THE CURIOSITY CUP 2025
A Global SAS® Student Competition

Unraveling Mass Killings: A Data-Driven Analysis of an American Epidemic of Death

Vedika Dahad, Swapna Ashish Patel, Northeastern University

Team Byte by Byte

ABSTRACT

This report presents an analysis of mass killings. Mass killing are incidents where multiple victims are either killed or injured in a single event. Using statistical methods such as the Kruskal-Wallis test, and Chi-Square test we examined correlations and trends in weapon types and fatalities. Additionally, we have assessed the statistical significance of fatality counts across different locations and weapon types.

INTRODUCTION

Mass shootings have increased and have become more frequent in the United States with deadly occurrences, which generates uncertainty, fear and concerns of public safety (Peterson et al., 2024). These events have led to high mortality rates and have a profound impact on the families of the victims.

We analyzed mass killings to identify patterns in incident locations, types, and weapon usage. While we anticipated certain trends in the data, our findings revealed unexpected insights, particularly in the relationships between victims and offenders. These findings aim to raise awareness and drive policy reforms that enhance public safety and protect communities from mass killing incidents.

PROBLEM STATEMENT

While mass killings have increased in the United States, researchers have not fully understood the factors influencing or predicting these incidents. Our goal is to uncover hidden patterns and trends, offering insights and recommendations that can aid in understanding and preventing these tragic events more effectively. We aim to highlight critical patterns in offender demographics, incident types, weapon usage, and victim-offender relationships. This will ultimately help slow the rise of this growing epidemic of violence in the United States.

DATA PREPARATION

DATASETS

In this project, we are using 4 datasets sourced from Data World named “Mass Killings in America, 2006-present" (*Mass Killings in America, 2006 - Present - Dataset by Associatedpress*, n.d.). It consisted of the incidents, victims, offenders, and weapons table.

DATA PREPROCESSING

During the initial review of the dataset, it became clear that the data was heavily skewed. We applied log transformation methods to reduce skewness; however, we later realized that we did not want to alter the real-world data points and scenarios. As a result, we pivoted to non-parametric statistical tests, which do not assume a normal distribution.

As part of our preprocessing activities, we ensured that all variables were assigned the appropriate data type. For example, the "incident_id" variable in the weapons dataset was originally in character format, but we converted it to a more suitable data type.

We employed various ways to address missing values based on missingness of the data. We substituted missing items in category variables with the label "Unknown”. To reduce bias in numeric variables, we imputed with either mean, median or mode of the appropriate columns.

We standardized date variables to a common format (YYYY-MM-DD) to ensure consistency.

In addition, we developed many derived measures to improve our analysis. They included:

- Total_victims: The sum of both dead and injured victims (num_victims_killed + num_victims_injured).
- Incident severity: A classification based on the overall number of victims, which aids in determining if an incident is low, medium, or high severity.
- Weapon lethality: A measure that connects the number of victims killed (num_victims_killed) to the weapon type, providing information about the relative lethality of various weapon types.
- Multiple_weapons_flag: A measure that distinguishes between single-weapon and multi-weapon incidents.
- Weapon_category: This variable groups weapon types into broader categories, such as "firearm," "melee," and "explosive," allowing for a more structured analysis of weapon usage.

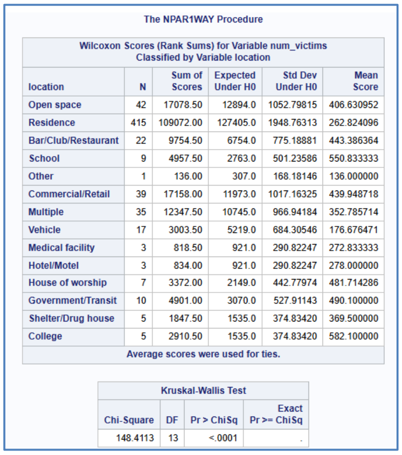
To streamline the analysis, we created subsets of the incidents table, including only the variables required for our specific analysis. This method reduced noise and enabled a more focused exploration into the variables causing mass killings in the United States.

