PREDICTORS OF HEIGHTENED CASUALTY NUMBERS IN MODERN DAY MASS SHOOTING EVENTS (1982-2022)

By

Wesnahika Charles

A Thesis submitted to the faculty of the Wilkes Honors College

in partial fulfillment of the requirements of the degree of

Bachelor of Science in Biological and Physical Sciences

with a concentration in Data Analytics

Wilkes Honors College of Florida Atlantic University

Jupiter, Florida

May 2023

PREDICTORS OF HEIGHTENED CASUALTY NUMBERS IN MODERN MASS SHOOTING EVENTS (1982-2022)

by

Wesnahika Charles

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Bharat Bhushan Verma, and has been approved by the members of their supervisory committee. It was submitted to the faculty of The Wilkes Honors College and was accepted in partial fulfillment of the requirements for the degree of Bachelor of Science in Biological and Physical Sciences.

SUPERVISORY COMMITTEE:

_____

Dr. Bharat Bhushan Verma

_____

Dr. Yaouen Fily

_____

Dean Justin Perry, Wilkes Honors College

_____

Date

ABSTRACT

Author:	Wesnahika B. Charles
Title:	PREDICTORS OF HIGHTENDED CASUALTY NUMBERS IN MODERN MASS SHOOTING EVENTS (1982-2022)
Institution:	Wilkes Honors College of Florida Atlantic University
Thesis Advisor:	Dr. Bharat Bhushan Verma
Degree:	Bachelor of Science in Biological and Physical Sciences
Concentration:	Data Analytics
Year:	2023

Gun violence remains an extremely important topic in American society with many individuals and groups trying to find the root causes of these events. This thesis seeks to analyze the aftermath, more specifically the casualty count of a given mass shooting, and extract predictor variables that may contribute to the number of those who are killed and injured. With the use of exploratory data analysis and regression models, variables such as location, bullying, paranoia, suicidality, adult trauma, and physical abuse have been shown to have both a negative and positive effect on the casualty rate. This study provides background information about mass shootings, the experimental design of the study, exploratory and regression analysis, as well as performance analysis of the model.

To my hardworking

mother

Table of Contents

Chapter 1 : Introduction 1

Opening and Purpose 1

Literature Review 2

Methods 5

Chapter 2 : Methods 8

Data Source 8

Design 8

Exploratory Data Analysis 9

Chapter 3 : Analysis 16

Predictor Variables 16

Inferences 17

Performance 19

Chapter 4 : Results and Conclusion 23

References 25

List of Figures

Figure 2-1: Box plot showing all of the outliers, especially the extreme 2017 Las Vegas Shooting outlier. 10

Figure 2-2: The sum of the number of casualties per year along with the average and trend lines 11

Figure 2-3: Percentage of Casualties per Region 12

Figure 2-4: Density Map by Casualty in the United States 13

Figure 2-5: Casualty Count by Mental Illness types 14

Figure 3-1: Part of the model summary showing the p value of each variable instance. 17

Figure 3-2 Broom table showing AIC and BIC metrics 20

Figure 3-3: ANOVA test for the full model 21

Chapter 1 : Introduction

Opening and Purpose

Mass shootings are a form of violent homicide that's typically characterized by the killing of multiple individuals (usually 4) within a small period. These events can come about due to numerous factors, such as gang warfare, terror or hate group violence, armed robberies or familicides. These events can and have taken place in many locations ranging from office buildings, places of worship, music concerts, sporting events, malls, or shopping centers, as well as schools and universities. These shootings can have a devastating effect on the psyche of individuals, the community as well as society. In recent years, mass shootings and how we can prevent or predict them have been a polarizing topic amongst American scholars, politicians and the public. This is due to the upsurge of gun violence we are experiencing with time. This phenomenon has sparked debate about the lack of gun control, mental health and poverty which contributes to such an event and how much weight those variables have. Many statisticians, psychologists and data scientists have developed models and works to gather the common variables that are involved in mass shootings to predict the profile of a mass shooter or mass shooting event, however, understanding the complexity of these shootings which includes the aftermath, is imperative in addressing this convoluted issue that affects Americans every day. We will overview the problem of gun violence and set the stage for an in-depth analysis of the issue using statistical methods.

Literature Review

In this section, we will discuss similar studies or works that can provide background information to this topic. The study "A systematic review of the causes and prevention strategies in reducing gun violence in the United States", isolates the factors which contribute to a mass shooting to guide lawmakers in reducing and getting to the root cause of gun violence. This study involves utilizing reference articles and studies (including on-line news reports) by typing in key phrases relating to firearm deaths here in America[1]. The studies were incorporated if they contain mentions of mental illness, firearm accessibility, and violent behavioral patterns with a focus on bullying, childhood trauma, socio-economic status, domestic abuse as well as others. The results from this study concluded that gun-violence is an extremely multi-factored phenomenon. Access to firearms in a given state or country contributes to the amount of gun violence incidents, with states such as California having a significantly reduced number of shootings and casualties per capita due to the act of open carrying being outlawed, as well as finding that young people who had more access to firearms were more prone to being violent or aggressive [2]. Childhood domestic violence, as well as bullying and gang related activity heavily contributes to higher levels of aggressive behavior which in turn is associated with heightened levels of gun violence. Mental or behavioral issues such as having a conduct disorder, unique

[1] Sanchez, Carol, Daniella Jaguan, Saamia Shaikh, Mark McKenney, and Adel Elkbuli. 2020. Review of *A Systematic Review of the Causes and Prevention Strategies in Reducing Gun Violence in the United States*. *American Journal of Emergency Medicine* 38 (10): 2169–78. https://www.sciencedirect.com/science/article/pii/S0735675720305507.

