Mass Shootings in the US: In Data

Shibali Mishra

Introduction

Mass shootings in the United States represent a growing and deeply concerning phenomenon. My research paper project seeks to explore patterns and uncover underlying trends related to mass shootings. By examining demographic, psychological, and contextual factors, this research aims to shed light on the characteristics of these tragic events and the individuals involved. Specifically, I aim to analyze temporal trends, shooter demographics, and key behavioral relationships to identify insights that could inform future prevention strategies.

To achieve this, I am utilizing the 7th version of the Violence Project mass shooters database. The Violence Project is a nonpartisan and nonprofit research center based at Hamline University in Saint Paul, Minnesota and it provides one of the most comprehensive and detailed datasets available on mass shootings. The organization was co-founded by psychologist Dr. Jillian Peterson and sociologist Dr. James Densley. They describe mass shooting as ” shooting [of] four or more people shot and killed, excluding the shooter, in a public location, with no connection to underlying criminal activity, such as gangs or drugs” (The Violence Project, n.d.).

The data set spans from 1966 to 2023 and has ongoing updates for 2024. It includes demographic details such as age, race, and religion, along with psychological, familial, and social background factors. Such a rich data set enables me to address important research questions about trends over time, shooter characteristics, and behavioral indicators, ultimately contributing to a better understanding of mass shooting in the United States.

Objectives and Research Questions

This project is guided by these exploratory questions:

- **Temporal Trends:**
  - Are there specific years when shooting trends were particularly high or low? What possible causes or events could explain these shifts?
  - Are there specific years that experienced a notably high or low number of fatalities due to shootings?
- **Shooter Demographics:**
  - What is the age distribution of mass shooters, and are there any notable patterns by age group?
  - How are mass shooters distributed by race and how does this distribution compare to the relative population in the U.S.?
- **Mosaic Plot Analysis:**
  - Does a history of abusive behavior correlate with the likelihood of having a family member as a victim?
  - Is there a relationship between showing signs of being in crisis and increased agitation?
  - How does parental substance abuse relate to being raised by a single parent?
  - Is there an association between psychiatric medication history and prior counseling?
  - Is employment status related to having an economic issue as a motive for the shooting?
  - Do signs of being in crisis correlate with evidence of leakage (warnings or threats before the incident)?

Visually Displaying the Data

Spatially Contexuatilizing the Data:

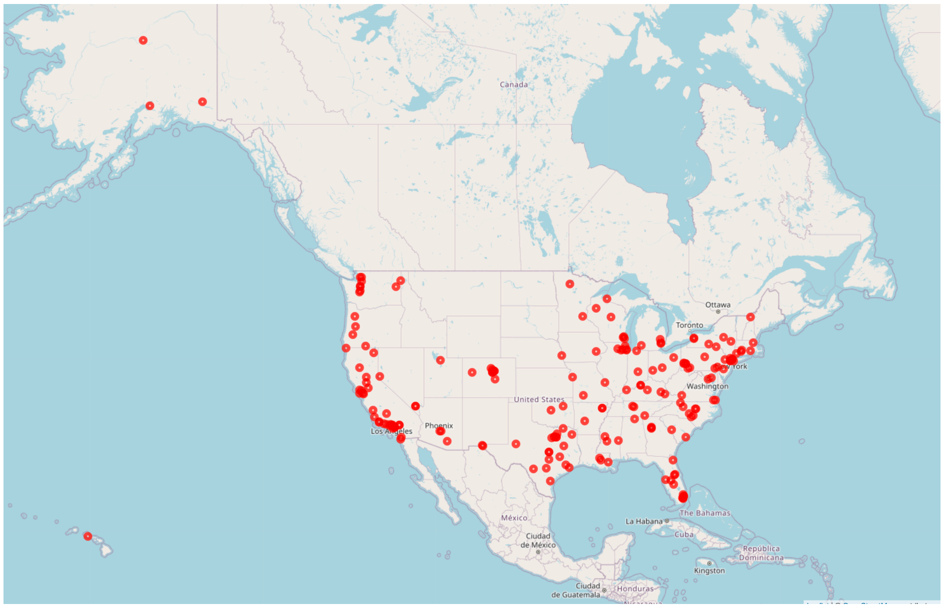
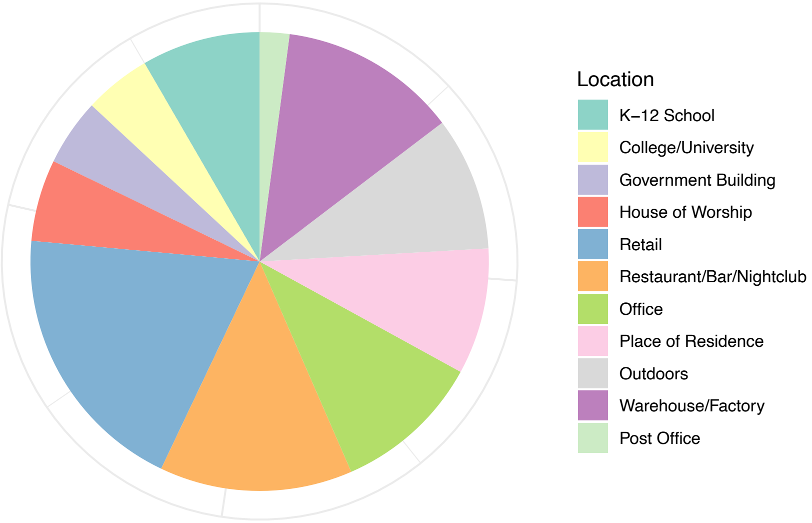