ANALYSIS

We performed all our statistical analyses using SAS® OnDemand for Academics (SAS Institute Inc., 2024) to examine the correlations in mass killing incidents across different locations, weapon types, and victim-offender relationships. Our primary objective was to determine whether significant differences exist in the number of fatalities, injuries, and offender relationships based on these key factors.

INCIDENT LOCATION AND VICTIM COUNTS

We used the Kruskal-Wallis test (*Kruskal Wallis Test - an Overview | ScienceDirect Topics*, n.d.) (Figure 1.1) to evaluate the variation in total victims (total_victims) by location, which indicated a statistically significant difference in victim counts among incident locations (Chi-Square = 148.4113, DF = 13, p < 0.0001). Subsequent analysis using the Wilcoxon Rank-Sum test (Haynes, 2013) revealed that incidents occurring in colleges, schools, governmental/transit areas, houses of worship, and commercial establishments generally yield higher victim counts. In contrast, the number of victims was the lowest in "Other" and vehicle-related incidents.



The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable num_victims
Classified by Variable location

location	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Open space	42	17078.50	12894.0	1052.79815	406.630952
Residence	415	109072.00	127405.0	1948.76313	262.824096
Bar/Club/Restaurant	22	9754.50	6754.0	775.18881	443.386364
School	9	4957.50	2763.0	501.23586	550.833333
Other	1	136.00	307.0	168.18146	136.000000
Commercial/Retail	39	17158.00	11973.0	1017.16325	439.948718
Multiple	35	12347.50	10745.0	966.94184	352.785714
Vehicle	17	3003.50	5219.0	684.30546	176.676471
Medical facility	3	818.50	921.0	290.82247	272.833333
Hotel/Motel	3	834.00	921.0	290.82247	278.000000
House of worship	7	3372.00	2149.0	442.77974	481.714286
Government/Transit	10	4901.00	3070.0	527.91143	490.100000
Shelter/Drug house	5	1847.50	1535.0	374.83420	369.500000
College	5	2910.50	1535.0	374.83420	582.100000

Average scores were used for ties.

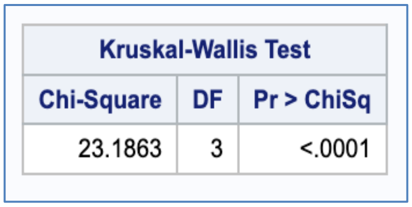
Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq	Exact Pr >= ChiSq
148.4113	13	<.0001	.

Figure 1.1: Total victims by location

A comparative analysis of the number of victims killed per incidence provided insights in location-specific variances. Residential locations recorded the highest number of incidents (N = 415); however, there were lower number of victims killed per incident. We used the Dwass, Steel, Critchlow-Fligner test (*SAS Help Center: Multiple Comparisons Based on Pairwise Rankings*, n.d.) (Figure 1.2, Appendix A) to further validate these differences, and the results showed that, in comparison to residential areas, victim counts per incident are much greater in open spaces, bars/clubs/restaurants, and commercial/retail enterprises. Furthermore, victim counts were continuously lowest in instances involving vehicles, whereas they were significantly higher at schools and universities. These results highlight the need for more study into location-specific risk factors and preventive measures by showing that certain areas are linked to greater risk levels.

WEAPON TYPE AND FATALITIES/INJURIES

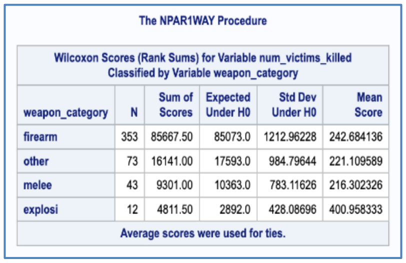


Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
23.1863	3	<.0001

Figure 2.1: Impact of weapon on fatalities

We used a Kruskal-Wallis test (Figure 2.1) to analyze the impact of weapon type on fatalities in mass killings across various weapon categories. The findings indicated significant differences in fatalities (Chi-Square = 23.1863, p < 0.0001) and injuries (Chi-Square = 44.9188, p < 0.0001) based upon weapon type. Explosive devices earned the highest mean score for both fatalities and injuries, signifying their tremendous lethality and capacity for mass casualties. Firearms, although still extremely dangerous, caused less fatalities and injuries than explosives. Melee weapons and other classifications exhibited the minimal impact.



