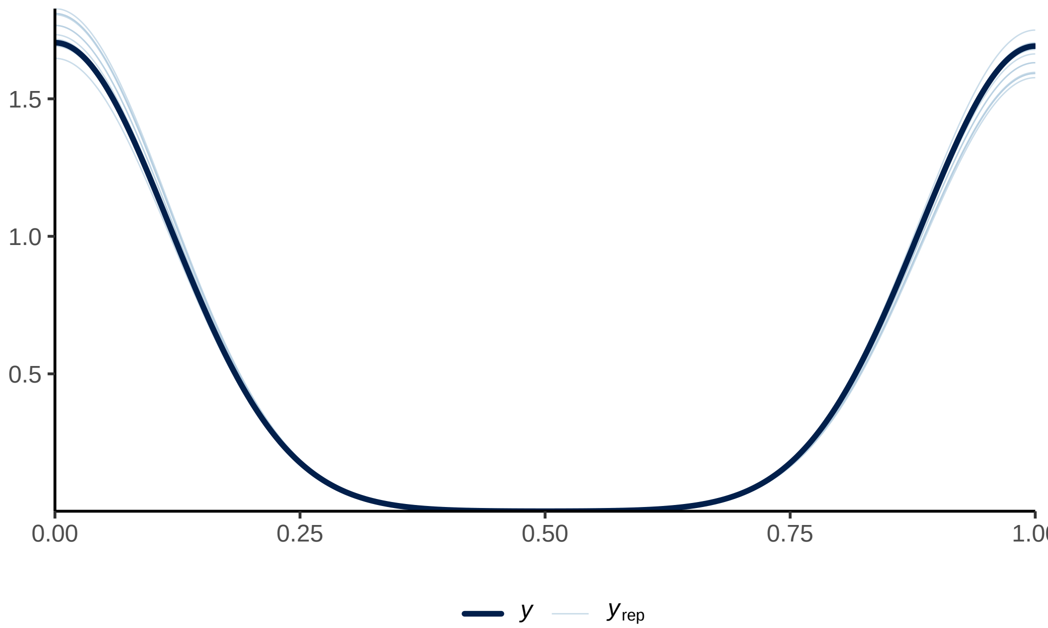

Figure 5: Posterior predictive check

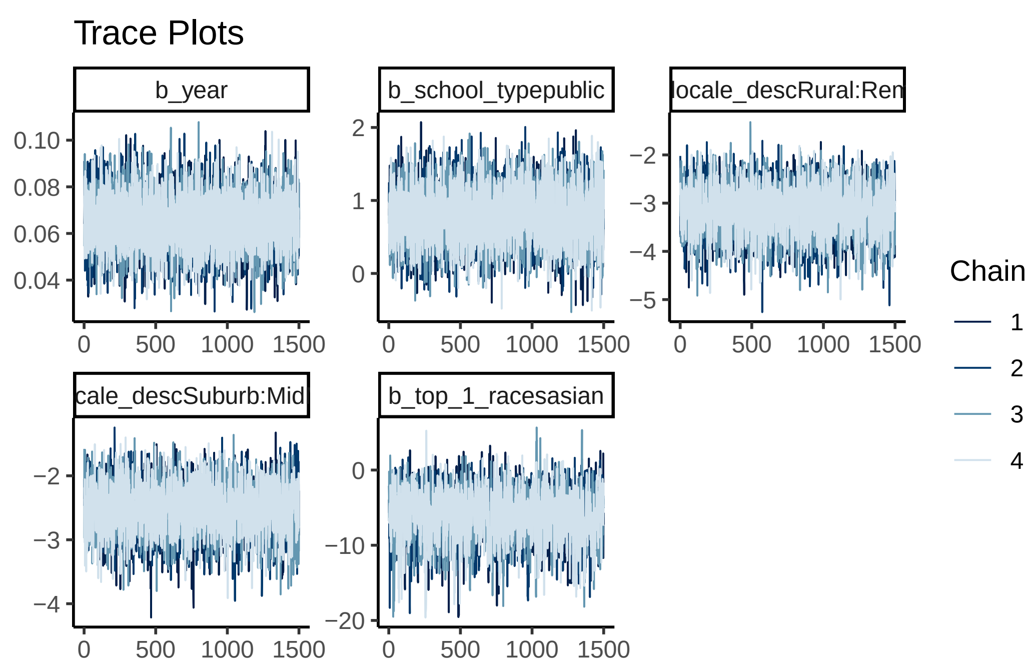

Figure 6: Trace plots for selected parameters

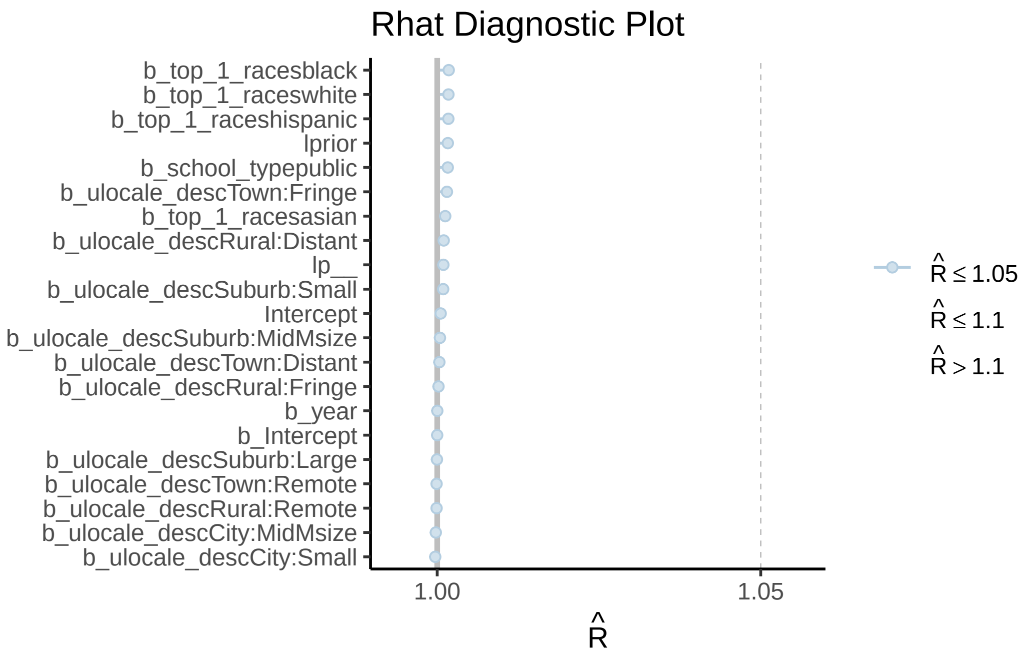