[2] IBID

phobias, or drug use also contributed to the increase of gun violence, as the youth who suffer from those conditions are more likely to carry weapons, specifically firearms[3].

Another similar study[4] is "Youth Violence: What We Know and What We Need To Know". This study focuses on youth violence given the fact that a disproportionate amount of the violent gun crimes are committed by people aged 15-24 years old, with these numbers being higher in American youth as compared to those in other developed countries. Here the authors divide youth gun violence into two factors, being violent rampages and street shootings, to distinguish the proper predictor variables that may be different or the same for each given scenario. When it comes to characterizing the shooters who target schools in rampage shootings, they typically target areas that are suburban and low in crime, with about 78% of the shooters being socially marginalized, 61% having some sort of depression and 78% of them contemplating or even attempting suicide. Most of these mass school shooters had access and experience with weapons and often acquired these weapons from their family members. It is also noted due to the massive media coverage that these mass shootings typically get; it provides almost a sort of "formula" for future mass shooters to follow and enact it in the future. In the case of street shootings, they mostly occur in urban areas with high crime rates with the general population having a lower amount of trust for one another, these shootings are almost never spontaneous and happen mostly due to gang disputes. 47% of these shooters typically get their firearms through unorthodox means such as

[3] IBID

[4] Bushman, Brad, Sandra Calvert, Mark Dredze, Nina Jablonski, Calvin Morrill, Daniel Romer, Katherine Newman, et al. 2016. "Youth Violence: What We Know and What We Need to Know." *American Psychologist* 71 (1): 17–39.

black markets or obtaining them from a relative or friend (38%)[5]. Exposure to violent media also appears to be a causal factor when it comes to the likelihood of getting involved in aggressive behavior, as well as access to guns, bullying, mental illness and poverty.

Another study that is closely related and uses regression analysis is "Predictors of firearm violence in urban communities: A machine-learning approach"[6]. This study aims to extract predictive variables through algorithms such as random forest and LASSO. In this study, the random forest algorithm is used for prediction by using regression trees and bagging techniques, and LASSO is used to reduce multicollinearity and dimensionality. These techniques are used to further understand the inner workings of gun violence that frequently plagues crime ridden communities, particularly in the state of California[7]. Goin et al. completed this study by first collecting geographical data on the cases of community gun violence, from here they also collect covariates that may give insight to the communities themselves and what may be causal to the urban violence. These variables include income, marital status, poverty, employment status, etc. When analyzing the predictive variables, they are ranked in terms of "importance", this is from the change in value of the mean squared error, which is mainly used during random forest evaluation. Using 10-fold cross validation on LASSO and random forest machine learning

[5] IBID

[6] Goin, Dana, Kara Rudolph, and Jennifer Ahern. 2018. "Predictors of Firearm Violence in Urban Communities: A Machine-Learning Approach." *Health Place* 51 (May): 61–67. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5985152/pdf/nihms949347.pdf.

[7] IBID

models, they were able to examine several possible predictors. The results of this experiment showed that during the years 2007 and 2011, the bay area as well as Compton and Los Angeles had high rates of gun related violence. The top five covariates that had the most importance included the black isolation index (the extent to which black individuals are exposed to one another or another group), the black segregation index, the percentage of men who are 65 and older with a high school education and the percentage of those who've been married. Of the top eighteen variables, six were related to education, marriage status, location and temperature. These results show that there may be an intergenerational factor to gun violence, as well as location variables which can possibly play a role in heightened gun violence.

Methods

Since the main idea of this study is to uncover the variables that may affect casualty numbers, regression techniques such as multiple linear regression would allow for the training and testing of a multitude of variables. Linear regression falls under the classification of supervised learners[8]. These types of learners operate using data that was already previously labeled to build a predictive model. This regression technique can be extremely simple with the argument only taking one independent variable and one dependent variable (simple linear regression). When graphing the linear relationship between the dependent and independent variable, a straight line is shown which is the result of the model attempting to plot the line that fits the data points most

[8] Kassambara, Alboukadel. 2003. "Simple Linear Regression in R - Articles - STHDA." Www.sthda.com. October 3, 2003. http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/.

accurately. A formula is also used to find this line as well, where y-hat represents the dependent variable, b0 represents the constant, bi is the slope and xi is the independent variable.

$$\hat{y} = b0 + b1 * x$$

Through this linear regression we will attempt to find the most accurate values for both b0 and bi. When evaluating the cost function of linear regression, mean squared error is typically used, this helps evaluate the best fit line for the data points. The strength and performance metrics of any linear regression model would include R-squared, which showcases the amount of variation that exists within the model. Another metric would be the root mean squared error, which calculates how close the actual data points are to the predicted values. The Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC) all measure how accurately the model fits with the data while keeping track of the number of parameters. The ANOVA table can also show us an array of metrics as well as the p-values of the variables all together.

