Christina Ling

IST 687 Introduction to Data Science  
April 2018

Final Project

Mass Shootings

# Table of Contents

|  |  |
| --- | --- |
| Introduction | 1 |
| Scope |  |
| Background |  |
| Business Questions | **2** |
| Data Acquisition, Cleansing, Transformation, Munging | **3** |
| Data Acquisition Process |  |
| Data Selection |  |
| Initial Assessment |  |
| Data Selection, Part II |  |
| Data Dictionary |  |
| Data Description |  |
| Data Cleansing |  |
| Discovery and Findings |  |
| Descriptive Statistics | **8** |
| Demographic |  |
| Statistical discovery |  |
| Visuals |  |
| Use of Modeling Techniques | **18** |
| Association Rules |  |
| Text Mining |  |
| Conclusion | **20** |
| Recommendations |  |
| Summary |  |
| Appendix | **21** |

# Introduction

Scope

This report is going to include data from one of Kaggle’s datasets on Mass Shootings. The data set includes all mass shootings in the United States of America from 1966 to 2017. Major data sources in this analysis are from Wikipedia, Mother Jones, Stanford, USA Today and other web sources. A mass shooting is defined any gun violence with three victims and above; however, the Federal Bureau of Investigation (FBI) defines mass shootings with four victims and above.

The object of this report is to determine whether there is a variable or factor that links the shootings. In addition, to provide advance modeling techniques to provide statistical insight.

Background

During the years 1966 to 2017, the United States of America witnessed 398 mass shootings resulting in 1,996 deaths and 2,488 injured. However, the worst mass shooting happened on October 1, 2017 with 59 deaths and 527 injured. On average, there are seven mass shootings per year that claim 39 lives and 48 injured.

# Business Questions

* What variables can be altered or controlled to minimize mass shootings?
* Is it possible to predict mass shootings? What are the variables?
* Is there any correlation with calendar dates?
* Is there a correlation between the variables?
* Do mass shootings cluster around cities or rural areas?
* Do shooters have a history of mental health issues?
* Do mass shootings happen more likely in open or closed locations?
* Are mass shootings better or worst recently compared to the past?

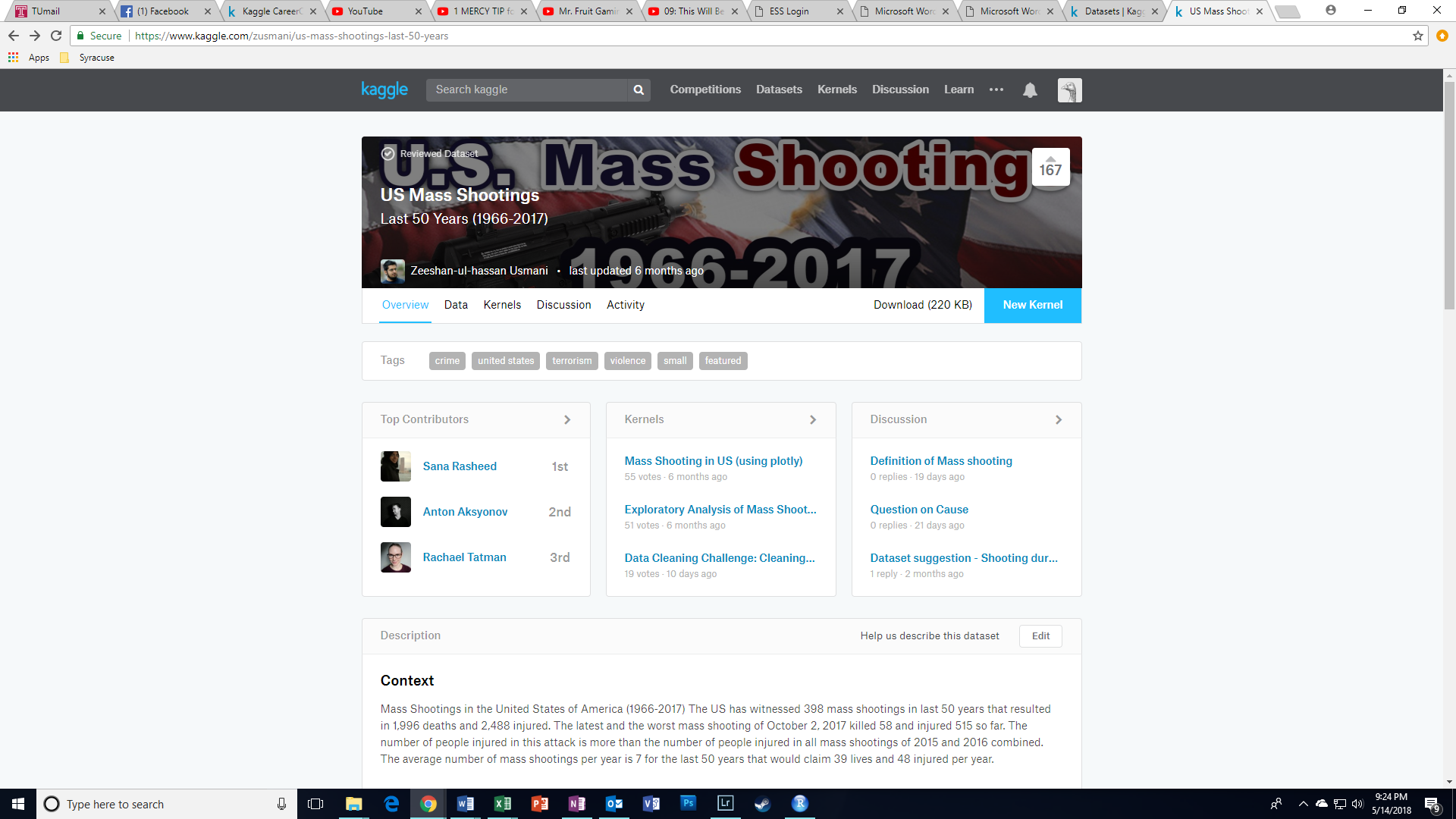
# Data Acquisition, Cleansing, Transformation, Munging

Data Acquisition Process

After exploring different datasets from Kaggle on gun violence and crime category, the US Mass Shootings seemed the most interesting. This reporting will be using the fifth version of the data set. The data set is compiled by user Zeeshan-ul-hassan Usmani who collected data from Wikipedia, Mother Jones, Stanford, USA Today and other web sources.

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

Figure 1. Data Source



Data Selection

Looking at the data set prior to any analysis, this report will be using a subset of the data provided. The analysis will be using the following variables:

* Title
* Location
* Date
* Incident Area
* Target
* Cause
* Summary
* Fatalities
* Injured
* Total Victims
* Age
* Employee (Y/N)
* Employed at
* Mental Health Issues
* Race
* Gender
* Latitude
* Longitude

These variables were chosen to help answer the business questions. Variables such as date, age, mental health issues, race, and gender were chosen to conduct analysis to determine correlations, relationships and to determine predictability. While variables such as title and summary will be used in text mining to determine any unexpected possible variables.

Initial Assessment

The data looks to be reliable, there are no duplicates in the recent version, version 5. Data cleanse will need to be done in R. For location, some used the abbreviated state name while others spell it out. As for Race, data cleansing will be required to make consistent factors (for example: Black vs. African America). Some data may need to be added, since latitude and longitude was provided, some null locations can be discovered with research.

In terms of analysis the data looks like it can provide some useful insight, possible on location, mental health, time of year, number day in between shootings, and severity of the mass shooting. The summary may be able to provide useful insight through text mining. However, the Las Vegas Strip mass shooting will need some more investigation, since it may be considered an outlier.

Data Selection, Part II

After performing the analysis, Table 1. describes where variables were used and the reason why.

