Executive Summary

School shootings have been a tragic part of our history for centuries. The earliest known United States shooting to happen on school property was the Pontiac's Rebellion school massacre on July 26, 1764.[1] Over the years, the frequency and impact of these incidents have increased, especially since the massacre at Colorado's Columbine High School in 1999.[2]

The Columbine High School shooting was particularly shocking because it occurred in an affluent suburb of Denver, Colorado, unlike many previous school shootings that had taken place in lower-income areas. Columbine is located in Jefferson County, which had a median household income significantly higher than the national average at the time of the shooting.

This aspect of the shooting challenged the prevailing notion that school violence was primarily a problem in disadvantaged communities. It brought to light the fact that school shootings could happen anywhere, even in seemingly safe, wealthy neighborhoods. The realization that no school or community was immune to such violence added to the widespread fear and concern generated by the incident.

Moreover, the perpetrators, Eric Harris and Dylan Klebold, came from relatively stable, middle-class families, further challenging stereotypes about the background of school shooters. This aspect of the case contributed to a broader conversation about the complex factors that could lead young people to commit such acts of violence, regardless of their socioeconomic status. Ever since Colorado's Columbine High School massacre, school shootings have continued to occur in communities of all socioeconomic backgrounds, reinforcing the lesson that no school or community is immune to this type of violence.

---

[1] Glavin, C. (2018, July 26).
[2] THE ASSOCIATED PRESS. (2022, May 25).

Between the years of 2001 and 2022 school years, there were 1,375 school shootings at public and private elementary and secondary schools, resulting in 515 deaths and 1,161 injuries. The highest number of school shootings and casualties occurred during the 2021-22 school year, with 327 incidents resulting in 81 deaths and 269 injuries. The parking lot was the most frequently reported site of school shootings (28.3% of cases), followed by any area immediately outside the school's front or side entrances (20.4%), and "elsewhere inside the school building," which refers to any area outside of the classroom, hallways, or basketball court (12.5%).[3]

Mass school shootings have emerged as a deeply concerning and tragic phenomenon, leaving non-removable scars on communities and igniting intense debates surrounding school safety measures, gun control policies, and mental health support systems. The alarming frequency and devastating impact of these incidents have captured public attention and underscored the pressing need for rigorous, multidisciplinary research to explain the complex web of factors that contribute to these heinous acts.

“Security theater” is a term that refers to the practice of implementing security measures that are considered to provide the feeling of improved security while doing little or nothing to achieve it.[4] In the context of schools, it often involves visible, tangible measures like metal detectors, cameras, and clear backpack policies. However, these measures often fail to actually make students safer. One school safety expert pointed out, "security technology is only as good as the weakest human link behind it".[5] Therefore, it’s crucial to ensure that the people implementing and managing these security measures are well-trained and that the school community as a whole is committed to maintaining a culture of safety.

---

[3] USAFacts. (2022, May 26).
[4] Wikipedia Contributors. (2024, April 11).
[5] Bankert, A. (2022, July 19).

By leveraging the power of empirical data and applying meticulous analytical techniques, we can unravel the intricate patterns, evolving trends, and potential risk factors associated with these heart-wrenching events. This research represents a vital step forward in understanding this multifaceted problem and holds the potential to guide meaningful change in safeguarding our educational institutions and protecting the lives of students, educators, and communities nationwide.

Examining the Characteristics and Trends of Mass School Shootings through Empirical Analysis

DS 490 Capstone Research Paper

By: Ryan Lee & Jake Klingler

Under the guidance of: Professor Dr. Christopher Briggs

Introduction

School safety is an important concern for educators, parents, and communities across the United States. As data scientists, we have a unique opportunity to contribute to the ongoing efforts to enhance school safety by analyzing patterns and trends in school shooting incidents. By identifying potential predictors related to community characteristics, we aim to provide valuable insights that can help schools better anticipate the nature of threats and implement effective safeguards.

One critical aspect of our research focuses on the possibility of two distinct types of mass shooting incidents at schools, categorized based on the final outcome for the perpetrator. We hypothesize that the target variable "Shooter Killed" may indicate different types of attacks, distinguishing between those where the perpetrator attempts to escape and those where the perpetrator is on a suicide mission. Additionally, we acknowledge the existence of a potential third category that does not fit into either of these two types referred to as “Uncertain or Mixed Motives”. This would be a scenario where the perpetrator's intentions are unclear or ambiguous, and their actions do not clearly indicate a definite attempt to escape or a deliberate suicide mission.

To investigate this hypothesis, we utilized a school shooting dataset compiled by David Riedman, the founder of the K-12 School Shooting Database, an independent research project started in February 2018. This extensive dataset consisted of 2,572 school shooting incidents in the United States, spanning from March 11, 1966, to November 3, 2023. The dataset includes incidents that occurred in public schools, both elementary and high schools, as well as public charter schools. By employing rigorous statistical methods, we aim to examine the characteristics, trends, and patterns that distinguish these different types of mass shootings.

Our preliminary analysis suggests two potential categories: "Type 1" incidents, where the fatality percentage is below 10%, and "Type 2" incidents, where the fatality percentage is above 25%. By identifying and analyzing these categories, we hope to uncover important differences in perpetrator psychology, planning, and execution of the attack, among other factors.

The insights gained from this study could significantly contribute to the development of more effective threat assessments, security measures, and intervention strategies. By understanding the complex nature of school shootings and the distinct characteristics of different types of incidents, we can provide valuable information to help schools better anticipate and respond to potential threats.

Through this research, we aim to make a meaningful contribution to the ongoing efforts to ensure the safety and well-being of students, educators, and communities across the United States. By leveraging the power of data science, we can work towards creating safer learning environments and reducing the devastating impact of school shootings on our society.

Analysis Approach

To gain a more comprehensive understanding of school shooting incidents, we embarked on a mission to expand the depth and breadth of our dataset. The original dataset, consisting of 40 columns, provided a solid foundation by capturing various aspects of each incident, such as the number of victims and the duration of the shooting. However, we recognized the need to incorporate additional characteristics related to the schools and their surrounding areas to uncover deeper insights and potential correlations.

Our initial focus was on the "LAT" and "LNG" columns, which held the latitude and longitude coordinates for each incident. These coordinates served as the key to unlocking a wealth of geographical information. By leveraging the power of reverse geocoding, we accessed the GPS Coordinates website to convert these coordinates into human-readable addresses. This process allowed us to associate each incident with a specific location, providing a more tangible context for our analysis.

To streamline the data extraction process and minimize manual effort, we utilized the capabilities of web automation. By employing Chrome and Firefox web browsers in conjunction with their respective WebDrivers (ChromeDriver and GeckoDriver), we created a seamless interface between our programming environment and the web pages we needed to interact with. The Selenium library, a powerful tool for web automation, enabled us to programmatically navigate to the GPS Coordinates website and simulate user actions, such as inputting coordinates, clicking buttons, and extracting relevant data from the rendered web pages.

As we delved deeper into the geographical context of each incident, we recognized the importance of incorporating zip code information. By parsing the retrieved addresses and

extracting the zip codes, we created a new column in our dataset, linking each incident to its corresponding zip code.

With the zip codes in hand, we embarked on a mission to collect demographic information associated with each incident's location. The Zip Atlas website emerged as a valuable resource, providing comprehensive demographic data based on zip codes. Once again, we leveraged the power of Selenium and WebDrivers to automate the scraping process, efficiently gathering relevant characteristics such as racial composition and average income earnings based on age and gender.