Figure 1: Shooting Locations Across the USA

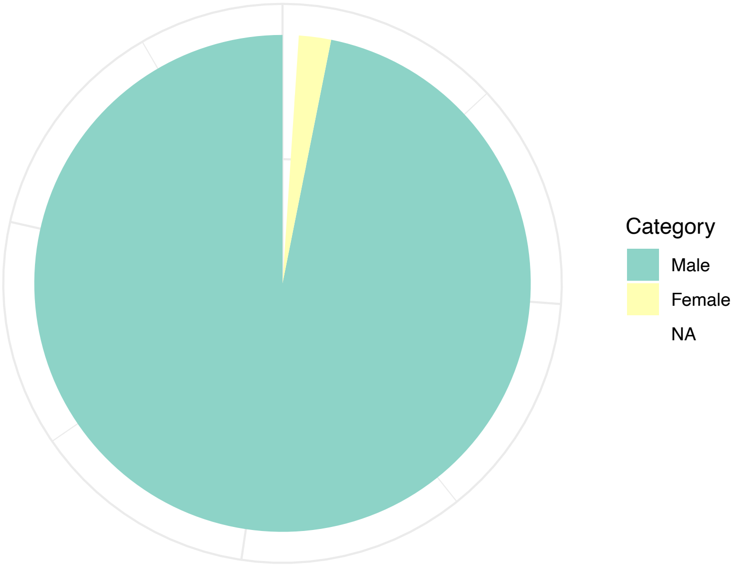
The spatial distribution in figure one is pretty consistent with the population distribution across the United States.

Where do these shootings take place?



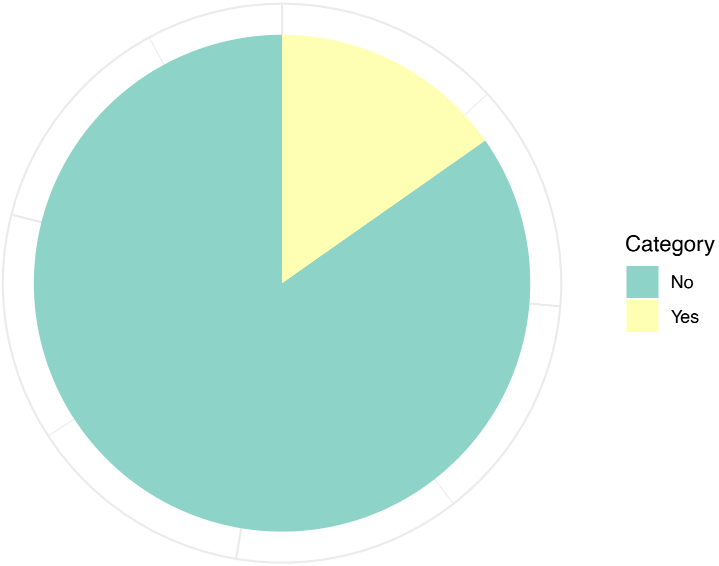
The pie chart shows the distribution of mass shooting across locations. Shooting in retail spaces and Restaurant/Bar/Nightclubs are most common. Government Buildings, and House of Worships are also targeted, highlighting how public educational, religious and civic spaces are vulnerable to these crimes.

Gender Identity of the Shooters:



The shooters are overwhelmingly male with some female, as well as NAs. NA constitute of people that identified with neither man or woman.

