**Project Proposal**

Tom Bagley, Hannah Hammond, Parker Short, Amber Oldroyd, Latham Jeppson

**Economic Question**

We will estimate the causal relationship between descriptive statistics of various mass shooters (region, year, background, mental health, etc.) and the number of victims killed or injured. Another topic of value would be the causal relationship between those same statistics and status as a perpetrator. This second topic would require more data related to individuals who were not mass shooters.

**Economic Data**

In order to better calibrate the data for the questions we will be asking, we need to create some new variables. An important one to create is a dummy variable based on years when the shooting happened. We can create a variable for every five years because this will allow us to see trends over time like if violence is increasing and if it is getting worse on average per incidence. Another important variable to include is a variable for yes or no to previous crime violence, previous trauma and childhood experiences, and signs of crisis. In our dataset, there are very specific results for each form of previous crime or trauma. For example, these include whether their parent was suicidal or if they experienced abuse as a child. It may be useful to combine these variables and see if they had any form of trauma. Another variable that we could create would be regions of the United States where these shootings happened. If we could see that different amounts of violence happened in different regions, we could look up whether policies in those regions are present and then test if those policies have an impact. This could also be done the other way around by looking at policies that we hypothesize may reduce violence and putting states with those policies as a categorical variable. An example of this is that some states require training in order to purchase guns. We could create a variable for the states that do and don’t require this and test whether it lowers or increases violence.

**Obtaining Data**

Fortunately, mass shootings are extremely well documented. As such, we will use data from a variety of sources. The K-12 School Shooting Database has data specifically about school shootings, including number of people shot, the situation during the shooting, how the incident ended, the location inside the school where the shooting occurred, time period, school level, etc. The Violence Project data covers 190 mass shootings from 1966 to 2021, with information about the shooter, the victims, the firearms, and relationship to the community of the shooter, but focuses specifically on the psychosocial characteristics of the shooter. We also have data for about 2000 gun violence incidents in which someone was injured or killed dating from 2014-2023 by location. Demographic characteristics about the counties / regions in which the shootings took place are available through the IPUMS Current Population Survey. Given the nature of shootings, most of the data is numerical, but some categorical variables include where in the school the shooting took place and situation of the shooting in our K-12 SS data.

**Methods of Data Generation**

Based on the data that we have, we want to generate new data to better understand school shootings that may happen in the future. One method we want to try is a neural network. A generative adversarial network, or GAN, is a neural network that can be used to create data. A GAN begins with using the data we have as the ‘real’ data. We then randomly generate values for new data for the desired columns. This data won’t be very accurate to start. We can classify the data points as ‘real’ or ‘fake’ and send feedback to the data generator. It will try to ‘beat’ the classifier by making data that can’t be distinguished as fake. This will be able to generate data that resembles the existing data, but has new features and qualities to help us uncover patterns in the data to uncover what trends are happening in school shootings.

Another method that we plan to use is random forests. Random forests are able to make predictions based on multiple previously observed averages within a dataset. They do this by implementing multiple learning algorithms for better predictive performance. Random forests are used for classification and regression purposes. We believe that our chosen dataset has enough variables and information to make a strong argument for the use of random forests. One limitation to random forests that we will weary of is covariate shift. Covariate shift is when there are differences in the distributions of the training data and those of the live data or in our case the new data that we will be producing. Random forests are susceptible to this issue.

**Problem Set 6**

**3. The following questions relate to your project**

**(a) With whom are you collaborating on the project?**

Tom Bagley, Hannah Hammond, Parker Short, Amber Oldroyd, Latham Jeppson

**(b) What is the research question your project will be answering?**

Does mental illness in the shooter increase the number of victims? (controlling for the type of gun and location of the shooting as well as other relevant variables)

**(c) What data are you using? Describe the source as well as the variables (outcome(s) and regressors)**

“Violence Project Mass Shooter Database - Version 6.1”

Peterson, J., & Densley, J. (2022). The Violence Project database of mass shootings in the United States (Version 6). <https://www.theviolenceproject.org>

“The database was constructed using public records and open-source data. Where available and applicable, we drew on first person accounts, such as the perpetrators’ diaries, ‘manifestos,’ suicide notes, social media and blog posts, audio and video recordings, interview transcripts, and personal correspondence. We also used secondary sources such as existing mass shooter databases, media coverage, documentary films and podcasts, biographies, monographs and academic journal articles, court transcripts, federal, state, and local law enforcement records, medical records, school records, and autopsy reports. Anything that could be requested or found on the internet was included. The community database only goes back to 1995 and draws on U.S. Census data for the closest census year, FBI Uniform Crime Reports, Google Maps, and other macro-level data sources.”

