US School Shooting Report

Jiaqi Sun

1. Introduction

In the past few decades, there have been at least 376 schools in the United States have experienced a shooting. and more than 348 thousand students have suffered from gun violence. It's important to note that students who have not been killed or injured also suffer from psychological trauma. These tragedies vary greatly in nature and the mournful can be difficult to express. In this report we use two datasets. The first dataset was collected since 1999 to 2023 and published by Washington post in 2023 and can be download from Washington post(2023). This dataset contains more information about the shooting but there are many missing data in shooter's race. And second dataset was collected during 2009 to 2018 and can be download from Kanggle. Because dataset2 has less NA value in race columns we will join the two table by school ID to create more detailed dataset.

The purpose of this report is to identify the major factors that contribute to the number of victims in school shootings and how they influence these tragedies. Additionally, we will provide advice on how to prevent these events from occurring. The response variable in this report will be "casualties", representing the number of people killed and injured in the shooting (excluding the shooter). Other predictor variables will be shown in the explanatory variables table.

2. Data preprocessing

Step 1

As a part of the data cleaning process, dataset1 has many empty cells we used the code $df[df == ””]$ to filter any empty cells in the dataset1. We then replaced these empty cells with the value 'NA' to enable better processing of the data. To obtain a dataset with more information, we first convert two date columns to formal date formate using $as.Date()$. Next, we used the $merge()$ function to perform a left join between dataset1 and dataset2, based on school id and date. We then used the $ifelse()$ function to fill in the race columns and standardized all race values[1] to the same format. By doing so we can combine the information from the two 'race' columns and fill in some missing values. To avoid duplicating columns, we dropped one of the 'race' columns after the merge process.

```
suppressMessages(library(dplyr))
suppressMessages(library(tidyverse))
suppressMessages(library(xtable))
suppressMessages(library(caTools))
suppressMessages(library(randomForest))
suppressMessages(library(Metrics))

df1<-read.csv("dataset1.csv")
```

[1] The documentation for dataset 1 did not provide clarification about the meaning of the "m" race. Therefore, we set the "m" race to NA to indicate that it is undefined.

```
df1<-df1[,c(2,3,6,8:21,31:35)]
df1[df1==""]<-NA
df1$date<-as.Date(df1$date,"%m/%d/%Y")

df2<-read.csv("dataset2.csv")
df2<-df2[,c(10,14,16)]
df2$date<-as.Date(df2$date,"%d-%b-%y")

#merge two database
df<- merge(df1,df2,by.x=c("nces_school_id","date"),by.y=c("NCESSCH","date"),all.x = TRUE)

#merge race columns
df$race<-ifelse(is.na(df$race),df$race_ethnicity_shooter1,df$race)

# See how many rows we filled
sum(!is.na(df1$race_ethnicity_shooter1))
```

```
## [1] 145
```

```
sum(!is.na(df2$race))
```

```
## [1] 180
```

```
sum(!is.na(df$race))
```

```
## [1] 159
```

```
#Replace the values
df$race<-ifelse(df$race == "w","WHITE",
                ifelse(df$race == "b","BLACK",
                  ifelse(df$race == "h","HISP",
                    ifelse(df$race == "a","Asian",
                      ifelse(df$race == "ai","Amecian indian",
                       ifelse(df$race == "m",NA,df$race))))))
# Remove replication column
df<- df[,-17]


#remove NA value
df<-na.omit(df)
```

Result

Prior to merging, dataset1 have 145 rows with filled race information, while dataset2 have 180 rows with filled race information. After merging the two race columns, we ended up with a total of 159 rows with filled race information.

Step2

We began by converting time to a 24-hour format using $strftime()$ and then converted it to a date-time object using $as.POSIXct()$. To further process the time variable, we converted it into a factor variable with three levels, “morning”, “midday”, and “afternoon”, using the $cut()$ function.