Figure 7: Rhat diagnostic plot for model convergence

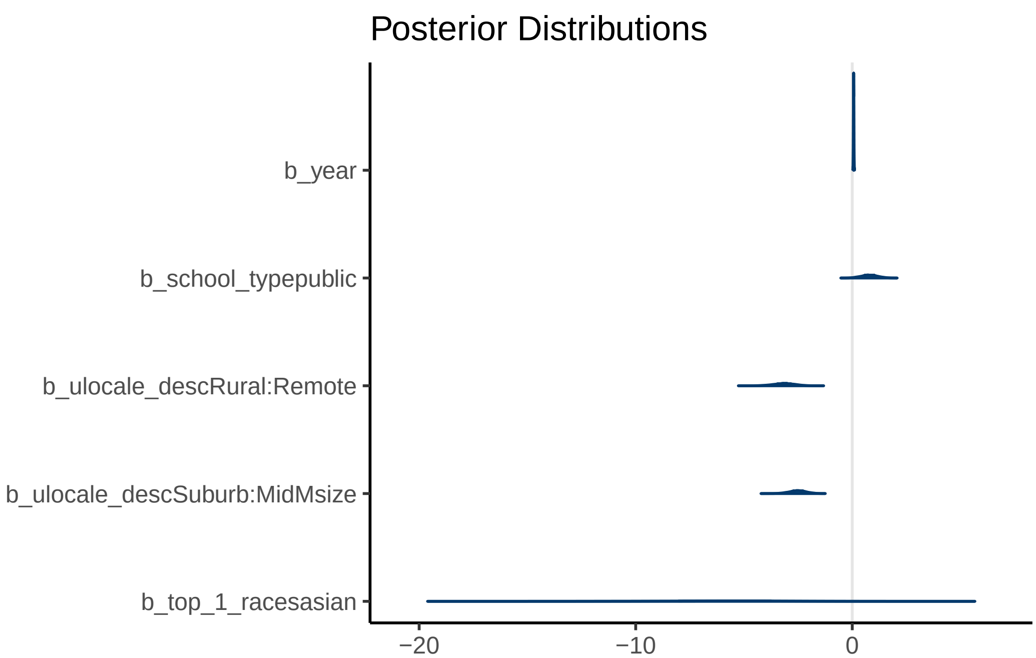

Figure 8: Comparison of posterior and prior distributions

5.3 C Survey and Sampling Methodology

5.3.1 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.