The model that will be used in this work will be multiple linear regression which is a variation of linear regression that aims to use two or more independent variables against a single dependent variable[9], this ultimately has the same characteristics and capabilities as simple linear regression with the formula being:

$$\hat{y} = b0 + b1 * x1 + ... + Bn * Xn$$

[9] ———. 2018. “Multiple Linear Regression in R - Articles - STHDA.” Www.sthda.com. October 3, 2018. http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

"Bn*Xn" is the coefficient of the final independent variable, with e representing the model's error. Just like simple linear regression one must consider the coefficients that can help reduce error, the t-statistic, as well as the p-value, which help us figure out if there is a significant relationship between the independent and dependent variables.

Chapter 2 : Methods

In this chapter, we will discuss the datasets, methodology and tools used for organizing and conducting this study. We will examine the data source that was chosen, how the experiment was designed based on our dataset, and the subsequent exploratory data analysis to make inferences about which variables may contribute to the casualty count.

Data Source

The data source that is being utilized is the Violence Project Database. This database[10] was compiled by The Violence Project founded by psychologist Dr. Jillian Peterson and sociologist Dr. James Densely. It is a nonprofit which aims to record, reduce and provide insight on gun violence in the United States. It holds over 180 variables and has recorded 185 mass shootings and is still updating with new variables and events periodically. The variables include the geographical data of the shootings, and the personal lives of the shooters, indicating factors such as mental illness, domestic violence, childhood trauma and more.

Design

This experiment was done mainly using the R language on Google Collab. Tidyverse and Hablar libraries were utilized to help organize the dataset and make visualizations and tables. For

[10] Peterson, Jillian, and James Densely. 2022. “The Violence Project.” https://www.theviolenceproject.org.

example, most of the variables in the dataset were “doubles", even though they are meant to be interpreted as number aliases or “characters”, the Hablar library helped convert said variables into characters to avoid confusion when doing exploratory or regression analysis. During the data exploration process, Tableau was used to create charts and graphs to help convey information and give insight to how certain variables may play a role in casualty numbers. Ggplot2 was also utilized in some of the exploratory data analysis to help create some of the graphs earlier in the process as well.

In the case of multiple linear regression, the built in regression function was used to build the model. To keep track of the performance and how well the model fits, the “broom” package was incorporated to monitor the AIC and BIC values of the model as I implemented different variables. Once my training model was created, the model was then used again on the test dataset to get a glimpse of how well the model would perform if given another set of data, I also kept note of the statistical significance of the variables that were being placed within the model, via the model summary.

Exploratory Data Analysis

Prior to creating the model, the data must first be explored to attempt to seek out patterns in the dataset. I had a rough idea about the variables to use in my model based on the reviewed studies and articles cited above. It was important to note that the dataset utilized in this study is completely different. One of the preliminary steps prior to any exploratory data analysis is to extract our period of interest so that we are only looking at the shootings that took place between

the year 1982 and 2022. This brought the number of instances down to 169. The first exploratory data analysis was filtering outliers. In this dataset, between the years of 1982 and 2022, the largest outlier was the 2017 Las Vegas shooting which had a casualty count of over 900. The source of this high casualty number can be explained with many different factors such as the shooter, Stephen Paddock, using a bump stock, which enabled a semi-automatic rifle to essentially replicate an automatic rifle by causing rounds to fire rapidly with the pull of one trigger. The ensuing panic, given the fact that the event happened during the Route 91 Harvest Music Festival, caused

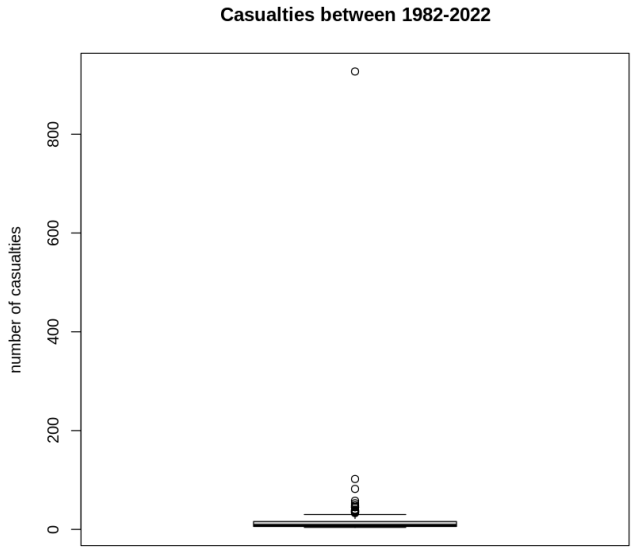

Figure 2-1: Box plot showing all the outliers, especially the extreme 2017 Las Vegas Shooting outlier.

trampling and stampeding. Paddock had no known motive for why he chose to do this act and this shooting only slightly surpassed the death toll for the 2017 Pulse Nightclub shooting in Orlando, Florida committed by Omar Mateen. Once we removed the outliers there were only 149 records.

The next exploratory analysis was looking at the number of casualties over time as well how frequent a given mass shooting is per year. On average, the sum of the casualties is about 35.15 per year and there seems to be a weak positive trend. There are massive dips in 2002 and 2020. For the frequency, once again there seems to be a weak positive correlation with its highest

point being 2021 and dips at point such as 1985-1987, 2002 and 2020, with 2020's frequency rate possibly being affected due to it corresponding with the height of the COVID-19 pandemic and its ensuing lockdowns, along with that the average frequency in this dataset is 3.61 shooting events per year.