Table 1. Data Selection & Reasoning

|  |  |
| --- | --- |
| Location | Nulls were filled by using data given by titles or from longitude and latitude. For the map, location was used to help provide the coordinates using geocode. This ensure that the data is accurate since it directly pools the data. |
| Date | Date was used to perform analysis over time. The date was used to determine days between shootings and to compare months and years. Descriptive statistics were used on the day between shootings. The date was also used to separate them into recent and old categories for the hypothesis test. |
| Open/Closed Location | Open/Closed Location was used to determine where is more likely for mass shootings to occur. This variable was used in the associative model analysis to help determine a pattern between the shootings. |
| Target | A bar chart was created of the top ten targets, since there are many targets, the bar chart was limited to the most frequent. This variable was used to help provide insight on who is likely to be a victim in relation to the shooter. |
| Cause | A bar chart was created of the causes. This chart provided information on what are motives and reasons behind these mass shootings. This allows society to help prepare by looking for early signs. |
| Summary | All summaries were compiled together for a text mining analysis. This determined what are likely words in the story. Additionally, a sentimental analysis was conducted to analyze the media’s responses. |
| Fatalities | As one of the few numeric variables, this help determined the severity of the mass shooting. A percentage of victims killed was also calculated to help compare shootings with few victims to higher victims. |
| Injured | This variable was used for descriptive statistics and for some visuals. As one of the few numeric variables, it was used for some visuals to add information. |
| Total Victims | Total victims was used to help determine how severe the mass shooting was. Since this is the combination of killed (excluding suicides of the shooter) and injured, this determines how many victims were affected by the shooting. |
| Policeman Killed | Policeman killed was used in the descriptive statistics and hypothesis test. In the descriptive statistics, analysis was performed on to see how effective the police force is. In the hypothesis test, it was used to determine if they are more effective recently or not. |
| Age | Age was used in visuals to see if there is a correlation between the age of shooters and the severity of the mass shootings. It was also used to determine if there is an age group that is more likely to become a shooter. This analysis will help determine who to look out for and for warning signs. |
| Mental Health Issues | This variable was used in the associative model to determine patterns between shooters. If there was a high pattern with shooter with mental health issues, then it would provide insight on early signs. |
| Race | Race was used to help determine the correlation between race and severity of the shooting. It was used to provide information on what to look for to help determine who might be a possible shooter. |
| Gender | Gender was used in some visuals to see if one gender of the other is more likely to become a shooter. This helps determine who might be a possible shooter and for profiling. |

Data Dictionary

Table 2. Data Dictionary

|  |  |
| --- | --- |
| Title | A brief distinct string of text that describes the shooting |
| Location | The city and state where the shooting took place |
| Date | When the shooting took place: year-month-day |
| Incident Area | A brief string of text that describe the area the shooting took place |
| Open/Close Location | A factor that describes if the incident area was an open or closed public space |
| Target | The target’s relationship to the shooter |
| Cause | Brief text that describes the shooter’s motive |
| Summary | A string of text that describes the shooting based on the news reports |
| Fatalities | Number of people killed from the shooting. This includes shooters who have killed themselves as a fatality. |
| Injured | Number of people injured from the shooting |
| Total Victims | Combine number of people killed (excluding shooter’s suicide) or injured from the shooting |
| Policeman Killed | Number of policeman on scene that was killed from the shooting (Null means there were no policeman active on the scene) |
| Age | The age of the shooter |
| Employeed (Y/N) | Whether the shooter worked at the location of the shooting |
| Employed at | Where the shooter worked at |
| Mental Health | Whether or not the shoot had any mental health issues. Unclear means the police/investigate has suspicion but not evidence |
| Race | The race of the shooter |
| Gender | Whether the shooter(s) was male or female |
| Latitude | The latitude coordinate of the shooting’s location |
| Longitude | The longitude coordinate of the shooting’s location |

Data Description

From the data set, there are 21 columns and 323 rows. This reported removed the first column (unique identifier).

After cleansing and formatting the data, the data was structured as shown in Table 3.

Table 3. Data Structure

|  |  |
| --- | --- |
| Title | Character |
| Location | Character |
| City | Character |
| State | Character |
| Date | Date |
| Incident Area | Character |
| Open/Closed Location | Character Factor |
| Target | Character |
| Cause | Character |
| Summary | Character |
| Fatalities | Number |
| Injured | Number |
| Total Victims | Number |
| Policeman Killed | Number |
| Age | Character |
| Employeed (Y/N) | Number Factor |
| Employed at | Character |
| Mental Health Issues | Character Factor |
| Race | Character Factor |
| Gender | Character Factor |
| Latitude | Number |
| Longitude | Number |

Data Cleansing

Table 4. describes which columned were data cleansed and how.

Table 4. Data cleansing

|  |  |
| --- | --- |
| Date | Time was remove from the date |
| Open/Closed Location | Capitalization was made to be consistent and changed the structure to a factor |
| Target | Changed all “&” to “+” for consistency and changed capitalization |
| Cause | Changed all null to “unknown” and changed to lower case for consistency |
| Employeed (Y/N) | Replaced nulls with ‘0’ and changed structured to factor |
| Mental Health Issues | Changed to lower case for consistency and change structure to factor |
| Race | Changed some entries to match others for consistency and changed structure to factor |
| Gender | Changed “M/F” to “M+F” for clarification and changed the structure to factor |
| Location | Separated column into city and state. Changed state to full name and all lower case for consistency and analysis. |

Added some data for location. Where data on longitude and latitude is provide but location is missing, we can determine the location based on that information using Google.

Discovery and Findings

Despite the latest news and shootings, there are more less severe mass shootings with the minimum number of victims, three. While people generally think of mass shootings with high victims, injured and killed, there are far more shootings with less than ten victims. Furthermore, despite the recent events, there are less victims in recent shootings (within the last ten years) than before. However, there are more occurrences of mass shootings recently with each decade exceeding the previous.

Mass shootings have occurred more frequently on the coastal regions and central. These are where more cities are located, so mass shootings are more likely to occur in cities than rurales area. In the East Coast, the South, and the Central regions there are the most mass shootings that have occurred.

The biggest common factors for profiling are male shooter between the ages of 16-25 or 35-45 of either the race of White or Black. While there are other shooters, these characteristics are the most common found in the data set. With the most common known motives being robbery, pyscho, and failing exams in the respective order. These shooters targeted party guest the most. According to the associative model, there is many patterns with closed locations, which helps support the party guest target.

# Descriptive Statistics

Demographics

This report includes all mass shootings that happened in the United States, including Alaska and Hawaii. For all map visualizations Alaska and Hawaii will be removed. Mass shootings can happen anywhere in the United States; therefore, all states are included in the data set.

Victims include anyone injured or killed during the mass shooting. Policeman killed on scene will separated from the victims. Any null in policeman kill mean there were no police on the scene during the shooting.

Statistical Discovery

By examining the descriptive statistics in Table 5. and Table 6. some meaningful insight can be concluded. By comparing the maximum and the 1st and 3rd quartile, there is at least one outlier in the data set. After further examination, the Nevada Shooting is considered an outlier. This will be considered during further evaluation of the data set and analysis.

Percentage killed was calculated by dividing fatalities by total victims and converted to a percentage. There are some cases were percentage killed is more than 100 percent. After further examination, these cases include the shooter who killed themselves as a fatality but not a victim. This provides more information about the data and what determines a fatality.

Table 5. Descriptive Statistics of given data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Fatalities | Injured | Total Victims | Policeman Killed |
| Mean | 4.437 | 6.176 | 10.26 | 0.1293 |
| 1st quartile | 1.00 | 1.00 | 4.00 | 0.00 |
| Median | 3.00 | 3.00 | 5.00 | 0.00 |
| 3rd quartile | 5.50 | 5.00 | 9.00 | 0.00 |
| Range | 59 | 527 | 582 | 5 |
| Min | 0 | 0 | 3 | 0 |
| Max | 59 | 527 | 585 | 5 |
| Mode | 1 | 0 | 4 | 0 |
| Variance | 33.4455 | 893.3632 | 1133.151 | 0.3724594 |
| Standard Deviation | 5.783208 | 29.88918 | 33.66231 | 0.6102945 |
| Skewness | 5.0609 | 16.43072 | 15.52947 | 5.504598 |

Figure 2. Histograms of Given Data

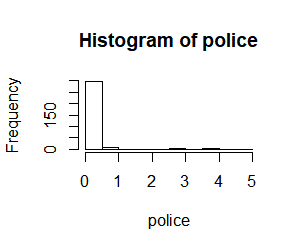
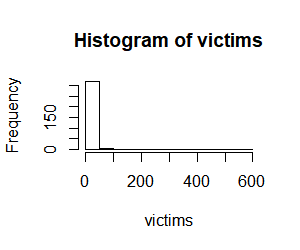
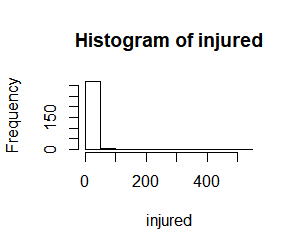
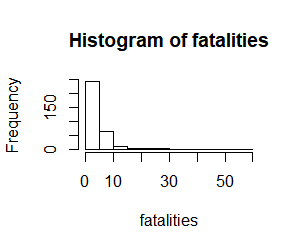
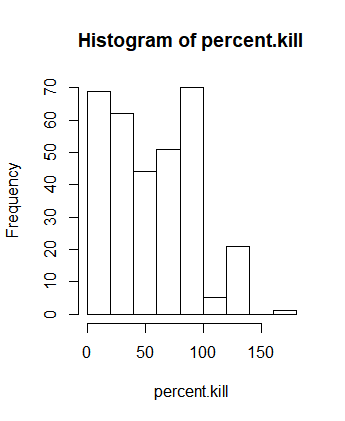
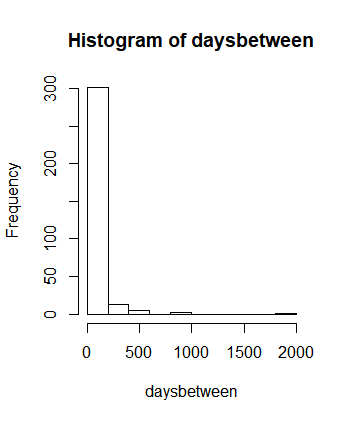


Table 6. Descriptive Statistics for Calculated fields

|  |  |  |
| --- | --- | --- |
|  | Day between Shooting | Percentage Killed |
| Mean | 58.15 | 56.81 |
| 1st quartile | 2.00 | 25.00 |
| Median | 10.50 | 54.55 |
| 3rd quartile | 50.75 | 89.44 |
| Range | 1825 | 166.67 |
| Min | 0.00 | 0.00 |
| Max | 1825 | 166.67 |
| Mode | 0 | 100 |
| Variance | 21,553.58 | 1,522.892 |
| Standard Deviation | 146.8114 | 39.02425 |
| Skewness | 7.144553 | 0.224862 |

Figure 3. Histograms of Calculated fields



Hypothesis tests were conducted to determine if recent shootings are better or worst. Recent shooting is defined as shootings that occurred within the last ten years (2008-2017) while old shootings will be in the remaining 40 years (1966-2007). A two-population hypothesis test was conducted using the population’s mean and standard deviation. This analysis uses α = 0.05 to determine statistically significance. For the calculations, the Nevada shooting was removed as an outlier from the recent data set. Table 7. shows the fields, their z score and p-value.

