Visual Analysis on School Shootings in the United States

Samantha Georgina Isaac Munoz

Abstract—This study provides a comprehensive visual analysis of school shootings in the United States, utilising a dataset, compiled by The Washington Post, that contains a total of 389 incidents from the years 1999 to 2023. The research tackled the complexity of this sensitive issue with a dual approach: advanced computational methods to handle data intricacies, and visual analytics to elucidate trends and insights. The main aim is to find patterns and trends in school shootings, with specific focus on the timing, geographical distribution, and perpetrator profile. The key findings revealed an increase in school shootings since 2012, with significant spikes in 2018 and 2022. Incidents were found to be more frequent during the spring time, and notably in urban settings like Los Angeles and Chicago. Demographic analysis indicated that the typical perpetrator profile is a 19-year-old male, predominatly white. And the most common type of shooting identifies was targeted attacks.

1 Problem Statement

School shootings have increased in the United States over the past few years. This trend shows how much this issue has become a concern since the Columbine High School incident in 1999. These incidents not on only endanger the lives of students and staff at these schools, but also have long-term effects on both victims and perpetrators families. The aim of this study is to analyse this phenomena to find any possible relevant findings or patterns. In researching this topic, this study will be guided by an ethical and respectful approach that fully recognises the seriousness and sensitivity of the matter.

The questions this research aim to answer are:

1. Has school shootings become more frequent in recent years?
2. Are there specific times during the day or week when these incidents tend to occur?
3. Which state has the highest number of incidents in the country?
4. What general characteristics (like age, race, and gender) are the most common among the perpetrators?

The dataset used for this analysis, which was compiled by The Washington Post [1], is highly suitable for addressing the research questions. It contains a total of 389 entries with 50 features each, including temporal, spatial, quantitative, and qualitative data from school shootings since 1999, all the way to 2023. This information can be used to provide significant and relevant insights.

2 State of the Art

The exploration of school shootings through visual analytics is a field marked by diverse and evolving methodologies. The four research papers obtained and reviewed demonstrate a range of approaches to analysing the issue of shootings, whether in school environments or in a more general setting. Each study offers a unique perspective, contributing to a more comprehensive understanding of this critical issue.

Shultz, et al. [1] employ an epidemiological approach, analysing national databases to discern patterns in firearm mortality, with a specific focus on school shootings. The research uses a range of statistical graphs and trend analyses like line graphs and histograms to track trends over time, offering valuable insights into the temporal patterns of these incidents. Maps are also used as a key analytical tool to visualize how these incidents are spread across the United States, highlighting areas with higher concentrations. Implementing a similar approach in this analysis will aid in understanding the prevalence and the spatial aspect of these incidents, this is mainly since the dataset that is going to be used has a considerable amount of temporal and geographical data.

Another study that presents an intriguing approach to shooting in the United States is that of Reeping and Hemenway [2], which adopts a temporal focus, specifically utilising data pertaining to the weather in Chicago. Their objective is to analyse how changes in temperature, humidity, and precipitation influence the frequency of shootings. Employing bat graphs and scatter plots, they strive to understand how shifts in weather patterns correlate with variations in daily shootings over a five-year period. This study's unique angle offers a nuanced view, exploring environmental factors that might intersect with human behaviour in the context of gun violence, providing a potentially ground breaking perspective on the triggers of such incidents.

This research also seeks to utilize statistical analysis to identify the demographics of the perpetrators, including age, gender, and race. In addition, it aims to examine the frequency and types of weapons used in shootings. These approaches are implemented through a quantitative analysis in Schildkraut's research [3], which focuses on analysing the trends and characteristics of mass shootings in the United States from 1966 to 2020. Pie charts and bar graphs are used to show the findings for demographic data and weapon use. Furthermore, the Schildkraut's research [3] uses temporal data, which is visualised using scatter plots and line plots, as well as bar charts.