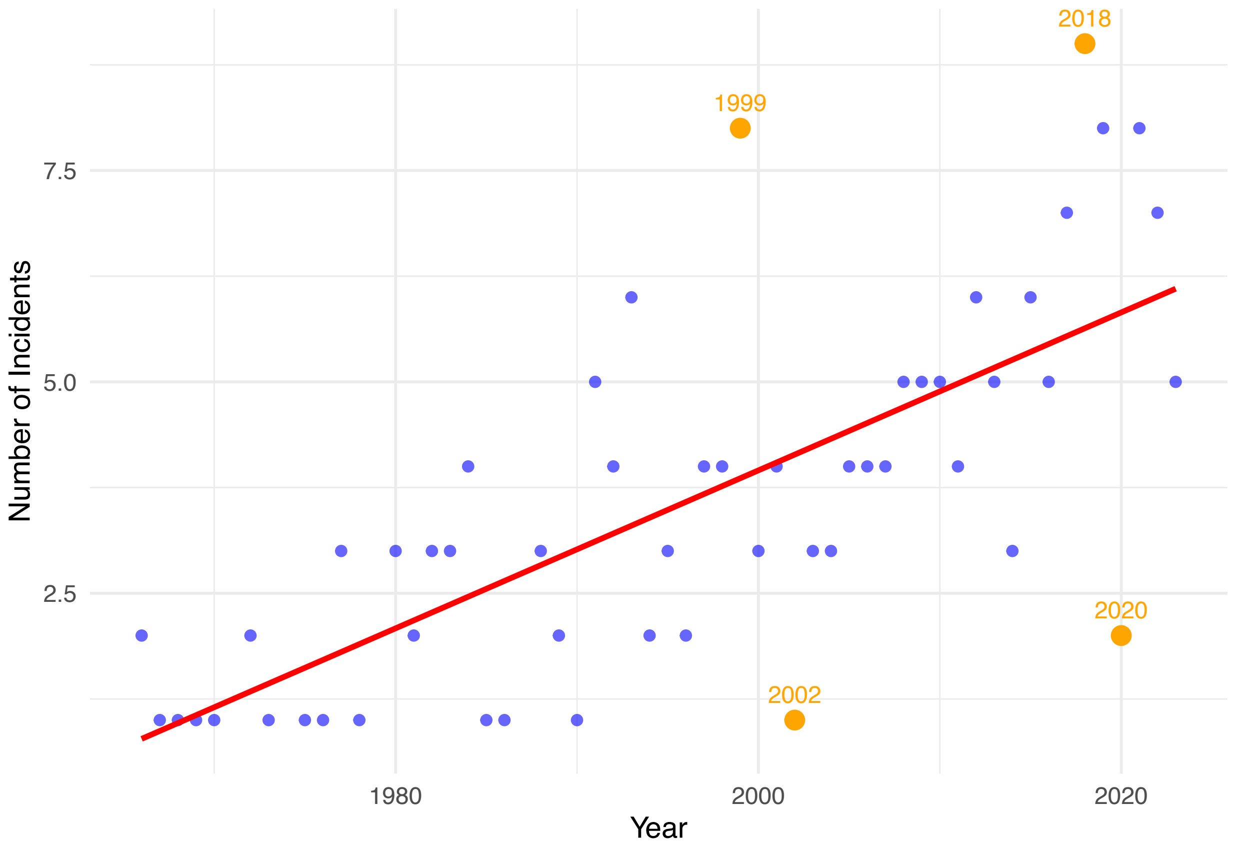
Immigrant Status of Shooters



Less than 25% of the shooters were immigrants.

Data Exploration

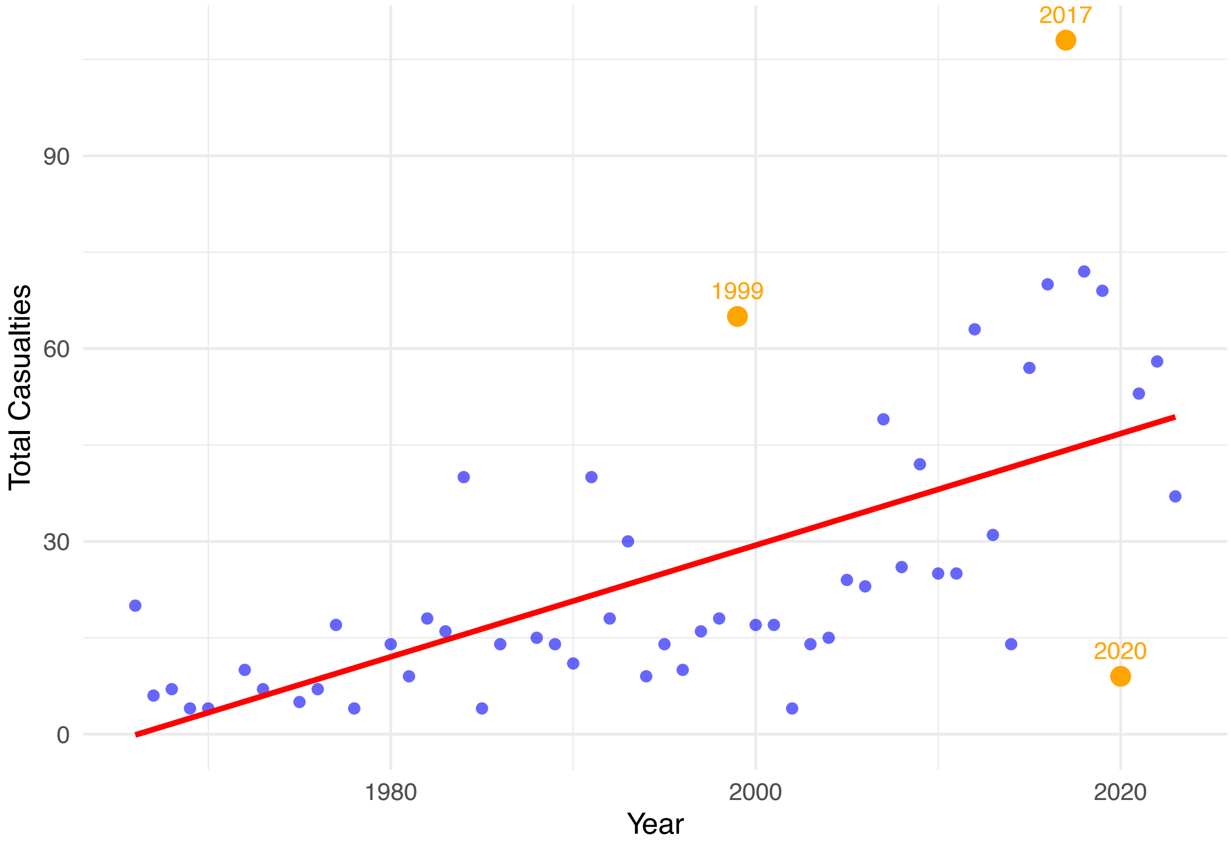
1.1 Temporal Trend: Number of Incidents Over Time



My plot indicates a general upward trend of annual number of mass shootings. My outlier analysis indicates that 1999 and 2018 experienced notably high numbers of mass shootings, while 2002 and 2020 had comparatively low incidents. Several factors may contribute to these variations. Upon some research, here are the reasons why I think this might be the case:

- 1999: I failed to find any reasonable justification for this, especially considering the enactment of Brady Law that required people to go through a mandatory background check before purchasing a firearm. “Overall, since 1993, the Brady Law has prevented more than 500,000 prohibited persons from acquiring firearms from licensed dealers,” so this unusual spike is interesting (U.S. Department of Justice, 1999).
- 2002: Following the September 11 attacks in 2001, there was heightened national security and law enforcement vigilance, which may have contributed to a reduction in mass shooting incidents in 2002.
- 2018: “In 2018, the number of concealed handgun permits in the United States exceeded 17.25 million, marking a 273% increase since 2007” (Concealed Nation, 2018). This significant rise reflects the expansion of concealed carry laws across various states, with could have led to more mass shootings.
- 2020: The COVID-19 pandemic led to widespread lockdowns and restrictions on public gatherings, resulting in fewer opportunities for mass shootings in public spaces.