The two populations are unlikely that they are from the same population, therefore the null hypothesis is rejected for fatalities, injured, and total victims. However, for policeman killed, the p-value is not low, and the null hypothesis was failed to be rejected. While the frequency of mass shootings has increased recently, the number of fatalities, injured, and total victims has decrease.

Table 7. Hypothesis Test Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Fatalities | Injured | Total Victims | Policeman Killed |
| Mean recent | 3.90566 | 4.00 | 7.575472 | 0.1601942 |
| Variance recent | 23.50765 | 51.24171 | 107.7241 | 0.4766517 |
| N recent | 212 | 212 | 212 | 206 |
| Mean old | 4.96363 | 5.636364 | 10.21818 | 0.063636 |
| Variance old | 25.15463 | 41.81151 | 94.41068 | 0.17023 |
| N old | 110 | 110 | 110 | 110 |
| Z-score | -1.82 | -2.08 | -2.26 | 1.55 |
| P-value | 0.0341 | 0.0188 | 0.0119 | 0.9394 |
| Significant? | yes | yes | yes | no |

Visuals

Visuals were created to help draw conclusions about shooters, locations and timing of mass shootings. For some of the visuals, the Nevada shooting was removed as an outlier. All visuals with the outlier removed has been noted on the title of the graph. When the outlier was removed from the data, the visuals show a more overview conclusion.

To explore the mass shooting over time, two line graphs were created; one to display the frequency and severity over the months and another to display the frequency and severity over the years. Figure 4. displays both line graphs. Frequency was calculated by counting the number of shooting in the category (month and year, respectively), while severity was calculated by averaging the number of total victims. Through using averages, frequency does not have an impact on how severe one is compared to another, while total victims might skew it towards higher frequency points.

Used counts for Frequency

Used averages of Total Victims for Severity

Figure 4. Freuency and Severity Time analysis

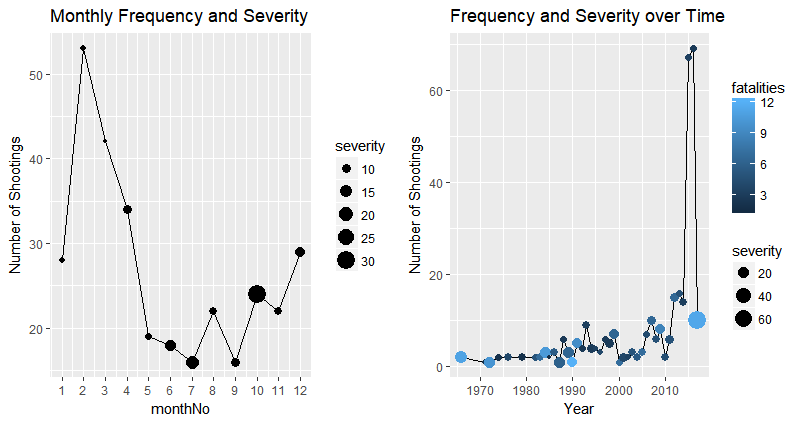
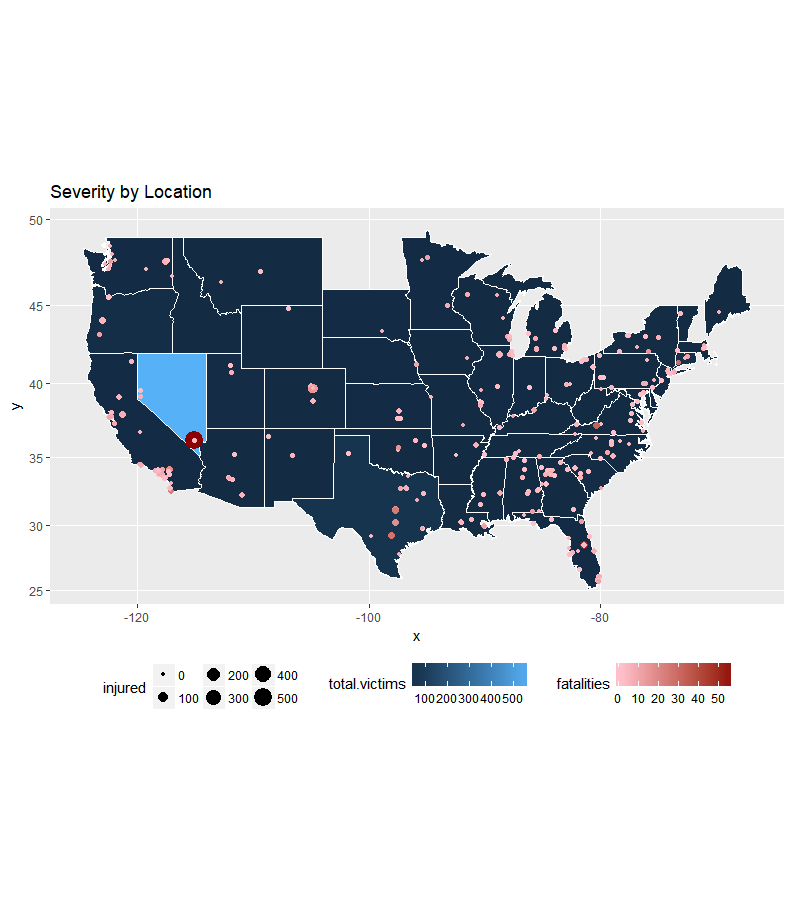


Figure 4. shows that February is the most frequent time for shootings, but October are the most severe. While the summer months have the less frequent shootings but overall more severe than average. The winter and early springs have shootings occurs the most often. The line chart on the right demonstrates that overall mass shootings frequency is hovering between zero to ten shootings a year; however, 2010s the amount of shootings has spiked up drastically. Over each decade there has been more and more mass shootings.

