Gunfire on School Grounds in the United States: Patterns, Trends, and Insights

Madiha Zara
School of Management
Wentworth Institute of Technology
Boston, MA, USA
zaram1@wit.edu

ABSTRACT

School gunfire incidents have become an increasingly critical issue in the United States, affecting students, educators, and communities across multiple decades. This project investigates patterns and trends in K–12 school gunfire incidents from 1966 to 2025 using data from the K–12 School Shooting Database (SSDB). Two key research questions guide this analysis: (1) How has the frequency and severity of school gunfire incidents changed over time? (2) Which states have experienced the highest number of incidents, and how do these trends differ regionally?

Using yearly aggregation, state-level grouping, and a Linear Regression model implemented through scikit-learn, the analysis quantifies long-term temporal trends and identifies geographic disparities. Results show a significant upward trend, with an estimated increase of approximately 2.7 incidents per year over the six-decade period. Spatial analysis indicates that states such as California, Texas, Illinois, and Florida experience disproportionately high numbers of incidents. The findings highlight the growing prevalence of school gunfire incidents and reveal region-specific concentrations that may inform targeted policy interventions and school safety strategies.

KEYWORDS

Insert 3-5 keywords for your project.

1 Introduction

Gunfire on school grounds has emerged as one of the most pressing school safety concerns in the United States over the past several decades. Although school shootings represent a relatively small proportion of overall gun violence, their psychological, social, and educational impacts are disproportionately severe. Prior research shows increasing complexity and frequency of school-associated gunfire incidents across multiple decades, as documented by national monitoring organizations and federal agencies [1]–[4].

This study was undertaken to quantitatively examine long-term trends in school gunfire incidents and identify geographic disparities that may inform prevention and policy efforts. The purpose of the research is to use historical data to understand how the frequency of these incidents has changed over time and to identify which regions of the country experience higher concentrations of incidents.

Two research questions guide this analysis:

1. How has the frequency and severity of school gunfire incidents changed over time in the United States?
2. Which states have experienced the highest number of incidents, and how do these patterns differ regionally?

The underlying hypothesis is that the number of school gunfire incidents has **increased over the past six decades** and that the distribution of incidents is **not uniform across states**, with certain states experiencing disproportionately higher event frequencies. To investigate these questions, this project uses the K–12 School Shooting Database (SSDB) as its primary data source and applies descriptive analytics, temporal modeling, and spatial aggregation techniques.

2 Data

The dataset used in this project was obtained directly from the **K–12 School Shooting Database (SSDB)**, maintained by David Riedman through the Center for Homeland Defense and Security (CHDS). The SSDB is not freely downloadable in raw form; instead, researchers are required to request access via email. In accordance with the usage guidelines printed on the dataset's cover page, I contacted the dataset owner at k12ssdb@gmail.com, provided my name, academic affiliation, and the intended educational purpose of this project, and received written permission to use the raw Excel file for academic analysis.

The SSDB is widely recognized as a credible and authoritative data source. According to the project's methodology documentation, the database is continuously maintained, verified, and updated. Incidents are compiled from law enforcement reports, open-source materials, court documents, government records, and media accounts to ensure completeness and accuracy. Each entry is reviewed before inclusion, and the dataset includes any incident in which a gun was **fired**, **brandished**, or a **bullet struck school property**, regardless of injuries, time of day, or day of the week.

The dataset provided to me represents the most recent public release as of the date of access (September 9, 2025). I did not create or modify the original dataset structure; my work is limited to analytic preprocessing for this project.

The dataset used in this project was provided in XLSX (Excel) format and had a file size of approximately 3.57 MB. It contains multiple worksheets, but this project specifically used the Incident worksheet, which documents each gunfire-related incident occurring on or near K–12 school grounds. This sheet includes detailed event-level information covering what happened, when and where it occurred, who was involved, and contextual descriptors of each incident.

The Incident sheet consists of 3,163 rows (one per incident) and 50 columns, each representing a specific event attribute. Table 1 provides an overview of the variables included in this worksheet.