To streamline the data integration process, we implemented a code solution that directly extracted the demographic information from the Zip Atlas website and incorporated it into the original dataset, aligning each zip code with its corresponding incident. This approach ensured that the newly acquired data was seamlessly merged with the existing information, enhancing the richness and depth of our dataset.

However, we encountered a challenge when dealing with zip codes that appeared more than once in the dataset, indicating multiple incidents within the same geographical area. To address this issue, we developed a customized code snippet that intelligently handled duplicate zip codes. The code was designed to identify and match the demographic data not only to the first occurrence of a zip code but also to any subsequent occurrences. This ensured that the demographic information was accurately associated with each incident, even when multiple incidents shared the same zip code.

To gain a better understanding of the school shooting incidents, we recognized the importance of gathering school-specific information. We utilized a Kaggle dataset that contained the names of the schools where the shootings occurred. To ensure accurate data collection, we

employed a fuzzy match Python library, which allowed us to obtain both the proper school names and their corresponding school ID numbers. With these school ID numbers and zip codes, we were able to access the National Center for Education Statistics (NCES) website, a reliable source for school-related data. By automating the navigation and data extraction process using Selenium and WebDrivers, we collected crucial data points such as the number of students in each grade level and the racial composition of the student body. This approach enabled us to establish a direct link between the school shooting incidents and the specific schools involved.

The final piece of the puzzle was the integration of all the collected data into our expanded dataset. We successfully incorporated the demographic information obtained from the Zip Atlas website into the original dataset, aligning each zip code with its corresponding incident. However, we still needed to merge the NCES dataset, which contained crucial school-specific information, with our expanded dataset.

To seamlessly integrate the NCES dataset with our expanded dataset, we employed the pd.merge() function, a pandas library in Python, which offers a streamlined approach to combine the two datasets, leveraging matching criteria for enhanced accuracy and precision. We strategically designed the merging process to rely on incident IDs to ensure that the relevant school-specific information from the NCES dataset was accurately aligned with the corresponding incidents in our expanded dataset.

By incorporating these additional data sources, we transformed our dataset into a rich tapestry of information, providing a multifaceted perspective on school shooting incidents. The demographic and school-specific data added valuable context, enabling us to explore the complexities and potential factors associated with these tragic events. With this expanded dataset

as our foundation, we were well-equipped to embark on a thorough analysis, uncovering patterns, correlations, and distinctions that would have otherwise remained hidden.

The integration of web automation, data scraping, and data enrichment techniques streamlined the entire process, allowing us to efficiently handle a large number of incidents and ensure the accuracy of the collected data. By minimizing manual effort and reducing the potential for human error, we created a robust and reliable dataset that served as a powerful tool for our subsequent analysis.

We applied K-means clustering to group school shooters based on their per capita income per zip code, with the aim of understanding the relationship between economic factors and the likelihood of a shooter being killed or not killed during the incident. To determine the optimal number of clusters, we employed silhouette analysis, which measures the quality of clustering by assessing the separation and cohesion of data points within and between clusters. Subsequently, we utilized logistic regression to model the binary outcome of whether a shooter is killed or not killed, using the per capita income per zip code as the predictor variable.

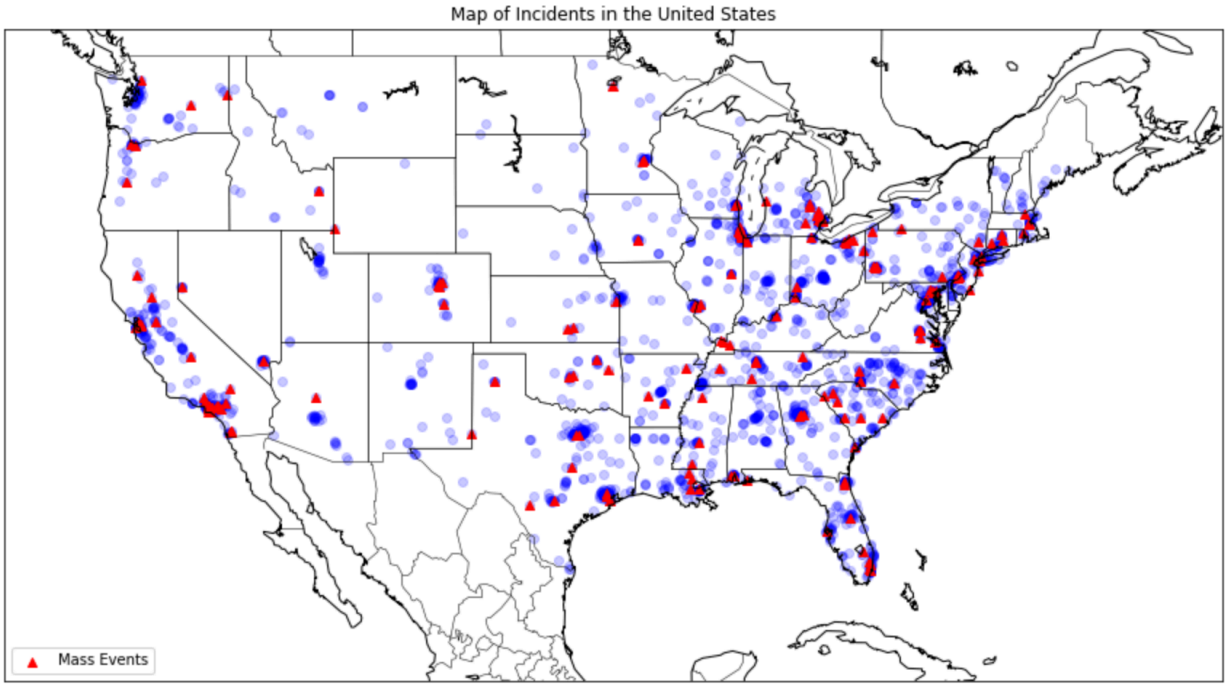
As for the machine-learning aspect of this project, we decided to utilize multiple different machine-learning models to see which model would perform the best and yield the greatest accuracy for predicting our target variables. The models that we chose to use were; multi-linear, multi-logistic, neural networks, random forests, decision trees, and XGBoost. Each of these models were split into regression and classification models to adhere to the different outputs of our target variables. We decided to use these models because we thought that it would be a good way to use different machine-learning algorithms to look at the same problem and, in turn, we would get the best results possible.

Exploratory Analysis

Upon completing the data collection process, we began our exploratory analysis with a comprehensive dataset consisting of 2,584 school shooting incidents, each characterized by 40 distinct attributes. However, our pursuit of a more in-depth understanding led us to expand the dataset by web scraping additional relevant information from various online sources. This exhaustive data-gathering effort resulted in a substantially enriched dataset, now comprising 2,584 rows and an impressive 1,248 columns, each representing a unique aspect of the incidents and their surrounding contexts.

As we delved into the intricacies of the dataset, we encountered instances where the school name associated with an incident was missing. Recognizing the critical importance of having complete and accurate information for our analysis, we made the decision to exclude these incidents from our study. By applying this filtering criterion, we narrowed down our focus to a refined dataset encompassing 2,572 school shooting incidents, each with a verified and recorded school name.

