Melissa Juarez

Data Analysis for the Social Sciences

Mike He

5 May 2023

Intervenable Events in School Shooting Victimhood Outcomes, from 1970-Present

Table of Contents

As I write this sentence, a school in my home state is under lockdown. Campbell Middle School is surrounded with heavy police presence, protecting against a local active shooter at large. Over the past few decades, school shootings have become an increasingly prevalent issue in the United States. Despite efforts to prevent these tragedies, they continue to occur with disturbing frequency, leaving in their wake devastating physical, emotional, and social impacts on the victims and their families.

This project seeks to understand the effects of intervenable events on school shooting victimhood outcomes. Using data obtained from the Center for Homeland Defense and Security, I analyze past instances of school shootings from 1970 to June 2022, limiting my analysis to single-shooter events. CHDS defines school shooting as any instance a gun "is brandished, is fired, or a bullet hits school property for any reason, regardless of the number of victims, time of day, or day of week."[1] With information about the incident, shooter, weapons, and victims, I look at two categories of intervenable events, gun behaviors and mental health behaviors, and establish relationships between multiple variables in each category and the outcome of victimhood.

Specifically, I consider the effects of multiple weapon use and weapon types used in a shooting incident on victimhood outcomes. Additionally, I consider the effects of a single shooter's history of being bullied and their sex on victimhood outcomes. Although I initially consider a Poisson regression model, I decide against it due to the lack of independence between the observed victim counts in one shooting incident. Thus, I use a categorical linear regression model, where most explanatory variables are binary and the response, measuring the victim count, is discrete.

I expect that incidents where multiple weapons are used will increase the expected number of victims in a shooting incident; I believe that the interaction between which weapons were used

[1] https://www.chds.us/sssc/data-map/

(whether rifle, shotgun, or handgun) and the quantity of weapons (whether single or multiple) will increase the expected victim count in significant ways, although it is currently unclear how. Additionally, I expect that a shooter having experienced bullying will increase the expected number of victims in an incident; I believe this relationship will be stronger when the shooter is male, rather than female[2].

I define an intervenable event as an action or behavior that has the potential to be prevented or modified through an intervention. Within the context of this project, I look at two categories of intervenable events: gun behaviors and mental health behaviors. The categories include events such as bullying history, accessing of weapons by minors, and other events that can be prevented or modified by interventions such as mental health screenings, background checks, gun restrictions, firearm safety trainings, etc. By focusing on intervenable events, the model results show how interventions in these areas might mitigate or prevent the effects on victims. They also allow us to direct the conversation to highlight actionable policies to do so.

Although I do include demographic control variables, my priority is to analyze them in a sociological context with respect to both mental health and gun behaviors, rather than on their own. From a sociological perspective, demographic variables are not simply innate characteristics of individuals, but rather, they are social constructs that are shaped by larger societal structures. So, while some might deduce identities such as race to a biological fact, we can also view them as a social construct that is informed by political and economic factors.

To put it another way: "there are no reasonable hypothetical interventions on race when race itself is the exposure,"[3]; viewing demographic factors through the broader social conditions

[2] The dataset includes 1 instance where the shooter is transgender. Because of the low number of these observations, the effect of transgender shooters on victimhood outcomes is unreliable and will not be included in my model.
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4125322/

that inform them is necessary. In doing so, we treat demographic behaviors not as simple individual-level characteristics, but as products of a larger societal mechanism.

Description of the Dataset

As previously mentioned, the dataset was collected from the Center for Homeland Defense and Security through their School Shooting Safety Compendium, an initiative that provides links, data, and resources to contribute to solutions for this string of tragedies. According to the CHDS's research methodology[4], their school shooting dataset was compiled using various sources including:

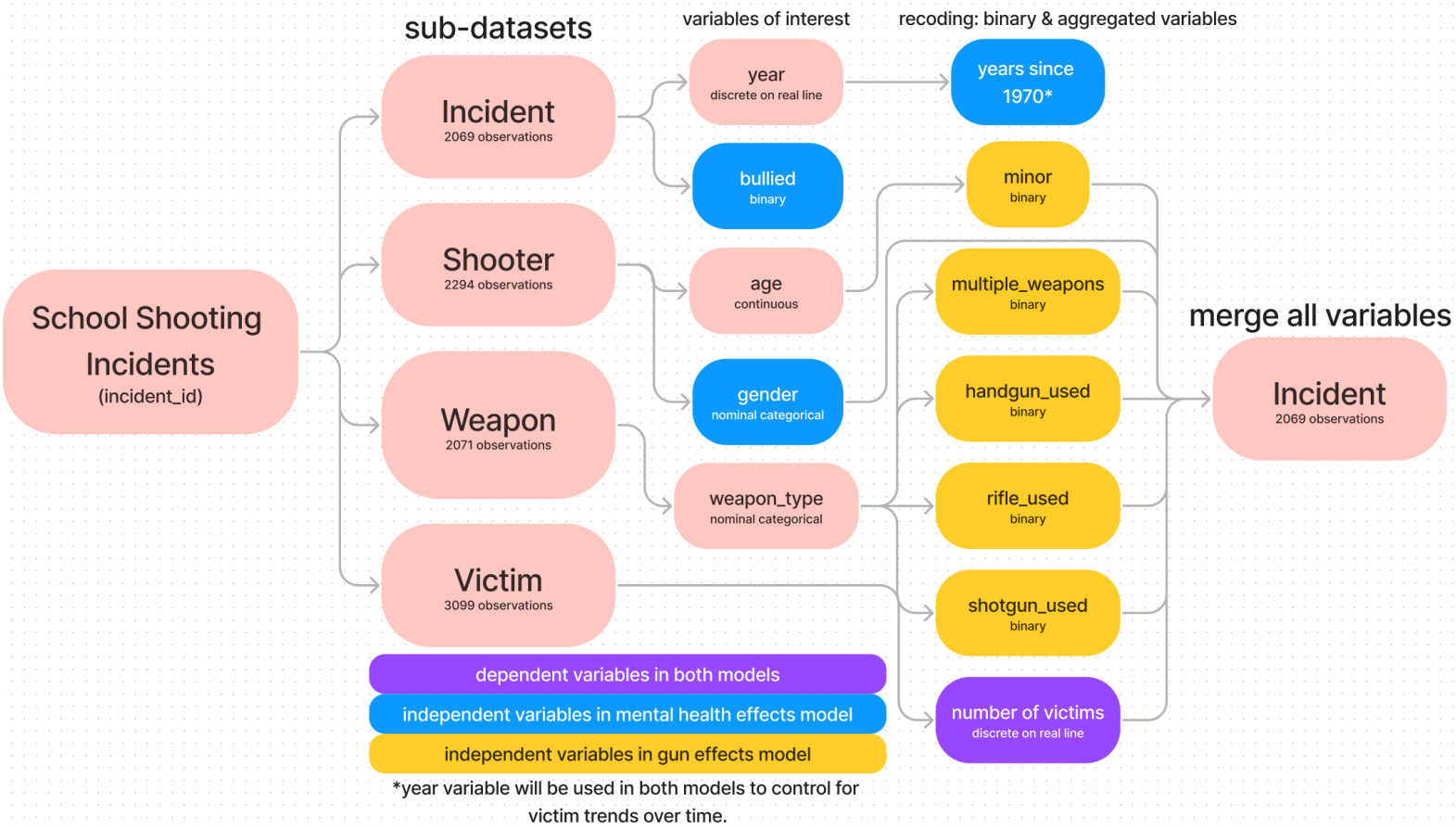
- government agencies such as the US Secret Service, FBI, and Department of Education.
- media or advocacy groups such as The Washington Post, CNN.
- websites or blogs including Wikipedia, schoolshootingdatabase.com, and schoolshootingtracker.com.

After reviewing the compiled data, the researchers deconflicted the data through cross referencing to avoid duplication. Although the dataset used extensive research and data collection, CHDS notes that the data is only as good as its sources. The many school shootings that do not gain media coverage or are not well and fully documented will forever be unknown to researchers and thus, they are not represented in this dataset. To combat this, police departments should develop better reporting protocols of shooting incidents, including investing in digitalization of paper records.

The entire dataset is comprised of 4 sub-datasets (Incident, Shooter, Weapon, Victim) which are related through a unique incident identification number given to each school shooting event.

[4] https://www.chds.us/sssc/methods/

The Incident dataset is aggregated at the level of unique shooting events; the Shooter dataset is aggregated at the level of unique shooters across all shootings. The Weapon and Victim datasets are aggregated in the same way. As each event can have multiple shooter, weapon, and victim counts, each dataset contains a different number of observations. The following diagram visualizes how each dataset is related and provides information on observations and variables of interest in this analysis.



