# Final Project Steps 1 & 2

Miles Peña

2024-02-18

# Step 1:

## Introduction:

Throughout the years, the United States of America has seen an increase in violence through mass shootings. Since the Sandy Hook Elementary School shooting, we have seen the numbers nearly triple in size. While, yes, the country should look at better gun control policies to solve this issue, as data scientists we can look at the data that we have collected from these shootings over the past 10 years to try to prevent these from happening in the future. We can look at different factors such as demographics, socioeconomic status, and the effect that mental illness can have on these shootings. This is important for individuals to be interested in because it is something that can potentially help reduce the number of shootings or hopefully eradicate them completely at some point in the future. With the amount of data that has been compiled since this great increase 10 years ago, it can now be a problem addressed through the help of data science.

## Research Questions:

1. What is the choice of weapon for these shootings?
2. Do the shooters have a criminal background or are they first time offenders?
3. Which gender is more prone to commit this type of crime?
4. Where do most shootings take place? What is the most frequent place?
5. Does race play a role in mass shootings?
6. Does mental illness play a role in mass shootings?

## Approach:

This approach will not fully address the problem but rather will serve as a means of attempting to predict criminal behavior based on past experiences. By gathering and compiling the data of previous mass shootings, we will aim to create a guide of what to look for or what to flag as suspicious activity. This will assist with helping to provide enough police presence in areas that are more prone to attacks. It will help create a log for legally obtained weapons and who purchased them in order to better track these. The biggest thing this approach will aim to accomplish is determine the correlation (or lack thereof) between mass shootings and individuals with mental illness in order to persuade and implement a program of mental health evaluation prior to firearm sales.

## Data:

1. https://www.kaggle.com/datasets/zusmani/us-mass-shootings-last-50-years

This dataset looks at mass shooting attacks in the United States of America between 1966 and 2017. The dataset contains Serial No, Title, Location, Date, Summary, Fatalities, Injured, Total Victims, Mental Health Issue, Race, Gender, and Latitude and Longitude information. The data set was posted to Kaggle in 2022 and has been modified multiple times to add new variables, add new data, and add missing data in order to help create visualizations and extract patterns. Similar to my project, the data set aims to aid in predictions for future events and prevention of such.

2. https://www.kaggle.com/datasets/carlosparadis/stanford-msa

This dataset attempts to facilitate research on gun violence in the US by making raw data more accessible. This dataset is comparable to the previous one except it has a bit more detail in the mental illness variable which can be very useful with the project.

3. https://www.kaggle.com/datasets/twinkle0705/mental-health-and-suicide-rates

The goal of using this dataset is to determine if a correlation exists between mental health and suicide rates. The majority of mass shootings end with the perpetrator being shot down by the police. Because of this, it stands to reason that mass shooters go into the situation not expecting to come out alive (i.e. go in as a suicide attempt while taking people out with them). This dataset compares countries, their resources for mental health, their mental health facilities as well as their suicide rates.

## Required Packages:

From what I know so far, the packages required to load and compare these datasets are as follows:

library(ggplot2)

library(dplyr)

library(readr)

library(tidyr)

library(stringr)

library(reshape2)

library(ggthemes)

library(lattice)

library(knitr)

## Plots and Table Needs:

Scatterplots, histograms, and boxplots. I would also like to incorporate a map graph to depict the rates per state and an area chart to show the change over time of mass shootings. As far as tables go, I would like to use the simple XY table as well as column tables and grouped table to go a bit more in depth for comparison purposes. A multiple variables table might also prove to be beneficial for displaying data and making it easier to understand for the reader.

## Questions for Future Steps:

The main thing that I do not right now that I will need in order to analyze and interpret the data is how to create the correct tables and plots as described above. I am sure that there will need to be other things to learn but right now, I do not know that that is. For the most part, I feel confident in the datasets I have collected as well as with the learnings we have covered so far to be able to compare the datasets and come to a conclusion that could possibly assist with predicting the who, what, where, how, and why of mass shootings in the United States.

# Step 2:

## How did you import and clean your data?

Data cleaning involves handling missing values, outliers, and incorrect data. Missing values can be handled by either dropping the rows/columns with missing values. For my dataset, I used na.omit() in order to remove any empty cells that would not give accurate information for the study. I also cleaned up and corrected the date format on all of the entries as they were not all the same.