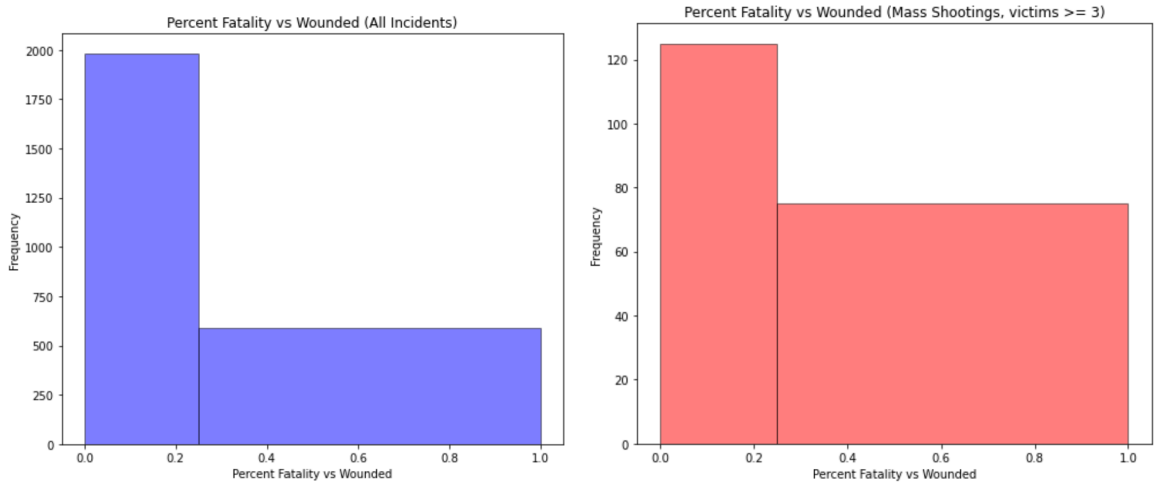
This refined dataset served as the foundation for our exploratory analysis, enabling us to dive deeper into the patterns, trends, and potential correlations within the data. With a comprehensive set of attributes spanning incident details, geographical information, demographic data, and school-specific characteristics, we were well-equipped to unravel the complexities surrounding school shooting incidents.



The map reveals a distinct pattern in the prevalence and severity of school shootings throughout the United States. A striking concentration of both blue and red dots, representing mass events, is evident in the Northeastern and Southeastern regions of the country. This clustering suggests that these areas have experienced a disproportionately high occurrence of school shootings compared to other parts of the nation. In contrast, the Western states exhibit a notably lower density of incident markers. This disparity indicates that school shootings have been less frequent in this region relative to the East Coast. The reduced frequency of both blue and red dots in the West suggests a lower overall incidence of these tragic events.

After analyzing the geographical distribution of school shootings across the United States, the next step in the exploratory analysis was to examine the percentage fatality associated with school shootings.

*Percentage Fatality = (Victims Killed / Number of Victims) * 100*



The two graphs presented show the relationship between percent fatality and percent wounded in school shooting incidents. The first graph on the left includes data from all incidents, while the second graph on the right focuses specifically on mass shootings, which are defined as incidents with three or more victims within our research.

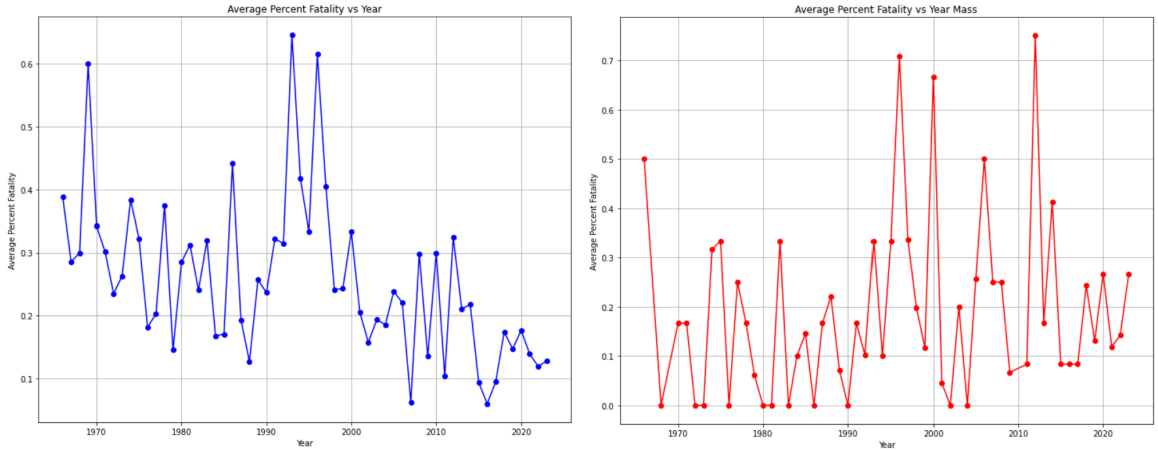
Both graphs utilize histograms to display the distribution of the data, with the x-axis representing the percent fatality divided into two bins: 0-25 percent fatality and 25-100 percent fatality. The y-axis indicates the frequency of incidents falling into each bin.

The graph on the left can be observed that the majority of the incidents fall into the first bin (0-25 percent fatality), with a frequency of reaching nearly 2,000 incidents. The second bin (25-100 percent fatality) has a significantly lower frequency of just above 500 incidents. This suggests that, overall, a larger proportion of school shooting incidents result in a lower fatality rate.

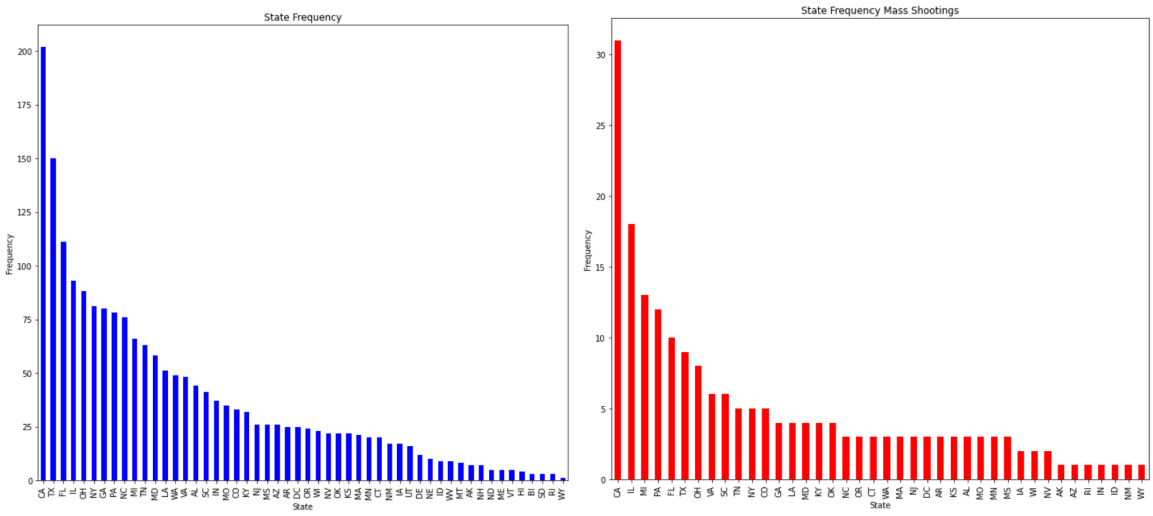
When we examine the second graph, which focuses on mass shootings, we notice a different distribution of the data. The frequency of incidents in the first bin (0-25 percent fatality)

is just above 120, while the frequency of incidents in the second bin (25-100 percent fatality) is approximately just under 80 incidents.