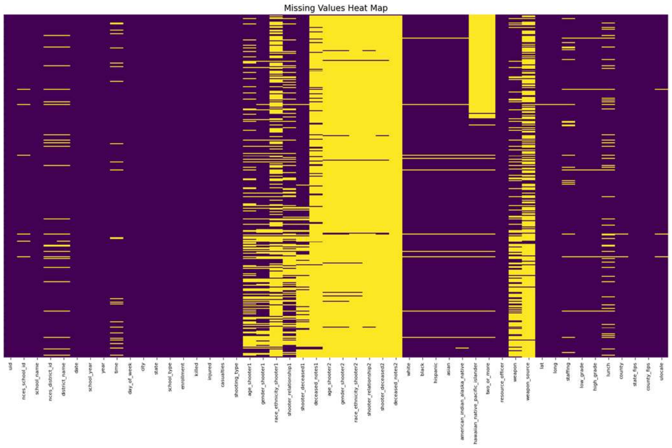
Finally, the article by The Washington Post [4], which is the source of the dataset for this research paper, was analysed.

The Washington Post primarily used bar graphs to represent both temporal and demographics data. Although the article does not explicitly detail the methods of analysis employed, it serves as a solid foundation for this study.

The reviewed papers demonstrate diverse methods in studying shootings in different settings, offering insights that are valuable for this research.

3 Properties of the Data

This dataset was compiled by journalists at The Washington Post, as their aim was to determine how many children have been impacted by school shootings and to produce a related article. To achieve this, they gathered information on every instance of gunfire in any primary or secondary school in the United States during school hours, starting from the Columbine High School massacre on April 20, 1999, and the last record is from 2023. As described in the dataset information [5], a variety of sources were used, including newspaper articles, open-source databases, law enforcement reports, information from school websites, and calls to schools and police departments. This dataset is available on GitHub [5] as an open-source, and is continually updated as new cases emerge.



The dataset contains a total of 387 entries and 49 features. The data types include numerical (integers and floats) and qualitative (strings). The features encompass a range of information, including identification fields, school-related information which details aspects such as the school's name, its district, longitude and latitude, and the type of school. There's also data on the incidents, covering details like the date, time, number of fatalities, and number of injured, type of shooting, among others. Additionally, the dataset contains information about the perpetrators, including gender, race, and age.

Fig. 1. Heat map that shows the quantity of missing values per feature (yellow represents the quantity of missing values).

Feature	Missing Values	Percentage of Missing Data
Deceased_notes2	386	99.7%
Shooter_Deceased2	381	98.4%
Shooter_Relationship2	381	98.4%
Race_ethnicity_shooter2	380	98.1%
Age_shooter2	375	96.9%
Gender_shooter2	375	96.9%
Deceased_notes1	349	90.1%
Weapon_source	282	72.8%
Race_ethnicity_shooter1	239	61.7%
weapon	138	35.6%
Shooter_deceased1	125	32.3%
Two_or_more	125	32.3%
Shooter_relationship1	121	31.2%
Age_shooter1	113	29.2%
Gender_shooter1	78	20.16
lunch	50	12.9%
Distric_name	25	6.46%
staffing	24	6.20%
High_grade	5	1.2%
Low_grade	5	1.2%
ulocale	3	0.7%

Fig. 2. Table showing the most relevant features with missing values and their respective percentage.

Upon analysing the dataset with Python, it was observed that a total of 33 features contained missing data. As shown in the Heat map from Figure 1, the features with over 90% of null values were related to demographic descriptions of a second shooter. Since many incidents did not involve a second shooter, the journalists compiling this dataset left these fields blank. To handle these features, a new column was created to indicate the presence of a second shooter in a binary format, and the original columns were deleted. Another feature, which was left blank on purpose, was the manner in which the perpetrator died. This column was removed as there is already another column that indicates whether the perpetrator died during the incident in a binary format. Similarly, the decision was made to eliminate several other features that were not relevant to the research questions, this was done also to avoid having unnecessary columns or those with a high number of null values.

