Statistical Analysis of Mass Shooting Incidents in the United States (2014 - 2024)

Kaylee Dehncke[a]

[a] *Computer Science / Data Science Student at CU Boulder*

Rachel Cox

1. Introduction & Background

Mass shootings have become a pressing public safety and social issue in the United States, drawing attention from policymakers, researchers, and the public. The central research question of this project is: **How do mass shooting incidents in the United States vary across states, seasons, and years, and what statistical evidence supports these differences?** This question is worth answering because understanding the temporal and geographic patterns of mass shootings can provide insight into when and where these incidents are most likely to occur, which in turn may inform prevention strategies, resource allocation, and broader discussions about gun violence.

For readers unfamiliar with the topic, a mass shooting is generally defined as an incident in which four or more people are shot or killed, not including the shooter. These events are not evenly distributed across the country or throughout the year. Prior descriptive research has suggested that certain states experience disproportionately high numbers of incidents, and that shootings tend to peak during warmer months. Studies in criminology and public health have also highlighted the importance of analyzing long-term trends, as the frequency of mass shootings has increased in recent years. However, much of the existing literature focuses on case studies or policy debates rather than statistical exploration of large-scale datasets. This project contributes by applying exploratory data analysis and statistical methods to examine variation across time and place systematically.

The dataset used in this project comes from the Gun Violence Archive (GVA), a nonprofit organization that collects and verifies information on gun-related incidents in the United States. The GVA compiles data from law enforcement reports, media coverage, and other public sources, making it an observational dataset rather than the result of a controlled experiment. The data was collected to provide a comprehensive, publicly accessible record of gun violence incidents, with the goal of supporting research, journalism, and policy discussions. For this project, the dataset includes mass shooting incidents from 2014 through 2024, with variables such as incident date, location, number of victims injured or killed, and suspect outcomes. Because the data are observational, they reflect real-world reporting practices and may be subject to limitations such as under-reporting or inconsistencies across jurisdictions, but they remain one of the most complete sources available for studying gun violence in the U.S.

One limitation of the dataset is that reporting practices vary across states and localities. Some jurisdictions may have more robust systems for documenting incidents, while others may rely more heavily on media coverage. This introduces potential bias into the dataset, as states with stronger reporting infrastructures may appear to have more incidents simply because they are better documented. Another limitation is that the definition of a mass shooting, while standardized by the GVA, may differ from definitions used in other contexts, such as law enforcement or academic studies. Despite these challenges, the dataset provides a valuable foundation for statistical analysis, particularly because it spans a full decade and includes thousands of incidents.

2. Methods

To address the research question, exploratory data analysis (EDA) was conducted on the Gun Violence Archive dataset spanning 2014–2024. Pre-processing steps included converting incident dates into year, month, and season variables and creating derived totals for victims and suspects. These transformations allowed for comparisons across time and geography. In addition, summary statistics were calculated to provide an overview of the distribution of incidents, including measures of central tendency and variability. This step was crucial for identifying skewness in the data, as mass shootings often involve a small number of victims but occasionally include extreme outliers that can distort averages.

Figure 1 shows the number of mass shootings per year from 2014 to 2024. The chart reveals a clear upward trend, peaking in 2021, followed by a slight decline. The dashed blue line marks the mean number of incidents per year (471.2), while the dotted green line marks the median (413), highlighting the overall elevation in frequency. The difference between mean and median suggests that certain years with exceptionally high counts, such as 2021, pull the average upward. This visualization provides context for the bootstrap confidence interval analysis, which formally quantifies uncertainty around the mean.

Figure 2 displays the number of mass shootings by season. With January–March months aggregated in Winter, April–June in Spring, July–September in Summer, and October–December in Fall, the seasonal imbalance becomes clear. Summer has the highest number of incidents, followed by Spring, Fall, and Winter. The reference lines showing the mean (1295.8) and median (1315.5) incident totals across seasons emphasize that the distribution is not uniform. This plot supports the hypothesis that warmer months may be associated with increased social activity, larger gatherings, or other contextual factors that elevate risk.

Figure 3 presents the number of mass shootings by state. Illinois, California, and Texas report the highest counts, with a long tail of states reporting fewer than 50 incidents. The dashed blue line (mean = 101.6) and dotted green line (median = 67) emphasize the skewed distribution. This visualization highlights geographic disparities and raises questions about state-level factors such as population size, urban density, gun laws, and socioeconomic conditions. While the plot does not directly address causality, it provides a foundation for future regression analyses that could explore predictors of state-level variation.