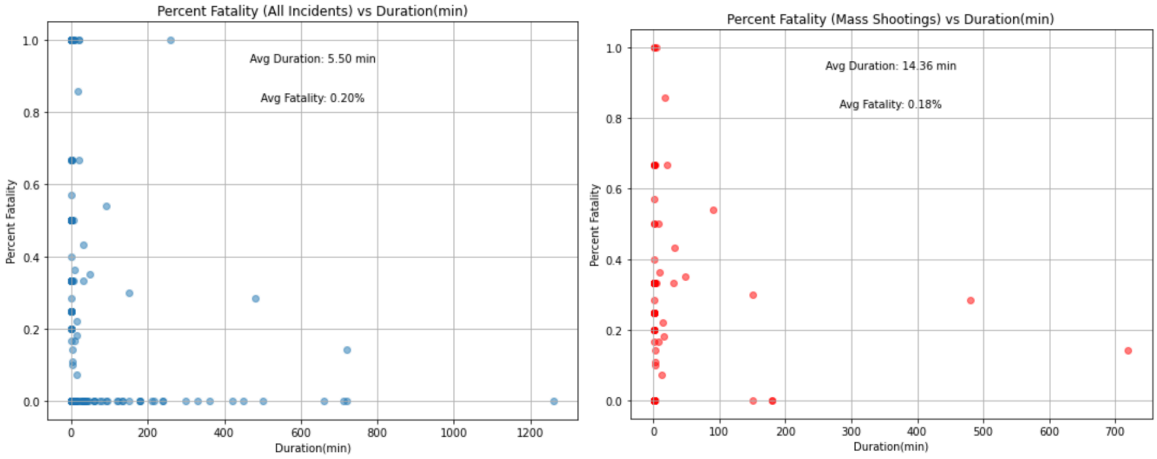
Both the graphs suggest that there is a positive relationship between the percent fatality and percent wounded in incidents, with mass shootings tending to be more severe in terms of both fatalities and injuries compared to incidents overall.



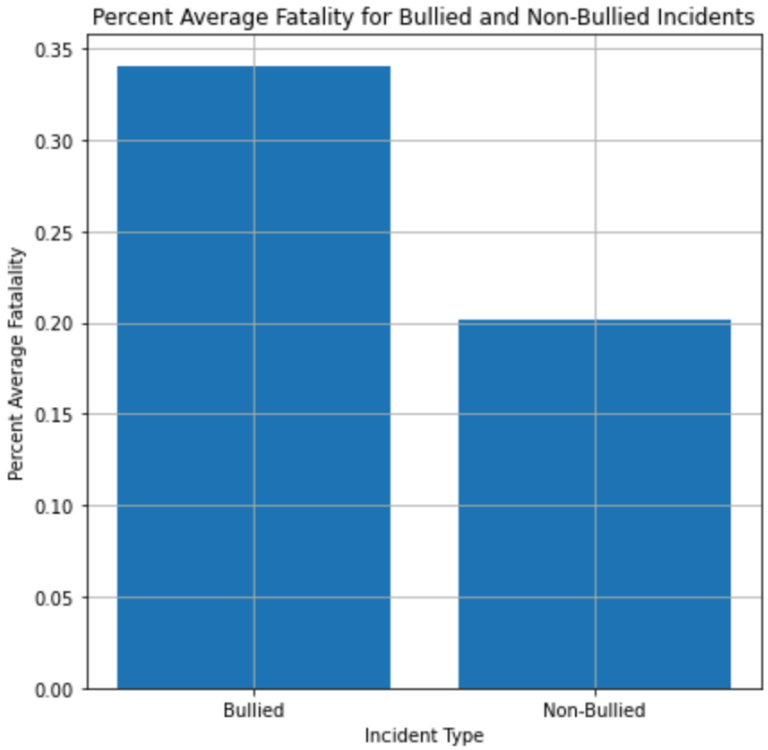
The line plot graph on the left illustrates the overall trend in the average percent fatality of school shootings over the specified time period. The data reveals a notable peak in fatalities prior to 1970, followed by two significant spikes between 1990 and the early 2000s. These peaks suggest periods of heightened severity in school shooting incidents. The right-hand line plot graph focuses specifically on the average percent fatality in school shootings classified as mass shootings. This graph highlights two pronounced peaks in fatalities, one spanning from approximately 1995 to the early 2000s, and another, more substantial spike occurring between 2010 and 2015. The later spike indicates a particularly deadly period for mass shootings within the context of school shootings.



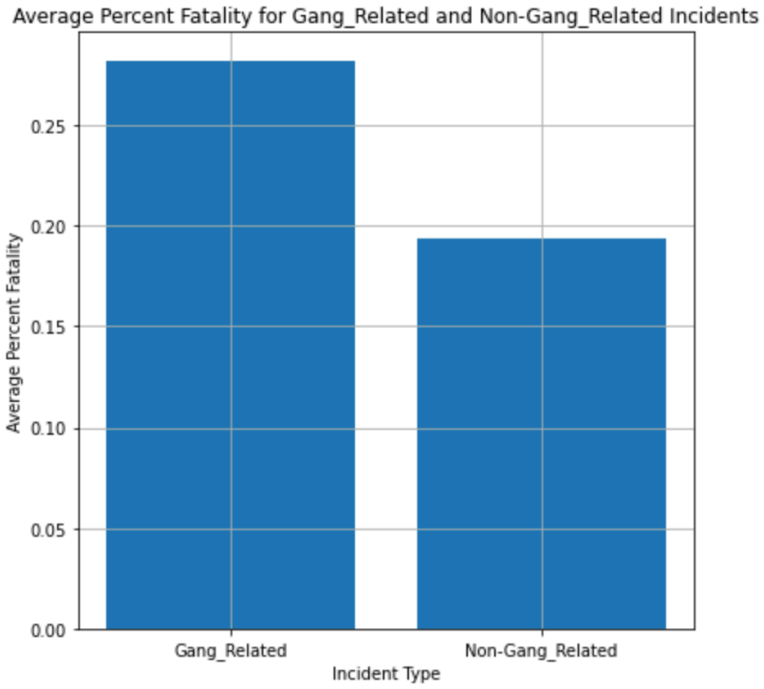
The bar graphs presented provide a comparative analysis of the frequency of school shooting incidents, both non-mass and mass shootings, across various states in the United States. The graph on the left, which focuses on non-mass shooting incidents, reveals that California experienced the highest number of such incidents, with nearly 200 cases reported. Texas closely follows California, with around 150 non-mass shooting incidents, while Florida ranks third, with over 110 reported cases. Shifting our attention to the bar graph on the right, which illustrates the frequency of mass shooting incidents, we observe a similar trend. California once again tops the list, with more than 30 mass school shooting incidents, a figure significantly higher than any other state. Illinois secures the second position, with a frequency exceeding 15 incidents, while Michigan ranks third, with just over 10 reported mass shootings. Notably, California consistently emerges as one of the top states for both non-mass and mass school shooting incidents.