For the remaining categorical features in the dataset, which contain a significant number of null values, a new value was used called 'Unknown'. This approach allows for the retention of these features in the analysis while acknowledging the gaps in the data. In the case of numerical features with missing values, the calculation of the average of the existing values for the given feature was employed to maintain the integrity of the dataset while minimizing the distortion that missing values might cause in the overall analysis.

In the end after completing these data cleaning processes, the dataset used for this research analysis was reduced to 31 features with 383 entries.

4 Analysis

4.1 Approach

To effectively address the research questions and to thoroughly understand the patterns, trends, and factors contributing to the school shooting incidents, the analysis that is going to be use adopts a comprehensive approach that merges visual analytics with computational methods, complemented by human analysis and judgement. This strategy is designed to obtain important conclusions and provide relevant information. This approach confronts the complexities inherent in school shootings by implementing a modified visual analytics framework. Based on the ideas presented by Keim et al. [6] this framework puts an emphasis on combining data transformation, computational processing, and iterative visualisation. These critical components are key in the pursuit of actionable insights and a deeper comprehension of the data at hand.

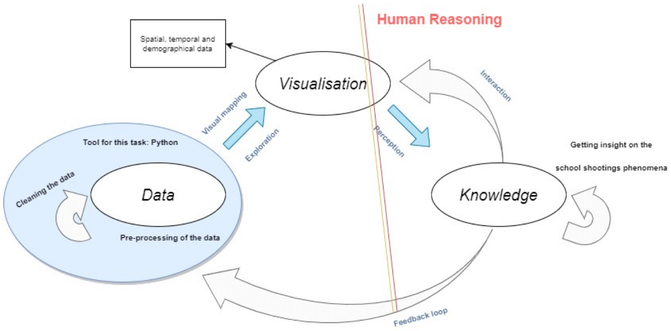

Human Reasoning
Spatial, temporal and demographical data
Visualisation
Interaction
Visual mapping
Exploration
Perception
Tool for this task: Python
Getting insight on the school shootings phenomena
Cleaning the data
Data
Knowledge
Pre-processing of the data
Feedback loop

Fig. 3. Diagram of the analysis work flow

The first step in the analysis will be data pre-processing, this is to prepare the dataset for the process of creating visualizations. Through human reasoning and the use of Python, a rigorous analysis will be conducted to determine if the data is sufficient and appropriate for answering the research questions, as well as for uncovering new information. Also, a decision was made regarding the most relevant features to retain, which will aid in achieving the research objective. Following this, a thorough data cleaning process was done. Human reasoning plays a crucial role in deciding the best approach to address issues related to missing values and to convert the temporal data into the correct format. The chosen method involves either imputing or addressing these concerns in a way that the data stays accurate. This approach ensures that the data not only maintains its integrity, but also becomes more conducive to generating insightful and actionable findings.

Following the pre-processing, the next step is a detailed examination of the temporal and spatial data, where human reasoning is integrated with visualisations in Tableau to interpret and learn from the data. Patterns of incidence related to specific days of the week and times of the day will be identified, providing a chronological view and enabling the identification of changes in the frequency and characteristics of school shootings. Heat maps, and line charts will facilitate this phase, with the selection of visualizations guided by human reasoning to ensure that the representations accurately reflect the data.

The spatial analysis will seek geographical patterns, using heat maps in Tableau to highlight areas of high incidence and allowing human to analyse and identify correlations. The decision to use a heat map is based by their ability to visually represent density and allow an intuitive interpretation of problem areas.