Figure 4 provides a heatmap of incidents by month and year. The darkest cells appear in June, July, and August of 2020 through 2023, indicating seasonal clustering during summer months and reinforcing the patterns seen in the seasonal bar chart. The heatmap adds granularity by showing not just seasonal totals but specific months where incidents spike. This visualization is particularly useful for identifying short-term surges that may be linked to social or political events, holidays, or other contextual factors.

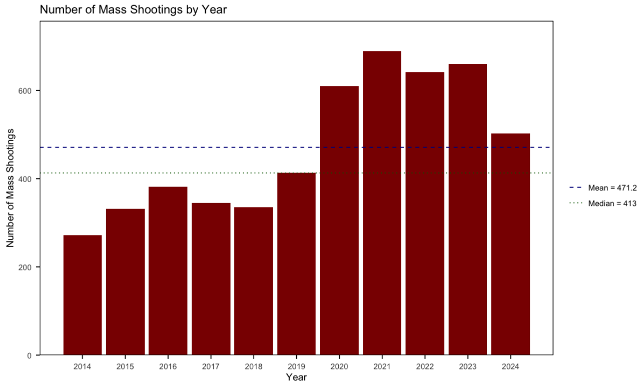

Figure 1. Number of Mass Shooting Incidents per Year

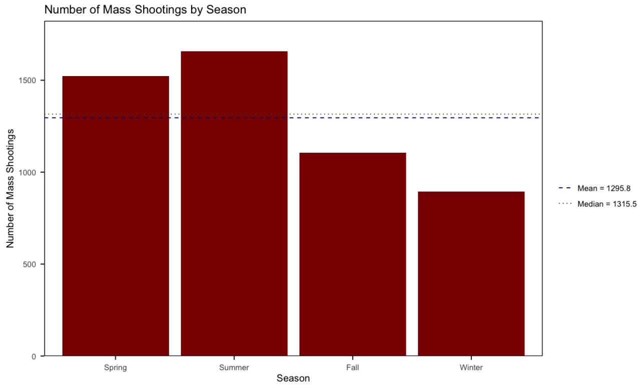

Figure 2. Number of Mass Shooting Incidents per Season

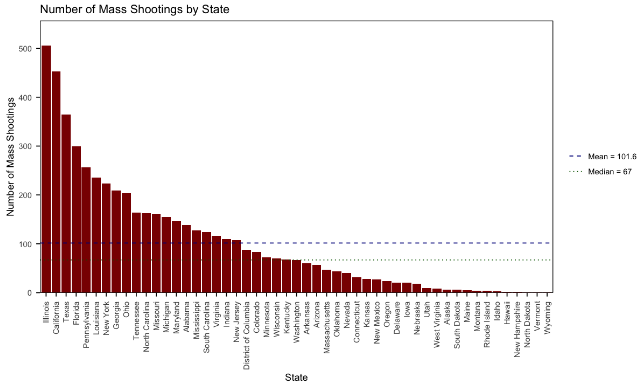

Figure 3. Number of Mass Shooting Incidents per State

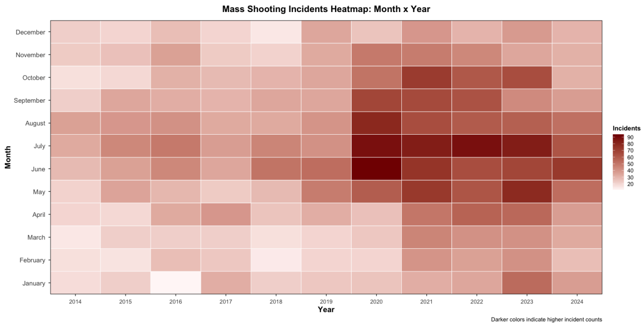

Figure 4. Number of Mass Shooting Incidents per Month and Year

3. Results

To formally test the patterns observed in the EDA, two statistical analyses were conducted: a bootstrap confidence interval for yearly incident counts and a chi-squared test for seasonal variation. These methods were chosen because they provide complementary insights: the bootstrap quantifies uncertainty around the mean, while the chi-squared test assesses whether observed seasonal counts deviate significantly from a uniform distribution.

Figure 5 shows the bootstrap distribution of mean yearly incidents, based on 10,000 resamples. The histogram is centered around a mean of 471.2, with a 95% confidence interval of (387.4, 558.3). This confirms that the average yearly frequency of mass shootings is statistically elevated and not due to random fluctuation. The relatively narrow confidence interval suggests that the mean is stable across resamples, reinforcing the conclusion that mass shootings occur at a consistently high rate.