I identified variables of interest by looking at how they related to gun and mental health behaviors and whether they had low rates of missingness. Visualizations of missingness[5] showed that the variables of interest had comparatively lower rates of missingness compared to other

[5] Appendix A

variables and would minimize the reduction in sample size when adding them to the regression models.

The diagram also illustrates the relationship between the sub-datasets' original variables and the recoding that took place to translate the shooter, weapon, and victim information to the incident-level dataset. For example, the Victim dataset had 3,099 observations corresponding to 3,099 victims across all shooting incidents. I grouped the victims by the unique incident identification numbers to find how many victims there were for each incident, which I then transferred to the Incident dataset. The same process was done for variables in the other datasets, Shooter and Weapon.

Some variables, such as `weapon_type`, were recorded into a group of binary variables that describe the quantity and type of weapons used at the incident. The reason for this recoding was due to the various categories of this variable: `Handgun, Rifle, Shotgun, Multiple Handguns, Multiple Rifles, Multiple Unknown, No Data, Other, Unknown.` Having over 6 categories as dummy variables would be hard to read in model output, and recoding the variable allowed for a deeper analysis over 2 effects (quantity and type) instead of just 1.

After filtering our dataset to only include single-shooter incidents[6], the sample size for our Incident data set, from which we will build our models, reduced to 1,869 observations of unique shootings.

[6] From now on, any mention of a 'shooting incident' is limited to single-shooter incidents.

Descriptive Statistics

Below are summary tables of variables used in the mental health and gun effects models.

Table 1 of Gun Model Variables

Gun Model Variables	
Variable	**N = 1,896**[1]
yr_since_1970	35 (15)
victim_count	1.41 (2.20)
multiple_weapons	
0	1,522 / 1,576 (97%)
1	54 / 1,576 (3.4%)
N/A	320
handgun_used	
0	254 / 1,576 (16%)
1	1,322 / 1,576 (84%)
N/A	320
rifle_used	
0	1,472 / 1,576 (93%)
1	104 / 1,576 (6.6%)
N/A	320
shotgun_used	
0	1,514 / 1,576 (96%)
1	62 / 1,576 (3.9%)
N/A	320
minor	
0	1,010 / 1,896 (53%)
1	886 / 1,896 (47%)

[1] Mean (SD); n / N (%)

We observe the following from the descriptive table:

- `yr_since_1970`: this is a continuous variable, describing the years passed since the start of the collection period. The center of distribution for the years for which shooting incidents occurred was 2005, 35 years after 1970.
- `victim_count`: the mean number of victims for a given shooting is 1.41, with standard deviation of 2.20.
- `multiple_weapons`: 3.4% of shooting instances involved a shooter using multiple weapons.
- `handgun_used`: 84% of shooting incidents involved use of a handgun.
- `rifle_used`: 6.6% of incidents involved use of a rifle.
- `shotgun_used`: 3.9% of incidents involved used of a shotgun.
- `minor`[7]: Almost **half** of all incidents were committed by children and teens under the age of 18.

[7] Includes shooters coded as ages 5 though 17, and those coded as `Children`, `Minors`, and `Teens`. The `Teens` category can technically include ages of 18 & 19, so the proportion of minors potentially overestimates the true parameter.

Table 1 of Mental Health Model Variables

We observe the following from this descriptive table, not mentioning repeat variables from the table above:

- `bullied`: 4.9% of incidents involved a shooter having been bullied by an intended victim, and the attack was not targeting victims at random.
- `sex`: Almost **all** instances of school shootings involved a male shooter, compared to 4.8% female shooters.

Mental Health Model Variables	
Variable	**N = 1,896**[1]
yr_since_1970	35 (15)
victim_count	1.41 (2.20)
bullied	
0	1,519 / 1,597 (95%)
1	78 / 1,597 (4.9%)
N/A	299
sex	
Female	75 / 1,571 (4.8%)
Male	1,496 / 1,571 (95%)
N/A	325

[1] Mean (SD); n / N (%)

On their own, the summary statistics already point to alarming trends among single-shooter school shooting incidents. A disproportionate number of shootings (95%) are committed by males, as opposed to females who commit less than 5% of school shootings. Additionally, 47% of shooters were minors, indicating that this large group of children illegally accessed weapon with intent.

We also identify that the most popular weapon of choice is a handgun (84%), by an incredible margin; handguns are followed by rifles (6.6%) and shotguns (3.9%). Most incidents involve the use of a single firearm, as opposed to multiple firearms.

From a sociological perspective, this initial screening of the data suggests that we must study the sociological factors that contribute to the alarming rate of male perpetrators of gun violence, given that school shootings are predominantly carried out by males, particularly young males who illegally access weapons. In research conducted by the Washington Post, Vanderbilt University psychiatrist Jonathan Metzl notes that in the context of young men not having fully

developed their prefrontal cortex—"which is critical to understanding the consequences of one's actions and controlling impulses"—a shooting "certainly feels like another kind of performance of young masculinity."[8] When young men experience sadness, loneliness, depression, and other destabilizing emotions, but cannot process them in healthy ways, they can resort to anger and aggression[9], the only societally accepted forms of masculine expression. In extreme cases, they commit violence against others as I study here.

Apart from research, this data also calls for public policy action to combat against irresponsible gun use. Specifically, there is a need for stricter regulations on firearm access for minors, particularly with regards to handguns. On a family level, gun owners need to practice safe storage protocols and always secure their firearms away from their children. Some states even impose criminal liability on parents whose firearms were accessed by minors due to negligent storage[10].

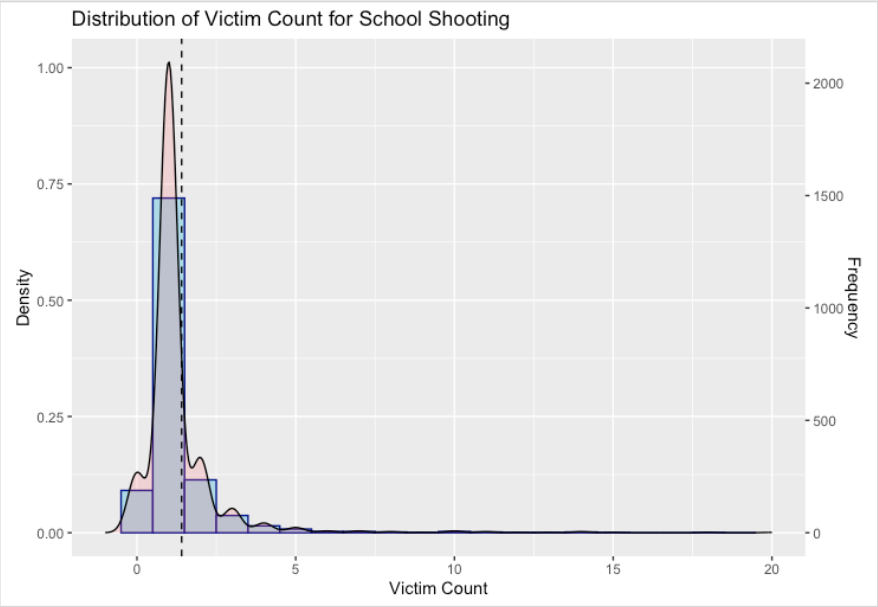
The following models examine the effects of gun and mental health behaviors on victimhood outcomes. First, I develop separate models for each behavior and provide a detailed interpretation with suggestions for model improvements. Finally, I present the final models for each behavior and provide an interpretation of the findings along with their implications.

Considering a Poisson Regression

My project seeks to understand how various intervenable events affect the outcome of victimhood, measured by the number of victims starting from 0. The distribution of the counts is displayed below. Initially, this seemed like a counting process that could be tackled through a

[8] https://www.washingtonpost.com/health/2022/06/03/why-so-many-mass-shooters-young-angry-men/
[9] https://sites.duke.edu/dukeidlab/files/2021/02/StanalandGaither2021_PSPB_PrePrint-1.pdf
[10] https://www.americanbar.org/groups/litigation/committees/childrens-rights/articles/2014/kids-and-gun-safety/

quasipoisson regression model. This is because according to our summary table, our mean for victim counts was 1.41 with a variance of 4.84; our model would be over-dispersed, which would be potentially mitigated through a quasipoisson regression.



However, our process violates one key assumption for Poisson regression: independence of events. In our case, the counts of victims in each incident are not independent of each other. For example, while the number of incidents themselves may be independently and identically distributed, the number of victims in given incident is not. This is because if there is already an existing victim, it is possible that the likelihood of subsequent victims is affected and enlarged.