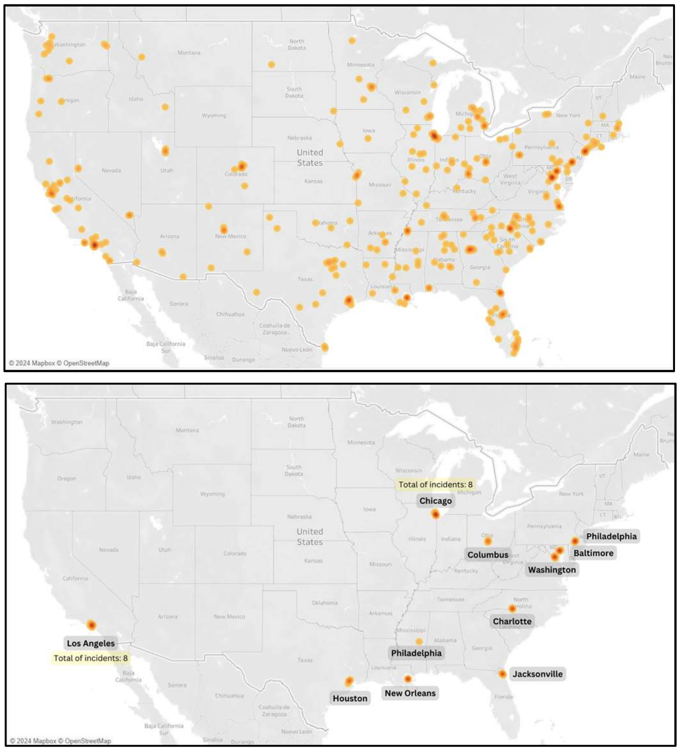
In the demographic analysis stage, the focus will be on uncovering trends related to the characteristics of the perpetrators. Bar charts will be used to show relevant insight regarding the age, gender, and race of the perpetuators. These types of charts are effective in representing distributions of categorical data, which will allow a quick and clear comparisons between different groups.

4.2 Process

Spatial Analysis

The analytical process began by applying spatial mapping of the latitude, longitude, and incident count per state, aiming to find geographical patterns and hotspots of incidents across the United States. This initial step in this part of the process was crucial for setting the context and guiding further in-depth analysis. The heat map showing these incidents (Figure 4, top map), served as a broad overview, highlighting regions with a higher frequency of shootings. Then, a more detailed visualisation (Figure 4, bottom map), focused on the cities was developed, offering a closer look at the distribution of incidents.

Fig. 4. Distribution of shootings across the United States and cities with the



highest number of incidents

The top map on Figure 4 displays the location of all recorded school shootings across the United States. This graph shows that there has been at least one incident in almost every state, with only a few exceptions. This map provides a clear depiction of how widespread these incidents are cross the country. The bottom map on Figure 4 focus only on the cities with the highest number of incidents across the country, with Los Angeles and Chicago being the most affected, each having a total of eight incidents. This visualisation shows that urban areas are major locations for these school shootings, suggesting a possible link between city settings and how often school shootings happen.

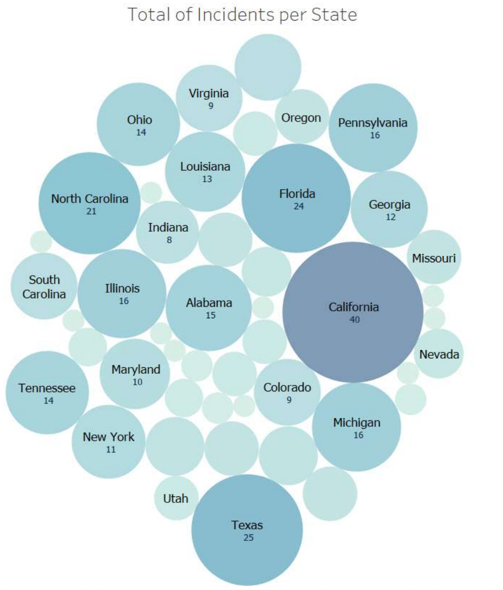


Fig. 5. Distribution of shootings across the United States and cities with the highest number of incidents

An additional chart was made to display each state with its respective total number of school shootings incidents. A bubble chart was chosen to provide a more complete presentation of the data. As shown in the chart (Figure 5), California leads with the highest number of incidents, totalling 40, followed closely by Texas, Florida, and then North Carolina.

Temporal Analysis

For the temporal analysis, the approach changed to examining the changing patterns of school shootings over time. First, a line chart (Figure 6) was made to show the yearly trends of casualties during these events across the United States. This graph detailed the count of both fatalities and injuries, providing a comprehensive comparison from 1999, the year of the Columbine shootings, up to 2023.