The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable num_victims_killed
Classified by Variable weapon_category

weapon_category	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
firearm	353	85667.50	85073.0	1212.96228	242.684136
other	73	16141.00	17593.0	984.79644	221.109589
melee	43	9301.00	10363.0	783.11626	216.302326
explosi	12	4811.50	2892.0	428.08696	400.958333

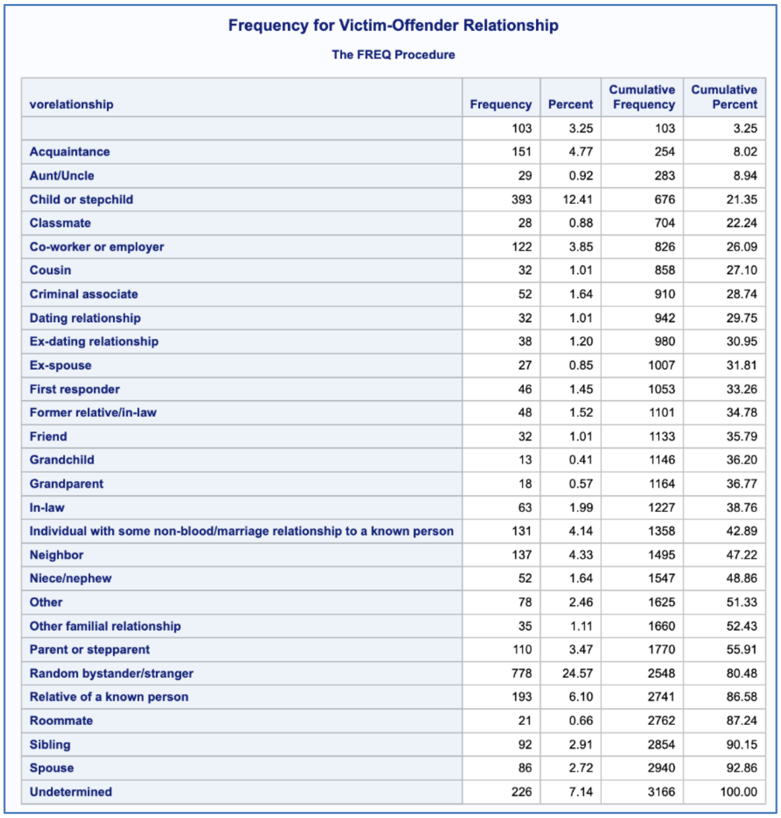
Average scores were used for ties.

Figure 2.2: Wilcoxon Rank-Sum test

The Wilcoxon Rank-Sum test (Figure 2.2) validated these tendencies, with boxplots (Figure 2.4, Appendix A) visually illustrating the severity of explosive-related occurrences. The injury distribution followed a similar pattern, with explosives (Mean Score = 417.13) and firearms (Mean Score = 251.66) inflicting the highest number of injuries. Melee weapons (Mean Score = 176.56) yielded significantly reduced casualty numbers. The correlation between fatalities and injuries across weapon categories highlights the significant influence of weapon choice on victim outcomes, stressing the necessity for further investigation into contextual elements including shooter intent and law enforcement reaction times.

VICTIM-OFFENDER RELATIONSHIPS

To analyze the prevalence of prior relationships between offenders and victims, we examined the victim-offender relationship variable using frequency distributions (Figure 3.1). A missing value analysis indicated that only 3.25% of records lacked relationship data, minimizing potential biases in interpretation.