Figure 5. Shootings across the United States



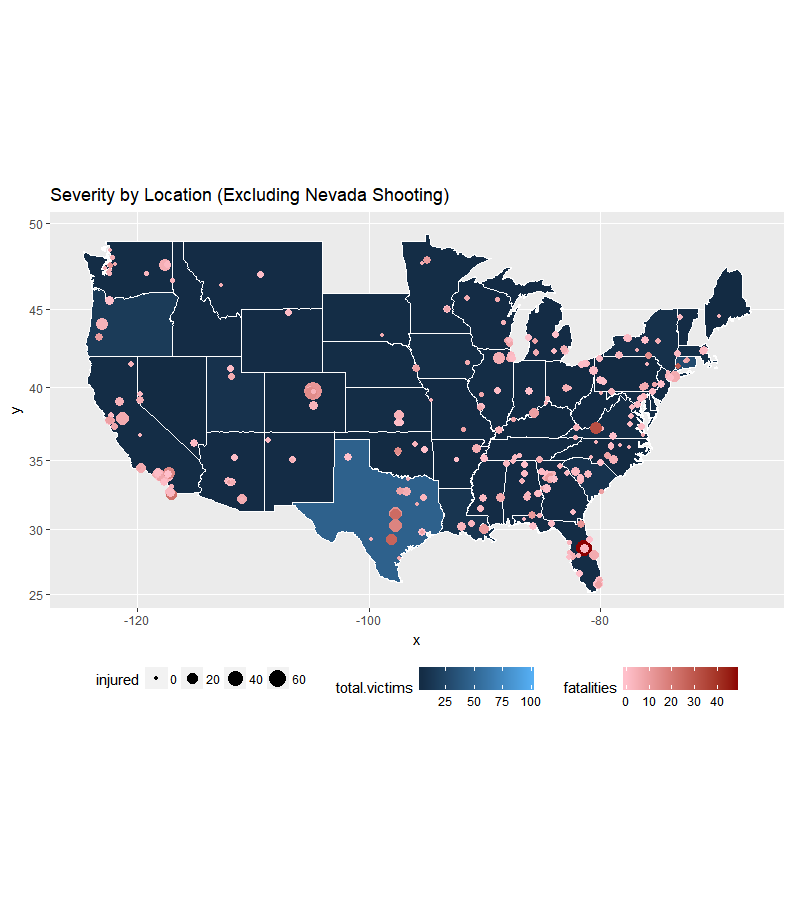
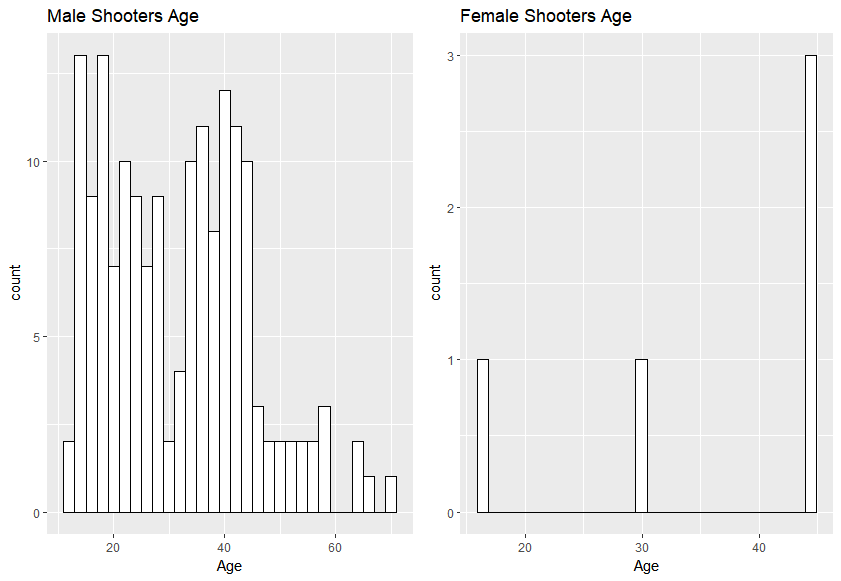


Figure 5. displays a map on where all the shootings within the past 50 years have occurred. Since the Las Vegas, Nevada Shooting skews the data and visual, it was removed for the bottom map to draw more normal conclusions. There was two shootings excluded from both maps, one in Hawaii and the other in Alaska, these were removed from the visual to make it more appealing and since the map does not have Hawaii or Alaska.

Based on the bottom map, more shootings occurred on the costal regions where more cities are located, while there are few in the midwest. The south region also has quite of bit of shootings, with the highest fatalities in Florida, but the high number of victims in Texas. The most shootings occurs in the on the East side (this includes East Coast, South, and Central Regions) of the United States.

Most frequent shootings occurs in the East Coast, the South, and the Central Regions

Figure 6. Number of Shooters by Age and Gender

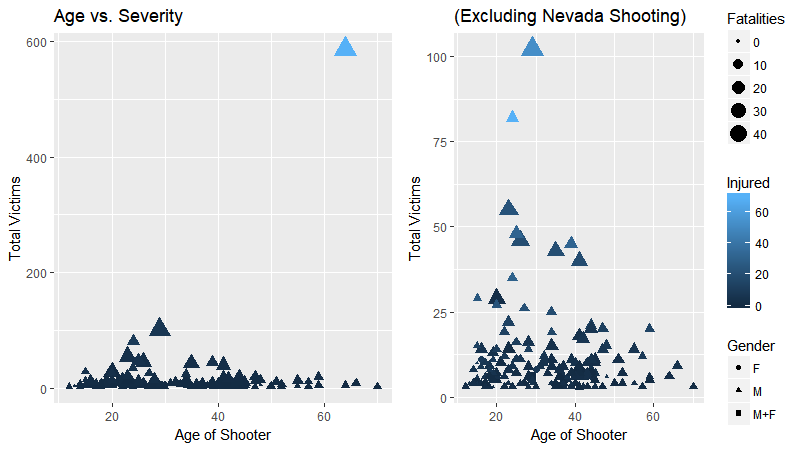


There are more males shooter than female; there were five cases where a female was the shooter. There is peak in male shooter around the ages of 16-25 and at 35-45. After the age of approximately 45, the number of male shooters drop off. In terms of profiling, young adults, late thirties, and early forties are the most common male shooters.

The scatter plot on the left in Figure 7. shows all mass shootings; however with the Las Vegas, Nevada shooting being an outlier, it is difficult to draw any meaningful conclusions. The scatter plot on the right has the outlier removed. Around the ages 18 to 30, there is more total victims with more fatalities. This also shows that most shooters are male compared to female. As expected the more total victims there are, the more fatalities and injured there are.

There is also a small number of cases where they were more than one shooter as a male and female. However, there are an overwhelm number of male shooters compared to female and the joint male and female shooters. While the bottom illustrates the linear model of age and total victims. The model shows that the slope is small and has a low correlation.

Figure 7. Age vs. Severity



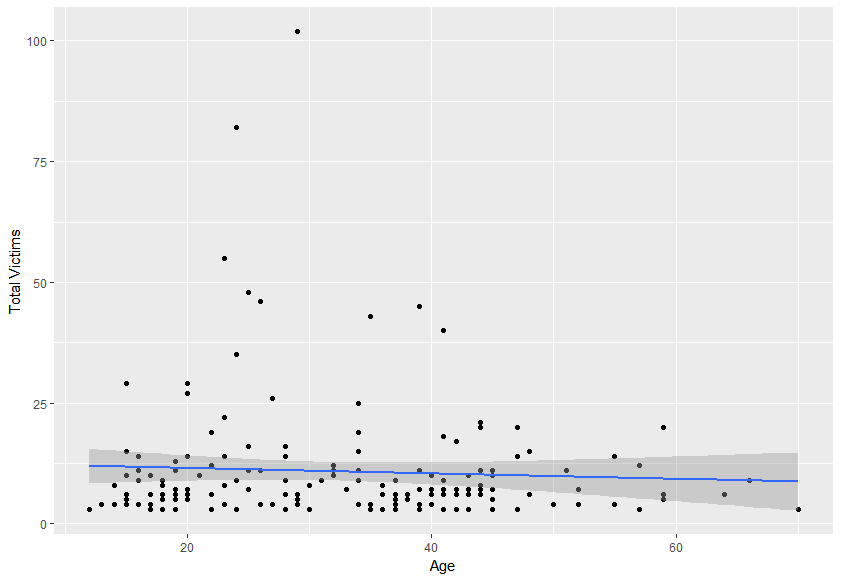
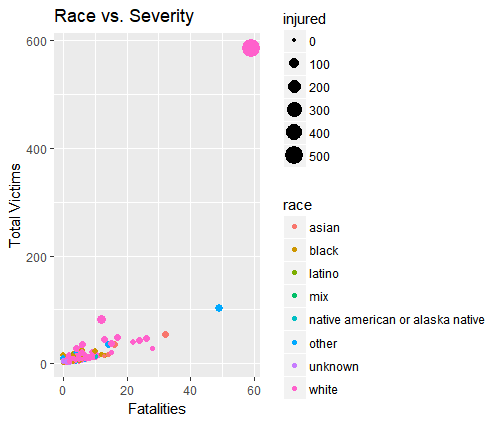


Figure 8a. Race vs. Severity of all Mass Shootings



Similar to the last two figures, Figure 8 is skewed from the Las Vegas, Nevada Shooting. To conduct a proper analysis, the outlier is removed. The most common race of shooters is White, followed by Black. White and Black races make up most of the shootings. There shootings with higher number of total victims generally had a White shooter, with few cases of Asian and Other race.