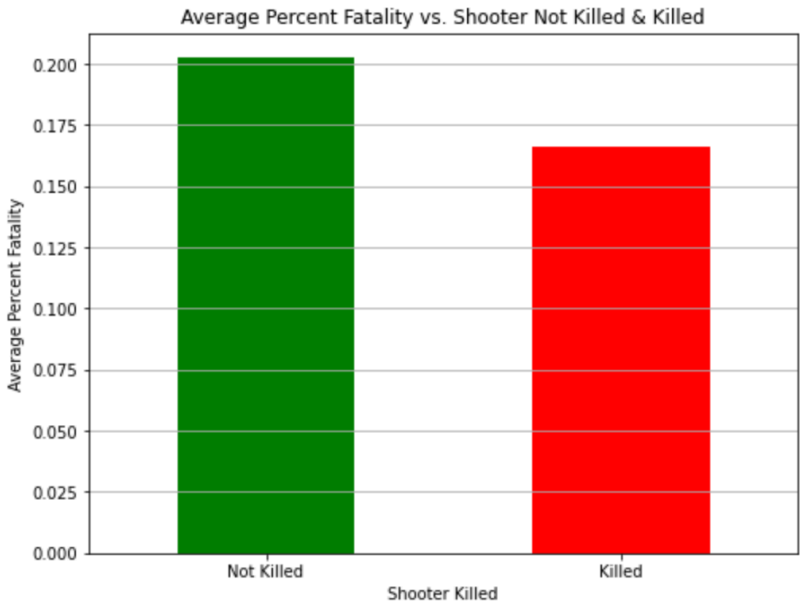
The two scatter plots show the relationship between percent fatality and incident duration in minutes. The graph on the left includes data from all shooting incidents, while the graph on the right focuses specifically on mass shootings. Both graphs indicate there is no strong correlation between incident duration and the percent of victims killed, suggesting that whether an incident results in a high or low percentage of fatalities does not depend strongly on how long the shooting lasts. However, the graphs do reveal that mass shootings tend to have a longer average duration (14.36 minutes) compared to the overall average across all incidents (5.50 minutes). Despite the longer duration though, the average percent fatality in mass shootings (0.18%) is actually slightly lower than the overall average (0.20%).



The graph shows the percent average fatality for bullied and non-bullied incidents. For bullied incidents, the average percentage fatality is around 34%. This means that on average, bullied incidents result in a fatality rate of about 34%. In contrast, for non-bullied incidents, the average percentage fatality is much lower at around 20%. So non-bullied incidents have an average fatality rate of approximately 20%. From this we can infer that bullied incidents have a substantially higher average fatality rate compared to non-bullied incidents. The percentage fatality for bullied incidents is over 1.7 times higher than for non-bullied incidents, highlighting the increased severity and danger associated with bullying situations that escalate to the point of an incident occurring.

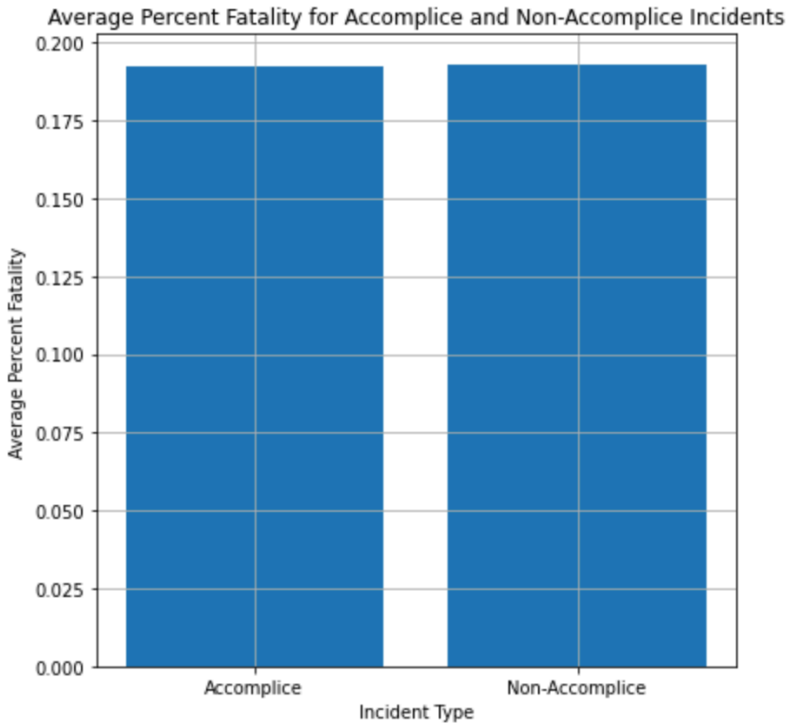


The graph above shows the average percent fatality for gang-related and non-gang-related incidents. Incidents involving gangs result in a fatality rate above 25%. Incidents not involving gangs have an average fatality rate of below 20%. We can infer that gang-related incidents tend to have a higher average fatality rate compared to incidents not involving gangs. The percentage fatality for gang-related incidents is roughly 1.5 times higher than for non-gang-related incidents, suggesting that the presence of gang activity or gang members in an incident is associated with an increased likelihood of fatalities occurring.

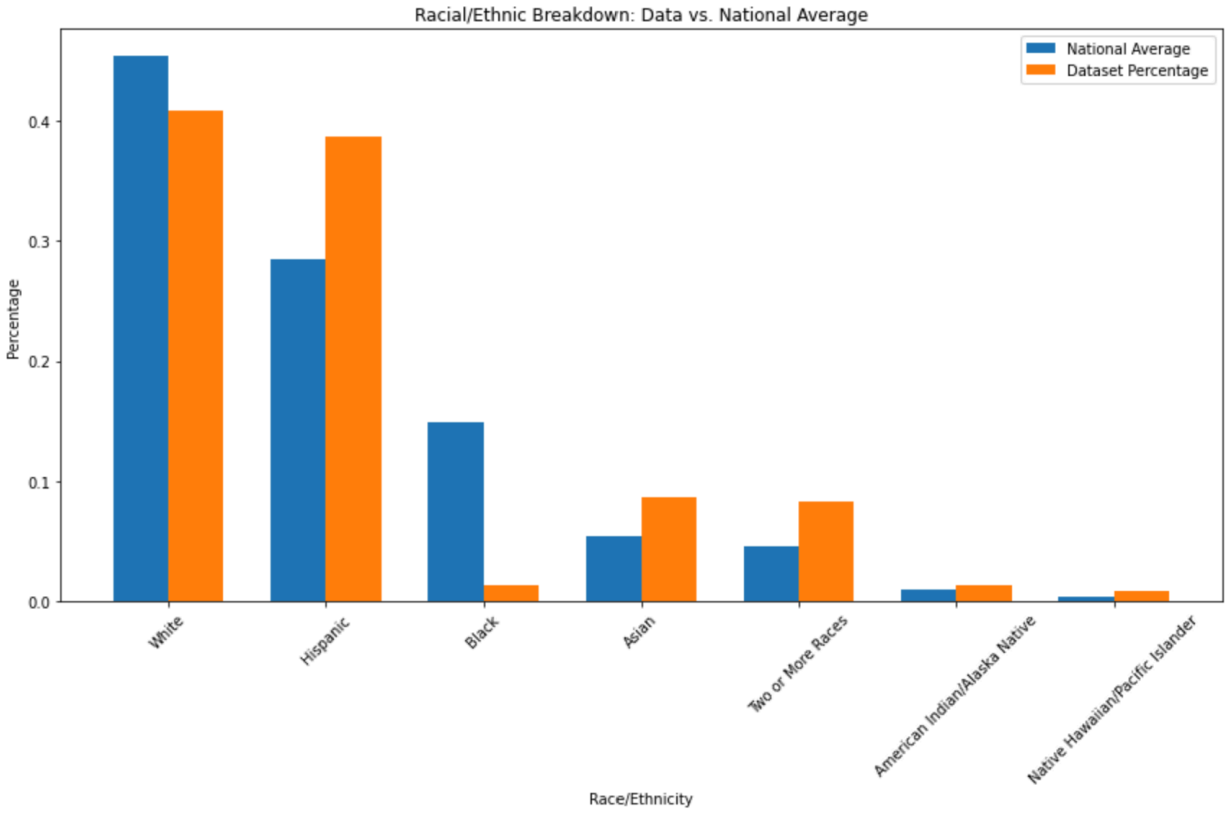


The graph above compares the average percent fatality between active shooter incidents where the shooter was not killed versus those where the shooter was killed. The green bars show that in incidents where the shooter was not killed, the average percent fatality was around 21%. In contrast, the red bars indicate that when the shooter was killed during the incident, the average percent fatality was much lower at roughly 16%.