Gun Effects on Victimhood Outcomes

Initial Models

In the preliminary model for the effect of gun behaviors on victim counts, I regressed victim counts on my primary independent variables, the types of weapons used:

$$victim\ count = \beta_0 + \beta_1 handgun\ use + \beta_2 rifle\ use + \beta_3 shotgun\ use + u$$

As shown above, weapon types are a categorical variable. After evaluating the linear regression model, we obtain the following OLS estimates:

$$\widehat{victim\ count} = 0.346 + 0.970\ handgun\ use + 2.976\ rifle\ use + 2.497\ shotgun\ use$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.3456     0.1975   1.750   0.0804 .
handgun_used   0.9697     0.2047   4.736 2.37e-06 ***
rifle_used     2.9762     0.2756  10.800  < 2e-16 ***
shotgun_used   2.4971     0.3349   7.456 1.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.263 on 1572 degrees of freedom
  (320 observations deleted due to missingness)
Multiple R-squared:  0.08899,   Adjusted R-squared:  0.08725
F-statistic: 51.18 on 3 and 1572 DF,  p-value: < 2.2e-16
```

Interpretation:

- 0.346 victims are expected for a shooting incident where the weapon used by the shooter is of an unspecified type.
- The use of handgun in a shooting incident increases the expected number of victims by 0.970, holding constant whether the shooter also used a rifle or a shotgun.
- The use of rifle in a shooting incident increases the expected number of victims by 2.976, holding constant whether the shooter also used a handgun or a shotgun.
- The use of shotgun in a shooting incident increases the expected number of victims by 2.497, holding constant whether the shooter also used a rifle or a handgun.
- 8.725% of the variation in the number of victims can be explained by the linear relationship with weapon type used.

This linear model shows us that out of the 3 weapon types, rifles are associated with the largest victimhood outcomes, followed by shotguns and handguns. Additionally, we see that all weapon types have a significant effect on the expected number of victims in a shooting incident.

However, the model does not consider the effects of weapon quantity (whether single or multiple) on victim count. I presume that the more weapons available at the incident, the more people might be hurt; considering the weapon quantity is necessary to better understand to what degree victims are impacted by gun use. Additionally, it does not control for the age of the shooter or the year[11] at which an incident occurred. On one hand, children in possession of weapons may be more irresponsible with gun use which could lead to more injuries and deaths. It is also possible that there may be trends over time for victimhood outcomes, whether due to new gun technologies and restrictions on gun access policies.

In the second iteration of the model, I incorporate weapon quantity, year, and minor status of the shooter and obtain the following outcome.

$$\widehat{victim\ count} = 0.663 + 0.749\ handgun\ use + 2.014\ rifle\ use + 1.849\ shotgun\ use$$
$$+ 4.527\ multiple\ weapons - 0.049\ minor - 0.005\ years\ since\ 1970$$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.662705   0.241539   2.744 0.006145 **
handgun_used      0.748625   0.192895   3.881 0.000108 ***
rifle_used        2.014069   0.267495   7.529 8.56e-14 ***
shotgun_used      1.848599   0.317910   5.815 7.34e-09 ***
multiple_weapons  4.526508   0.304567  14.862  < 2e-16 ***
minor            -0.048817   0.109420  -0.446 0.655553
yr_since_1970    -0.004976   0.003553  -1.401 0.161518
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.12 on 1569 degrees of freedom
  (493 observations deleted due to missingness)
Multiple R-squared:  0.2017,    Adjusted R-squared:  0.1987
F-statistic: 66.08 on 6 and 1569 DF,  p-value: < 2.2e-16
```

The regressor coefficients `minor` and `yr_since_1970` were not significant, but all weapon type variables and `multiple_weapons` were statistically significant. Further, we see that weapon quantity (`multiple_weapons`) has the strongest effect on the victim count;

[11] In my analyses, I control for time using years relative to the starting year of data collected, 1970. This is so that the scale of our model is more easily interpretable.

when multiple weapons are used in a shooting incident, we can expect an increase of 4.527 victims, all other variables held constant. Additionally, adding the new variables increases the explained variation by over 10%. Our adjusted R-squared tells is that in the model considering weapon type and weapon quantity, 19.87% of variation in the number of victims can be explained by this linear relationship.

We also see that the introduction of new variables decreases the effect of rifle and handgun use on expected victim count outcomes, more than the decrease in the effect of handgun use. This is because those weapon type variables are correlated with the weapon quantity variables. In my final models section, I run a diagnostic check to see whether any of my final variables are collinear or highly correlated. Additionally, I remove the non-significant regressors.

The last change to my gun model is to look at the potential for interaction between the weapon types and the weapon quantity. Is it possible that the weapon quantity amplifies (or diminishes) the effects of weapon on victim count outcomes?

Final Models

After checking for interaction and removing non-significant variables, my final model is

$$\begin{aligned}\widehat{victim\ count} = \ & 1.239 + 3.021\ multiple\ weapons + \ 0.666\ rifle\ used \\ & + 5.513\ multiple\ weapons * rifle\ used \\ & - 1.511\ multiple\ weapons * handgun\ used \\ & + 5.970\ multiple\ weapons * shotgun\ used\end{aligned}$$

This interaction model includes the main effects of `multiple_weapons` and `rifle_used` and the interaction effects between `multiple_weapons` and all weapon types. After accounting for the interaction effect between weapon types and weapon quantity, the main effect of both handgun and

shotgun use became non-significant. In other words, the effects of shotgun and handgun use only impact expected victim count when they are among the multiple weapons used, rather than when used as the sole weapon.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.23853    0.05407  22.905  < 2e-16 ***
multiple_weapons               3.02110    0.49511   6.102 1.32e-09 ***
rifle_used                     0.66624    0.23017   2.895  0.00385 **
multiple_weapons:rifle_used    5.51332    0.63574   8.672  < 2e-16 ***
multiple_weapons:handgun_used -1.51089    0.58889  -2.566  0.01039 *
multiple_weapons:shotgun_used  5.96990    0.79071   7.550 7.34e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.051 on 1570 degrees of freedom
Multiple R-squared:  0.2528,    Adjusted R-squared:  0.2504
F-statistic: 106.2 on 5 and 1570 DF,  p-value: < 2.2e-16
```

The output of the final linear model can be interpreted as follows:

- When a single, unknown weapon is used, 1.239 victims are expected in a shooting incident.
- When multiple weapons are used, an increase of 3.021 victims is expected in a shooting incident, controlling for all other variables. This effect:
    - Increases by an expected 5.513 victims when a rifle is among the weapons used, controlling for other variables.
    - Decreases by an expected 1.511 victims when a handgun is among the weapons used, controlling for other variables.
    - Increases by an expected 5.970 victims when a shotgun is among the weapons used, controlling for other variables.
- When a rifle is used, an increase of 0.666 victims is expected in a shooting incident, controlling for all other variables. This effect:
    - Increases by an expected 5.513 victims when the rifle was among multiple weapons used, controlling for other variables.
- This model explains 25% of the variation in victim counts.

This model tells us that (1) multiple weapons are associated with an increase in victim count outcomes and (2) rifles and shotguns have the strongest association with increases in victim counts, when compared to handguns. Additionally, when multiple weapons are in use, shotguns are the weapon associated with the highest victim counts.

Specifically, for a single-shooter shooting where multiple weapons are used, having a shotgun increases the expected victim count by 0.46 more victims than when having a rifle and by 7.48 more victims than when having a handgun. However, when multiple weapon types are used to together, the impacts are far greater. In a worst-case scenario, when shotguns and rifles are present together, the model expects there to be around 16 victims. When looking back at our original dataset, incidents when shotguns and rifles were used had an average of 14 victims per shooting.

To identify this model as the stronger model compared to previous iterations, I used ANOVA to compare the fit between the main effects and interaction models. My first model contains no interaction:

$$victim\ count = \beta_0 + \beta_1\ handgun\ use + \beta_2\ rifle\ use + \beta_3\ shotgun\ use + \beta_4\ multiple\ weapons + u$$

My second model contained main effects of weapon type and quantity and the interactions between them:

$$\begin{aligned} victim\ count = {} & \beta_0 + \beta_1\ handgun\ use + \beta_2\ rifle\ use + \beta_3\ shotgun\ use + \beta_4\ multiple\ weapons \\ & + \beta_5\ handgun\ use * multiple\ weapons + \beta_6 rifle\ use * multiple\ weapons \\ & + \beta_7 shotgun\ use * multiple\ weapons + u \end{aligned}$$

I used an ANOVA test to compare the fit between the simpler main effects model and the more complex interaction model. If the p-value form the test is less than 0.05, we can conclude that the more complex model is a significantly better fit for our data.