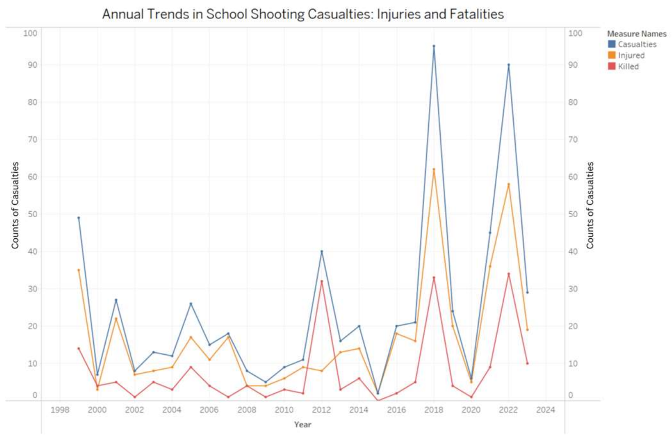


Fig. 6. Annual trends in shool shootings casualties with a breakdown between quantity of injured or killed

2018 stood out as a critical year in terms of the number of incidents recorded across the whole timeline, with a worrying total of 100 school shootings in the United States, indicating the highest point in the history of these incidents. 2022 recorded the second highest number of incidents, totalling 90, and notably, this year also had the highest number of fatalities. Additionally, there was a significant drop in school shootings in 2020, this could possibly be because of the COVID-19 pandemic.

It is also noteworthy to mention that the year with the lowest total number of incidents was 2015. This finding may require further analysis to better understand why that particular year saw a lower incidence of school shootings compared to other periods.

The analysis of yearly data revealed variations in the frequency of events which led to a more detailed analysis segmented by month and hour. A heat map was chosen for this (Figure 7) since it facilitates the identification of peak periods of incidents and allows for a quick, intuitive understanding of complex data.

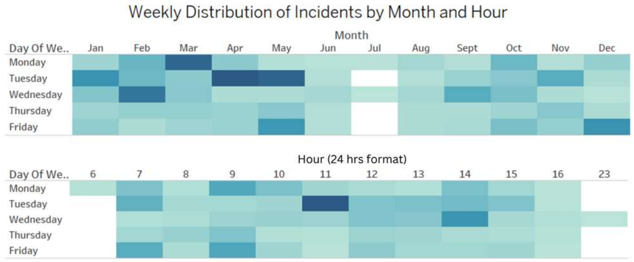


Fig. 7. Weekly distribution of incidents by month and hour (darker color indicates more incidents)

This heat map (Figure 7) shows that the months of March, April, and May have the highest number of incidents, with April being the month with the highest number of record. On the contrary, June, July and August have the fewest count of incidents, likely due to the summer holidays. Another interesting finding is a trend that shows that there is a higher frequency of incidents in the early months of the year, as well as for the beginning of the week. Regarding the time were these shootings were more prone to happen, is at 11 am, probably around lunchtime. Is worth mention that a higher concentration of these incidents happens in the morning, but 2 pm also has a significant number of incidents.

Demographic Analysis

For the last part of the analysis process, the focus was on the demographic characteristics get a more comprehensive understanding of the profile of those responsible for the school shootings to. The main data used for this sections was the age, gender, and race of the shooters, with the implementation of lethality of the incidents.