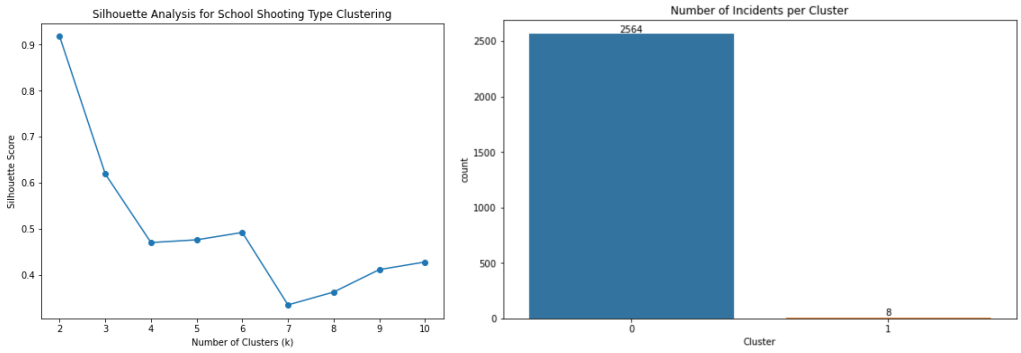
One could infer from this graph that active shooter incidents tend to result in a substantially higher fatality rate when the shooter is not killed compared to when the shooter is killed. The percentage fatality is over 1.3 times higher in cases where the perpetrator is not killed, suggesting that neutralizing or stopping the shooter quickly, even if it results in the shooter's death, may significantly reduce the number of victims killed in active shooting situations.



The bar plot above depicts the average percent fatality for incidents where the shooter had an accomplice and when they didn't have an accomplice. The reason that we chose to look at this plot was because we thought that if there was an accomplice to the shooter, that meant that the attack was more likely to be planned and could've led to a higher fatality rate. Also, if there was an accomplice that could mean that there was an additional;y shooter in the incident which we also assumed would lead to a more fatal event. However, after looking at the plot, we see that there was no significant difference between the two categories, and our hypotheses were incorrect.

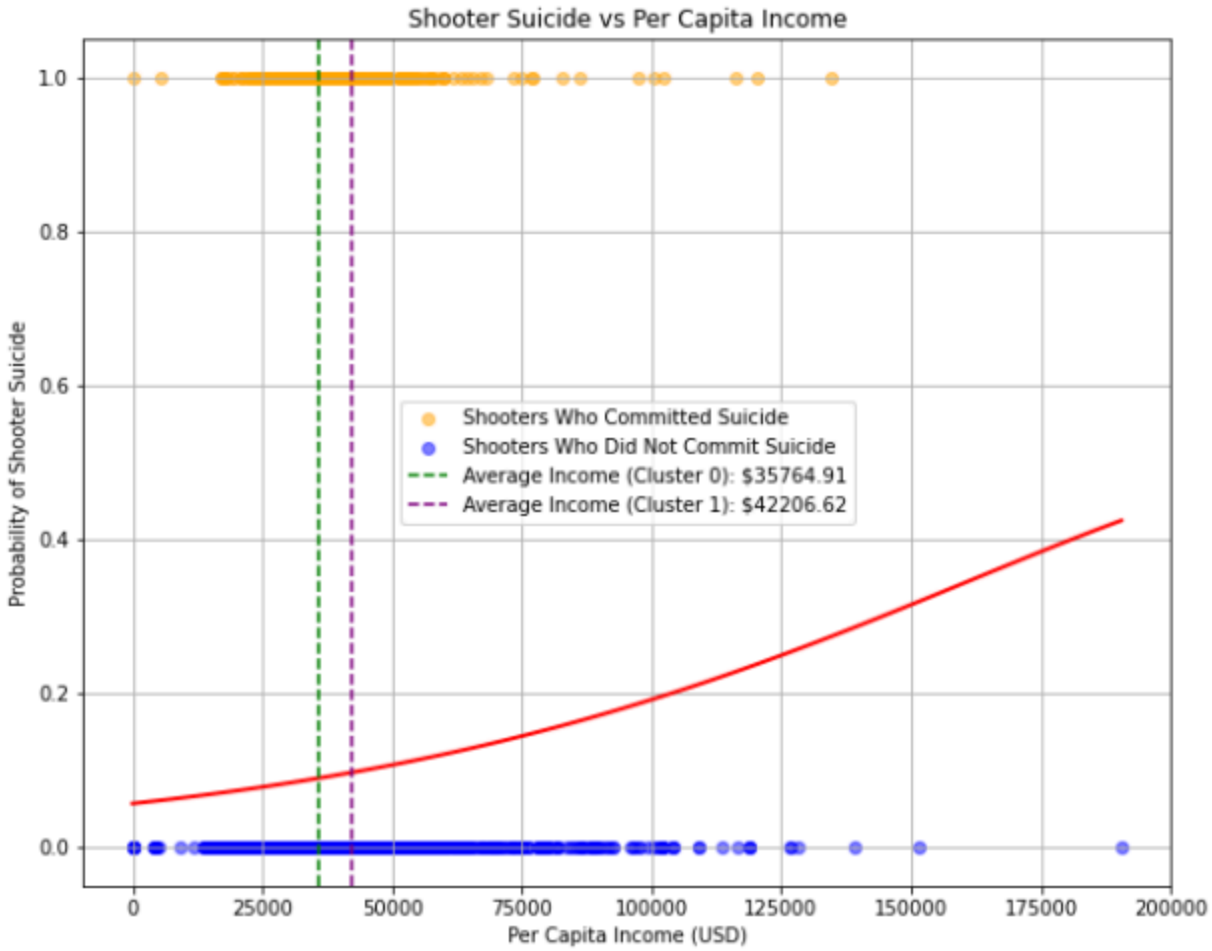


The plot above is a double bar plot that shows the ethnicity breakdown of the schools in our dataset compared to the nationally average. With a quick look at this plot, we can see that our dataset has a higher percentage of hispanics, asians, and students with a mixed ethnic background.