```
Analysis of Variance Table

Model 1: victim_count ~ handgun_used + rifle_used + shotgun_used + multiple_weapons
Model 2: victim_count ~ handgun_used * multiple_weapons + rifle_used *
    multiple_weapons + shotgun_used * multiple_weapons
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1571 7061.4
2   1568 6585.9  3     475.5 37.736 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

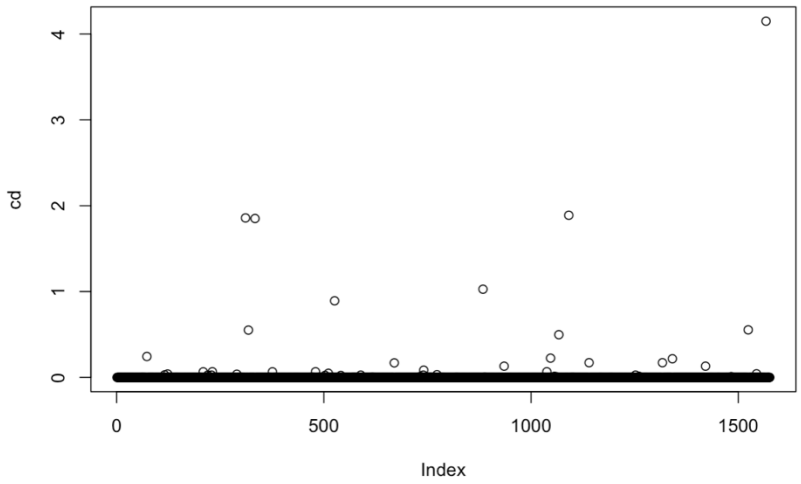
As shown above, the output form the test showed that there was a significant difference between the fits of the two models, favoring the more complex one. Thus, I decided to keep the interaction terms in my final model. As previously mentioned, the interaction model resulted in the handgun and shotgun main effects being non-significant, so I removed them from the final model as shown at the beginning of this section.

Diagnostic Tests

Outliers

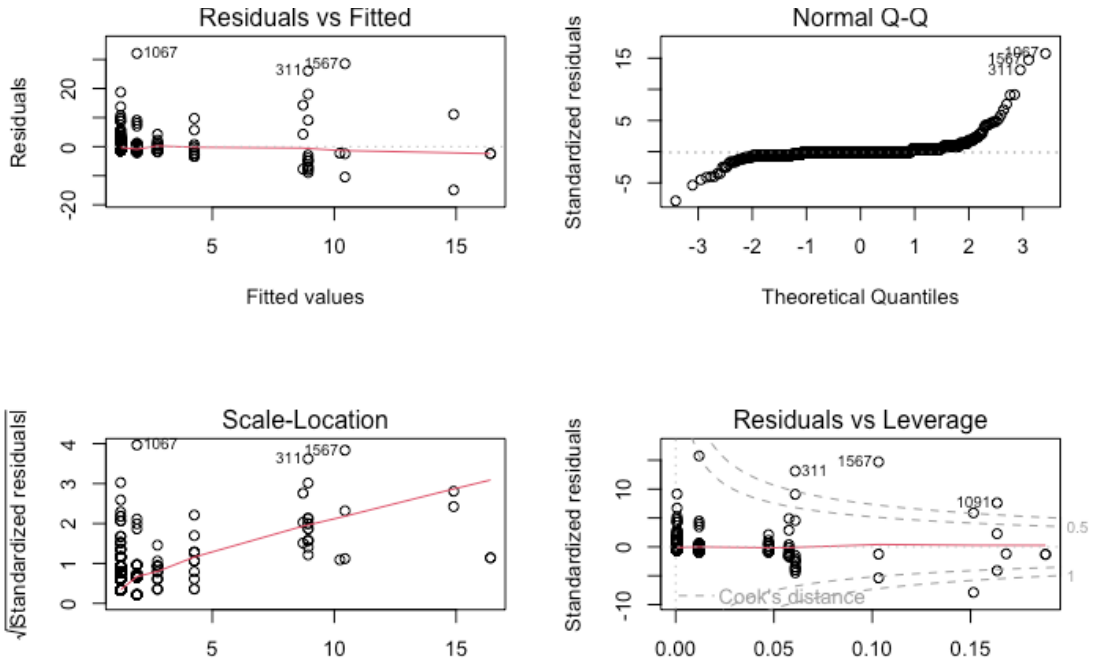
As shown in the distribution of victim counts on page 10, most victims counts are below 5. However, there are various shooting incidents that result in extreme victim counts; the largest count in the dataset is 39, representing the Robb Elementary School Shooting in Uvalde, TX in 2022. To understand the impact of high-victim school shootings, I visualized the Cook’s distances of each of my points and ran an outlier test.

Row.names	victim_count	multiple_weapons	rifle_used	handgun_used	shotgun_used	rstudent	p	bonf.p
1061	20	0	0	1	0	9.404296	1.789582e-20	2.820381e-17
1067	34	0	1	0	0	17.154082	1.379039e-60	2.173366e-57
1091	23	1	0	1	1	7.757202	1.552699e-14	2.447053e-11
1524	0	1	1	0	0	-5.424456	6.723044e-08	1.059552e-04
1567	39	1	1	0	0	15.835065	1.659077e-52	2.614706e-49
311	35	1	1	1	0	13.899893	1.645889e-41	2.593921e-38
334	0	1	1	1	1	-8.046051	1.671140e-15	2.633716e-12
526	27	1	1	1	0	9.340609	3.169339e-20	4.994878e-17
572	15	0	0	1	0	6.809919	1.386464e-11	2.185067e-08
884	26	1	1	1	1	5.941473	3.473062e-09	5.473546e-06



We identify 10 outliers among our sample of 1,576 incidents used in our target model. Most of them seem to be flagged due to their extremely high victim count, or because of low victim count despite having multiple weapons at the incident.

The plot below reiterates this information. Specifically, we see that the Normal QQ plot has points that "fall along a line in the middle of the graph but curve off in the extremities"[12]. This means that our data has "more extreme values than would be expected if they truly came from a normal distribution" and therefore violates the normality assumption.



[12] https://data.library.virginia.edu/understanding-q-q-plots/

Although the inclusion of these outliers has a considerable effect on our model, their presence reflects a true event rather than a mistake. Therefore, I would not remove them from the model even if it improves the model fit. Instead, this suggests a need to analyze mass or high-victim shootings separately to understand the factors that influence this outcome. It also points to using a different model type to analyze the relationships between the variables in question.

Collinearity

```
GVIFs computed for predictors
                    GVIF Df GVIF^(1/(2*Df))                        Interacts With          Other Predictors
multiple_weapons 1.000000  5        1.000000 rifle_used, handgun_used, shotgun_used                       --
rifle_used       2.959706  3        1.198233                       multiple_weapons handgun_used, shotgun_used
handgun_used     1.742153  2        1.148872                          handgun_used   rifle_used, shotgun_used
shotgun_used     2.951366  2        1.310707                          shotgun_used   rifle_used, handgun_used
```

To detect multicollinearity between the variables, I used a Variance Inflation Factor to see whether any of the independent variables could be explained by other independent variables in the model. Using the `car::vif()` function for my final model, the output did not alert to any high VIF values. This means that there were no variables in the dataset that could be predicted by others.

Mental Health Effects on Victimhood Outcomes

In this section, I evaluate the effects of mental health behaviors on victim count outcomes. I use the same process as described with the gun behaviors model, meaning that this section will be shorter in terms of method reasoning since it was already previously described.

Initial Models

In the preliminary model for the effect of mental health behaviors on victim counts, I regressed victim counts on my primary independent variable, `bullied`:

$$victim\ count = \beta_0 + \beta_1 bullied$$

`bullied` is binary variable where 1 indicates that the shooter was bullied by a victim, and that the attack was not targeting victims at random. After evaluating the model using linear regression, we obtain the following OLS estimates:

$$\widehat{victim\ count} = 1.293 + 1.284\ bullied$$

The following model can be interpreted as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.29296    0.04556  28.377  < 2e-16 ***
bullied      1.28397    0.20617   6.228 6.04e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.776 on 1595 degrees of freedom
  (299 observations deleted due to missingness)
Multiple R-squared:  0.02374,   Adjusted R-squared:  0.02313
F-statistic: 38.79 on 1 and 1595 DF,  p-value: 6.035e-10
```

- When shooters are not bullied, shooting incidents are expected to result in 1.293 victims.
- When shooters are bullied, there is an expected increase of 1.284 victims.
- Only 2.313% of the variation in victim count outcomes can be explained by the linear relationship with bullying history.

The linear model surprisingly shows that when shooters are bullied by a victim and thus had a specific target, the shooting incident is expected to result in more victims compared to if the

shooting was at random. Specifically, non-bullied shooters are expected to physically harm 1.293 victims whereas bullied shooters are expected to physically harm 2.577 victims. This implies that some sort of targeted emotional hurt on the shooter's part relates to an increased threat beyond the targeted victims themselves.

However, the model does not consider the effects that gender have on victim count. Previously, we mentioned that almost all school shootings are carried out by males, and I wonder if gender plays a part in predicting the quantity of the victims in a shooting. Additionally, as with the gun behavior model, this model does not control for the year[13] at which an incident occurred. It is possible that there may be trends over time for victimhood outcomes, as society responds to the mental health needs of their community.

In the second iteration of the model, I incorporate gender and year and obtain the following outcome.

$$\widehat{victim\ count} = 1.639 + 1.219\ bullied - 0.009\ years\ since\ 1970 - 0.111\ sex$$

The model can be interpreted as follows:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.639041   0.127690  12.836  < 2e-16 ***
bullied        1.219187   0.220590   5.527 3.92e-08 ***
yr_since_1970 -0.008641   0.003408  -2.535   0.0113 *
sex           -0.110542   0.233609  -0.473   0.6362
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 1327 degrees of freedom
  (738 observations deleted due to missingness)
Multiple R-squared:  0.0279,    Adjusted R-squared:  0.0257
F-statistic: 12.69 on 3 and 1327 DF,  p-value: 3.498e-08
```

-At a baseline, when shooters are not bullied, are male, and the incident occurs in 1970, there is an expected 1.639 victims for the shooting incident.