When exploring some of the individual variables, we realize that there are indeed higher casualty and frequency rates when a shooter or shooting is positive for a specific trait or

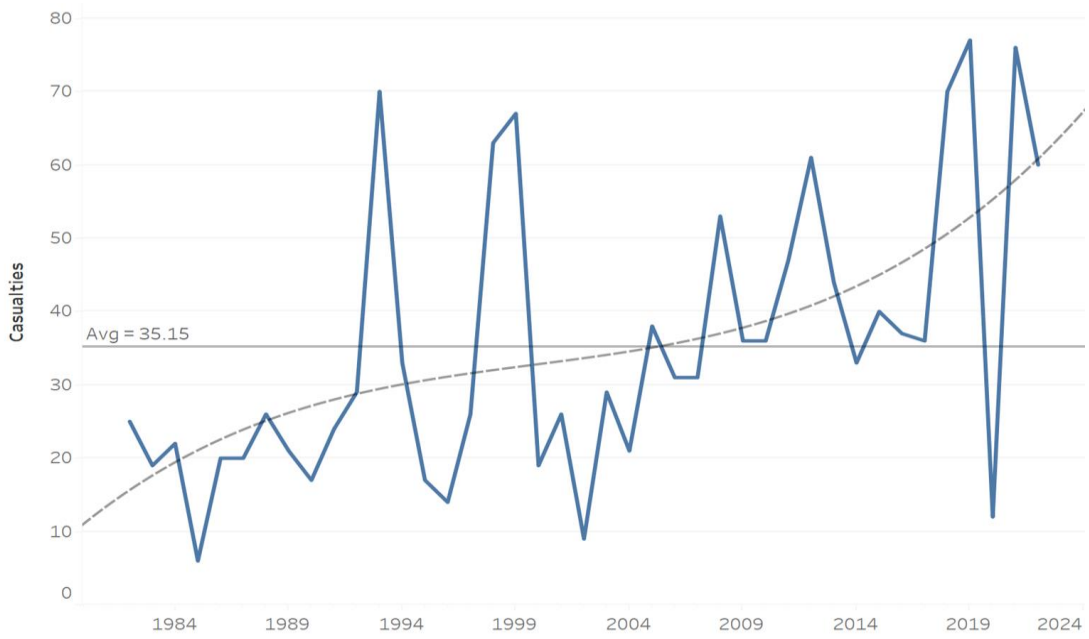

Figure 2-2: The sum of the number of casualties per year along with the average and trend lines

characteristic. When looking at external examples such as location, we begin to uncover patterns and trends between casualty count, frequency and area. For example, when examining the casualty count nationally, California and Texas have the highest number of casualties without accounting for population, and this also reflects when we look at the casualty numbers for regions with the south and the west accounting for more than half of the total casualty in the available dataset. The high casualty rate for the south

could be explained by the fact that the southern United States has the highest number of gun ownership by region, with about 36% of the residents owning at least one firearm[11], as compared to the northeast which is at just 16%. The southern United States also has some of the most unrestricted gun laws in the country with states such as Mississippi having only five gun provisions compared to Massachusetts having 100 enacted[12]. Looking at city wide numbers nationally, the mass shootings typically occur near high populated areas in the given state, for example, there exists larger densities in areas such as the zone that surrounds the Miami Metropolitan area,

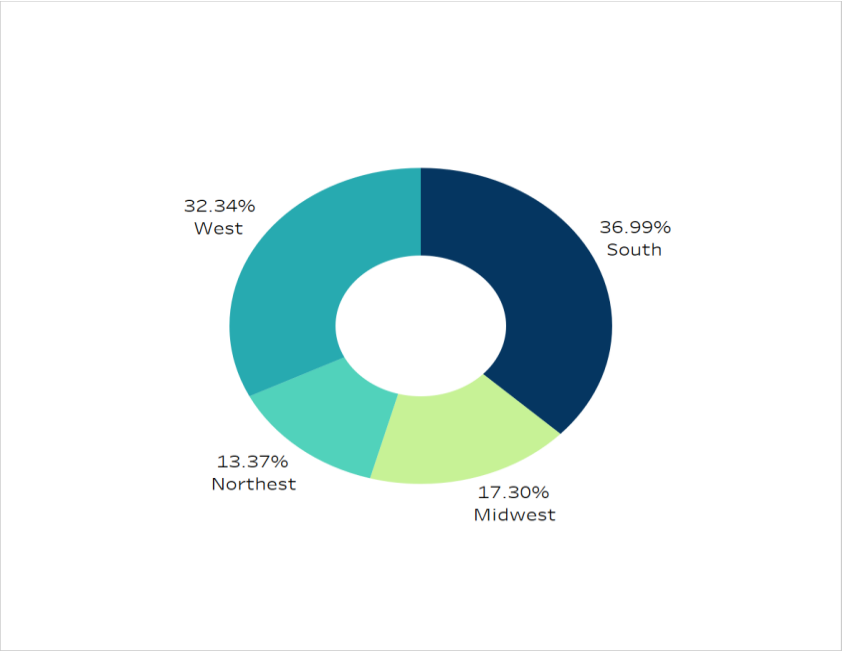

Figure 2-3: Percentage of Casualties per Region

Southern California, the area around the Bronx in New York, Dallas and Fort Worth, Texas, as

[11] "Gun Ownership by State 2020." 2022. Worldpopulationreview.com. 2022. https://worldpopulationreview.com/state-rankings/gun-ownership-by-state.