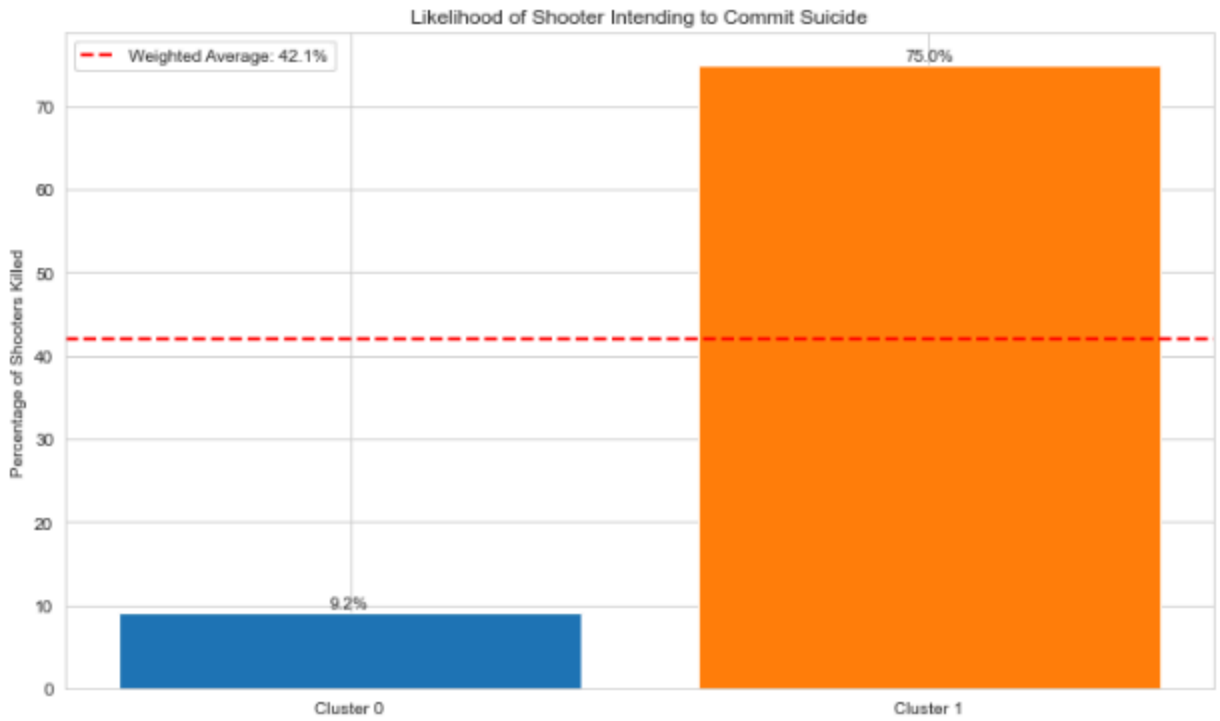


The silhouette analysis, conducted as part of the K-means clustering approach, revealed that the optimal number of clusters for grouping school shooters based on their per capita income

per zip code is two. The resulting clusters exhibit a striking disparity in size and economic characteristics. The first cluster, comprising an overwhelming majority of 2,564 incidents, represents shooters from areas with a similar average income per capita. In contrast, the second cluster consists of only 8 incidents, signifying a much smaller group of shooters who share a distinctly different average income per capita compared to the first cluster. This imbalance suggests that the majority of school shooters in the dataset share similar economic characteristics, while a small minority of shooters have distinctly different economic profiles. The clear separation and imbalance in cluster sizes suggest that economic factors, such as per capita income, may have a strong influence on the occurrence and characteristics of school shooting incidents.



The logistic regression model above suggests that shooters from areas with a higher per capita income are more likely to engage in suicide missions, where they do not expect to survive the attack. Conversely, shooters from areas with lower per capita income are more likely to attempt to escape or surrender after the shooting, indicating a lower inclination for suicide missions.



To account for the disparity in the number of incidents between the two clusters and make the percentages more comparable, we introduced a weighted average approach. This method gives equal weight to each cluster, regardless of the number of incidents, to provide a balanced representation of the overall likelihood of shooter suicide intention.

We assigned equal weights of 0.5 to both Cluster 0 and Cluster 1. The choice of 0.5 as the weight for each cluster ensures that both clusters contribute equally to the weighted average, irrespective of the number of incidents in each cluster. This allows us to compare the clusters on a more equal footing and prevents the cluster with a larger number of incidents from dominating the analysis.

The weighted average percentage of shooters killed is calculated using the following equation:

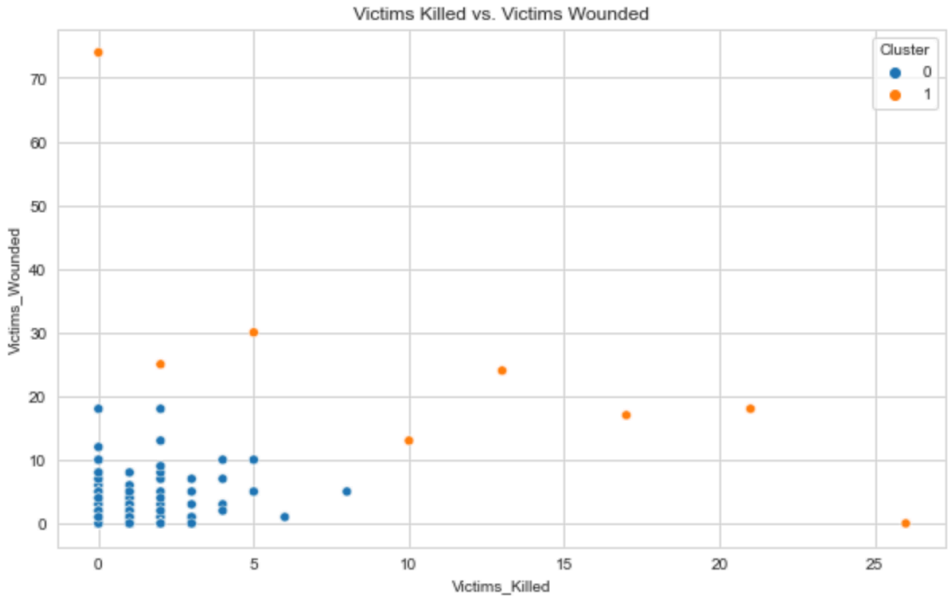
Weighted Average = (Percentage of Shooters Killed in Cluster 0 * Weight of Cluster 0) + (Percentage of Shooters Killed in Cluster 1 * Weight of Cluster 1)

Weighted Average = (9.2% * 0.5) + (75.0% * 0.5) = 42.1%

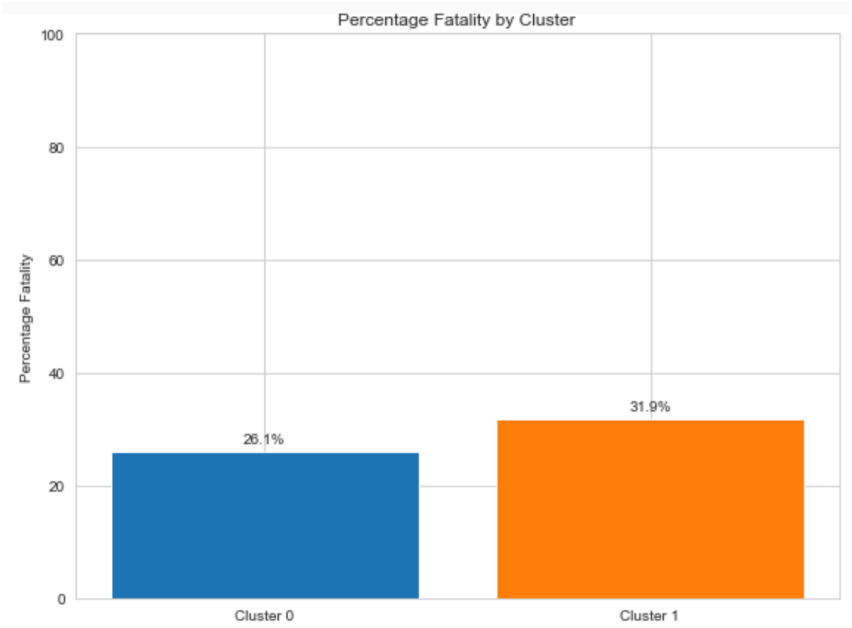
The weighted average percentage of shooters killed across both clusters is 42.1%. This means that when giving equal importance to each cluster, the overall likelihood of a shooter intending to commit suicide is estimated to be 42.1%.