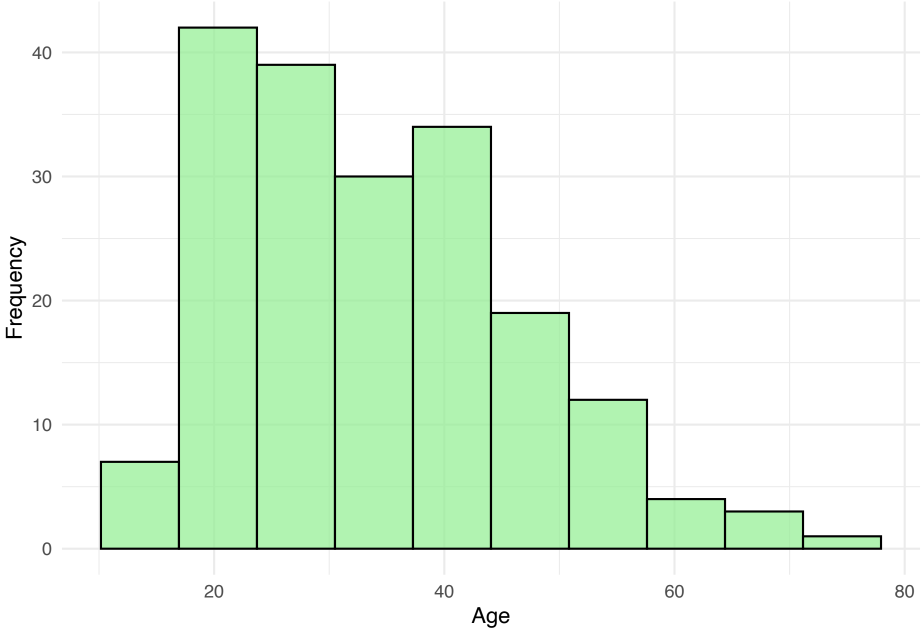
1.2 Temporal Trend: Number Casualties Over Time



The plot indicates a general upward trend in mass shooting casualties over time, with significant outliers in 1999, 2017, and 2020. The year 1999 corresponds to the Columbine High School massacre, one of the deadliest school shootings in US history and the year 2017 includes the Las Vegas shooting, the deadliest mass shooting in US history. In contrast, 2020 shows a drop in casualties, likely due to COVID-19 lock downs reducing public gatherings and opportunities for mass shootings.

From the two temporal analysis above, I can conclude that 1999 saw a spike in gun violence, with uncharacteristically high number of shootings & fatalities.

2.1 Shooter Demographics: Age Distribution of Mass Shooters



The histogram shows the age distribution of mass shooters, with the majority concentrated between 20 and 40 years old. The mean age is approximately 33.8, and the median age is 32, indicating a slightly right-skewed distribution. This suggests that while younger individuals dominate, there is a significant number of shooters in older age groups, particularly up to their 50s.

2.2 Shooter Demographics: Race

This section examines whether the racial demographics of mass shooters are consistent with the racial composition of the broader US population.

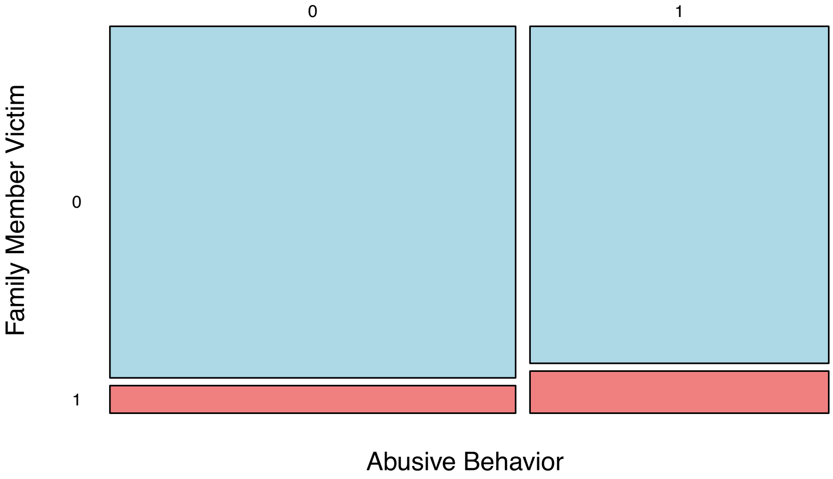
Using the code provided in the appendix, we calculated the proportion of each racial group in the shooter data set and compared it to the corresponding proportions in the US population obtained from the US Census Bureau. The results of this comparison, along with the null hypothesis test results, are summarized in Table 1 below.