Frequency for Victim-Offender Relationship

The FREQ Procedure

vorelationship	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	103	3.25	103	3.25
Acquaintance	151	4.77	254	8.02
Aunt/Uncle	29	0.92	283	8.94
Child or stepchild	393	12.41	676	21.35
Classmate	28	0.88	704	22.24
Co-worker or employer	122	3.85	826	26.09
Cousin	32	1.01	858	27.10
Criminal associate	52	1.64	910	28.74
Dating relationship	32	1.01	942	29.75
Ex-dating relationship	38	1.20	980	30.95
Ex-spouse	27	0.85	1007	31.81
First responder	46	1.45	1053	33.26
Former relative/in-law	48	1.52	1101	34.78
Friend	32	1.01	1133	35.79
Grandchild	13	0.41	1146	36.20
Grandparent	18	0.57	1164	36.77
In-law	63	1.99	1227	38.76
Individual with some non-blood/marriage relationship to a known person	131	4.14	1358	42.89
Neighbor	137	4.33	1495	47.22
Niece/nephew	52	1.64	1547	48.86
Other	78	2.46	1625	51.33
Other familial relationship	35	1.11	1660	52.43
Parent or stepparent	110	3.47	1770	55.91
Random bystander/stranger	778	24.57	2548	80.48
Relative of a known person	193	6.10	2741	86.58
Roommate	21	0.66	2762	87.24
Sibling	92	2.91	2854	90.15
Spouse	86	2.72	2940	92.86
Undetermined	226	7.14	3166	100.00

Figure 3.1: Frequency for Victim-Offender Relationship

Our distribution analysis identified random bystanders/strangers (24.57%) and children/stepchildren (12.41%) as the most frequently recorded victim-offender relationships. Other notable relationships included relatives of known persons (6.10%), acquaintances (4.77%), and neighbors (4.33%).

Overall, the analysis illustrates the complexity of mass violence, demonstrating how location, weapon type, and victim-offender dynamics collectively influence incident severity. These findings reinforce the importance of interdisciplinary research to better understand the underlying patterns and inform targeted prevention strategies.

SUGGESTIONS FOR FUTURE STUDIES

Future research should explore the psychological characteristics of offenders, including aspects like mental health history, social isolation, and ideological reasons (*MASS SHOOTINGS IN THE UNITED STATES*, N.D.) . Additionally, a longitudinal study comparing mass killings in different countries could shed light on how cultural, legal, and economic factors influence event rates. Integrating qualitative data, such as interviews with law enforcement and survivors, may provide a more comprehensive contextual understanding. Finally, machine learning approaches could be used to forecast high-risk circumstances and inform proactive preventative strategies.

CONCLUSION

This study provides a data-driven analysis of mass killings, highlighting critical patterns in offender demographics, incident types, weapon usage, and victim-offender relationships. Our findings highlight the impact of location and weapon choice on fatality rates, as well as the complexities of the offender-victim interaction. By leveraging statistical methods, we uncover insights that challenge conventional narratives and contribute to a more nuanced understanding of mass violence. In the end, this study shows how important it is to use a variety of methods to stop these tragedies, such as combining law enforcement strategies, mental health interventions, and policy changes, to lessen their effects.

REFERENCES

Haynes, W. (2013). Wilcoxon Rank Sum Test. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 2354–2355). Springer. https://doi.org/10.1007/978-1-4419-9863-7_1185

Kruskal Wallis Test—An overview | ScienceDirect Topics. (n.d.). Retrieved February 21, 2025, from https://www.sciencedirect.com/topics/medicine-and-dentistry/kruskal-wallis-test

Mass Killings in America, 2006—Present—Dataset by associatedpress. (n.d.). Data.World. Retrieved February 21, 2025, from https://data.world/associatedpress/mass-killings-public

Mass shootings in the United States: Population health impacts and policy levers | Health affairs brief. (n.d.). Retrieved February 21, 2025, from https://www.healthaffairs.org/do/10.1377/hpb20220824.260250/full/

SAS Help Center: Multiple Comparisons Based on Pairwise Rankings. (n.d.). Retrieved February 21, 2025, from https://documentation.sas.com/doc/en/statug/15.2/statug_npar1way_details23.htm