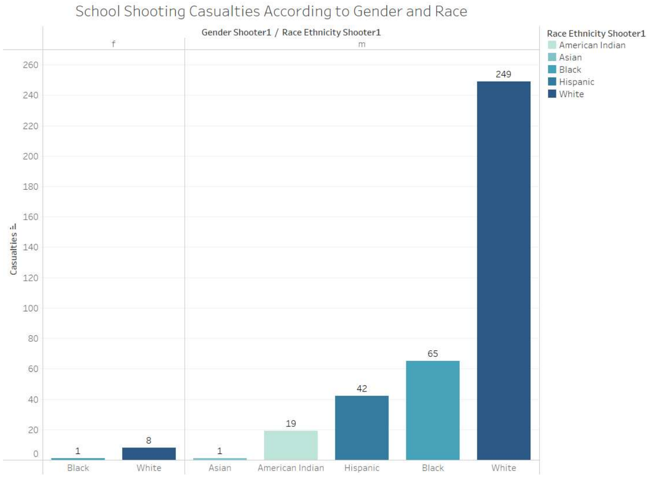


Fig. 8. School shooting casualties by the shooters gender and race

A bar chart (Figure 8) was made to compare the number of casualties associated with the gender, and the race of the shooters. During the analysis of the graph, it was observed that school shootings are predominantly carried out by males, with only a few cases involving females. Additionally, both genders showed a tendency towards one specific racial group being more prevalent in these incidents.

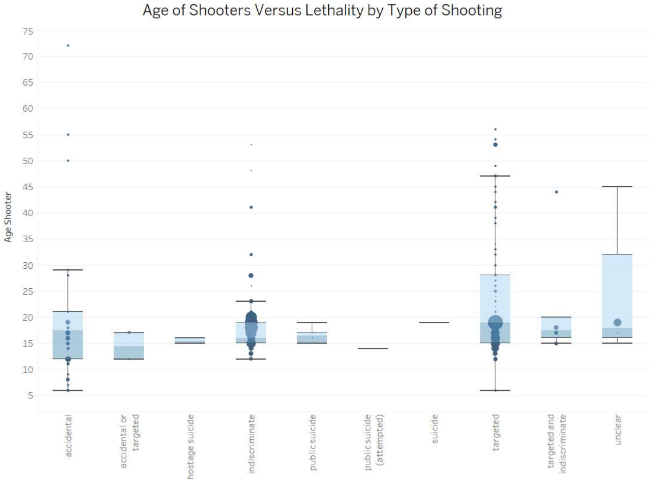


Fig. 9. Age of shooters versus the lethality according to each type of shooting (the size of the circles represent the amount of casualties)

It was decided to create a boxplot chart to gain a more in-deep understanding of the profile of the shooters. This approach is aimed at providing a clearer view of the age ranges and motivations behind these incidents. It was concluded that the average age in the shooter profile is 19 year, and the majority of incidents were targeted. The age range for this type of shooting varies greatly, with cases ranging from 6 to 56 years old. The second most common type was indiscriminate shooting, where the age range is mainly concentrated between 12 and 23 years. The third common type is accidental shootings, which show an age range from 6 to 76 years, and these incidents had a relatively high number of casualties.

An important point to highlight is the discovery of two recorded instances where a 6 year-old child was the perpetrator of the shootings. This alarming detail raises critical questions about accessibility to firearms by very young individuals and the circumstances that lead to their involvement in such serious incidents at such a young age.

4.3 Results

In the initial phase of the analysis, which was focused on the temporal data, it was found that California leads in the number of incidents, with a total of 40. Most of these were concentrated in the city of Los Angeles.

For the temporal analysis, it was concluded that there was an increase in incidents since 2012, with significant spikes in 2018 and 2022. This indicates that there has been an increase in school shootings in recent years, but this is not gradual; it involves significant spikes and falls. There is a trend in the days of the week and the month in which this incidents tend to occur, particularly during the spring season on the months of March, April, and May, mostly on Tuesdays before noon.

Finally, the demographic analysis provided a profile of the most common characteristics of the perpetrators. Finally, the demographic phase of the analysis revealed that the typical age of the shooters is 19 year, with the majority being male. Furthermore, the most prevalent race or ethnicity, according to the records, is white. Additionally, it was found that the most common type of shootings is targeted.

5 Critical reflection

With the large number of features in The Washington Post's dataset, the first challenge of this analysis was presented. This quantity of data initially made it difficult to narrow down the focus of the analysis to a few key areas and discoveries, requiring a careful and deliberate process to formulate relevant and unique research questions that were neither too broad nor overly simplistic.