Table 1: Proportion of Racial Groups in Mass Shooter Dataset vs. U.S. Population and Null Hypothesis Results

Race	**U.S. Proportion**	**Database Proportion**	**CI Lower**	**CI Upper**	**Null Hypothesis**
White	0.584	0.523	0.34	0.70	Accepted (Consistent)
Black	0.137	0.207	0.08	0.36	Accepted (Consistent)
Latinx	0.195	0.083	0.00	0.20	Accepted (Consistent)
Asian	0.064	0.067	0.00	0.16	Accepted (Consistent)
Native American	0.013	0.016	0.00	0.08	Accepted (Consistent)

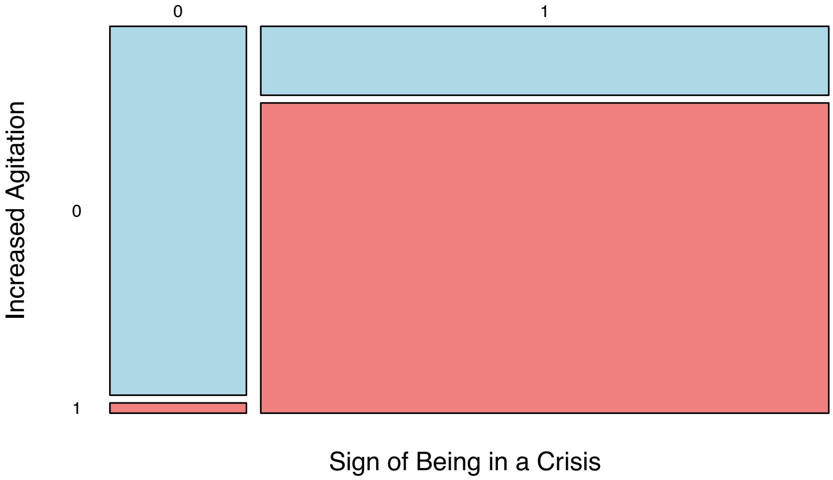
As shown in Table 1, the null hypothesis was accepted for all racial groups, indicating no statistically significant differences. This suggests that the racial composition of mass shooters does not differ significantly from the broader US population.

Independance testing 3.1: Family Member Victim vs Abusive Behavior



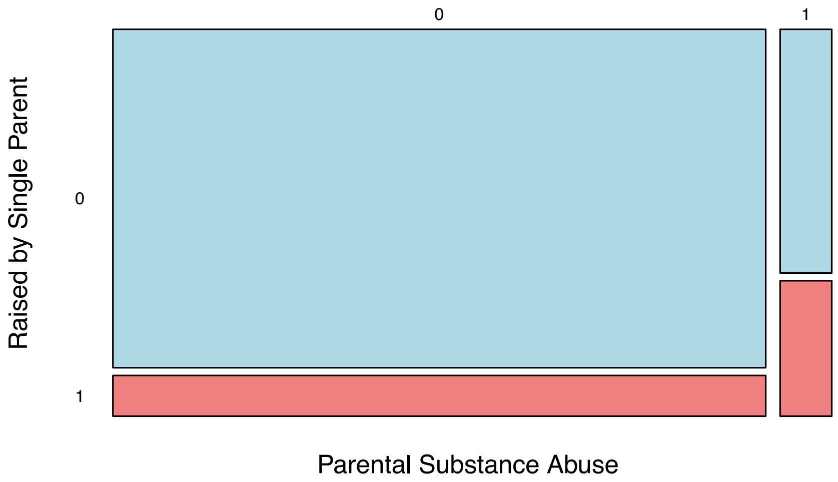
Most individuals without a history of abusive behavior did not target family members, as indicated by the large blue area. However, among individuals with a history of abusive behavior, there is a slight increase in the proportion of family member victims. This points to a potential relationship between abusive behavior and family-targeted violence, though such incidents remain infrequent within the broader data set.

Independance testing 3.2: Signs of Being in Crisis vs. Increased Agitation



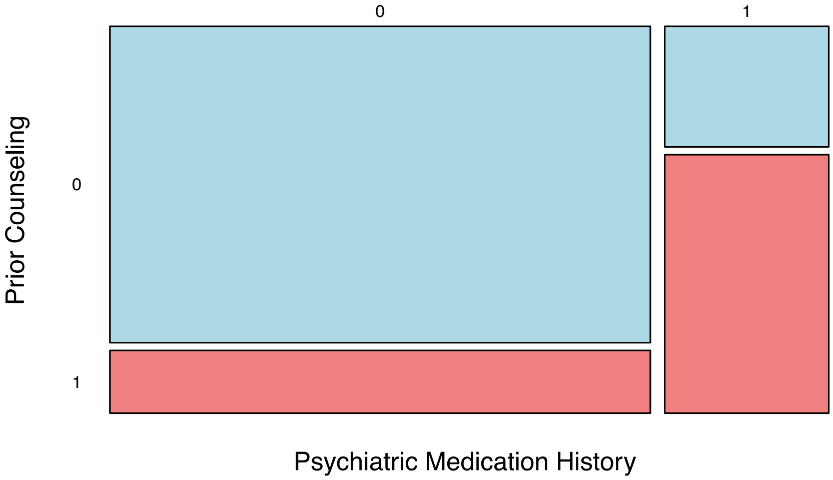
This mosaic plot shows a strong overlap between increased agitation and sign of being in a crisis. Individuals displaying signs of being in crisis are expected to have a higher proportion of increased agitation compared to those not in crisis. This relationship highlights that behavioral agitation may be a key indicator of an underlying crisis, enforcing the need to identify and address such signs early as part of intervention strategies.