[12] McClenathan, Jane , Molly Pahn , Michael Siegal , Department of Community Health Services , and Boston University School of Public Health. 2016. "The Changing Landscape of U.S. Gun Policy: State Firearm Laws, 1991-2016." Edited by Rebecca Sherman .

well as Chicago, Illinois. These areas are also known to be high-crime areas, which may explain the frequency and severity of casualty cases. The specific locations of where a given mass shooting may occur can also have a varied range of casualty count as well, with specific locations such as retail spaces, warehouses or factories and K-12 schools being the most prominent spaces for mass shooting casualties.

Certain variables about the mass shooters themselves have been shown to also have a higher number of casualties when applicable. An instance of this would be how people between the ages of 24 and 54 make up 69% of the mass shooters and cause approximately 63% of the casualties listed. This could possibly be because at such a mature age, it would be easier to obtain a firearm as compared to those who fall below that age range. Interestingly enough, those who are between the ages of 55 and 64, only commit 4% of the mass shootings; however, this age group caused 30% of the mass shootings in this data set, making these two age groups the most dominant amongst perpetrators. When investigating mental illness it is noted that a large number of the

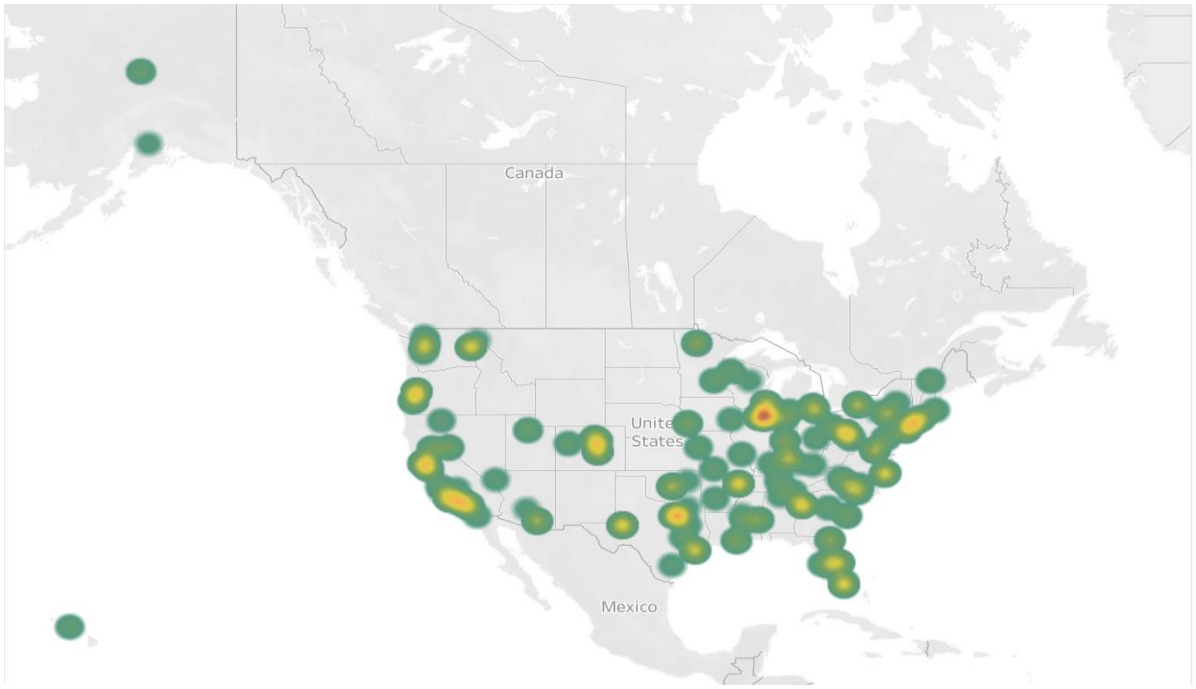

Figure 2-4: Density Map by Casualty in the United States

shooters did not have any recorded evidence of being diagnosed with any psychiatric or psychological problem at 32%, however, 14% recorded having a mood disorder (examples being bipolar or major depressive disorder), 9% recorded having multiple disorders (usually a mixture of thought or mood condition), 17% had a thought disorder (when someone has trouble creating intelligible sentences through writing or speech), 3% had an unspecified mental disorder and 23% had an indication of a mental illness but just wasn't formally diagnosed with similar numbers reflecting the casualty count.