Column Name	Description
Incident_ID	Unique identifier for each incident.
Month	Month the incident occurred (numeric or categorical).
Day	Day of the month the incident occurred.
Year	Year of occurrence (as provided in the dataset).
Date	Full date of incident (MM/DD/YYYY or similar).
School	Name of the school where the event occurred.
Victims_Killed	Number of individuals killed in the incident.
Victims_Wounded	Number of individuals wounded.
Number_Victims	Total victims (killed + wounded).
Shooter_Killed	Whether a shooter was killed (binary or categorical).
Source	Source of information used for documentation.
Number_News	Count of news sources referencing the incident.
Media_Attention	Level of media attention reported.
Reliability	Data reliability score assigned by the dataset curator.
Quarter	Quarter of the year (Q1–Q4).
City	City where the incident occurred.
State	U.S. state where the incident occurred.
School_Level	Level of school (elementary, middle, high, K–12, etc.).
Location	Where the event occurred on/near campus.
Location_Type	Indoor, outdoor, bus, parking lot, etc.
During_Classes	Whether the incident occurred during school hours.
Time_Period	Time-of-day descriptor (morning, afternoon, etc.).
First_Shot	Where the first shot occurred.
Duration_min	Duration of incident in minutes.
Summary	Short textual summary of the incident.
Narrative	Full narrative description.
Situation	Classification of incident type (accidental, dispute, targeted, etc.).
GV_Type	Type of gun violence (e.g., drive-by, suicide, etc.).
Involves_Students_Staff	Whether students/staff were involved.
Targets	Intended target(s) if any.
Accomplice	Whether an accomplice was involved.
Accomplice_Narrative	Details about accomplices.
Hostages	Whether hostages were taken.
Barricade	Whether barricades were used.
Officer_Involved	Whether law enforcement fired shots or intervened.
Bullied	Whether bullying was a factor.
Domestic_Violence	Whether domestic violence was involved.
Gang_Related	Whether the incident was gang-related.
Active_Shooter_FBI	Active shooter classification per FBI criteria.
Multiple_Location	Whether the incident occurred across multiple locations.
Preplanned	Whether the incident was pre-planned.
SRO_School	Whether a School Resource Officer was present.
Security_Screening	Security/screening measures at the site.
Screening_Outcome	Screening result, if applicable.
Shots_Fired	Number of shots fired.
School_Lockdown	Whether the incident triggered a lockdown.
LAT	Latitude of incident location.
LNG	Longitude of incident location.
Campus_Type	Urban, suburban, or rural campus classification.
Zipcode	Postal code of the incident location.

For this project, only the **Incident** worksheet was used because it directly aligns with the two research questions:

1. How has the frequency and severity of school gunfire incidents changed over time?
2. Which states have experienced the highest number of incidents, and how do these trends differ regionally?

Minimal preprocessing was needed because the dataset is well structured. The following steps were applied:

2.1 Date parsing

The **Date** column was converted to a datetime format using: **pd.to_datetime(df["Date"], errors="coerce")**. Any rows with invalid or missing dates were excluded

2.2 Creation of the Year Variable
Although the dataset includes a "Year" column, a new **Year** variable was created from the parsed Date field to ensure accuracy: **Year = Date_parsed.dt.year**

This variable was used for:

- yearly aggregation
- temporal modeling with scikit-learn

2.3 Filtering and Type Correction

- Only rows with valid dates and states were retained.
- The Year column was cast to an integer type.

2.4 No Merging or External Sources
Although no new categorical variables or external datasets were introduced, one derived feature was created for analysis: the **Year** variable extracted from the parsed Date column. This represents a basic form of feature engineering and was necessary for computing yearly incident totals and fitting the temporal Linear Regression model.

3 Methodology
The dataset used in this analysis comes from the K–12 School Shooting Database (SSDB), which documents incidents in which a gun was fired, brandished, or a bullet struck school property. The dataset includes detailed information, including the incident date, state, and unique incident identifiers. The analysis drew exclusively from the 3,163 documented gunfire incidents contained in the Incident worksheet of the K–12 School Shooting Database. These incidents represent the "study population" used for all temporal modeling and state-level spatial analysis.

3.1 Data Preprocessing and Preparation
Before applying analytical methods, the following preprocessing steps were performed:

- **Date parsing:** The Date column—containing text-based dates—was converted into Python datetime objects using **pandas.to_datetime()**. Any invalid or missing dates were removed.
- **Year extraction:** A Year variable was created from the parsed dates to enable temporal aggregation.
- **Column validation:** The script automatically verifies the presence of required fields (Incident_ID, Date, State). If the incident identifier column appears under a different name, the program automatically detects and renames it.

These preprocessing operations ensure the dataset is correctly formatted for modeling and aggregation.