5.3.2 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.

5.3.3 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.

5.4 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.

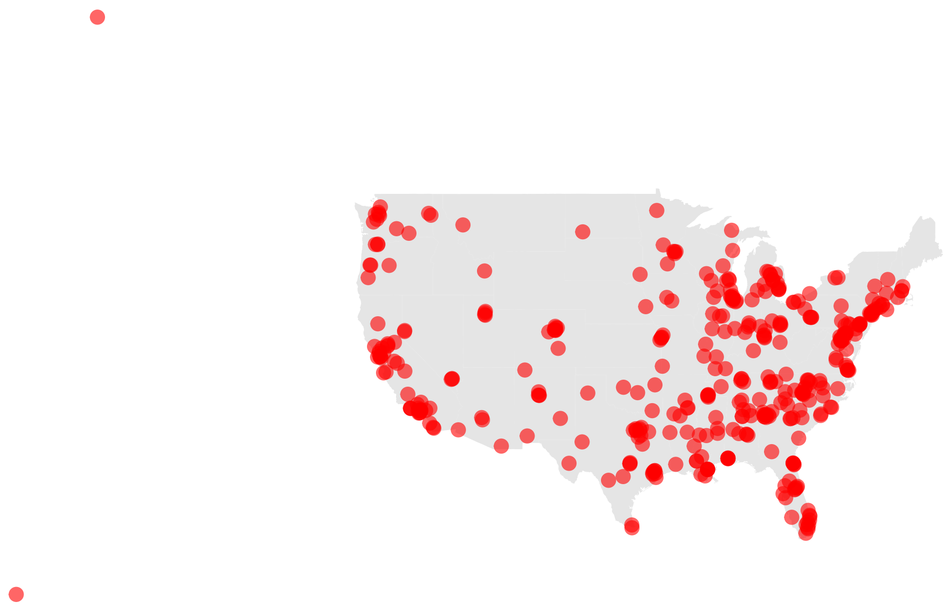

Figure 9: School Shooting Occurrence

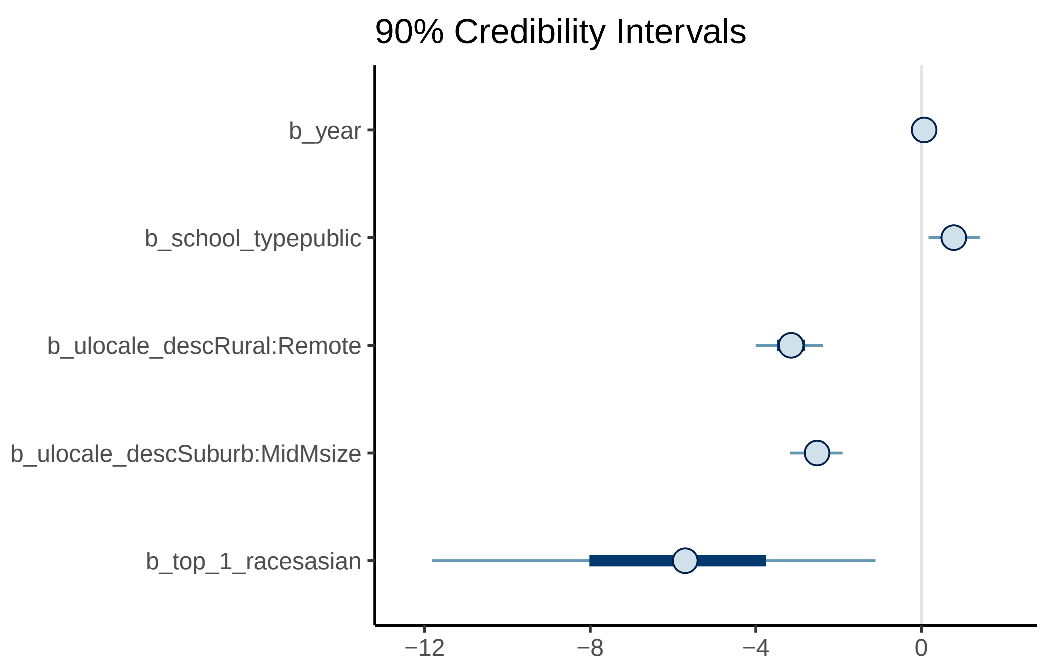

Figure 10: 90% credibility intervals for selected predictors

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.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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.