A chi-squared goodness-of-fit test was conducted to assess whether incidents are evenly distributed across seasons. The observed counts were: Spring = 1524, Summer = 1657, Fall = 1107, Winter = 895. These differ significantly from the expected uniform distribution (1296 per season), with $\chi^2(3) = 292.36$ and $p < 0.001$. Standardized residuals revealed that Summer (+11.6) and Winter (-12.9) contributed most strongly to the test statistic, indicating that summer incidents are substantially more frequent than expected, while winter incidents are substantially fewer. Spring (+7.3) also showed an excess, and Fall (-6.1) a deficit, though their contributions were smaller. This analysis provides strong statistical evidence that mass shootings are not evenly distributed across seasons.

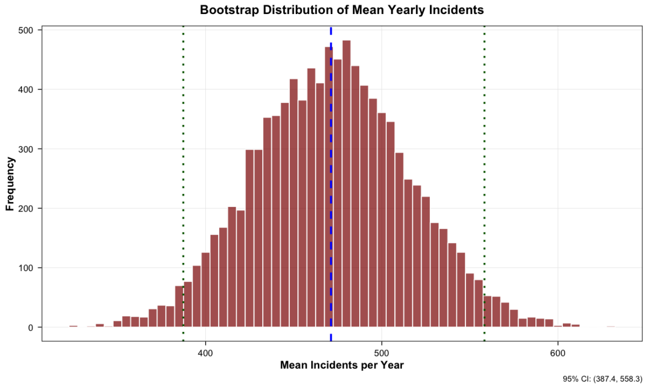

Figure 5. Bootstrap Confidence Interval of Mean Yearly Incidents

Table 1. Season Result Data

Season	Frequency	Standardized Residuals
Spring	1524	7.321828
Summer	1657	11.588217
Fall	1107	-6.054743
Winter	895	-12.855302

Note: χ-squared = 292.36, df = 3, p-value < 2.2e-16

4. Conclusion

This project investigated how mass shooting incidents in the United States vary across states, seasons, and years. The results provide clear evidence of non-uniformity: incidents are disproportionately concentrated in certain states, peak during summer months, and have remained consistently elevated in frequency over the past decade. The bootstrap confidence interval confirmed that the average number of mass shootings per year is well above 400, while the chi-squared test demonstrated strong seasonal variation driven by excess summer incidents and deficits in winter.

These findings highlight the importance of considering both temporal and geographic variation when analyzing gun violence. While the dataset is observational and subject to reporting limitations, it remains one of the most comprehensive sources available. Future work could extend this analysis by incorporating per-capita rates, regression modeling, or spatial clustering to further understand the drivers of variation. For example, adjusting state-level counts by population size could reveal whether high totals in states like California and Texas reflect sheer population or disproportionate risk. Similarly, regression models could test whether socioeconomic indicators, gun ownership rates, or policy differences explain variation across states. Spatial clustering methods could identify regional hotspots that are not apparent in state-level aggregates.

Another important direction for future research is the integration of qualitative data with quantitative analysis. While statistical methods can reveal patterns of frequency and distribution, qualitative approaches—such as case studies of specific incidents or interviews with affected communities—can provide insight into the social dynamics that numbers alone cannot capture. Combining these perspectives would allow for a more holistic understanding of mass shootings as both statistical phenomena and lived experiences. Ethical considerations also play a role: researchers must remain sensitive to the fact that each data point represents real human suffering, and analyses should be framed in ways that respect victims and communities rather than reducing them to abstract figures.

Policy implications are equally significant. Evidence of seasonal clustering, for instance, suggests that prevention efforts could be strategically timed to coincide with periods of heightened risk, such as summer months. Geographic disparities highlight the need for localized interventions, as states with consistently high incident counts may require targeted resources or policy reforms. Moreover, the consistent elevation in yearly averages underscores that mass shootings are not isolated anomalies but a persistent public safety challenge. Recognizing this persistence is crucial for policymakers, who must move beyond reactive responses to individual tragedies and instead develop sustained strategies informed by long-term statistical evidence.

Overall, the statistical evidence underscores that mass shootings are not evenly distributed across time and place, and that these differences are both practically and statistically significant. By combining exploratory visualization with formal inference, this project demonstrates the value of statistical analysis in understanding complex social phenomena. The findings contribute to ongoing discussions about gun violence and provide a foundation for future research that can inform policy and prevention strategies. In doing so, this work emphasizes the role of data-driven approaches in addressing one of the most urgent public safety issues facing the United States today.

Note

The full source code, datasets, and visualizations used in this analysis are available on GitHub:
https://github.com/kayleedehncke/stat5000_final_project

References

[1] G. V. Archive, *Gun violence archive dataset*, 2024. [Online]. Available: https://www.gunviolencearchive.org/.