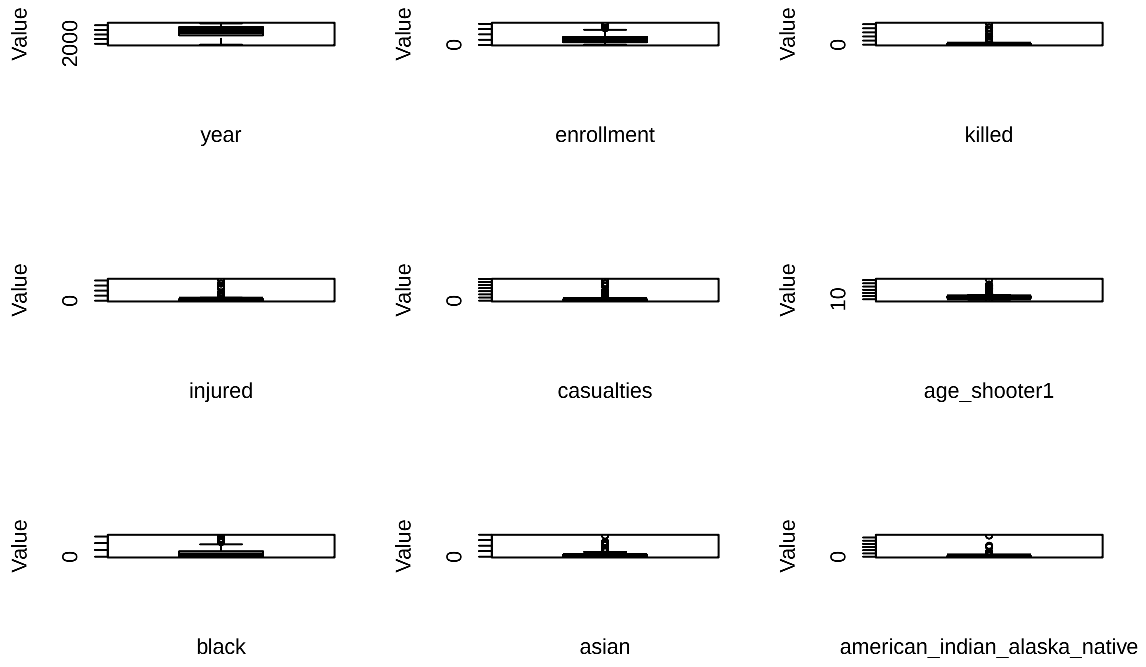
```
#Convert time to 24 hour format
df$time <- strftime(strptime(df$time,"%I:%M %p"),"%H:%M")
df$time <- as.POSIXct(df$time,format = "%H:%M")


#Categorize time variable
df$time <- cut(df$time,
               breaks = c(as.POSIXct("06:00", format = "%H:%M"),
                                as.POSIXct("11:00", format = "%H:%M"),
                                as.POSIXct("14:00", format = "%H:%M"),
                                as.POSIXct("20:00", format = "%H:%M")),
                labels = c("morning", "midday","afternoon"))
```

Step 3

In this step, we identify outliers by using boxplot. After ploting the data we found that the columns “Asian”, ““american_indian_alaska_native” contains outliers.

```
# Identify outliers using boxplot
# Select only the numeric variables
numeric_vars <- df[, sapply(df, is.numeric)]
# Create a boxplot for each numeric variable
par(mfrow=c(3,3))
col_name<-colnames(numeric_vars)
  for (col in col_name){
    boxplot(numeric_vars[col],
          xlab = {{col}}, ylab = "Value")
  }
```



```
max_asian<-df[which.max(df$asian),]
max_ai<-df[which.max(df$american_indian_alaska_native),]
```

Figure 1. Box plot for all numeric variables

Result

It is unnecessary to remove these outliers since the school are located in the districts with higher proportion of certain races, so it is reasonable for these school to have higher relative number of students from those racial backgrounds. Removing these outliers may result in misleading and inaccurate conclusion.