3.2 Yearly Temporal Analysis and Linear Regression Model
To answer Research Question 1—How has the frequency of gunfire incidents changed over time?—The dataset was aggregated by year, with the total number of incidents per calendar year computed.

3.2.1 Model Selection
A Linear Regression model was selected to capture long-term temporal trends and quantify whether incidents are increasing, decreasing, or stable over time. This model estimates a straight-line relationship between the predictor variable (Year) and the response variable (Incident Count).

The model takes the form: $\text{Incidents}=\beta_0+\beta_1(\text{Year})$
where:
β_0: Intercept
β_1: Slope (rate of change per year)

3.2.2 Assumptions of Linear Regression
Linear Regression relies on several assumptions:

- Linearity: There is a linear relationship between year and incident count.
- Independence of errors: Error terms are independent across observations.
- Homoscedasticity: The variance of residuals is constant across years.
- Normality of residuals: Residuals are normally distributed.

Given that this analysis seeks a *long-term directional trend* rather than highly precise predictive modeling, these assumptions are acceptable and provide meaningful high-level insight into temporal dynamics.

3.2.3 Advantages of Linear Regression

- Simple and interpretable
- Provides a quantitative rate of change
- Robust for long-term directional trends
- Suitable for decades-long historical data

3.2.4 Disadvantages of Linear Regression

- Does not capture nonlinear cycles or structural breaks
- Moderately low R^2 values are expected for social-behavioral data
- Sensitive to outliers in extreme years

3.2.5 Python Implementation
The model was implemented using:
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

The notebook and script perform the following steps:

1. Extract Year and Incident Count.
2. Train LinearRegression on:
    - X = Year (reshaped to n × 1)

    - y = Incidents per year
3. Generate predictions.
4. Compute R^2.
5. Add predictions to the yearly DataFrame.
6. Save the results to CSV.
7. Plot actual vs. fitted trendline.

This provides both numerical and visual confirmation of long-term temporal trends.

3.3 State-Level Spatial Analysis

To answer Research Question 2—Which states have experienced the highest number of incidents?—the dataset was grouped by **State**, and counts were computed for each state across all years.

This aggregation reveals:

- Geographic concentration
- Regional disparities
- Outliers with unusually high incident counts

3.3.1 Why State-Level Aggregation?

Unlike the temporal analysis, geographic analysis does not require predictive modeling. Instead, **descriptive aggregation** is the most appropriate method because:

- The goal is to examine **distribution**, not prediction.
- State-level counts are direct and interpretable.

A model might obscure rather than clarify geographic variation.

3.3.2 Python Implementation

State-level grouping was conducted using:

state_counts = df.groupby("State")["Incident_ID"].count()

3.4 Visualization Methods

Two primary visualizations were generated using **matplotlib**:

1. **Yearly trend plot**
    - Displays actual incident counts per year
    - Overlaid with Linear Regression predictions
    - Provides a clear view of increasing trends
2. **State-level bar chart**
    - Shows top 15 states by incident count
    - Allows easy comparison of geographic patterns

Visualization reinforces the numerical findings and communicates results clearly.

3.5 Why This Methodological Approach Was Chosen

This methodology was selected because:

- Linear Regression is ideal for quantifying long-term directional change.
- Temporal aggregation provides clarity on trends over 60 years.
- State-level aggregation directly answers the regional comparison question.
- Visualizations are intuitive and accessible to broad audiences.
- These methods align perfectly with the project's learning objectives (Python-based analytics, scikit-learn modeling, and reproducible workflow).

The approach balances interpretability with analytical rigor, producing results that are both meaningful and easy to communicate.

REFERENCES

[1] A. Cox, S. Rich, and R. Chason. 2023. *The Washington Post school shootings database.* The Washington Post. Retrieved from https://www.washingtonpost.com/graphics/2018/local/us-school-shootings-database/.

[2] National Center for Education Statistics. 2023. *Indicators of School Crime and Safety: 2023.* U.S. Department of Education. Retrieved from https://nces.ed.gov/.

[3] Robert F. Kennedy School of Government. 2024. *K–12 School Shooting Database (SSDB), Public v4.1.* Retrieved from https://www.chds.us/ssdb/.

[4] U.S. Department of Justice, Bureau of Justice Statistics. 2023. *Indicators of School Crime and Safety.* Retrieved from https://bjs.ojp.gov/.