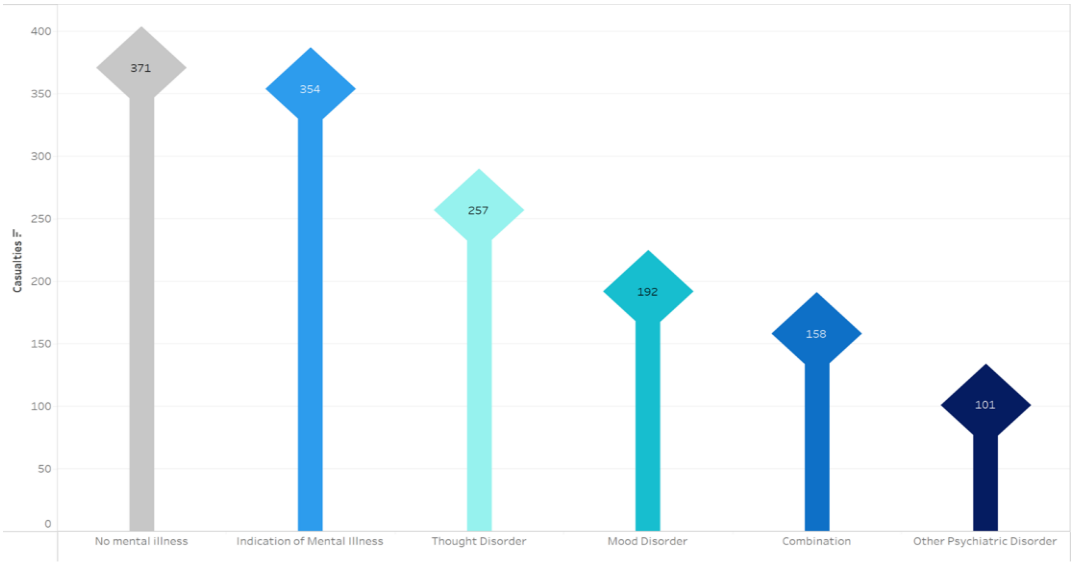

Figure 2-5: Casualty Count by Mental Illness types

Within the same vein of mental health, conditions such as paranoia or suicidality may also be discussed separately, as many of those who may have known the perpetrator may or not been aware they were plagued with such conditions prior to the event at hand and may have come to light via a manifesto or letter/note that was left behind after the fact. In this dataset, most of the mass shooters had an element of suicidality surrounding the event, with 31.08% being suicidal before or around the time of the shooting, and 36.49% who didn't show any signs of suicidality but planned on ending their lived during the event. This reflects the casualty statistics which shows that those who were suicidal caused 38.72% of the deaths and injuries, and those who

only planned on taking their own lives amounted to 33.10%. As mentioned before paranoia is another variable that can be used to measure the role that mental health plays in the event of a mass shooting. The role of paranoia when looking at the amount of those that displayed it seems low, as many of the perpetrators didn't display signs of paranoia at 72.97% and 65.51% of the casualties were committed by those who showed no signs of paranoia or paranoid activities or behavior, furthermore, this mirrors the casualty quantity with 65.88% of the casualties committed by those who didn't suffer from paranoia, and 34.12% resulting from those afflicted.

Chapter 3 : Analysis

This section summarizes the findings of the regression models, such as the viable variables that were uncovered, the relationship between the variables and the casualty count, as well as the performance for both the training and testing model.

Predictor Variables

Of the over 180 variables that are available in the dataset, only six have been chosen as viable predictors of casualty count due to their statistical significance at p-value (< 0.05). These variables include specific locations of the shootings, however as a way to uniformly gather the data for the source, the creator listed the different types of locations by number so in this instance locations such as government building (2), house of worship (3), retail (4), restaurant/bar/nightclub (5), office (6), residence (7), outdoors factory (9), and post office (10) were applicable as well as other variables which includes whether or not the perpetrator had any form of paranoia (1) during or around the shooting, if the shooter was suicidal (2), whether or not the shooter had adult trauma to which the different types include a death of a parent (1), the death of a child (2), the death of a family member (3), trauma from war (4), trauma from an accident (5), and unspecified (6), with those dealing with the death of a child, and the death of a family member being statistically significant. Other variables include the presence of physical abuse (1) in their life, as well as if they were bullied (1) in school. The overall P-value of the model is 1.19e-10, which means, as we