Independance testing 3.2: Parental Substance Abuse vs. Raise by Single Parent



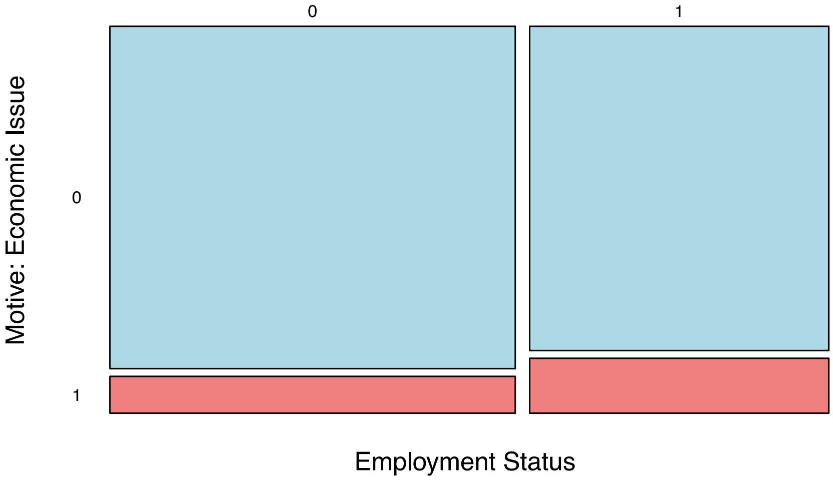
The majority of individuals were not raised by single parents and did not have a history of parental substance abuse, as indicated by the dominant blue area. However, among those raised by single parents, there is a noticeable overlap with parental substance abuse, suggesting a potential link. This highlights that parental substance abuse may contribute to or coexist with single-parent households, though such cases remain less frequent overall.

Independance testing 3.3: Psychiatric Medication History vs. Prior Counseling



Most individuals without a history of psychiatric medication also did not receive prior counseling, indicating limited access or usage of mental health services. However, among those with psychiatric medication history, a significant proportion had received prior counseling, showing overlap between these two factors. This suggests that individuals with psychiatric treatment histories are more likely to engage with counseling services, though counseling access remains underutilized overall.

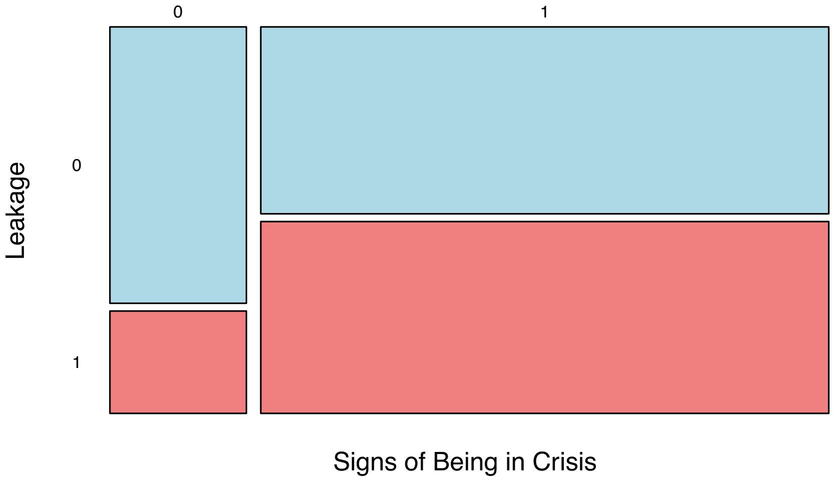
Independance testing 3.4: Employment Status vs. Motive: Economic Issue



The majority of individuals, regardless of employment status, did not cite economic issues as a motive for their actions. For those motivated by economic issues, the proportions between employed and unemployed groups appear similar but remain relatively small overall. This suggests that while economic issues may play a role in some cases, they are not a dominant motive, even among unemployed individuals.

Independance testing 3.5: Signs of Being in Crisis vs. Leakage

*Leakage: warnings or threats before the incident



A significant portion of individuals who showed signs of being in crisis also exhibited leakage, such as warnings or threats, before their actions. Conversely, those who did not show signs of crisis rarely exhibited leakage. This highlights that signs of being in crisis can serve as an important early indicator of potential threats, emphasizing the need for timely intervention and monitoring.

Conclusion

My analysis of the Violence Project's mass shooter database reveals patterns and relationships that offer deeper insight into the dynamics of mass shootings in the United States. Key temporal trends highlight a general increase in the frequency and casualties of mass shooting incidents over the years, with notable spikes in 1999 and 2018. Possible explanations for these anomalies relate to major events, such as the Columbine High School shooting and changes in firearm permit laws. Conversely, a significant decline in 2020 is linked to COVID-19 lock downs, which reduced public gatherings. As seen in our location pie chart, public spaces are most vulnerable to these attacks.

Demographic analysis shows that most shooters are aged between 20 and 40, with the mean age being around 33. The race distribution of the shooters supports the null hypothesis that the racial composition of mass shooters is representative of the general population.

The project further investigates several behavioral and contextual factors using mosaic plots. Notable findings include the association between a history of abusive behavior and the likelihood of having family members as victims, as well as the link between signs of being in crisis and the occurrence of leakage. Additionally, it was interesting to observe that economic motives were rarely indicated as a driving factor for mass shootings, even among unemployed individuals. This suggests that economic distress is not a driver for such acts of violence (at least in this dataset which excludes gang violence). These insights underscore the importance of early detection of behavioral signals, which could support intervention efforts to prevent future incidents.