Peterson, J. K., Densley, J. A., Hauf, M., & Moldenhauer, J. (2024). Epidemiology of Mass Shootings in the United States. *Annu Rev Clin Psychol*, *20*(1), 125-148. https://doi.org/10.1146/annurev-clinpsy-081122-010256

SAS Institute Inc. (2024, February 16). *SAS® OnDemand for Academics: SAS® Enterprise Guide® 8.3 installation and connection.* SAS Help Center. https://welcome.oda.sas.com

ACKNOWLEDGEMENT

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author(s) at:

1. Vedika Dahad

   Northeastern University

   +1 (617) 953 9442

   dahad.v@northeastern.edu

2. Swapna Ashish Patel

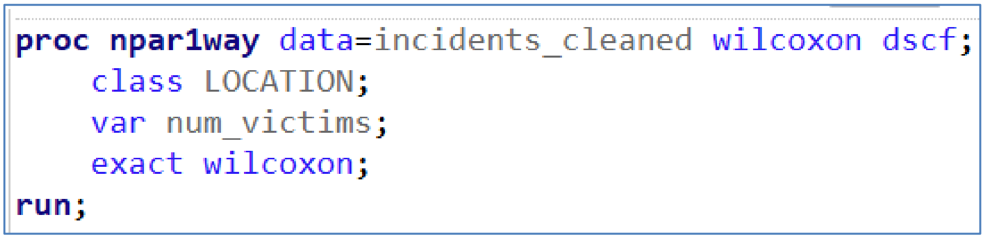
   Northeastern University

   +1 (913) 387 9839

   patel.swapn@northeastern.edu

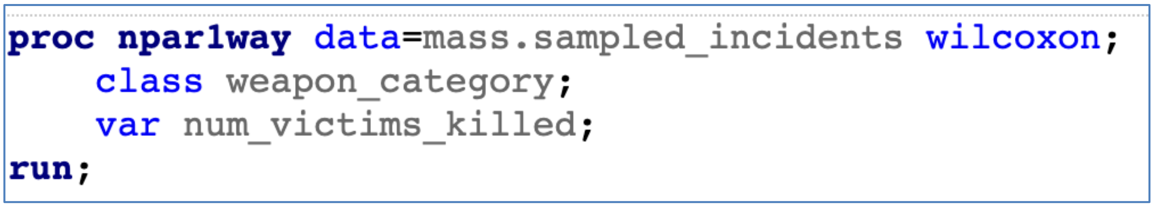
APPENDIX A

Figure 1.2: A Wilcoxon rank-sum test was performed to compare the number of victims across different locations using exact p-values.



```
proc npar1way data=incidents_cleaned wilcoxon dscf;
    class LOCATION;
    var num_victims;
    exact wilcoxon;
run;
```

Figure 2.3: A kruskal-wallis test and Wilcoxon Rank-Sum was performed to analyze the impact of weapon type on fatalities



```
proc npar1way data=mass.sampled_incidents wilcoxon;
    class weapon_category;
    var num_victims_killed;
run;
```