Outcome variable: number of victims (killed / injured)

Regressors: type of gun, location of the shooting, whether or not the shooter had mental illness, motivation of the shooting, whether the shooter had a previous criminal record, etc.

**(d) Give a table of summary statistics for your data, including the mean, standard deviation, and number of observations. Make the table presentable! (not just copy-pasted output from a software package)**

| **classification** | **Frequency** | **Percent** |
| --- | --- | --- |
| Handgun | 223 | 55.61 |
| Shotgun | 45 | 11.22 |
| Rifle | 48 | 11.97 |
| Assault weapon | 85 | 21.20 |
| Total | 401 | 100.00 |

|  | Month | Number Killed | Number Injured | Age | Gender | Height | Education |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | 189 | 189 | 189 | 188 | 188 | 56 | 138 |
| mean | 6.60 | 7.21 | 11.30 | 33.61 | 0.04 | 70.88 | 1.62 |
| std | 3.44 | 6.63 | 63.40 | 12.37 | 0.26 | 3.85 | 1.15 |
| min | 1 | 4 | 0 | 11 | 0 | 62 | 0 |
| 25% | 3 | 4 | 1 | 23 | 0 | 68 | 1 |
| 50% | 7 | 5 | 3 | 32 | 0 | 71 | 2 |
| 75% | 10 | 7 | 7 | 43 | 0 | 73.25 | 2 |
| max | 12 | 60 | 867 | 70 | 3 | 84 | 4 |

|  | School Perform-ance | Total Firearms Brought to the Scene | Race\_1 | Race\_2 | Race\_3 | Race\_4 | Race\_5 | Race\_Bosnian |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 88 | 188 | 189 | 189 | 189 | 189 | 189 | 189 |
| mean | 1.07 | 2.23 | 0.21 | 0.08 | 0.06 | 0.04 | 0.02 | 0.01 |
| std | 0.88 | 2.21 | 0.41 | 0.28 | 0.23 | 0.20 | 0.13 | 0.07 |
| min | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50% | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75% | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| max | 2 | 24 | 1 | 1 | 1 | 1 | 1 | 1 |

|  | Criminal Record\_1 | Criminal Record\_1` | Mental Illness\_1 | Mental Illness\_1, 2 | Mental Illness\_2 | Mental Illness\_3 | Mental Illness\_4 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | 189 | 189 | 189 | 189 | 189 | 189 | 189 |
| mean | 0.56 | 0.01 | 0.15 | 0.09 | 0.16 | 0.05 | 0.24 |
| std | 0.50 | 0.07 | 0.36 | 0.29 | 0.37 | 0.21 | 0.43 |
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50% | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75% | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| max | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

|  | Number of Siblings | History of Animal Abuse | History of Domestic Abuse | Bully | Bullied | Sexual Orientation |
| --- | --- | --- | --- | --- | --- | --- |
| count | 140 | 188 | 188 | 187 | 187 | 187 |
| mean | 2.19 | 0.06 | 0.59 | 0.21 | 0.18 | 0.03 |
| std | 1.97 | 0.24 | 0.93 | 0.41 | 0.39 | 0.16 |
| min | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 1 | 0 | 0 | 0 | 0 | 0 |
| 50% | 2 | 0 | 0 | 0 | 0 | 0 |
| 75% | 2.25 | 0 | 1 | 0 | 0 | 0 |
| max | 11 | 1 | 3 | 1 | 1 | 1 |

**(e) Which machine learning methods will you be using?**

We will use regularized regression and/or decision trees as steps in determining the causal relationship. In the next few weeks as we learn how to evaluate causal relationships, we will narrow in on how we will fully answer the question.