Appendix: Code Blocks

Pie Chart of Location Distribution

```r
library(ggplot2)
data <- read.csv("mass_shooter_database.csv")
data$Location <- as.numeric(data$Location)
data_clean <- data[!is.na(data$Location) & data$Location %in% 0:10, ]

location_labels <- c(
  "K-12 School", "College/University", "Government Building", "House of Worship",
  "Retail", "Restaurant/Bar/Nightclub", "Office", "Place of Residence",
  "Outdoors", "Warehouse/Factory", "Post Office"
)

data_clean$Location <- factor(data_clean$Location, levels = 0:10, labels =
location_labels)
location_summary <- as.data.frame(table(data_clean$Location))
colnames(location_summary) <- c("Location", "Count")
ggplot(location_summary, aes(x = "", y = Count, fill = Location)) +
  geom_bar(stat = "identity", width = 1, color = NA) +
  coord_polar("y", start = 0) +
  labs(fill = "Location") +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.ticks = element_blank(), axis.title =
  element_blank()) +
  scale_fill_brewer(palette = "Set3")
```

Pie chart of Gender Identity & Immigrant Status

```r
# Function to create pie chart for a binary column!
create_pie_chart <- function(column, labels) {
  summary_data <- as.data.frame(table(column))
  colnames(summary_data) <- c("Category", "Count")
    summary_data$Category <- factor(summary_data$Category, levels = c(0, 1),
    labels = labels)
    ggplot(summary_data, aes(x = "", y = Count, fill = Category)) +
    geom_bar(stat = "identity", width = 1, color = NA) +
    coord_polar("y", start = 0) +
    labs(fill = "Category") +
    theme_minimal() +
    theme(axis.text.x = element_blank(), axis.ticks = element_blank(),
    axis.title = element_blank()) +
    scale_fill_brewer(palette = "Set3")
}

# Gender Pie Chart
create_pie_chart(data$Gender, c("Male", "Female"))
```

```r
# Immigration Pie Chart
create_pie_chart(data$Immigrant, c("No", "Yes"))
```

1.1 Temporal Trend: Number of Incidents

```r
library(ggplot2)
library(dplyr)

data <- read.csv("mass_shooter_database.csv")
data$Year <- as.numeric(data$Year)
data <- data %>% filter(!is.na(Year))

# Aggregate data to count the number of incidents per year
yearly_data <- data %>%
  group_by(Year) %>%
  summarize(Incidents = n(), .groups = 'drop')

model <- lm(Incidents ~ Year, data = yearly_data)
yearly_data$Predicted <- predict(model, newdata = yearly_data)
yearly_data$Residuals <- yearly_data$Incidents - yearly_data$Predicted

# Identify outliers
threshold <- 2 * sd(yearly_data$Residuals, na.rm = TRUE)
outliers <- yearly_data[abs(yearly_data$Residuals) > threshold, ]

# Plot scatterplot with regression line, outliers, and labels for outliers
p <- ggplot(yearly_data, aes(x = Year, y = Incidents)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_line(aes(y = Predicted), color = "red", size = 1) +
  geom_point(data = outliers, aes(x = Year, y = Incidents),
             color = "orange", size = 3, shape = 21, fill = "orange") +
  geom_text(data = outliers, aes(x = Year, y = Incidents, label = Year),
            vjust = -1, size = 3, color = "orange") + # Outlier labels
  labs(title = "Mass Shooting Incidents Over Time",
       x = "Year",
       y = "Number of Incidents") +
  theme_minimal()

print(p)
```

1.2 Temporal Trend: Casualties

```r
library(ggplot2)
library(dplyr)

data$Year <- as.numeric(data$Year)
data$Number.Killed <- as.numeric(data$Number.Killed)
data <- data %>% filter(!is.na(Year), !is.na(Number.Killed))
```

```r
# Aggregate data to calculate total number of casualties per year
yearly_data <- data %>%
  group_by(Year) %>%
  summarize(TotalCasualties = sum(Number.Killed, na.rm = TRUE), .groups = 'drop')

model <- lm(TotalCasualties ~ Year, data = yearly_data)
yearly_data$Predicted <- predict(model, newdata = yearly_data)
yearly_data$Residuals <- yearly_data$TotalCasualties - yearly_data$Predicted

# Identify outliers
threshold <- 2 * sd(yearly_data$Residuals, na.rm = TRUE)
outliers <- yearly_data[abs(yearly_data$Residuals) > threshold, ]

# Plot scatterplot with regression line, outliers, and labels for outliers
p <- ggplot(yearly_data, aes(x = Year, y = TotalCasualties)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_line(aes(y = Predicted), color = "red", size = 1) +
  geom_point(data = outliers, aes(x = Year, y = TotalCasualties),
             color = "orange", size = 3, shape = 21, fill = "orange") +
  geom_text(data = outliers, aes(x = Year, y = TotalCasualties, label = Year),
            vjust = -1, size = 3, color = "orange") +
  labs(title = "Mass Shooting Casualties Over Time",
       x = "Year",
       y = "Total Casualties") +
  theme_minimal()

print(p)
```

2.1 Shooter Demographics: Age

```r
library(ggplot2)
library(dplyr)

data <- data %>% filter(!is.na(Age))
p <- ggplot(data, aes(x = Age)) +
  geom_histogram(fill = "lightblue", color = "black", alpha = 0.7, bins = 10) +
  labs(title = "Age Distribution of Mass Shooters", x = "Age", y = "Frequency") +
  theme_minimal()

print(p)
```

Independance testing 3.1: Family Member Victim vs Abusive Behavior

```r
x <- "Signs.of.Being.in.Crisis"
y <- "Increased.Agitation"

# Select and clean the relevant binary columns dynamically (OpenAI, 2024)
data_clean <- data %>%
  filter(!is.na(!!sym(x)), !is.na(!!sym(y))) %>%
  mutate(
    !!sym(x) := as.factor(!!sym(x)),
    !!sym(y) := as.factor(!!sym(y))
  )

tab <- table(data_clean[[x]], data_clean[[y]])

# Generate the mosaic plot
mosaicplot(tab,
           main = paste("Mosaic Plot: Sign of Being in a Crisis vs Increased Agitation"),
           xlab = "Sign of Being in a Crisis",
           ylab = "Increased Agitation",
           col = c("lightblue", "lightcoral"),
           las = 1)
```