- When shooters are bullied by the victims, there is an expected increase of 1.219 victims in the shooting incident, controlling for year and sex.

[13] In my analyses, I control for time using years relative to the starting year of data collected, 1970. This is so that the scale of our model is more easily interpretable.

- For every year after 1970, there is an expected decrease of 0.009 victims in the shooting incident, controlling for bullying and sex.
- When a shooter is female, there is an expected decrease of 0.111 victims (n.s.) in the shooting incident, controlling for year and bullying.
- When including year and sex information, this linear model explains 2.6% of variation in victim outcome.

We see that the effect of the gender of the shooter, when controlling for bullying and year, is not significant. Thus, we can remove it from our model. While the effect is extremely small, we see that over time, victims counts per shooting incident have decreased by 0.009 every year after 1970, holding other variables constant.

The last change to my mental health model is to look at the potential for interaction between bullying and year. Is it possible that the bullying is experienced more or less severely over time? When running the interaction model, I found that both the interaction effect and the main bullying effect became insignificant, while the effect of years since 1970 became more significant than before. Although confusing, this suggests that the interaction is not an important predictor for victim count, and we should move forward without it. Still, in the following section I show the ANOVA comparison between the fits of the main effects and interaction model.

Final Models

My final model is the main effects model of year and bullying on victim count outcomes.

$$\widehat{victim\ count} = 1.688 + 1.244\ bullied - 0.011\ years\ since\ 1970$$

In this model, we can predict victim outcome in the following way:

- At a baseline, when shooters in are not bullied and the incident occurs in 1970, there is an expected 1.688 victims for the shooting incident.
- When the shooter is bullied by the victims they are targeting, there is an expected increase of 1.244 victims in the shooting incident, controlling for year.
- For every year after 1970, there is an expected decrease of 0.011 victims in the shooting incident, controlling for bullying.
- This linear model explains 3.14% of variation in victim outcome.

This model tells us that shooters who experienced bullying are associated with greater victim counts, despite them not targeting victims at random. While this model does not explain causality, it would be interesting to explore the conditions that differ between bullied and non-bullied shooters that lead to the statistical difference in victimhood outcomes. I wonder whether this could be explained by identity-based violence; for example, if a shooter is bullied by a person of a specific identity, are they more likely to carry out attacks against other people who fit those descriptions?

To identify this model as the stronger model compared to previous iterations, I used ANOVA to compare the fit between the main effects and interaction models. My first model contains no interaction:

$$victim\ count = \beta_0 + \beta_1\ bullied + \beta_2 years\ since\ 1970 + u$$

My second model contained main effects of bullying and year and the interactions between them:

$$victim\ count = \beta_0 + \beta_0 + \beta_1\ bullied + \beta_2 years\ since\ 1970 + \beta_3 bullied * years\ since\ 1970 + u$$

I used an ANOVA test to compare the fit between the simpler main effects model and the more complex interaction model. If the p-value form the test is less than 0.05, we can conclude that the more complex model is a significantly better fit for our data.

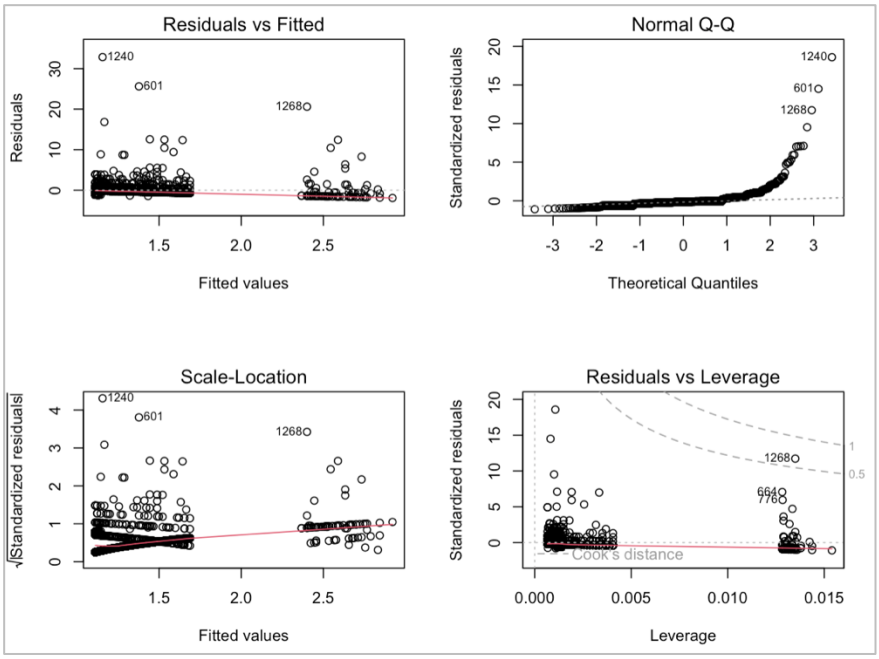
```
Analysis of Variance Table

Model 1: victim_count ~ bullied + yr_since_1970
Model 2: victim_count ~ bullied * yr_since_1970
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1594 4983.8
2   1593 4975.7  1    8.1282 2.6023 0.1069
```

The results of the test show that there is not enough statistical evidence to say that the complex model has a better fit compared to the main effects model. Therefore, I chose the main effects model for the final model.

Diagnostic Tests: Outliers

As in the previous model, I look at the Cook's distances of each observation in our model and did an outlier test. Compared to the gun behaviors model, there are less points with large Cook's distances. Additionally, most of outliers tend to be low victim count incidents from over 30 years ago.

Row.names	victim_count	bullied	yr_since_1970
1215	1	0	50
1240	1	0	50
1268	3	0	50
231	1	0	16
238	4	0	17
415	1	0	24
601	1	0	33
664	1	0	35
776	1	0	37
83	6	0	5



Again, the normal QQ plot suggests that our data violates the normality assumption due to the points falling along the line and curving off at the extremities.

Conclusion and Implications

My goal through this analysis was to look at the relationships between multiple variables related to mental health behaviors and gun behaviors and the outcome of victimhood. I initially posited that multiple weapons would have an increased effect on victim count outcomes, and the analysis provided further clarity onto how the interactions between weapon type and quantity play out. The analysis found that multiple weapons and the use of shotguns and rifles are strongly associated with an increase in victim count outcomes. Specifically, in cases where multiple weapons are used, having a shotgun increases the expected victim count more than having a rifle or a handgun.

Additionally, I believed that the mental health effects model would show a positive association between bullying history and victimhood outcomes in addition to a significant interaction effect between gender and bullying. The analysis of the mental health model suggests that bullying is associated with greater victim counts; there was also a weak, negative association with victim counts and the passing years. However, there was no significant interaction effect between bullying and sex. Although most of the school shootings in our dataset were carried out by males, there is no indication that gender plays a role into how many victims are expected from a shooting incident.

One possible explanation, and limitation, is that there were very few observations of female shooters and very few mental health related variables in the dataset. Overall, the mental health model was weaker in comparison to the gun effects model, just by looking at the difference between R-squared values. If our dataset had included more information on mental health records of shooters, prior incidents, or any other contextualization of the lives of the shooter, the mental health model would have a broader scope of what mental health encompasses. Currently, I am not

confident that I can call my model a "mental health model" as the variables having to do with mental health are extremely limited. And as it currently stands, the model does not provide any reliable insight on gender-specific mental health behavior and outcomes, which was a key initial interest in this project.