Figure 2.4: Wilcoxon Rank-Sum test validated these tendencies, with boxplots

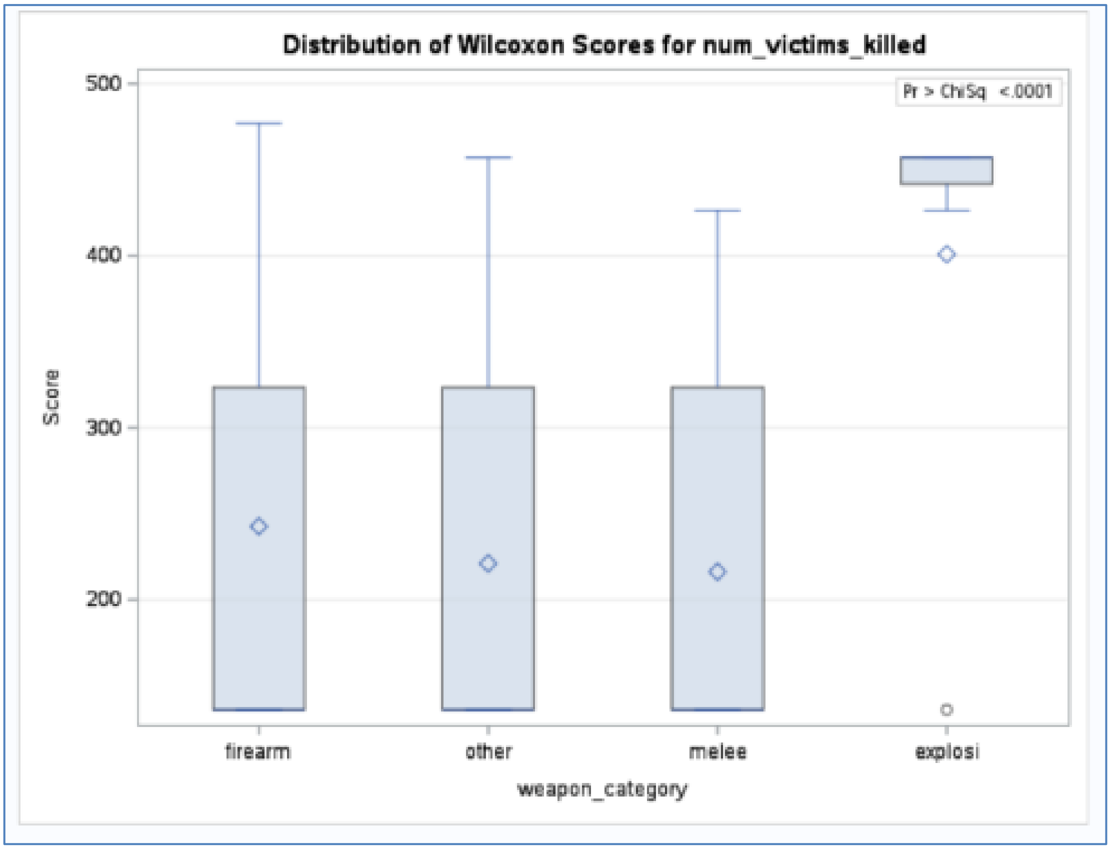
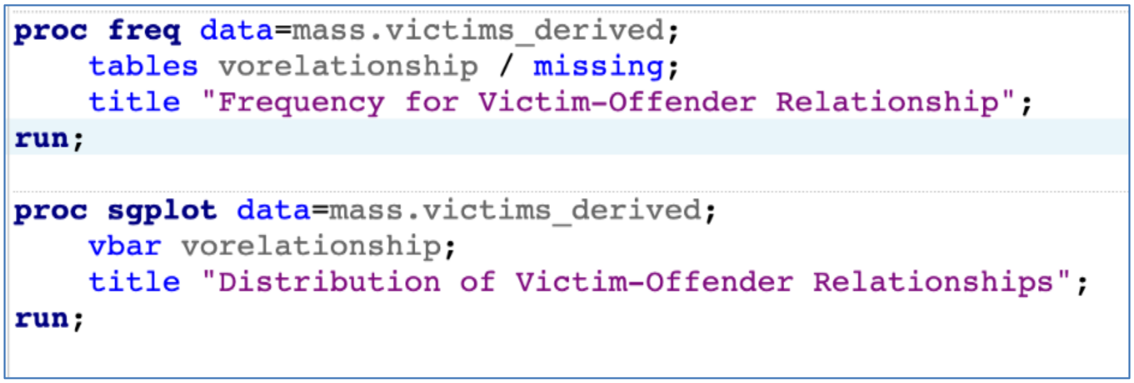


Figure 3.1: We analyzed victim-offender relationships using frequency distributions.



```
proc freq data=mass.victims_derived;
    tables vorelationship / missing;
    title "Frequency for Victim-Offender Relationship";
run;

proc sgplot data=mass.victims_derived;
    vbar vorelationship;
    title "Distribution of Victim-Offender Relationships";
run;
```