Step 4

To improve the data quality, we began by inspecting the data type of each column using the function $str(df)$ or $typeof()$ for individual columns. We then used functions $as.integer()$ and $as.factor()$ to convert columns. Next,we removed all the missing values from the dataset using $no.omit$. To ensure that all missing values were successfully removed, we can use the function $colSums(is.na(df))$ Where $is.na()$ Identifies any missing value as $TRUE$ and non-missing values as $FALSE$, and then sums the number of $TRUE$ values for each column(with each $TRUE$ value counted as 1) using $colSums()$.

```
str(df)

df$white<-as.integer(df$white)
df$hispanic<-as.integer(df$hispanic)

# convert all categorical into factor
df[]<-lapply(df, function(x) {
  if (is.character(x)) {
    return(as.factor(x))
  } else {
    return(x)
  }
})

df<-na.omit(df)
colSums(is.na(df))
```

3. Main Data Analysis

3.1 Find out the most killed/casualties shooting

We can determine the shooting event with the highest number of killed by applying the $which.max()$ function to the “killed” column. This function returns the index of the row that contains the highest value of killed. Once we have this information, we can use the $xtable$ package which provides functions for creating formatted tables from data frams of matrix in R and output as LaTex or HTML. The tables allow us to easily present the key details of the event, such as the number of killed and other relevant information.

```
cas_table<-xtable(df[which.max(df$casualties),],caption = "\\label{tab:tab1}Shooting even with the max
print(cas_table, include.rownames = FALSE, floating = FALSE)
```

```r
kill_table<-xtable(df[which.max(df$killed),],caption = "\\label{tab:tab2}Shooting even with the max kil
print(kill_table, include.rownames = FALSE, floating = FALSE)
```

Results and Analysis

Table1 (Maximum Number of Casualties in Shooting Incidents)

Nces_School_id	**School_name**	**School Type**	**City**	**State**
080480000707	Columbine High School	public	Littleton	Colorado
Date	**Year**	**Time**	**Day_Of_Week**	
4/20/1999	1999	11:19 AM	Tuesday	
Enrollment	**Killed**	**Injured**	**Casualties**	**Shooting Type**
1965	13	21	34	Indiscriminate
Age_shooter1	**Gender_shooter1**	**Race**		
18	male	white		
White	**Black**	**Hispanic**	**Asian**	**American indian**
1783	16	112	42	12

Table2 (Maximum Number of Killed in Shooting Incidents)

Nces_School_id	**School_name**	**School Type**	**City**	**State**
090291000617	Sandy Hook Elementary School	public	Newtown	Connecticut
Date	**Year**	**Time**	**Day_Of_Week**	
12/14/2012	2012	9:35 AM	Friday	
Enrollment	**Killed**	**Injured**	**Casualties**	**Shooting Type**
454	26	2	28	indiscriminate
Age_shooter1	**Gender_shooter1**	**Race**		
20	male	white		
White	**Black**	**Hispanic**	**Asian**	**American indian**
389	15	22	23	0

Result Based on the above table we can observe some common between the two shooting incidents. Both incidents happened in the morning and involved indiscriminate shooting. The shooters in both cases were young white males. Additionally, both schools had higher proportion of white students. From these observations, we can form a hypothesis that the number of killed and casualties may related to several factors, including the shooting type, the time of the day, the age, gender, and race of the shooter, as well as the racial composition of the school.

3.2 Testing our hypothesis

3.2.1 Fitting linear regression model

Methods

To test the hypothesis that the number of casualties in shooting incidents is related to various factors, we can start by fitting a linear model using the lm() function in R. In this model, we used the number of casualties as the response variable, and include the following independent variables: time, shooting type, age of the shooter, gender of the shooter, race, and the number of white students.