White and Black are the most common race of shooters

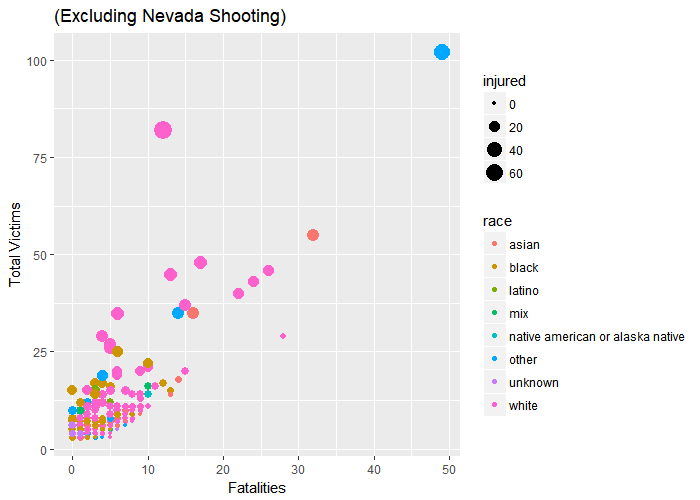
Figure 8b. Race vs. Severity Exlcluding the Las Vegas, Nevada Shooting

Figure 9. Causes/Motives of Shooters

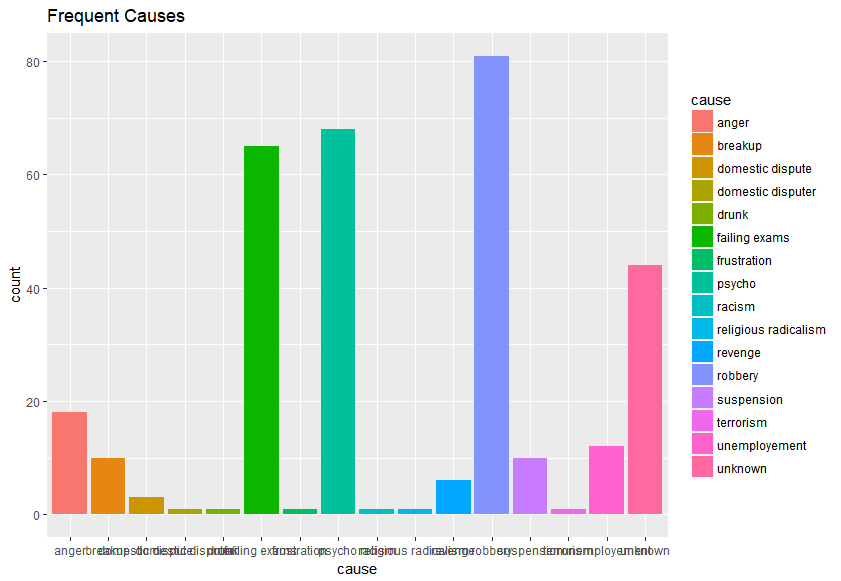
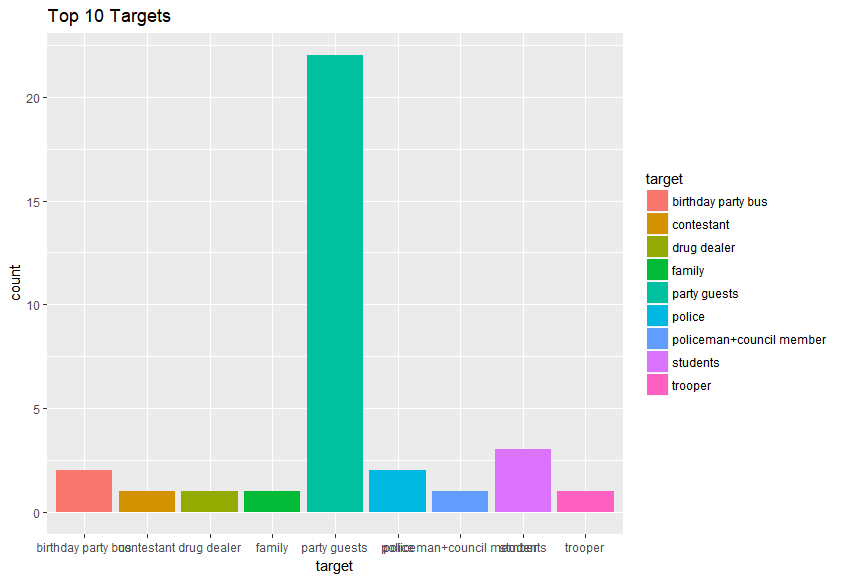


Figure 10. Top ten Targets in relations to the Shooter



Robbery, Pyscho, and Faliing Exams are the most common motives behind the mass shootings, followed by unknown. The failing exams supports the months were shooting occur. Since finals are generally in December for college and midterms in January for high school. The reasoning of failing exams supports the month of February being the most frequent time of the year for shootings. By February midterms are gradeds, and college students have returned to campus.

Other notable motives are angry, unemployment, suspension, and breakup. While these motives are not as common, there are still quite a few cases of them. Some of the motivations have a common theme of failure – failing exams, unemployement, suspension, and breakup. Professional thereapy and pyschologist should pay attention and follow up with patiences under these situations or patiences with similar feelings. Teachers should not be afraid to fail students, but instead provide support when their students do fail.

Robbery, Psycho, and Failing Exams are   
the most known common motives

Figure 10. displays the top ten targets shooter relationship. While there are many more relationships, looking at the top ten shows that party guests are by far the most common. The graph was reduced to top ten in order to minimize the smaller unique mass shootings cases.

# Use of Modeling Techniques

Association Rules

Variables that are factors were consider for the associative mode, such as Open/Closed Location, Mental Health Issues, Race, Gender, and Employeed (Y/N). Ultimately Employeed (Y/N) was not included since it did not affect to results too much and there were not significant results with it added to it. Table 8. shows the most significant results from the association rules based on support, confidence and lift.

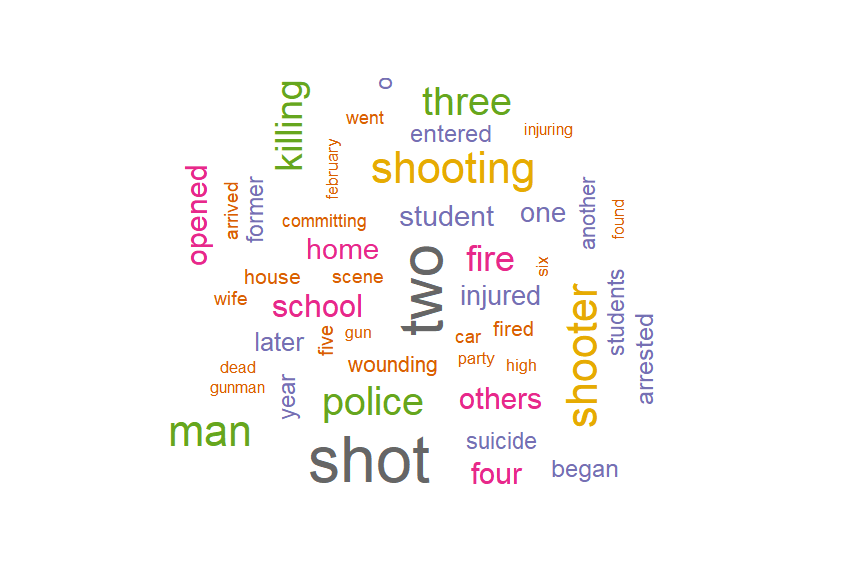
There is a common association with male shooting at closed locations, generally a private location. There is no real common association with mental health since there are high support levels for both mental health issues and not mental issues. Common races determined by the associative rules are both White and Black. With lift values greater than one, these variables are likely dependent on each other.

Table 8. Significant Results from Associative Rules

|  |  |  |  |
| --- | --- | --- | --- |
|  | Support | Confidence | Lift |
| Open/closed Location=Closed =>  Gender=M | 0.5759 | 0.9442 | 1.0444 |
| Open/Closed Location=Closed, Mental=Yes =>  Race=White | 0.1548 | 0.6944 | 1.5577 |
| Open/Closed Location=Closed, Race=Black =>  Gender=M | 0.1362 | 1.0000 | 1.1062 |
| Open/Closed Location=Closed, Mental=No =>  Gender=M | 0.1889 | 0.9384 | 1.0381 |
| Race=white =>  Gender=M | 0.4241 | 0.9514 | 1.0524 |
| Race=Black =>  Gender=M | 0.2632 | 1.0000 | 1.1062 |
| Mental Health=No =>  Open/Closed Location=Closed | 0.2102 | 0.6989 | 1.1460 |
| Mental Health=No =>  Gender=M | 0.2724 | 0.9462 | 1.0467 |
| Mental Health=Yes =>  Race=white | 0.2198 | 0.6698 | 1.5024 |
| Mental Health=Yes =>  Gender=M | 0.3189 | 0.9717 | 1.0749 |
| Mental health=No, Gender =M =>  Open/Closed Location=Closed | 0.1889 | 0.6932 | 1.1365 |
| Mental Health=Yes, Race=White =>  Open/Closed Location=Closed | 0.1548 | 0.7042 | 1.1546 |
| Gender=M =>  Open/Closed Location=Closed | 0.5759 | 0.6370 | 1.0444 |

Text Mining

Figure 11. Word cloud of the summaries

The most frequently-used words in the summary are shot and two. Unsurprisingly the words shooting and shooter are common words. Some words that jump out that does provide insight is suicide, home, school, and student. Student is interesting since Figure 9. showed that failed exams are one of the top causes for mass shootings. The suicide also jumps out since from Table 6. there are cases where the shooter killed themselves afterwards. The word cloud shows that it is more common than the descriptive statistics showed. This word cloud helps support other analysis conducted in this report.

AFINN total sentimental value = -138

Table 9. Sentimental Analysis Ratios

|  |  |
| --- | --- |
| Positive/Negative Words | 0.3333 |
| Percentage of Positive Words | 0.0175 |
| Percentage of Negative Words | 0.0524 |

The ratios confirm that the summaries are generally more negative than positive, as it should be expected when something tragic occurs. There are three times as much negative words than positive. Out of all the words about two percent are positive and about five percent are negative. According to the AFINN list, the overall value is -138, showing that the tone is very negative in general.