It's important to note that both models do not explain causality, although they do offer important insights for possible school shooting interventions:

1. Weapon Control Policies: Because handguns are the most common weapon in school shootings and mass shootings in general[14], they are often posited as the most dangerous weapon. However, this model makes clear the devastating effects of shotgun and rifle use, which have the greatest impact on increasing victim counts. Governments should be creating stricter gun policies to limit who can get access to these weapons, and those deemed responsible enough to own then should be held accountable for safe weapon storage.
2. Mental Health and Education Programs: Our analysis showed that a shooter's bullying history has a significant association with greater victim counts; in other words, a shooting where the shooter was bullied by a targeted victim is expected to have more victims that a shooting where the shooter was not bullied by a targeted victim. Schools can intervene on bullying though social and emotional learning programs[15] and anti-bullying programs. Additionally, schools can hire more mental health counselors and case workers to help students experiencing mental health issues and identify and intervene on behaviors that could lead to violence.

[14] https://www.statista.com/statistics/476409/mass-shootings-in-the-us-by-weapon-types-used/
[15] https://doi.org/10.1111/j.1467-8624.2010.01564.x

3. Expanded Research: Finally, while significant gender differences were not present in the model, the fact that 95% of shooting were committed by males suggests that there is more to be researched about how young boys are socialized and how they learn to navigate complex emotions.

Future Directions

For an expanded and more comprehensive study on gun behaviors, data collection could be expanded to include governance related variables that detail the local and state gun restrictions at the time of the incident. An analysis on this data, specifically how gun restrictions impact victim counts, shooting frequency, etc., would provide useful information on how government policies play a role in mitigating the effects and occurrences of gun violence in schools.

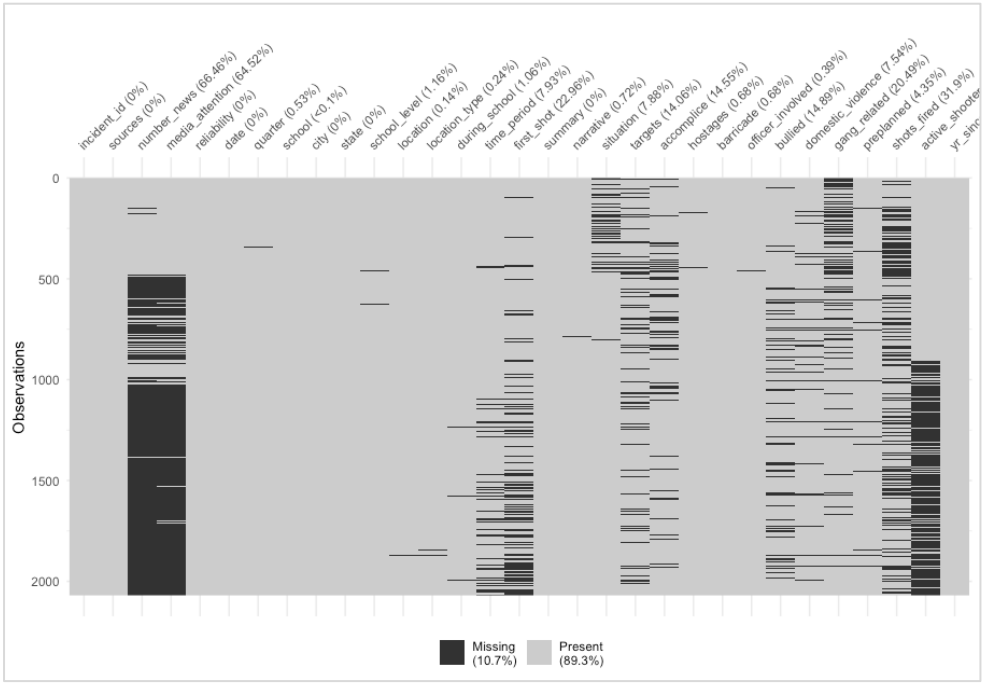
Another avenue to analyze government accountability would be through case studies designed to compare states with progressive and regressive gun reform policies. By examining the implementation and outcomes of these policies in different states, we can assess the extent to which government actions create effective measures in reducing gun violence in schools.

We could even take a casual approach in studying gun policies. For example, when given dates of major gun law reforms in a specific state or locality, can we see a difference in number of shootings, number of victims, number of weapons used? A causal approach would allow us to establish a clearer understanding of the impact of specific gun policies on the prevalence and outcomes of school shootings. We would be able to determine whether changes in gun laws directly influence the number of shootings, the number of victims, and the types of weapons used.