After fitting the linear model, we can generate a summary table to identify which variables are statistically significant. This table will provide us with information on the relationship between these variables and the number of casualties in shooting incidents, allowing us to better understand the factors that contribute to the severity of such incidents.

```r
ts_df<-df[,c(5,10,13,14,15,16,17,22)]

model<-lm(casualties~ .,ts_df)
summary(model)
xtable(summary(model),caption = "\\label{tab:tab3}Summary table for the fitting model")
```

Results and Analysis

From the Table3 we see p value of discriminate shooting type and number of white students are both statistically significant. We can reject the null hypothesis that those two variables are independent from the response variable and conclude that there are evidences of association between the variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.5761	5.4130	0.85	0.3997
timemidday	-0.3231	1.2385	-0.26	0.7947
timeafternoon	1.3600	1.4981	0.91	0.3660
enrollment	-0.0012	0.0010	-1.18	0.2388
shooting_typeaccidental or targeted	-0.4180	4.4883	-0.09	0.9260
shooting_typehostage suicide	-0.7748	6.0210	-0.13	0.8978
shooting_typeindiscriminate	5.1932	1.9812	2.62	0.0100
shooting_typepublic suicide	-0.5513	4.5681	-0.12	0.9042
shooting_typetargeted	0.6022	1.8209	0.33	0.7415
shooting_typetargeted and indiscriminate	3.1258	4.5296	0.69	0.4916
age_shooter1	0.0005	0.0484	0.01	0.9924
gender_shooter1m	-1.1429	2.7160	-0.42	0.6747
white	0.0053	0.0019	2.76	0.0067
raceAsian	-5.2556	7.2371	-0.73	0.4692
raceBLACK	-3.4969	4.3213	-0.81	0.4201
raceHISP	0.3017	4.7687	0.06	0.9497
raceWHITE	-4.9945	4.3104	-1.16	0.2491

Table 3: Summary table for the fitting model

3.2.2 Random Forest

Random Forest is one of the most popular supervised mechine learning technique in Data Analysis invented by Leo Breiman and Adele Cutler (2001). It can handle the both categorical and continuous variables. Random Forest works by building multiple decision tress on different samples and then combining their outputs through through ensemble methods such as bagging, boostrap aggregation and boosting.The majority vote of the individual decision trees is used for classification or regression tasks. In 1996, Leo Breiman introduced bagging method which used to improve the performance of machine learning models by reducing variance and increasing accuracy. It works by generating multiple bootstrap samples from the training set and fitting a model to each sample. The final prediction is obtained by aggregating the individual predictions of all models.