# Conclusion

Recommendation

Since mass shootings generally occur on the east side of the United States, the police force should have proper training to handle the situation. Especially since policeman kill is low, they should be trained and be able to get to the scene within a reasonable amount of time. Profilers should know these common traits between shooters and who they are likely to target. If profilers can help identify these shooters before the event happen, it might be possible to get them the proper help beforehand.

From the associative model analysis, there is a link between mental health to the race white and male gender. Therapist and psychologist should know the warning signs and follow up with patience that fit into the profile. Professionals in this fields should learn the warning sign and know when to contact the police. They should be careful since within the last decade there has been a spike in mass shootings.

Summary

It will be difficult to predict who will likely become a shooter, especially since not everyone who fit this description will be shooters. For example, not every male with mental health issues will become a shooter. Additionally, people and society are not entire black and white or easy to predict in patterns, making it nearly impossible to accurate predict who will be the next shooter. Instead this analysis provides location, general time of the year shootings is likely to occur, and profiling traits.

It is difficult to one hundred percent say if recent shootings for better or worst than before. The analysis has proven that there are less victims, fatalities, and injured; however, it has always proven that mass shootings are occurring more often. Overall, in today’s society, professional who work with such individuals (police, teachers, therapist, parents, etc.) should look for early sign and help predict such tragedies before they occur.

# Appendix

Code

# ------- libraries ---------

install.packages("stringr")

library("stringr")

install.packages("readxl")

library("readxl")

install.packages("tibble")

library("tibble")

install.packages("lubridate")

library("lubridate")

install.packages("tidyr")

library("tidyr")

install.packages("ggmap")

library("ggmap")

install.packages("modeest")

library("modeest")

install.packages("moments")

library("moments")

install.packages("ggplot2")

library("ggplot2")

install.packages("dpylr")

library("dplyr")

install.packages("gridExtra")

library("gridExtra")

install.packages("arules")

library("arules")

install.packages("arulesViz")

library("arulesViz")

install.packages("e1071")

library("e1071")

install.packages("tm")

library("tm")

install.packages("wordcloud")

library("wordcloud")

# -------- import data ---------

Mass\_Shootings <- read\_excel("My Documents/School/Syracuse/April 2018/IST 687/Project/Mass Shootings.xlsx")

# Dataframe of original data

Mass\_Shootings

#change to data frame

#will do all changes to data on this data frame

df <- Mass\_Shootings

df

# ------ data cleanse ---------

# remove column S# column

df <- df[,-1]

df

#look at structure of data

str(df)

#Date Column

#remove time from date

df$Date <- as.Date(as.POSIXct(df$Date))

#Open/Close Location Column

#look at all responses

unique(df$`Open/Close Location`)

#change to consistent lower cases

df$`Open/Close Location` <- tolower(df$`Open/Close Location`)

#change to factor

df$`Open/Close Location` <- as.factor(df$`Open/Close Location`)

#Target Column

#look at all responses

unique(df$Target)

#change to consistent lower cases

df$Target <- tolower(df$Target)

#change and symbols to consistent symbol

df$Target <- gsub("ex-wife & family", "ex-wife+family", df$Target)

df$Target <- gsub("ex-girlfriend & family", "ex-girlfriend+family", df$Target)

#Cause Column

#look at all responses

unique(df$Cause)

#fill in NAs with Unknown

df$Cause[is.na(df$Cause)] <- "unknown"

#change to consistent lower cases

df$Cause <- tolower(df$Cause)

#Age Column

#look at all responses

unique(df$Age)

# Employeed (Y/N) Column

#fill in NAs with 0

df$`Employeed (Y/N)` [is.na(df$`Employeed (Y/N)`)] <- 0

#change structure to factor, Y/N (1/0)

df$`Employeed (Y/N)` <- as.factor(df$`Employeed (Y/N)`)

#Mental Health Issues Column

#look at all responses

unique(df$`Mental Health Issues`)

#consistent captalization

df$`Mental Health Issues` <- gsub("unknown", "Unknown", df$`Mental Health Issues`)

#change to factor

df$`Mental Health Issues`<- as.factor(df$`Mental Health Issues`)

#Race Column

#look at all responses

unique(df$Race)

#change to consistent wording between rows

df$Race <- gsub("Black American or African American", "Black", df$Race)

df$Race <- gsub("White American or European American", "White", df$Race)

#fill NAs with unknown

df$Race[is.na(df$Race)] <- "Unknown"

df$Race <- gsub("Some other race", "Other", df$Race)

#captilization fixing

df$Race <- gsub("black", "Black", df$Race)

df$Race <- gsub("white","White",df$Race)

#change to consistent wording between rows

df$Race <- gsub("Two or more races", "Mix", df$Race)

df$Race <- gsub("Black/Unknown", "Black", df$Race)

df$Race <- gsub("White/Some other Race", "White", df$Race)

df$Race <- gsub("Asian American/Other", "Asian", df$Race)

df$Race <- gsub("Native American or Alaska ative", "Native American", df$Race)

df$Race <- gsub("asian american", "asian", df$Race)

#captilization fixing

df$Race <- tolower(df$Race)

#change to factor

df$Race <- as.factor(df$Race)

#Gender Column

#look at all responses

unique(df$Gender)

#Change and symbols for consistency

df$Gender <- gsub("M/F", "M+F", df$Gender)

df$Gender <- gsub("Male/Female","M+F", df$Gender)

#change to consistent wording between rows

df$Gender <- gsub("Male", "M", df$Gender)

df$Gender <- gsub("Female","F", df$Gender)

#change to factor

df$Gender <- as.factor(df$Gender)

# ------- adding data to complete the set

#adding missing locations

df[16,2] <- "Forestville, Maryland"

df[17, 2] <- "Halifax County, VA"

df[18,2] <- "Baltimore, MD"

df[19,2] <- "Chicago, IL"

df[20,2] <- "Houston, Texas"

df[21,2] <- "Blountsville, AL"

df[22,2] <- "Long Beach, CA"

df[23,2] <- "Albuquerque, NM"

df[24,2] <- "Memphis, TN"

df[25,2] <- "Chicago, IL"

df[26,2] <- "Albuquerque, NM"

df[30,2] <- "Greenhill, AL"

df[35,2] <- "Atlanta, GA"

df[36,2] <- "Fort Myers, FL"

df[37,2] <- "Elberton, Georgia"

df[38,2] <- "Trenton, NJ"

df[39,2] <- "Detroit, MI"

df[41,2] <- "Wilkinsburg, PA"

df[42,2] <- "Kansas City, Kansas"

df[43,2] <- "Lafayette, LA"

df[44,2] <- "Kansas City, Kansas"

df[46,2] <- "Roswell, GA"

df[47,2] <- "Wichita, Kansas"

df[48,2] <- "Detroit, MI"

df[49,2] <- "Riverside, CA"

df[54,2] <- "Belfair, WA"

df[56,2] <- "Hazelwood, MO"

df[57,2] <- "Houston, Texas"

df[60,2] <- "Tampa, FL"

df[61,2] <- "Kalamazoo, MI"

df[62,2] <- "Vallejo, CA"

df[63,2] <- "Muskegon, MI"

df[65,2] <- "Rochester, NY"

df[66,2] <- "Tampa, FL"

df[67,2] <- "Los Angeles, CA"

df[68,2] <- "Uvalde, TX "

df[69,2] <- "New Orleans, LA"

df[70,2] <- "Glendale, AZ"

df[71,2] <- "Bowling Green, VA"

df[73,2] <- "Perris, CA"

df[74,2] <- "Crestview, FL"

df[75,2] <- "Los Angeles, California"

df[76,2] <- "Gloucester County, VA"

df[77,2] <- "Wilmington, DE"

df[78,2] <- "Memphis, Tennessee"

df[79,2] <- "Lakeland, Florida"

#Location Column

# investigate & fix error

df[147,2]

df$Location <- gsub("Pennsburg, Souderton, Lansdale, Harleysville, Pennsylvania", "Momtomery County, Pennsylvainia", df$Location)

df[176,2]

df$Location <- gsub("South Valley, Albuquerque, New Mexico", "South Valley, New Mexico", df$Location)

df[225,2]

df$Location <- gsub("Nickel Mines, Lancaster, Pennsylvania", "Lancaster, Pennsylvania", df$Location)

df[241,2]

df$Location <- gsub("Santee, San Diego, California", "San Diego, California", df$Location)

# add column: city and state

#separating into City and State

Location <- separate(df, Location, c("City","State"), sep = ",", remove=FALSE)

df<-Location

#State Column

unique(df$State)