## What does the final data set look like?

```
setwd("/Users/milespena/Documents/R")
shootingData <- read.csv("shooting-1982-2023.csv")
shootingData.No.NA <- na.omit(shootingData)
summary(shootingData.No.NA)
```

```
##    i..case            location             date              summary
##  Length:80          Length:80          Length:80          Length:80
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    fatalities       injured          total_victims       location.1
##  Min.   : 3.00   Length:80          Length:80          Length:80
##  1st Qu.: 5.00   Class :character   Class :character   Class :character
##  Median : 6.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 9.25
##  3rd Qu.: 9.00
##  Max.   :58.00
##  age_of_shooter  prior_signs_mental_health_issues mental_health_details
##  Min.   :11.00   Length:80                        Length:80
##  1st Qu.:25.00   Class :character                 Class :character
##  Median :34.00   Mode  :character                 Mode  :character
##  Mean   :34.11
##  3rd Qu.:42.00
##  Max.   :66.00
##  weapons_obtained_legally where_obtained     weapon_type
##  Length:80                Length:80          Length:80
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
##
##
##
##  weapon_details         race              gender             latitude
##  Length:80          Length:80          Length:80          Min.   :21.32
##  Class :character   Class :character   Class :character   1st Qu.:33.40
##  Mode  :character   Mode  :character   Mode  :character   Median :37.80
##                                                           Mean   :37.22
##                                                           3rd Qu.:41.71
##                                                           Max.   :48.05
##    longitude          type               year
##  Min.   :-157.88   Length:80          Min.   :1982
##  1st Qu.:-117.10   Class :character   1st Qu.:1996
##  Median : -92.42   Mode  :character   Median :2006
##  Mean   : -97.29                      Mean   :2004
##  3rd Qu.: -82.36                      3rd Qu.:2012
##  Max.   : -71.08                      Max.   :2018
```

```
head(shootingData.No.NA)
```

| i..case <chr> | location <chr> | date <chr> |
|---|---|---|
| 40 Fifth Third Center shooting | Cincinnati, Ohio | 9/6/1918 |
| 43 Waffle House shooting | Nashville, Tennessee | 4/22/1918 |
| 45 Marjory Stoneman Douglas High School shooting | Parkland, Florida | 2/14/1918 |
| 48 Texas First Baptist Church massacre | Sutherland Springs, Texas | 11/5/1917 |
| 51 Las Vegas Strip massacre | Las Vegas, Nevada | 10/1/1917 |
| 52 San Francisco UPS shooting | San Francisco, California | 6/14/1917 |

6 rows | 1-4 of 22 columns

```
tail(shootingData.No.NA)
```

| i..case <chr> | location <chr> | date <chr> |
|---|---|---|
| 136 ESL shooting | Sunnyvale, California | 2/16/1988 |
| 137 Shopping centers spree killings | Palm Bay, Florida | 4/23/1987 |
| 138 United States Postal Service shooting | Edmond, Oklahoma | 8/20/1986 |
| 139 San Ysidro McDonald's massacre | San Ysidro, California | 7/18/1984 |
| 140 Dallas nightclub shooting | Dallas, Texas | 6/29/1984 |
| 141 Welding shop shooting | Miami, Florida | 8/20/1982 |

6 rows | 1-4 of 22 columns

## What information is not self-evident?

The information that is not self-evident will be the relationship between the variables. In order to uncover new information in the data, I plan to perform exploratory data analysis (EDA) involving visualizing the data using plots, finding correlations, and identifying patterns and trends in the data. There are also two questions for which I may need to do my own research for and add columns to represent this such as previous criminal history as well as generalized location of where they shooting took place.

## What are different ways you could look at this data?

By performing descriptive statistics to get a sense of the distribution of each variable, or by creating visualizations like histograms or scatter plots to understand the relationships between variables.

## How do you plan to slice and dice the data?

I plan to explore the data by grouping it based on variables or possibly filtering to focus on a subset. The groupby() function that we used previously I feel may come in handy for this assignment. It is important to look at the data for each variable as each variable will play a key role in answering the questions of this study.

## How could you summarize your data to answer key questions?

In order to summarize the data, I plan on calculating measures of central tendency like the mean, median and mode as well as measures of dispersion like the standard deviation or interquartile range.

## What types of plots and tables will help you to illustrate the findings to your questions?

I will use bar graphs, box plots, histograms, line graphs and scatter plots to illustrate my findings. I am still unsure of which visualization will work best for the dataset and the data manipulation I perform but I intend to experiement until I find whatever best explains the point I am trying to get accross. Correlation matrices will also aid in demonstrating relationships between variables.

## Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Since my research questions involve finding patterns in the data, I will be incorporating machine learning techniques to answer these. I again I am unsure if this will work or help with what I am trying to do but I aim to use regression models for now and depending on research and knew knowledge within the course, I may incorporate other techniques as well.

## What questions do you have now, that will lead to further analysis or additional steps?

The questions that I have at this point are mostly pertaining to which of the methods, techniques, and/or concepts that we have covered thus far could work for this project. For all previous assignments, we have been given the data and what to do with it so now having to determine which to use is overwhelming. As far as the data goes, my questions are solely in relation to the correlation of variables as well as attempting to find the appropriate answers to the questions I presented in part 1.