Methods

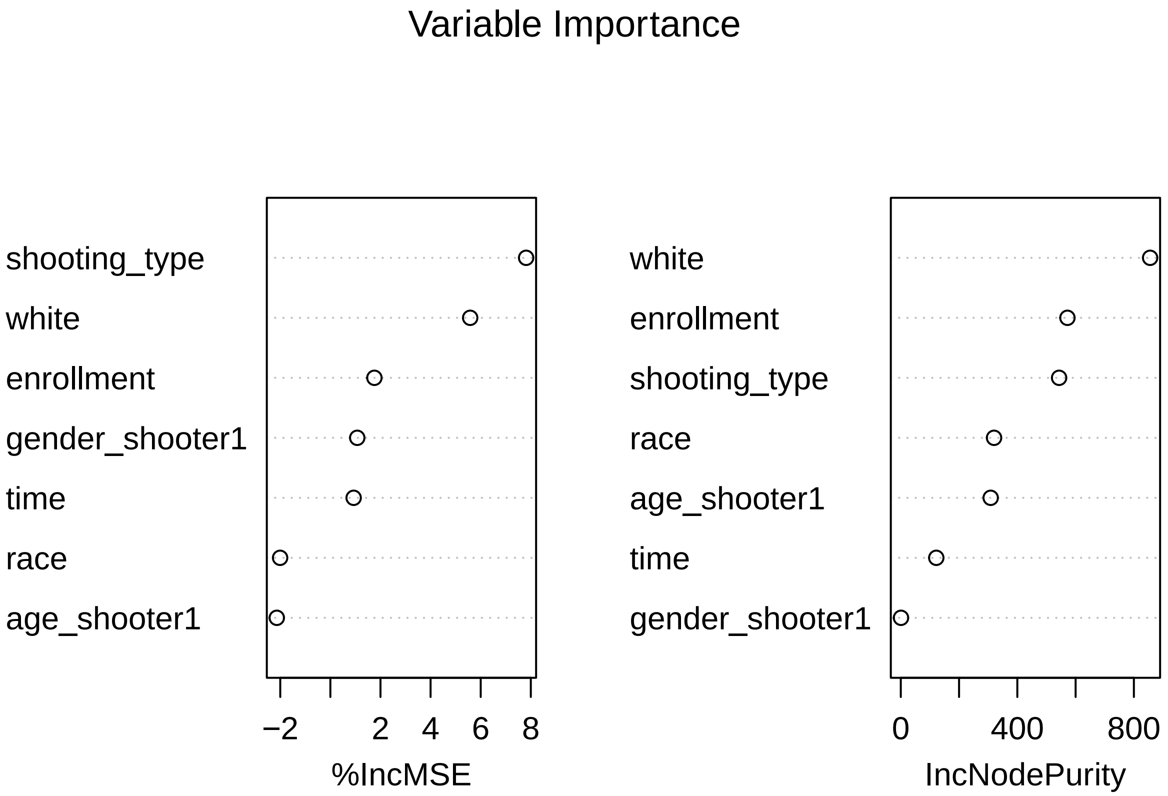
We will use “caret” package to split the data into 70% training and 30% testing data. And randomforest” package to fit randomforest model and plot feature importance using “varImpPlot”.

```r
split_index <- sample.split(ts_df$casualties, SplitRatio = 0.7)
train_data <- subset(ts_df, split_index == TRUE)
test_data <- subset(ts_df, split_index == FALSE)

#Create the random forest model:
```

```r
rf_model <- randomForest( casualties~., data = train_data, importance = TRUE)

#Plot feature importance:
varImpPlot(rf_model, main="Variable Importance")
```



Result and Analysis

The above plot shows variable importance where stands for percent increase in mean squared error. It quantifies how much accuracy would be decreased from removing the predictor variable. The larger %IncMSE value indicate the more important that variable is to the model's performance.

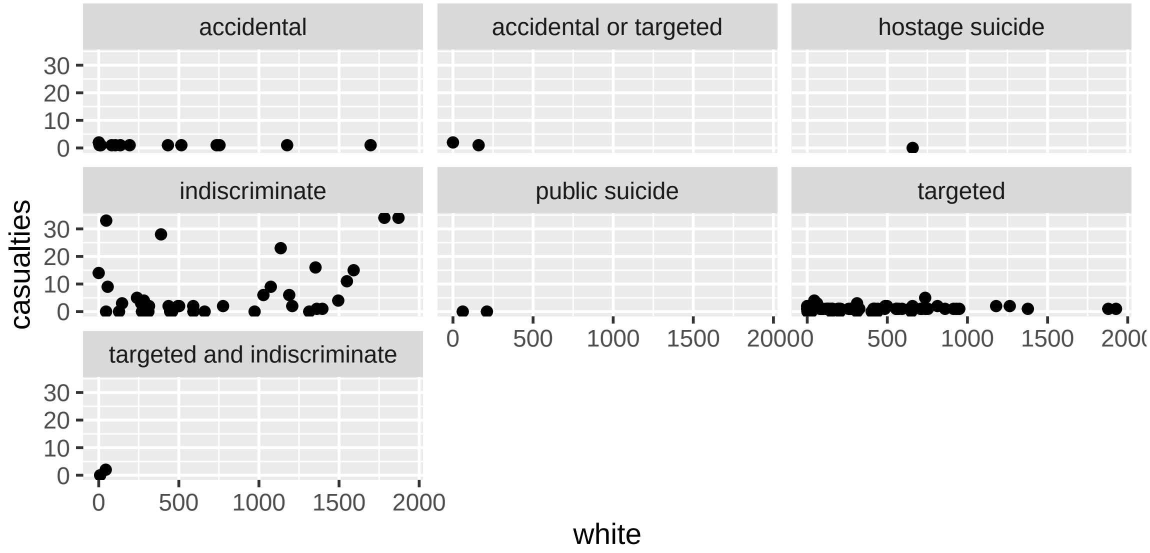
The plot on the right *IncNodePurity* stands for increase in node impurity. It measures how much improvement in the homogeneity of the tree nodes when add the predictor variable in the decision making process. The higher the value for a variable the more important that variable is in the decision-making process in random forest.