stated before, at least one of our variables is highly significant. The specific P-values of each significant variable are shown on figure 3-1.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           14.7492     1.7063   8.644 2.54e-14 ***
Paranoia1              4.3476     0.9443   4.604 1.02e-05 ***
Location1             -4.6802     2.5848  -1.811 0.072649 .
Location10            -7.6420     2.8792  -2.654 0.009008 **
Location2             -7.3820     2.2631  -3.262 0.001435 **
Location3             -8.3165     2.1485  -3.871 0.000176 ***
Location4             -9.9513     1.8194  -5.470 2.44e-07 ***
Location5            -10.6971     1.9451  -5.500 2.13e-07 ***
Location6             -9.6946     2.0021  -4.842 3.81e-06 ***
Location7             -9.9941     2.0211  -4.945 2.46e-06 ***
Location8             -6.2428     2.0888  -2.989 0.003388 **
Location9             -9.2593     1.8276  -5.066 1.46e-06 ***
Suicidality1           3.9087     1.0304   3.793 0.000233 ***
Suicidality2           1.4619     1.0050   1.455 0.148341
Adult_Trauma1          2.3760     1.4779   1.608 0.110494
Adult_Trauma2         -8.1874     3.4202  -2.394 0.018198 *
Adult_Trauma2 or more -5.8784     5.0705  -1.159 0.248586
Adult_Trauma3          7.8833     2.7343   2.883 0.004656 **
Adult_Trauma4         -3.6065     3.3463  -1.078 0.283279
Adult_Trauma5          0.2934     4.6887   0.063 0.950202
Adult_Trauma6          3.5584     3.3623   1.058 0.292005
Physically_Abused1    -4.3792     1.3853  -3.161 0.001982 **
Bullied1               3.8643     1.1800   3.275 0.001376 **
```

Figure 3-1: Part of the model summary showing the p value of each variable instance.

Inferences

As a result of this model, the regression equation is also used to look at how each variable contributes to the casualty count. The equation of the model is:

'casualties = 14.749 + 4.348 * Paranoia1 - 4.680 * Location1 - 7.642 * Location10 - 7.382 * Location2 - 8.317 * Location3 - 9.951 * Location4 - 10.697 * Location5 - 9.695 * Location6 - 9.994 * Location7 - 6.243 * Location8 - 9.259 * Location9 + 3.909 * Suicidality1 + 1.462 * Suicidality2 + 2.376 * Adult_Trauma1 - 8.187 * Adult_Trauma2 - 5.878 * Adult_Trauma2 or

more + 7.883 * Adult_Trauma3 - 3.606 * Adult_Trauma4 + 0.293 * Adult_Trauma5 + 3.558 * Adult_Trauma6 - 4.379 * Physically_Abused1 + 3.864 * Bullied1'

When looking at the equation, the location variable seems to always account for a loss in casualty points, with residence and restaurants, bars or night clubs having the lowest coefficient estimate variable amongst the statistically significant variables. Compared to other areas such as K-12 schools, colleges or universities and other recreational areas, places of residence may have a lower number of people around as residential casualties were amongst the lowest, and the nature of these mass shootings may be more of a personal altercation, for example, in Kentucky 2010, Stanley Neace shot his wife, her daughter, the daughter's boyfriend, and two other neighbors over an altercation with his wife, resulting in a total of 5 victims, not including himself[13]. Other variables such as paranoia can cause for a risen casualty count with many of those afflicted seeming to believe that there were people coming to harm or destroy their lives. For example, the 1999 Atlanta Day Trading Firm shootings occurred due to the perpetrator Mark Orrin Barton growing distrustful of his wife as well as many others around him and making it a mission to stop or kill the people who "sought for his destruction"[14]. Other variables such as bullying or physical abuse can be tied to suicidality or other mental health issues, hence why those variables can have a

[13] Press, The Associated. 2010. "Enraged over Breakfast, Kentucky Man Kills 5." *The New York Times,* September 12, 2010, sec. U.S. https://www.nytimes.com/2010/09/12/us/12kentucky.html.

[14] *The New York Times*. 1999. "SHOOTINGS in ATLANTA: THE NOTES; 'There Is No Reason for Me to Lie Now . . . '," July 31, 1999, sec. U.S. https://www.nytimes.com/1999/07/31/us/shootings-in-atlanta-the-notes-there-is-no-reason-for-me-to-lie-now.html.

heightened or lower effect on the quantity of those physically affected by mass shootings. Adult trauma, especially the death of family member is a variable that also looks to have an additive affect on the casualty rate, once again, possibly having ties to a weakened mental state.

Performance

Investigating the models includes looking at the different metrics that can be used to indicate performance. To individually look at the performance of each model, the R-squared value, also known as the coefficient of (multiple) determination is a sound metric which measures the scatter of regression points around a regression line, as well as the variation of the dependent variable by percentage[15]. The R-squared metric is characterized by the equation:

R^2 = 1 - RSS (variance explained by the model)/ TSS (total variance)

The value of R-squared is between 0%, meaning that none of the variation of the dependent variable can be explained by the variation of the independent variable(s), and 100% meaning that all variations of the dependent variable are able to be explained by the independent variable. What is considered a "good" R-squared value varies from the background of the model as well. In the model as seen in figure 3.5, the R-squared value is 50.64%, which means that more than half of

[15] Ballard, Callum. 2019. "An Ode to R-Squared." Medium. July 8, 2019. https://towardsdatascience.com/an-ode-to-r-squared-804d8d0ed22c.

the model's variation can be explained via the model's inputs. Since the topic of this paper is based on human behavior and activities, an R-squared of 50% or more is considered a relatively good model[16].

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.50645	0.4174492	4.517765	5.690398	1.19012e-10	22	-411.8869	871.7739	943.2155	2490.045	122	145

Figure 3-2 Broom table showing AIC and BIC metrics.

AIC/BIC compare the performance of different models. The AIC (Akaike's Information Criteria) metric essentially penalizes the addition of extra variables to the model and adds this penalty when the error increases as well. The BIC (Bayesian information criteria) metric is quite similar to the AIC however it incorporates a stronger penalty[17]. The model