Note: Every else in section 3 follows the same code with changes in x & y

Generating a map plotting shotoings as points

(OpenAI, 2024)

```
library(ggplot2)
library(dplyr)
library(readr)

data <- read.csv("mass_shooter_database.csv")
library(leaflet)

data <- read.csv("mass_shooter_database.csv")
data <- data[!is.na(data$Latitude) & !is.na(data$Longitude), ]

# Create the map
leaflet(data = data) %>%
 # Using OpenStreetMap as a base layer
  addProviderTiles(providers$OpenStreetMap) %>%
  addCircleMarkers(
    ~Longitude, ~Latitude,
    radius = 4, color = "red",
    opacity = 0.8
  ) %>%
  setView(lng = mean(data$Longitude, na.rm = TRUE),
          lat = mean(data$Latitude, na.rm = TRUE),
          zoom = 4)
```

Function used to calculate proportion of races in database

```
# Function to calculate the proportion of a specific value in a given column
calculate_proportion <- function(data, column_name, value) {
  # Filter out NA values and 0s from the column
  non_na_values <- data %>%
    filter(!is.na(.data[[column_name]])) %>%
    pull(.data[[column_name]])

  # Count the occurrences of the specified value in the column
  count_value <- sum(non_na_values == value)

  # Calculate the total number of non-NA and non-0 values in the column
  total_count <- length(non_na_values)

  # Calculate the proportion of the specified value
  proportion <- count_value / total_count

  return(proportion)
}

# Example: Proportion of shooters who identified as Christian (Religion = 1)
value <- 1
column <- 'Race'
proportion_result <- calculate_proportion(data, column, value)

cat("Proportion of value", value, "in the" , column, "column:", round(proportion_result, 3))
```

Code used to do null hypothesis testing

Race: White = 0 Black = 1 Latinx = 2 Asian = 3 Middle Eastern = 4 Native American = 5

```
knitr::opts_chunk$set(echo = TRUE)

# Now doing the sampling test to see if shooter race proportion is significant or
# consistent with US population proportions for the race

# Function to generate confidence intervals for proportions
conf_interval <- function(m, n, p) {
  storage <- rep(NA, m)

  for (i in 1:m) {
    s <- sample(c(0, 1), n, replace = TRUE, prob = c(1 - p, p))
    storage[i] <- mean(s)
  }

  return(storage)
}

# Function to test null hypothesis for each race
test_null_hypothesis <- function(m = 10000, n = 50, conf_level = 0.99) {

  race_proportions <- read.csv("race_proportions - Sheet1.csv")

  results <- data.frame(
    Race = character(),
    prop_us = numeric(),
    prop_database = numeric(),
    CI_Lower = numeric(),
    CI_Upper = numeric(),
    Null_Hypothesis = character()
  )

  for (i in 1:nrow(race_proportions)) {
    race <- race_proportions$race[i]
    p_us <- race_proportions$prop_us[i]
    p_database <- race_proportions$prop_database[i]
    answer <- conf_interval(m, n, p_database)
    ci <- quantile(answer, c((1 - conf_level) / 2, 1 - (1 - conf_level) / 2))

    # Determine if the U.S. proportion lies within the confidence interval
    if (p_us >= ci[1] && p_us <= ci[2]) {
      null_hypothesis <- "Accepted (Consistent)"
    } else {
      null_hypothesis <- "Rejected (Not Consistent)"
    }

    # Store the results
    results <- rbind(results, data.frame(
      Race = race,
      prop_us = p_us,
      prop_database = p_database,
      CI_Lower = ci[1],
      CI_Upper = ci[2],
      Null_Hypothesis = null_hypothesis
    ))
  }

  return(results)
}

results <- test_null_hypothesis(m = 10000, n = 50, conf_level = 0.99)
print(results)
```

```
##                    Race prop_us prop_database CI_Lower CI_Upper
## 0.5%              White   0.584         0.523     0.34     0.70
## 0.5%1            Black    0.137         0.207     0.08     0.36
## 0.5%2           Latinx    0.195         0.083     0.00     0.20
## 0.5%3             Asian   0.064         0.067     0.00     0.18
## 0.5%4 Native American    0.013         0.016     0.00     0.08
##             Null_Hypothesis
## 0.5%  Accepted (Consistent)
## 0.5%1 Accepted (Consistent)
## 0.5%2 Accepted (Consistent)
## 0.5%3 Accepted (Consistent)
## 0.5%4 Accepted (Consistent)
```

References

The Violence Project: The Violence Project. (n.d.). *Mass shooter database*. Retrieved from https://www.theviolenceproject.org/mass-shooter-database/

Brady Law 1999: U.S. Department of Justice. (1999). *Brady Handgun Violence Prevention Act 1999*. Retrieved from https://www.justice.gov/archive/opd/

Concealed Carry Reciprocity Act: Wikipedia contributors. (n.d.). *Concealed Carry Reciprocity Act*. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Concealed_Carry_Reciprocity_Act

Concealed Nation SSRN Paper: Concealed Nation. (2018). *Concealed Carry Reciprocity Act: A policy analysis*. Retrieved from https://concealednation.org/wp-content/uploads/2018/08/SSRN-id3233904.pdf

U.S. Census Quick Facts (2023): United States Census Bureau. (2023). *QuickFacts: United States*. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045223

Religion Census 2020: Public Religion Research Institute (PRRI). (2020). *2020 Census of American Religion*. Retrieved from https://www.prri.org/research/2020-census-of-american-religion/

ChatGPT: OpenAI. (2024). *ChatGPT (Dec 2024 version)* [AI language model]. Available from https://openai.com

Stack Overflow: Stack Overflow. (n.d.). *Stack Overflow: Developer community and Q&A*. Retrieved from https://stackoverflow.com