Hence, we can infer that the shooting type, proportion of white students, and enrollment have higher scores compared to other variables, indicating that they are the most important variables. This conclusion is consistent with the one we reached by fitting the lm model.

3.2.3 Visualization the relationship between the important variables

Methods As we learned in class facet_wrap can create separate plots for each group.

```r
ggplot(ts_df, aes(white, casualties)) + geom_point() + facet_wrap(~shooting_type)
```



Result and Analysis

Based on the plot, we can see that the casualties tend to be higher for indiscriminate shooting types, while target shooting types often have fewer than 5 casualties. Additionally, there is a cluster of target shooting types at lower white student proportion values.

3.3 Shooting incidents by year

Method To analyze the trend of shooting incidents, we can create a tibble that shows the count of incidents by year. We can then use ggplot to generate a line plot that visualizes the trend over time.

```
year_tab<-df %>%
      group_by(year) %>%
      summarise(
        count =n()
      )
ggplot(year_tab,aes(x=year,y=count))+geom_line()
```

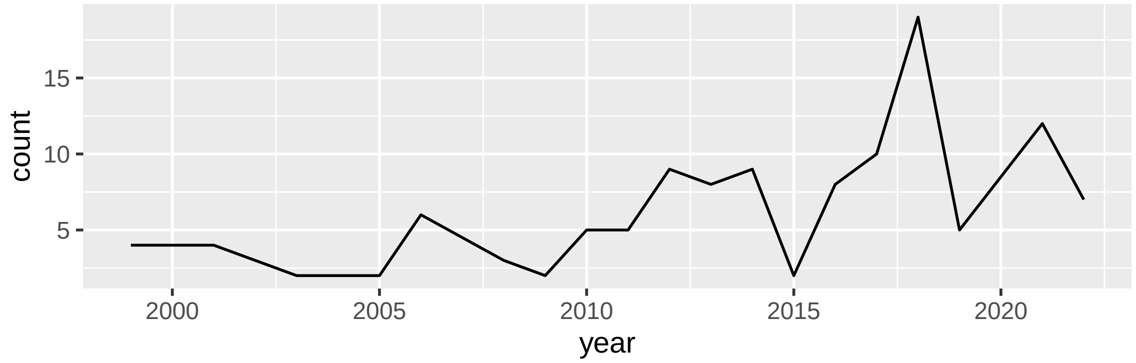


Figure 2. Kernel Density Plot- shooting trend by year

Results and Analysis

Based on the plot, we can observe that there are two peaks in the incidence of shooting events. The first peak occurred from 2016 to 2018, while the second peak occurred during 2021 and 2022.

3.5 Analyzing location variable

Methods

To analyzing location variable such as 'city' and 'state' we plot a map of the United States using 'usmap' package from 'ggplot2' library and used *aggregate*() function from 'dplyr' package to creat a data frame that counts the number of shooting incidents by stats. To visualize the data, we colored the map scaled by number of shooting in each stats, with color 'red' representing high number of shooting incidence and 'white' representing low count.