- Cluster 0 (lower income): 9.2% < 42.1% (weighted average)
- Cluster 1 (higher income): 75.0% > 42.1% (weighted average)

We can see that Cluster 1 (higher income) is more likely to have shooters intending to commit suicide compared to Cluster 0 (lower income) and the overall weighted average.



The scatter plot above showcases the relationship between the number of victims killed and wounded in school shooting incidents, with data points categorized into two clusters. Cluster 0 (lower income) represents the majority of incidents and demonstrates a proportional trend between fatalities and injuries. Cluster 1 (higher income) consists of a smaller subset of incidents with a higher ratio of victims killed to wounded, suggesting more lethal attacks.



Two distinct types of incidents based on the fatality percentage:

Type 1 incidents: Clusters with a fatality percentage below 10%

Type 2 incidents: Clusters with a fatality percentage above 25%

Type 1 incidents (low fatality percentage) may involve perpetrators who have different objectives, such as seeking attention, making a statement, or causing general harm, rather than specifically aiming to maximize fatalities.

Type 2 incidents (high fatality percentage) may involve perpetrators who have a clear intention to cause a high number of casualties, possibly driven by a more aggressive or violent mindset, better planning and preparation, or access to more lethal weapons.

To determine the percentage fatality for each cluster, we used the following equation:

Percentage Fatality = (Total Fatalities in Cluster / Total Victims in Cluster) * 100

Cluster 0 (lower income) is categorized as a Type 1 incident, with a percentage fatality of 26.1%. This suggests that incidents in Cluster 0 (lower income) have a relatively low proportion of fatalities compared to the total number of victims. In these incidents, the perpetrator may have had different motivations or objectives, such as causing injury or creating chaos, rather than maximizing fatalities.

Cluster 1 (lower income) is categorized as a Type 2 incident, with a percentage fatality of 31.9%. This indicates that incidents in Cluster 1 (lower income) have a higher proportion of fatalities compared to the total number of victims. In these incidents, the perpetrator may have had a clear intention to cause a high number of fatalities, possibly driven by a more lethal mindset or better planning and execution of the attack.

Results

Below are the results from the models that were trained and tested. The first table holds the results for the regression model and the second table shows the results for the classification models. The regression model aims to predict the percent fatality of an incident and the classification model aims to predict whether or not the shooter was killed in the incident. For regression, the Mean Squared Error(MSE) and R-squared (R^2) scores are what the models were evaluated on. A lower MSE indicates a better performance, while an R^2 score closer to 1 signifies a better model. As for the classification models, they were assessed based on a built-in sklearn accuracy score function, which calculates the fraction of correct predictions, and the MSE of the model. The missing MSE in the second table is for the neural network model where you would typically use a loss function, which doesn't have the same scale or meaning as the MSE so we decided to leave it off the table.

As you can see in the table there are two rows for each evaluation method. This is because we tested on two different sets of features. The first feature set (orig) contains only data from the original data combined with the NCES data and the second feature set (Final) consists of the first feature set as well as financial data for specific zip codes that were selected with feature selection methods. The reason why we decided to do this was because we wanted to see the difference in scores for school-specific features and features that contain data surrounding the schools. In the tables, the better models are highlighted in green and the best model for each target variable is highlighted in blue. Across all regression models, it's evident that the first feature set marginally outperformed the second, with many of the model MSE scores differing by a mere hundredth. The machine learning regression model that performed the best was the XGBoost model with an MSE of 0.0002 and an R-squared score of 0.9983. When looking at the

classification results, the XGBoost and the Random Forest models yielded very good and similar results with the XGBoost producing a slightly higher accuracy score of 0.9047 and the Random Forest recording a slightly better MSE of 0.1051.

Regression	NN	Multi-Linear	Decision Tree	XGBoost	Random Forest
MSE (orig)	0.0175	0.0839	0.0003	0.0002	0.0009
MSE (Final)	0.0562	0.0843	0.0019	0.0008	0.001
R^2 (orig)	0.8689	0.3706	0.9977	0.9983	0.9931
R^2 (Final)	0.5788	0.3676	0.9856	0.9941	0.9927

Classification	NN	Multi-Logistic	Decision Tree	XGBoost	Random Forest
Accuracy(orig)	0.6868	0.4086	0.749	0.8852	0.9008
Accuracy(Final)	0.7023	0.5895	0.677	0.9047	0.8969
MSE(orig)		1.1342	0.2743	0.1206	0.1051
MSE(Final)		0.6206	0.3288	0.107	0.1089

Conclusion/Challenges and Future Ideas

In conclusion, the strong performance of our models points out the potential of leveraging school and city data to predict fatal incidents on school campuses. The consistent accuracy and reliability exhibited by our models indicate that the features derived from such datasets play a crucial role in effectively forecasting these tragic events. This finding not only highlights the importance of data-driven approaches in enhancing campus safety but also underscores the significance of proactive measures in mitigating potential risks.

The biggest challenge of this project was the process of collecting additional data. We started with a dataset containing school safety incidents and we went out to scrape new data on the schools involved in the original dataset, as well as data on the surrounding areas. The majority of the time spent on this project was in the collection stage, as we kept running into obstacle after obstacle when trying to gather our data. Another challenge that we faced was making sure we were on the same page. Since we had to use separate notebooks when conducting EDA on our data, we would constantly edit and create new data frames and it would get a little messy when trying to share work with one another. Feature selection was also a challenging part of this project. We had so many features to choose from it was a little overwhelming. Finally, time, like always, was a challenge. We had to decide which things we should focus on for the project since there were so many ideas brought up about what we could do, so understanding our time constraints was challenging.

Some future ideas for this project would be to gather more data, the more data to run our models on the better. Also, the real end goal of a project like this would be to make a difference by preventing such events, so, if we had more time and resources, we could try to come up with a

way to use our data and models to show at-risk schools and hopefully prevent incidents surrounding school safety from happening.

Acknowledgements

We'd like to thank Dr. Christopher Briggs for the support, guidance, and instruction throughout this capstone project as well as the help and feedback received from our peers.

References

Abhishek Bagwan. (2023). *School dataset csv-file*. Kaggle.com. https://www.kaggle.com/datasets/abhishekbagwan/school-dataset/data

THE ASSOCIATED PRESS. (2022, May 25). *List of deadliest US school shootings*. AP News; AP News. https://apnews.com/article/list-of-deadliest-us-school-shootings-f25dad31e68c8acbdbcb952352df9249

Bankert, A. (2022, July 19). *School expert: We spend too much on "security theater."* NewsNation; NewsNation. https://www.newsnationnow.com/morninginamerica/answersforamerica/school-expert-school-safety-security-theater/

Glavin, C. (2018, July 26). *History of School Shootings in the United States | K12 Academics*. K12academics.com. https://www.k12academics.com/school-shootings/history-school-shootings-united-states

GPS Coordinates - Latitude and Longitude Finder. (2024). Gps-Coordinates.org. https://gps-coordinates.org/

National Center for Education Statistics (NCES) Home Page, part of the U.S. Department of Education. (2022). Ed.gov; National Center for Education Statistics. https://nces.ed.gov/

Riedman, David (2023). K-12 School Shooting Database

USAFacts. (2022, May 26). *The latest government data on school shootings*. USAFacts; USAFacts. https://usafacts.org/articles/the-latest-government-data-on-school-shootings/

Wikipedia Contributors. (2024, April 11). *Security theater*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Security_theater