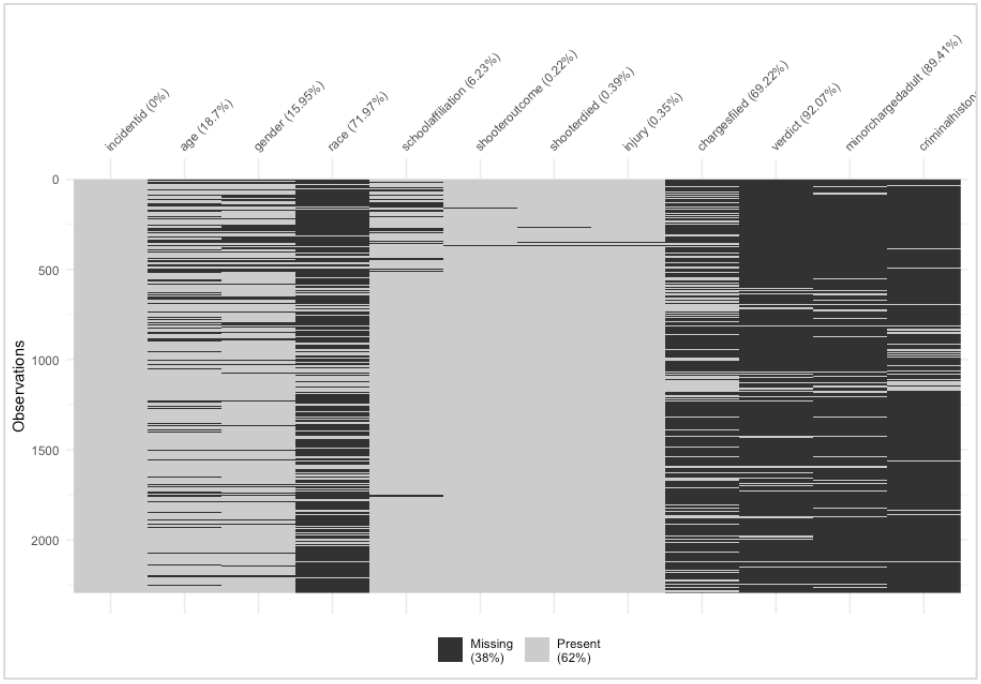
Appendix

A. Visualization of Dataset Missingness

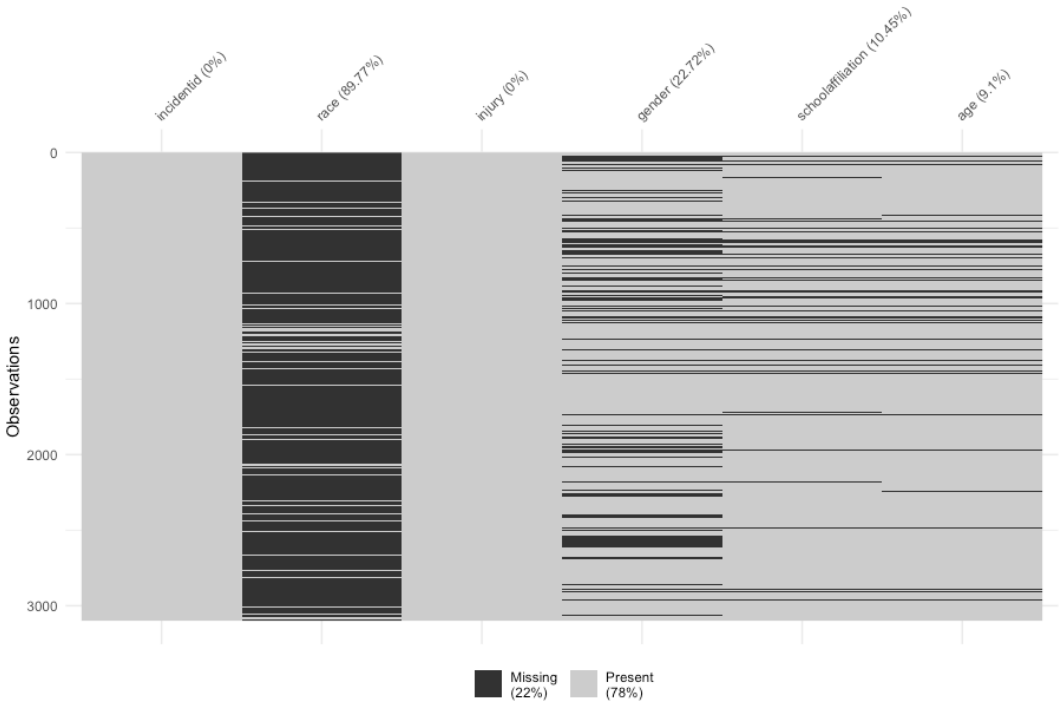
Incident dataset



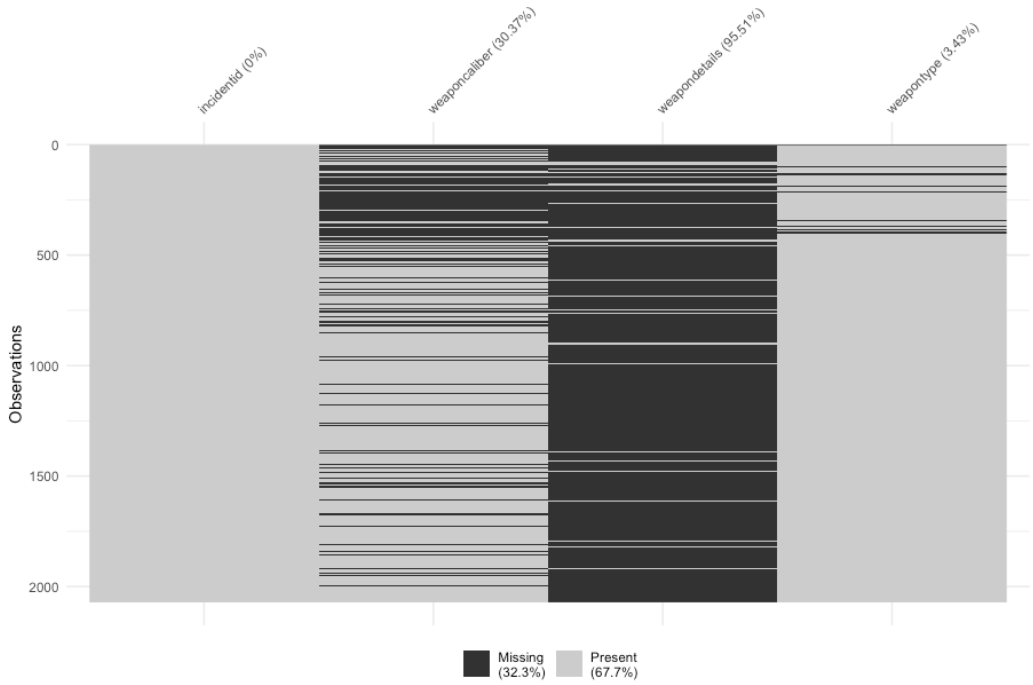
Shooter dataset



Victim dataset



Weapon dataset



B. Code

file: 01_tidy.R

purpose: read and clean the data, transfer shooter, victim, weapon data into one dataset

```
# Read data
# school-shooting-analysis/code/
# Melissa Juarez

## Research Question: Using a logit model, how do X, Y, and Z variables affect the
## likelhihood that there will be a victim in a school shooting incident?

library(here)
library(readxl)
library(dplyr)
library(ggplot2)

######### READ DATA ###########
## data has 4 tabs: INCIDENT, SHOOTER, VICTIM, WEAPON
incident_df <- read_excel(here::here("data/SSDB_Raw_Data_2022.xlsx"), sheet = "INCIDENT") %>%
janitor::clean_names()
shooter_df <- read_excel(here::here("data/SSDB_Raw_Data_2022.xlsx"), sheet = "SHOOTER") %>%
janitor::clean_names()
victim_df <- read_excel(here::here("data/SSDB_Raw_Data_2022.xlsx"), sheet = "VICTIM") %>%
janitor::clean_names()
weapon_df <- read_excel(here::here("data/SSDB_Raw_Data_2022.xlsx"), sheet = "WEAPON") %>%
janitor::clean_names()

######### CLEAN DATA ###########

## replace 'null' string values with N/As
incident_df[incident_df == 'null'] <- NA
shooter_df[shooter_df == 'null'] <- NA
victim_df[victim_df == 'null'] <- NA
weapon_df[weapon_df == 'null'] <- NA

incident_df[incident_df == 'N/A'] <- NA
shooter_df[shooter_df == 'N/A'] <- NA
victim_df[victim_df == 'N/A'] <- NA
weapon_df[weapon_df == 'N/A'] <- NA

incident_df$domestic_violence[incident_df$domestic_violence == 'NO'] <- 'No'


## fix date formats to date
incident_df$date <- as.Date(incident_df$date)
incident_df$yr_since_1970 <- as.numeric(format(incident_df$date,'%Y')) - 1970

############################ AGGREGATION ##############################
# The goal of this is to create one dataframe used for our analysis
# that contains relevant information on the victims, weapons, and shooters,
# for each incident. currently, each of the victims, weapons, and shooters
# dataframes contains information on the victim, weapon, and shooter level,
# rather than at the incident level.

######### VICTIMS AGGREGATION ###########

## For each shooting incident, how many total victims were there? How are the number of victims
per incident distributed?
victims_aggregation <- victim_df %>%
  group_by(incidentid) %>%
  mutate(victim_count = n()) %>%
  distinct(incidentid, .keep_all = TRUE) %>%
  select(incidentid, victim_count) %>%
  mutate(any_victims = case_when(
    victim_count == 0 ~ 0,
    victim_count > 0 ~ 1
  ), multi_victim = case_when(
```

```
    victim_count <= 1 ~ 0,
    victim_count > 1 ~ 1
  )
  )

## merge victim aggregation with incident_df
model_df <- incident_df %>% merge(victims_aggregation, by.x = "incident_id", by.y = "incidentid",
all.x=TRUE)
model_df$victim_count[is.na(model_df$victim_count)] <- 0
model_df$any_victims[is.na(model_df$any_victims)] <- 0


table(model_df$victim_count, useNA = "ifany")

######### SHOOTERS AGGREGATION ###########
# the goal is to retain information on if there were multiple shooters at the incident,
# and how many shooters of a certain gender were the incident

table(shooter_df$gender, useNA = "ifany")
## (1) For each shooting incident, how many total shooters were there & what were their genders?
shooters_agg <- shooter_df %>%
  group_by(incidentid) %>%
  mutate(shooter_count = n(),
         shootersexes = paste(gender, collapse = ", "),
         shooterages = paste(age, collapse = ", ")
         ) %>%
  distinct(incidentid, .keep_all = TRUE) %>%
  select(incidentid, shooter_count, shootersexes, shooterages)

# look at distribution of number of shooters and ages per incident
table(shooters_agg$shooter_count, useNA = "ifany")

# filter to only look at instances of single-shooters
single_shooters_agg <- shooters_agg %>%
  filter(shooter_count == 1) %>%
  mutate(sex = case_when(
    stringr::str_count(shootersexes, "Male") == TRUE ~ 0,
    stringr::str_count(shootersexes, "Female") == TRUE ~ 1
  ))

#shooters_agg$shootersexes <- gsub(', NA','',shooters_agg$shootersexes)
single_shooters_agg$shootersexes <- ifelse(stringr::str_detect(single_shooters_agg$shootersexes,
"NA"),
                                   NA, single_shooters_agg$shootersexes)
single_shooters_agg$shooterages <- ifelse(stringr::str_detect(single_shooters_agg$shooterages,
"NA"),
                                          NA, single_shooters_agg$shooterages)

table(single_shooters_agg$shootersexes, useNA = "ifany")
table(single_shooters_agg$shooterages, useNA = "ifany")

# For each incident, was the shooter underage and therefore in illegal possession of a gun?
# For these purposes, anyone under 18 possessing of a gun is considered unlawful.

single_shooters_agg <- single_shooters_agg %>%
  mutate(minor = case_when(
    shooterages %in% c('5', '6', '7', '8', '9', '10','11','12','13','14','15','16','17',
                       'Child', 'Minor','Teen') ~ 1,
    !(shooterages %in% c('5', '6', '7', '8', '9', '10','11','12','13','14','15','16','17',
                         'Child', 'Minor','Teen')) ~ 0
  ))

## merge victim aggregation with model_df
model_df <- model_df %>% merge(single_shooters_agg, by.x = "incident_id", by.y = "incidentid",
all.x=TRUE) %>%
  filter(!is.na(shooter_count) & shooter_count == 1)


######### WEAPONS AGGREGATION ###########
# the goal is to retain information on if there were multiple weapons at the incident, and what
category of weapons there were
```

```
table(weapon_df$weapontype, useNA = "ifany")

weapons_agg <- weapon_df %>%
  group_by(incidentid) %>%
  mutate(weapon_entries = n(),
         weapontypes = paste(weapontype, collapse = ", ")) %>%
  distinct(incidentid, .keep_all = TRUE) %>%
  select(incidentid, weapon_entries, weapontypes)

# code string NAs  and "No Data" entries as <NA>
weapons_agg$weapontypes <- ifelse(stringr::str_detect(weapons_agg$weapontypes, "NA|No Data"),
                                  NA, weapons_agg$weapontypes)

# merge weapon aggregation to model_df
model_df <- model_df %>% merge(weapons_agg, by.x = "incident_id", by.y = "incidentid",
all.x=TRUE)
table(incident_df$weapontypes, useNA = "ifany")

# create columns `handgun_used`, `rifle_used`, `shotgun_used`, `multiple_weapons`, based on
weapontypes list & length of list
model_df <- model_df %>%
  mutate(handgun_used = case_when(
    is.na(weapontypes) ~ NA_real_,
    stringr::str_detect(weapontypes, "Handgun") ~ 1,
    !(stringr::str_detect(weapontypes, "Handgun")) ~ 0,
  ),
  rifle_used = case_when(
    is.na(weapontypes) ~ NA_real_,
    stringr::str_detect(weapontypes, "Rifle") ~ 1,
    !(stringr::str_detect(weapontypes, "Rifle")) ~ 0,
  ),
  shotgun_used = case_when(
    is.na(weapontypes) ~ NA_real_,
    stringr::str_detect(weapontypes, "Shotgun") ~ 1,
    !(stringr::str_detect(weapontypes, "Shotgun")) ~ 0,
  ),
  multiple_weapons = case_when(
    is.na(weapontypes)  ~ NA_real_,
    weapon_entries > 1 | stringr::str_detect(weapontypes, "Multiple") ~ 1,
    !(weapon_entries > 1 | stringr::str_detect(weapontypes, "Multiple")) ~ 0
  )
  )


### Recoding model_df for no = 0, yes = 1

model_df <- model_df %>% mutate(
   during_school = case_when(
     during_school == "No" ~ 0,
     during_school == "Yes" ~ 1
   ),
   bullied = case_when(
     bullied == "No" ~ 0,
     bullied == "Yes" ~ 1
   ),
   preplanned = case_when(
     preplanned == "No" ~ 0,
     preplanned == "Yes" ~ 1
   )
)
```