Results and Analysis

Based on the plot, we can observe that most states have had at least one school shooting incident, with higher incidence in the eastern part of the United States, particularly in the southeastern region. In contrast, the western part of the United States has relatively fewer states with school shooting incidents. Additionally, North California, Texas, California, Alabama, and Florida have relatively higher numbers of shooting incidents compared to other states. Notably, North California has the highest number of school shooting incidents among all states.

```
library(usmap)
us_df <- aggregate(. ~ state, data = df,FUN=length)
names(us_df)[2] <- "count"

plot_usmap(regions = "states",
           data = us_df,
           values = "count") +
  scale_fill_continuous(
    low = "white", high = "red", name = "School Shooting", label = scales::comma
  ) + theme(legend.position = "right")
```

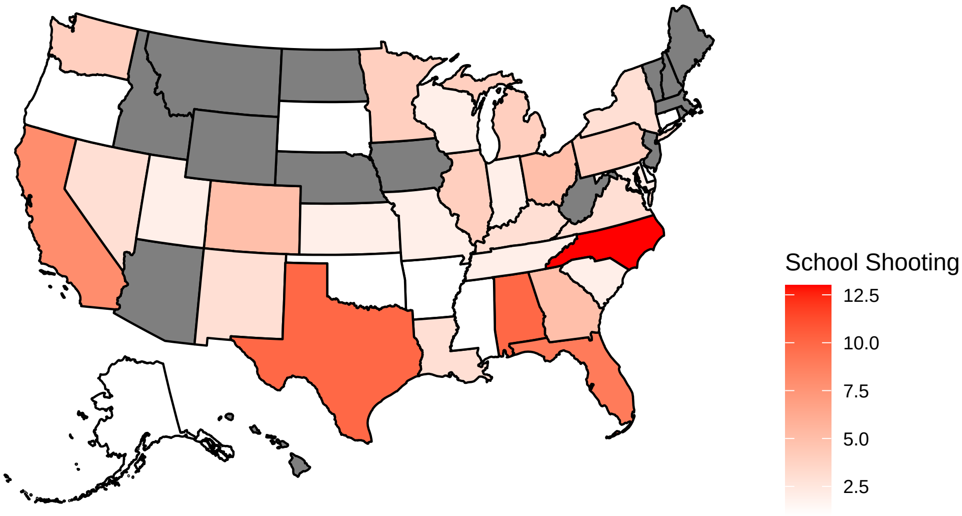


Figure 3. Colored map by shooting count

3.6 Analyzing shooter

Method

The dataset contains information on the shooter's age, gender, and race. We used ggplot to display the frequency of shooting incidents for each age and grouped by gender. The frequency plot provides insight into the age groups most commonly involved in shooting incidents.

```
ggplot(df, aes(age_shooter1, fill = gender_shooter1)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~gender_shooter1, ncol = 1)
```