[16] IBID

[17] Kassambara, Alboukadel . 2018b. "Regression Model Accuracy Metrics: R-Square, AIC, BIC, Cp and More - Articles - STHDA." Www.sthda.com. November 3, 2018. http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/#:~:text=AICc%20is%20a%20version%20of.

had an AIC of 871.77 and a BIC of 943.22. Another performance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Paranoia	1	373.1078	373.10783	18.280458	3.807315e-05
Location	10	1165.5657	116.55657	5.710701	5.429546e-07
Suicidality	2	273.9551	136.97753	6.711229	1.717187e-03
Adult_Trauma	7	328.6075	46.94393	2.300023	3.089118e-02
Physically_Abused	1	194.9819	194.98191	9.553159	2.473360e-03
Bullied	1	218.9099	218.90986	10.725512	1.375594e-03
Residuals	122	2490.0446	20.41020	NA	NA

Figure 3-3: ANOVA test for the full model

metric used was ANOVA, or Analysis of Variance, in figure 3-3, we are shown the overall p-values of the variables, F-value, Mean Squared and Sum Squared. Here we notice once again that our p-values are all lower than 0.05, thus making it significant. We also notice our F-values which is another way of showing whether the variation is caused by the independent variable, in this table, Paranoia, and Bullied have the highest F-values, thus making it least likely that these variations are due to chance. Mean squared error is another common performance metric to calculate an error value, it is another way to calculate how close a regression line is to the data

points; however, it works like a risk function coinciding with the value of the squared error loss[18].

The MSE is calculate as:

$$(1/n)\text{-}\sum (\text{actual} - \text{forecast})^{2}.$$

In this model the MSE was found to be 17.17.

[18] Gupta , Aryan . 2023. “Mean Squared Error : Overview, Examples, Concepts and More | Simplilearn.” Simplilearn.com. February 14, 2023. https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error.

Chapter 4 : Results and Conclusion

It is reiterated from this analysis that gun violence,mass shootings, are extremely complex and multifaceted with many variables in place to predict certain outcomes of these events. When it comes to casualty count, one can safely say that both external and internal factors of the mass shooting events can affect its number.

When investigating the internal variables, one can argue that mental health and the status of one's wellbeing can very well affect the quantity of those who are killed or injured. Looking at the overall mental health analysis, many mass shooters have had some sort of mental health disorder, specified or not, and those with said disorders cause more of a mortality rate than those who do not have any evidence of mental illness. As stated before, specific afflictions such as paranoia can cause someone to want to kill off as many individuals due to obsessive delusions about people who are close around them, and other factors such as suicidality, physical abuse, bullying and adult trauma can contribute positively to the casualty rate as well.

External variables such as location have a role in the quantity as well, with regional areas such as the south amounting to many mass shootings and people injured as well as densely populated cities across the United States especially those who have a higher crime index. When looking at the specific types of locations, it is shown that mass shootings that are committed in residential areas can lower the casualty rate as compared to those that were committed in universities or colleges, a possible reason being due to the volume and density of people typically at these areas.

The key takeaway from this study is that there is never solely one reason or cause for a mass shooting or how catastrophic the event may be as there are many variables at play that can change the magnitude of a given shooting.

References

Ballard, Callum. 2019. “An Ode to R-Squared.” Medium. July 8, 2019. https://towardsdatascience.com/an-ode-to-r-squared-804d8d0ed22c.

Bushman , Brad , Sandra Calvert, Mark Dredze , Nina Jablonski , Calvin Morrill , Daniel Romer , Katherine Newman , et al. 2016. “Youth Violence: What We Know and What We Need to Know.” *American Psychologist* 71 (1): 17–39.

Goin , Dana , Kara Rudolph, and Jennifer Ahern . 2018. “Predictors of Firearm Violence in Urban Communities: A Machine-Learning Approach.” *Health Place* 51 (May): 61–67. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5985152/pdf/nihms949347.pdf.

“Gun Ownership by State 2020.” 2022. Worldpopulationreview.com. 2022. https://worldpopulationreview.com/state-rankings/gun-ownership-by-state.

Gupta , Aryan . 2023. “Mean Squared Error : Overview, Examples, Concepts and More | Simplilearn.” Simplilearn.com. February 14, 2023. https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error.

Kassambara, Alboukadel. 2003. “Simple Linear Regression in R - Articles - STHDA.” Www.sthda.com. October 3, 2003. http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/.

Kassambara, Alboukadel . 2018b. “Regression Model Accuracy Metrics: R-Square, AIC, BIC, Cp and More - Articles - STHDA.” Www.sthda.com. November 3, 2018. http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/#:~:text=AICc%20is%20a%20version%20of.

McClenathan, Jane , Molly Pahn , Michael Siegal , Department of Community Health Services , and Boston University School of Public Health. 2016. “The Changing Landscape of U.S. Gun Policy: State Firearm Laws, 1991-2016.” Edited by Rebecca Sherman .

Peterson, Jillian , and James Densely . 2022. “The Violence Project.” https://www.theviolenceproject.org.

The Associated Press. 2010. “Enraged over Breakfast, Kentucky Man Kills 5.” *The New York Times*, September 12, 2010, sec. U.S. https://www.nytimes.com/2010/09/12/us/12kentucky.html.

Sanchez, Carol , Daniella Jaguan , Saamia Shaikh , Mark McKenney, and Adel Elkbuli . 2020. “A Systematic Review of the Causes and Prevention Strategies in Reducing Gun Violence in the United States.” *American Journal of Emergency Medicine* 38 (10): 2169–78. https://www.sciencedirect.com/science/article/pii/S0735675720305507.

The New York Times. 1999. “SHOOTINGS in ATLANTA: THE NOTES; 'There Is No Reason for Me to Lie Now . . . ',” July 31, 1999, sec. U.S. https://www.nytimes.com/1999/07/31/us/shootings-in-atlanta-the-notes-there-is-no-reason-for-me-to-lie-now.html.