file: 02_analysis.R

purpose: analyze the data and create models

```
# Analyze Data & Create Models
# school-shooting-analysis/code/
# Melissa Juarez

library(here)
library(readxl)
library(dplyr)
library(ggplot2)

## source functions and variables from prior scripts
source(here::here("code/01_tidy.R"))

##########  Visualizing Missingness ##########
# to select variables of interest
library(naniar)
vis_miss(incident_df)
vis_miss(shooter_df)
vis_miss(victim_df)
vis_miss(weapon_df)

# visualize our target dataframe, which contains recoded and aggregated variables
# at the incident-level of analysis

model_df <- model_df %>%
  select(yr_since_1970, bullied, minor, sex, shootersexes, multiple_weapons, handgun_used,
         rifle_used, shotgun_used, victim_count, multi_victim)
vis_miss(model_df)

## we see that our dataframe has 12.5% missingness


########## GUN EFFECTS MODEL ##########

# describe variables
model_df %>%
  select(yr_since_1970, victim_count, multiple_weapons, handgun_used,
         rifle_used, shotgun_used, minor) %>%  # keep only the columns of interest
  tbl_summary(type = list(yr_since_1970   ~ "continuous",
                          victim_count     ~ "continuous",
                          multiple_weapons      ~ "categorical",
                          handgun_used      ~ "categorical",
                          rifle_used      ~ "categorical",
                          shotgun_used      ~ "categorical",
                          minor      ~ "categorical"
  ),
  statistic = list(all_continuous() ~ "{mean} ({sd})",
                   all_categorical() ~ "{n} / {N} ({p}%)"),
  missing_text = "N/A") %>%
  modify_header(label = "**Variable**") %>%
  modify_caption("Gun Model Variables") %>%
  bold_labels()

# create initial linear model
attach(model_df)
g1 <- lm(victim_count ~ handgun_used + rifle_used + shotgun_used)
summary(g1)

## handgun, rifle, and shotgun are expected to be more deadly compared to cases with
unknown/other weapon used
## among the three, rifle has the strongest effect on victimhood, handgun & shotgun held constant

## what happens when I control for instances where the shooter has multiple weapons?
g2 <- lm(victim_count ~ handgun_used + rifle_used + shotgun_used + multiple_weapons)
summary(g2)

## minors & year
```

```
g3 <- lm(victim_count ~ handgun_used + rifle_used + shotgun_used + multiple_weapons + minor +
yr_since_1970)
summary(g3)

## interaction between quantity and type of gun?
g4 <- lm(victim_count ~ handgun_used*multiple_weapons + rifle_used*multiple_weapons +
shotgun_used*multiple_weapons)
summary(g4)


## which model reduces RSE and fits our data better?
## the fact that g4 reduces to g2 means that we can test a hypothesis where the null is that the
model
## g2 is an adequate fit for the data and the alternative is the full model g4, as we'll test
below.
anova(g2, g4)

## Based on the lack of fit test of the two models, we get an F-statistic of 70.258 and a p-value
of less
## than 0.05. Thus, there is compelling evidence for us to reject the hypothesis that the reduced
g2 model
## fits our data well.


## model with only significant terms
guns_df <- model_df %>%
  select(victim_count, multiple_weapons, rifle_used, handgun_used, shotgun_used) %>%
  na.omit()

rownames(guns_df) <- 1:nrow(guns_df)

g5 <- lm(victim_count ~ multiple_weapons + rifle_used*multiple_weapons +
           handgun_used:multiple_weapons + shotgun_used:multiple_weapons, data = guns_df)
summary(g5)


## Diagnostics
cd <- cooks.distance(g5); cd
plot(cd)
ot <- outlierTest(g5)

ot_df <- as.data.frame(do.call(cbind, ot))
outliers <- merge(guns_df, ot_df, by = 0)
outliers$Row.names <- as.numeric(outliers$Row.names)

par(mfrow=c(2,2))
plot(g5)

# VIF
car::vif(g5, type = 'predictor')

## considering quasipoisson?
# visualize the outcome
ggplot(model_df, aes(x=victim_count)) +
  geom_histogram(aes(y=..density..), color="darkblue", fill="lightblue", binwidth = 1) +
  scale_y_continuous(

    # Features of the first axis
    name = "Density",

    # Add a second axis and specify its features
    sec.axis = sec_axis( trans=~.*2069, name="Frequency")
  ) +
  geom_density(alpha=.2, fill="#FF6666") +
  geom_vline(aes(xintercept=mean(victim_count)),
             color="black", linetype="dashed", size=0.5) +
  xlim(-1,20) +
  labs(title="Distribution of Victim Count for School Shooting",
       x ="Victim Count", y = "Density")
```

```
g6.poi <- glm(victim_count ~ rifle_used*multiple_weapons + shotgun_used*multiple_weapons, data =
model_df, family = quasipoisson())
summary(g6.poi)

########## MENTAL HEALTH EFFECTS MODEL ##########

# describe variables
library(gtsummary)
# describe variables
model_df %>%
  select(yr_since_1970, victim_count, bullied, shootersexes) %>%  # keep only the columns of
interest
  tbl_summary(type = list(yr_since_1970   ~ "continuous",
                          victim_count     ~ "continuous",
                          bullied      ~ "categorical",
                          shootersexes      ~ "categorical"),
              statistic = list(all_continuous() ~ "{mean} ({sd})",        # stats and format for
continuous columns
                               all_categorical() ~ "{n} / {N} ({p}%)"),
              missing_text = "N/A",
              label = list(shootersexes  ~ "sex"))  %>%
  modify_header(label = "**Variable**") %>%
  modify_caption("Mental Health Model Variables") %>%
  bold_labels()


# create initial linear model
attach(model_df)
m1 <- lm(victim_count ~ bullied)
summary(m1)

## what happens when I control for gender and year?
m2 <- lm(victim_count ~ bullied + yr_since_1970 + sex)
summary(m2)

## take away sex bc not significant
m3 <- lm(victim_count ~ bullied + yr_since_1970)
summary(m3)

## interaction term?
m4 <- lm(victim_count ~ bullied *yr_since_1970 )
summary(m4)

## which fit is better?
anova(m3,m4)
## not enough evidence to say that the interaction effects model is a better fit for the data.

mentalhealth_df <- model_df %>%
  select(victim_count, bullied, yr_since_1970) %>%
  na.omit()

rownames(mentalhealth_df) <- 1:nrow(mentalhealth_df)


## Diagnostics
cd2 <- cooks.distance(m3); cd
plot(cd2)
ot2 <- outlierTest(m3)

ot2_df <- as.data.frame(do.call(cbind, ot2))
outliers2 <- merge(mentalhealth_df, ot2_df, by = 0)
outliers2$Row.names <- as.numeric(outliers2$Row.names)

par(mfrow=c(2,2))
plot(m3)

# VIF
car::vif(m3)
```