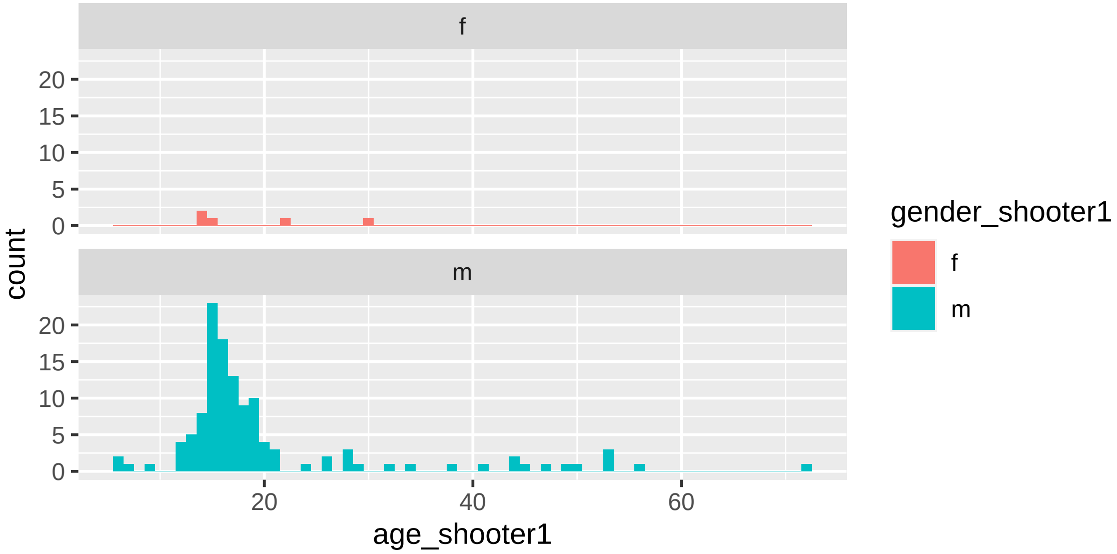


Figure 6. Barplot by shooter age and gender

Results and Analysis

Based on the plot, we can observe that school shooting incidents are more frequent among younger age groups, especially among those aged 15 to 22. Additionally, majority of school shooting incidents are commited by male shooters.

4. Conclusion

After conducting the above analysis, we discovered that the number of casualties is associated with the type of shooting, with indiscriminate shootings being more likely to have higher casualties. Surprisingly, the proportion of white students also played a significant role in shooting incidents. Additionally, we observed two peaks in shooting incidents during the periods of 2016-2018 and 2021-2022.

Using the plot_usmap() function in ggplot2, we created a colored US map which revealed that most states have experienced at least one shooting incident, and that the southeastern region had a higher frequency of shooting incidents.

Upon analyzing the shooters, we discovered that a majority of them are quite young, between the ages of 15 and 22. Additionally, the majority of school shooting incidents were committed by male shooters.

Hence, in order to prevent school shooting incidents, it is crucial to provide psychological support to teenagers, especially those between the ages of 15 to 20. Additionally, it is recommended that schools offer shooting incident simulations and provide corresponding equipment, such as bulletproof vests and emergency shelters, to prevent unnecessary harm and fatalities.

Appendix

Table 4: explanatory variables

Name	Type	description
enrollment	Numeric	Enrollment at school at time of shooting
nces school id	Numeric	National Center for Education Statistics unique school ID
year	Numeric	Year of shooting
school name	Categorical	Name of school
state	Categorical	State where school is located
city	Categorical	City where school is located
killed	Numeric	Number of Number killed in shooting (excludes shooter)
injured	Numeric	Number injured in shooting (excludes shooter)
date	date	Date of shooting
school type	Categorical	Type of school (public or private)
day of week	Numeric	Day of week of shooting
time	Numeric	Approximate time of shooting
shooting type	Categorical	Type of shooting
age shooter1	Numeric	Age of first shooter
gender shooter1	Numeric	Gender of first shooter
race ethnicity shooter1	Categorical	Race or ethnicity of first shooter
white	Numeric	Enrollment of white students at time of shooting
black	Numeric	Enrollment of black students at time of shooting
Asian	Numeric	Enrollment of Asian students at time of shooting
Hispanic	Numeric	Enrollment of hispanic students at time of shooting
american indian alaska	Numeric	Enrollment of Ameican Indian and Alaskan native students at time of shooting

Reference

Aggregate: Compute summary statistics of data subsets. RDocumentation. (n.d.). Retrieved April 17, 2023, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2022). randomForest (Version 4.7-1.1) [Computer software]. Retrieved October 14, 2022, from https://CRAN.R-project.org/package=randomForest

John Woodrow Cox, S. R. (2023, April 3). There have been 377 school shootings since Columbine. The Washington Post. Retrieved April 17, 2023, from https://www.washingtonpost.com/education/interactive/school-shootings-database

Kapadnis, S. (2023, February 26). School Shooting Data. Kaggle. Retrieved April 17, 2023, from https://www.kaggle.com/datasets/sujaykapadnis/school-shooting-data

Random forests - University of California, Berkeley. (2001). Retrieved April 17, 2023, from https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Lorenzo, P. D. (2022, November 12). Mapping US. Mapping in the US. Retrieved April 17, 2023, from https://cran.r-project.org/web/packages/usmap/vignettes/mapping.html

Strptime & strftime in R: 5 example codes (Year, day, hour & time zone). Statistics Globe. (2022, March 16). Retrieved April 17, 2023, from https://statisticsglobe.com/strptime-strftime-r-example