#change State to full name

df$State <- gsub("TX", "Texas", df$State)

df$State <- gsub("CO", "Colorado", df$State)

df$State <- gsub("MD", "Maryland", df$State)

df$State <- gsub("NV", "Nevada", df$State)

df$State <- gsub("CA", "California", df$State)

df$State <- gsub("PA", "Pennsylvania", df$State)

df$State <- gsub("WA", "Washington", df$State)

df$State <- gsub("LA", "Louisianna", df$State)

df$State <- gsub("VA", "virginia", df$State)

df$State <- gsub("IL", "Illinois", df$State)

df$State <- gsub("AL", "Alamaba", df$State)

df$State <- gsub("NM", "New Mexico", df$State)

df$State <- gsub("TN", "Tennessee", df$State)

df$State <- gsub("GA", "Georgia", df$State)

df$State <- gsub("FL", "Florida", df$State)

df$State <- gsub("NJ", "New Jersey", df$State)

df$State <- gsub("MI", "Michigan", df$State)

df$State <- gsub("MO", "Missouri", df$State)

df$State <- gsub("AZ", "Arizona", df$State)

df$State <- gsub("DE", "Delaware", df$State)

df$State <- gsub("NY", "New York", df$State)

df$State <- gsub(" ", "", df$State)

df$State <- gsub("NorthCarolina", "North Carolina", df$State)

df$State <- gsub("SouthDakota", "South Dakota", df$State)

df$State <- gsub("NewYork", "New York", df$State)

df$State <- gsub("SouthCarolina", "South Carolina", df$State)

df$State <- gsub("NewJersey", "New Jersey", df$State)

df$State <- gsub("NewMexico", "New Mexico", df$State)

df$State <- gsub("WestVirginia", "West Virginia", df$State)

#correcting spelling

df$State <- gsub("pennsylvirginiania", "pennsylvania", df$State)

#lower case

df$City <- tolower(df$City)

df$State <- tolower(df$State)

#---------Descriptive Statistics ---------

summary(df)

#Fatalities

fatalities <- df[complete.cases(df$Fatalities),11]

fatalities <- c(fatalities)

fatalities <-unlist(fatalities)

skewness(fatalities)

var(fatalities)

sd(fatalities)

mfv(fatalities)

hist(fatalities)

#injured

injured <- df[complete.cases(df$Injured),12]

injured <- c(injured)

injured <- unlist(injured)

skewness(injured)

var(injured)

sd(injured)

mfv(injured)

hist(injured)

#Total Victims

victims <- df[complete.cases(df$`Total victims`),13]

victims <- c(victims)

victims <- unlist(victims)

skewness(victims)

var(victims)

sd(victims)

mfv(victims)

hist(victims)

#Policeman Killed

police <- df[complete.cases(df$`Policeman Killed`),14]

police <- c(police)

police <- unlist(police)

skewness(police)

var(police)

sd(police)

mfv(police)

hist(police)

#Calculating days between shootings

length(df$Date)

date1 <- df$Date[-323]

date2 <- df$Date[-1]

daysbetween <- difftime(date1, date2, units=c("days"))

daysbetween <- as.numeric(daysbetween)

summary(daysbetween)

skewness(daysbetween)

var(daysbetween)

sd(daysbetween)

mfv(daysbetween)

hist.days <- hist(daysbetween)

#percent of victims killed

percent.kill <- (df$Fatalities/df$`Total victims`)\*100

summary(percent.kill)

skewness(percent.kill)

var(percent.kill)

sd(percent.kill)

mfv(percent.kill)

hist.percent <- hist(percent.kill)

# hypothesis test

#compare recent to old

time <- year(df$Date)

time

#shooting within the last 10 years

recent <- data.frame(df,time)

recent <- recent[recent$time > 2007,]

#shooting that happened more than 10 years ago

old <- data.frame(df,time)

old <- old[old$time <= 2007,]

#removing outlier

recent <- recent[-4,]

#function for z score in hypoth test

hypo.test.zscore <- function(vector1, vector2)

{

mean1 <- mean(vector1)

var1 <- var(vector1)

n1 <- length(vector1)

mean2 <- mean(vector2)

var2 <- var(vector2)

n2 <- length(vector2)

mean.diff <- mean1 - mean2

var.n1 <- var1/n1

var.n2 <- var2/n2

var.combine <- var.n1 + var.n2

demon <- sqrt(var.combine)

z <- mean.diff/demon

return(z)

}

#Fatalities, injured, total victims z score

hypo.test.zscore(recent$Fatalities, old$Fatalities)

hypo.test.zscore(recent$Injured, old$Injured)

hypo.test.zscore(recent$Total.victims, old$Total.victims)

#policeman killed

recent.police <- recent[complete.cases(recent$Policeman.Killed),]

old.police <- old[complete.cases(old$Policeman.Killed),]

hypo.test.zscore(recent.police$Policeman.Killed, old.police$Policeman.Killed)

# ------Visuals

# ------ line graph severity/frequency by time of year

#pulling month

month <- data.frame(df, months(df$Date), month(df$Date))

month <- month[order(month$month.df.Date.),]

colnames(month)

month1 <- unique(month$months.df.Date)

month1 <- month1[order(month1)]

#getting frequency

month1 <- data.frame(month1, tabulate (month$months.df.Date.))

#changing column names

colname <- c("Month", "NoShootings")

colnames(month1) <- colname

monthNo <- c(4, 8, 12, 2, 1, 7, 6, 3, 5, 11, 10, 9)

#combining

month1 <- data.frame(month1, monthNo)

month1 <- month1[order(month1$monthNo),]

#grouping to get average total victims "severity"

groupedMonth <- group\_by(month, months.df.Date.) %>%

summarize(severity=mean(Total.victims))

groupedMonth = merge(groupedMonth, month1, by.x='months.df.Date.', by.y='Month')

#line graph

monthLine <- ggplot(groupedMonth, aes(x=monthNo, y=NoShootings))+

geom\_line() +

geom\_point(aes(size=severity))

#change axis lables

monthLine <- monthLine +

scale\_x\_continuous("Month(Number)") +

scale\_y\_continuous("Number of Shootings")

#change x axis scale

monthLine <- monthLine +

scale\_x\_continuous(breaks = c(1:12))

#Title

monthLine <- monthLine +

ggtitle("Monthly Frequency and Severity")

monthLine

#Pulling year

year <- data.frame(df, year(df$Date))

year <- data.frame(year$year.df.Date, year$Total.victims)

#changing col names

colnames <- c("year", "Total.Victim")

colnames(year) <- colnames

#severity by year

year.severity <- group\_by(year, year$year) %>%

summarize(severity=mean(Total.victims))

#frequency by year

year.frequency <- tabulate(match(year$year,unique(year$year)))

#reverse order

year.frequency <- rev(year.frequency)

#average killed by year

year.killed <- data.frame(year$year, df$Fatalities)

#changing column names

colnames <- c("year", "fatalities")

colnames(year.killed) <- colnames

#grouping by year

year.killed <- group\_by(year.killed, year.killed$year) %>%

summarize(fatalities=mean(fatalities))

#changing column names

colnames(year.killed) <- colnames

#combining variables

year <- data.frame(year.severity, year.frequency, year.killed$fatalities)

#changing col names

colnames <- c("year", "severity", "frequency", "fatalities")

colnames(year) <- colnames

#line graph

line.year <- ggplot(year, aes(x=year, y=frequency)) +

geom\_line() +

geom\_point(aes(size=severity, color=fatalities))

#changing x & y lables

line.year <- line.year +

scale\_x\_continuous("Year") +

scale\_y\_continuous("Number of Shootings")

#adding title

line.year <- line.year +

ggtitle("Frequency and Severity over Time")

line.year

#combining line graphs into one window

grid.arrange(monthLine, line.year, nrow=1)

# --------- Maps

Location <- df[,-21:-22]

#combining city and state

Location$Location <- paste(Location$City, ",", Location$State)

#change to lower case

Location$Location <- tolower(Location$Location)

latlon <- geocode(Location$Location)

geocode(Location$Location[322])

latlon[322,1] <- -111.8315

latlon[322,2] <- 33.41518

Location <- data.frame(Location, latlon)

#selecting data

Location <- data.frame(Location$Location,

location$State,

Location$Fatalities,

Location$Injured,

Location$Total.victims,

Location$lon,

Location$lat)

#changing column names

colnames <- c("location", "state", "fatalities", "injured", "total.victims", "lon", "lat")

colnames(Location) <- colnames

#removing alaska & hawaii

Location <- Location[Location$state != "hawaii",]

Location <- Location[Location$state != "alaska",]

#making the map

us <- map\_data("state")

map.severity <- ggplot(Location, aes(map\_id = state))

map.severity <- map.severity +

geom\_map(map=us, color="white", aes(fill=total.victims))

map.severity <- map.severity +

expand\_limits (x = us$long, y=us$lat)

map.severity <- map.severity +

coord\_map()

#adding points on location

map.severity <- map.severity +

geom\_point(aes(x=lon, y=lat, size=injured, color=fatalities))

#changing color scale

map.severity <- map.severity +

scale\_colour\_gradient(low="pink", high="darkred")

#legend position

map.severity <- map.severity +

theme(legend.position="bottom")