The study's approach integrated visual analytics with computational methods. This helped to understand the importance of meticulous data handling and the interpretation during the analysis. The high quantity of missing data, particularly in the demographics features, was a substantial obstacle. However, the strategy to address these issue through careful data cleaning and with the introduction of the "Unknown" categories for missing values was effective. This approach may not have been ideal, as it could have led to a potential skewing of the results. However, given the extensive number of records, reaching this conclusion was challenging. Ultimately, the results appeared to be accurate, which not only preserved the integrity of the dataset but also enabled meaningful analysis despite the inherent limitations.

Reflecting on potential improvements, the incorporation of additional data source or more advanced statistical methods could provide a deeper insight. For an instance, exploring the psychological and sociological factors potentially correlated with these incidents could provide a more insightful understanding. This would involve an integration of the findings from this study's dataset with the insights from these disciplines, thereby enriching the analysis. However, such approach would require a multidisciplinary methodology.

In summary, while this study addressed the initial research questions and offered significant insights, it also brought to light the challenges and limitations inherent in data-driven research, particularly when dealing with large, and complex datasets. The lessons learned here emphasize on the importance of rigorous data cleaning, thoughtful analysis, and the recognition of data constraints. This information is crucial for authorities for them to be able to detect and prevent the continuous growing of such incidents. Furthermore, this study serves as a valuable reference for researchers engaged in similar data-driven inquiries, particularly in complex and sensitive domains like school shootings, laying a baseline for future research.

Table of word counts

Problem statement	224
State of the art	474
Properties of the data	486
Analysis: Approach	477
Analysis: Process	940
Analysis: Results	191
Critical reflection	362

References

[1] J. M. Shultz, A. Cohen, G. W. Muschert, and R. F. De Apodaca, "Fatal school shootings and the epidemiological context of firearm mortality in the United States," Disaster Health, vol. 1, no. 2, pp. 84–101, Apr. 2013, doi: 10.4161/dish.26897.

[2] P. M. Reeping and D. Hemenway, "The association between weather and the number of daily shootings in Chicago (2012–2016)," *Injury Epidemiology*, vol. 7, no. 1, Jun. 2020, doi: 10.1186/s40621-020-00260-3.

[3] J. Schildkraut, "Can Mass Shootings be Stopped?: To Address the Problem, We Must Better Understand the Phenomenon," Rockefeller Institute of Government, Jul. 2021. https://rockinst.org/wp-content/uploads/2021/07/Public-Mass-Shootings-Brief.pdf

[4] J. W. C. S. R. Ulmanu Linda Chong, Lucas Trevor, John Muyskens, Monica, "There have been 392 school shootings since Columbine," Washington Post, Dec. 06, 2023. [Online]. Available: https://www.washingtonpost.com/education/interactive/school-shootings-database/

[5] J. W. Cox, S. Rich, A. Chiu, H. Thacker, and L. Chong, "GitHub - washingtonpost/data-school-shootings: The Washington Post is compiling a database of school shootings in the United States since Columbine.," *GitHub*. https://github.com/washingtonpost/data-school-shootings

[6] Keim, D., Kohlhammer, J., Ellis, G. & Mansmann, F., 2010. Mastering the Information age: solving problems with visual analytics. Germany, Druckhaus: Eurographics Association.

[7] (Software used) "Tableau: Business intelligence and analytics software," Tableau. https://www.tableau.com/en-gb

[8] (Software used) Google Colab, "Colab.google," colab.google. https://colab.google/

[9] Samantha-Isaac, "GitHub - samantha-isaac/School-Shootings-Visual-Analysis: This study embarked on a visual analysis of school shootings in the United States, leveraging a comprehensive dataset from The Washington Post. Covering incidents from 1999 to 2023, the dataset's rich temporal, spatial, and demographic data facilitated a deep dive into patterns and characteristics of these tragic events.," *GitHub*. https://github.com/samantha-isaac/School-Shootings-Visual-Analysis