#title

map.severity <- map.severity +

ggtitle("Severity by Location")

map.severity

#removing outlier - nevada shooting

Location1 <- Location[-4,]

#map without outlier

map.severity1 <- ggplot(Location1, aes(map\_id= state))

map.severity1 <- map.severity1 +

geom\_map(map=us, color="white", aes(fill=total.victims))

map.severity1 <- map.severity1 +

expand\_limits (x= us$long, y=us$lat)

map.severity1 <- map.severity1 +

coord\_map()

#adding points on location

map.severity1 <- map.severity1 +

geom\_point(aes(x=lon, y=lat, size=injured, color=fatalities))

#changing color scale

map.severity1 <- map.severity1 +

scale\_colour\_gradient(low="pink", high="darkred")

#legend position

map.severity1 <- map.severity1 +

theme(legend.position="bottom")

#title

map.severity1 <- map.severity1 +

ggtitle("Severity by Location (Excluding Nevada Shooting)")

map.severity1

#-------histogram

#histogram age

age <- df[complete.cases(df$Age),]

#remove the one with two shooters

age <- age[-157,]

age <- age[-49,]

age <- age[-14,]

age$Age <- as.numeric(age$Age)

hist.age <- ggplot(age, aes(x=Age, color=Gender)) +

geom\_histogram(fill="white")

#Adding a title

hist.age <- hist.age +

ggtitle("Age of Shooters by Gender")

hist.age

age.male <- age[which(age$Gender == "M"),]

age.female <- age[which(age$Gender == "F"),]

#male age histogram

hist.age.male <- ggplot(age.male, aes(x=Age)) +

geom\_histogram(fill="white", color="black")

hist.age.male <- hist.age.male +

ggtitle("Male Shooters Age")

hist.age.male

#female age histogram

hist.age.female <- ggplot(age.female, aes(x=Age)) +

geom\_histogram(fill="white", color="black")

hist.age.female <- hist.age.female +

ggtitle("Female Shooters Age")

#one window

grid.arrange(hist.age.male, hist.age.female, nrow=1)

# -------- scatter plots

#scatter plot

#age by total victim

scatter.age <- ggplot(age, aes(x=Age, y=age$`Total victims`, shape=Gender)) +

geom\_point(aes(color=Injured, size=Fatalities))

#axis lables

scatter.age <- scatter.age +

scale\_x\_continuous("Age of Shooter") +

scale\_y\_continuous("Total Victims")

#title

scatter.age <- scatter.age +

ggtitle("Age vs. Severity")

scatter.age

#removing outlier

age1 <- age[-4,]

#scatter plot without outlier

scatter.age1 <- ggplot(age1, aes(x=Age, y=age1$`Total victims`, shape=Gender)) +

geom\_point(aes(color=Injured, size=Fatalities))

#axis lables

scatter.age1 <- scatter.age1 +

scale\_x\_continuous("Age of Shooter") +

scale\_y\_continuous("Total Victims")

#title

scatter.age1 <- scatter.age1 +

ggtitle("Age and Severity (Excluding Nevada Shooting)")

scatter.age1

#combining scatter plots to one window

#removing one legend

scatter.age <- scatter.age +

theme(legend.position="none")

#changing title

scatter.age1 <- scatter.age1 +

ggtitle("(Excluding Nevada Shooting)")

grid.arrange(scatter.age, scatter.age1, nrow=1)

#scatter plot for race

#Pulling the data

race <- data.frame(df$Fatalities,

df$Injured,

df$`Total victims`,

df$Race)

#changing column names

colnames <- c("fatalities", "injured", "total.victims", "race")

colnames(race) <- colnames

unique(race$race)

#scatter plot

scatter.race <- ggplot(race, aes(x=fatalities, y=total.victims, color=race)) +

geom\_point(aes(size=injured))

#axis lables

scatter.race <- scatter.race +

scale\_x\_continuous("Fatalities") +

scale\_y\_continuous("Total Victims")

#title

scatter.race <- scatter.race +

ggtitle("Race vs. Severity")

#legend to bottom

scatter.race <- scatter.race +

theme(legend.position="right")

scatter.race

#removing outlier

race1 <- race[-4,]

#scatter plot

scatter.race1 <- ggplot(race1, aes(x=fatalities, y=total.victims, color=race)) +

geom\_point(aes(size=injured))

#axis lables

scatter.race1 <- scatter.race1 +

scale\_x\_continuous("Fatalities") +

scale\_y\_continuous("Total Victims")

#title

scatter.race1 <- scatter.race1 +

ggtitle("(Excluding Nevada Shooting)")

#legend to bottom

scatter.race1 <- scatter.race1 +

theme(legend.position="right")

scatter.race1

#combining to one window

grid.arrange(scatter.race, scatter.race1, nrow=1, widths=1:2)

#bar graphs

#causes

#pulling just the cause data

cause <- df$Cause[complete.cases(df$Cause)]

#changing to factor

cause <- as.factor(cause)

#count for each factor

tabulate(cause)

df.cause <- data.frame(unique(cause), tabulate(cause))

#looking at the structure

str(df.cause)

#column names

colname <- c("cause", "count")

colnames(df.cause) <- colname

#changing to numeric

df.cause$count <- as.numeric(df.cause$count)

str(df.cause)

#creating a bar graph

bar.cause <- ggplot(df.cause, aes(x=cause, y=count, fill=cause)) +

geom\_bar(stat="identity")

#adding title

bar.cause <- bar.cause +

ggtitle("Frequent Causes")

bar.cause

#target

#pulling just the target data

target <- df$Target[complete.cases(df$Target)]

#change to factor

target <- as.factor(target)

#count for each factor

df.target <- data.frame(unique(target), tabulate(target))

#changing the column names

colname <- c("target", "count")

colnames(df.target) <- colname

#looking at the structure

str(df.target)

#change to numeric

df.target$count <- as.numeric(df.target$count)

#taking only the top 10

df.target <- df.target[sort(df.target$count, decreasing=TRUE),]

df.target <- df.target[-1,]

df.target <- head(df.target, 10)

#creating a bar graph

bar.target <- ggplot(df.target, aes(x=target, y=count, fill=target)) +

geom\_bar(stat="identity")

#adding title

bar.target <- bar.target +

ggtitle("Top 10 Targets")

bar.target

# -------- Use of Modeling Techniques -------

#associative model

associative.data <- data.frame(df$`Open/Close Location`,

df$`Mental Health Issues`,

df$Race,

df$Gender)

associative.data

#changing column names

colnames <- c("open/closed.location",

"mental.health.issues",

"race",

"gender")

colnames(associative.data) <- colnames

#looking at structure, ensure factors

str(associative.data)

#model

rules <- apriori(associative.data, parameter=list(support=0.005, confidence=0.5))

#inspecting the results

inspect(rules)

#text mining

#changing to word string

file <- df$Summary

#AFINN data

afinn <- "/Users/Christina/Documents/My Documents/School/Syracuse/April 2018/IST 687/Project/AFINN.txt"

afinn <- readLines(afinn)

#separate into word and value

afinn <- data.frame(afinn)

afinn <- separate(afinn, afinn, c("word", "value"), sep=", ", remove=FALSE)

afinn <- afinn[,-1]

#looking at structure

str(afinn)

#changing value to number

afinn$value <- as.numeric(afinn$value)

str(afinn)

#cleansing the words

words.vec <- VectorSource(file)

words.corpus <- Corpus(words.vec)

words.corpus <- tm\_map(words.corpus, content\_transformer(tolower))

words.corpus <- tm\_map(words.corpus, removePunctuation)

words.corpus <- tm\_map(words.corpus, removeNumbers)

words.corpus <- tm\_map(words.corpus, removeWords, stopwords("english"))

tdm <- TermDocumentMatrix(words.corpus)

m <- as.matrix(tdm)

wordCount <- rowSums(m)

words <- names(wordCount)

#word cloud

wordCount <- sort(wordCount, decreasing=TRUE)

cloudFrames <- data.frame(words=names(wordCount), freq=wordCount)

wordcloud(names(wordCount), wordCount, min.freq=5, max.words=50, rot.per=0.35, colors=brewer.pal(8, "Dark2"))

#sentiment analysis

#matching words

matched <- match(words, afinn$word, nomatch=0)

#getting index numbers

afinn.index <- matched[which(matched !=0)]

#getting the assigned values

afinn.value <- afinn[(afinn.index),2]

#summing the sentimal values

sum(afinn.value)

#counting positive and negative words

#splitting list into positive and negative

p <- afinn[which(afinn$value >0), -2]

n <- afinn[which(afinn$value <0), -2]

#getting the positive word count

matched.p <- match(words, p, nomatch=0)

matched.p <- matched.p[which(matched.p !=0)]

matched.p <- length(matched.p)

matched.p

#getting the negative word count

matched.n <- match(words,n,nomatch=0)

matched.n <- matched.n[which(matched.n != 0)]

matched.n <- length(matched.n)

matched.n

#ratios

matched.p/matched.n

matched.p/length(words)

matched.n/length(words)