US\_Mass\_shootings

Aditi Deokar

1/10/2020

## DATASET DESCRIPTION

### Source-

* <https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/>

### Background-

* Since 1982, there have been at least 117 public mass shootings across the country, with the killings unfolding in 35 states, from Massachusetts to Hawaii.
* They are occurring more often: An analysis of this database by researchers at Harvard University, further corroborated by a different study from the FBI, determined that mass shootings have tripled in frequency in recent years.

### Dataset-

* The dataset contains detailed information of mass shootings from the year 1982 to 2019.
* No. of columns= 24
* No. of rows= 118
* Variables- case,location,date,summary,fatalities,injured,total\_victims,location,age\_of\_shooter,prior\_signs\_mental\_health\_issues, mental\_health\_details,weapons\_obtained\_legally,where\_obtained,weapon\_type,weapon\_details,race,gender,sources,mental\_health\_sources, sources\_additional\_age,latitude,longitude,type,year

### Characteristics-

* Number of categorical data columns:18
* Number of numerical data columns:6

### Reference-

* <https://www.kaggle.com/denisafonin/data-cleaning-and-dealing-with-missing-values/#data>

### IMPORT LIBRARIES

library(tibble)   
library(lubridate)  
library(tidyr)  
library(stringr)  
library(ggmap)  
library(RColorBrewer)

### IMPORT DATA

mass\_shootings<-read.csv("Assignment1-Deokar-US\_Mass\_Shootings.csv", stringsAsFactors = FALSE)

### EXPLORE DATA

glimpse(mass\_shootings)

## Observations: 117  
## Variables: 24  
## $ case <chr> "Jersey City kosher market sh...  
## $ location <chr> "Jersey City, New Jersey", "P...  
## $ date <chr> "12/10/19", "12/6/19", "8/31/...  
## $ summary <chr> "David N. Anderson, 47, and F...  
## $ fatalities <int> 4, 3, 7, 9, 22, 3, 12, 5, 3, ...  
## $ injured <int> 3, 8, 25, 27, 26, 12, 4, 6, 1...  
## $ total\_victims <int> 7, 11, 32, 36, 48, 15, 16, 11...  
## $ location.1 <chr> "Other", "Military", "Other",...  
## $ age\_of\_shooter <chr> "-", "-", "36", "24", "21", "...  
## $ prior\_signs\_mental\_health\_issues <chr> "-", "-", "yes", "-", "-", "T...  
## $ mental\_health\_details <chr> "-", "-", "\"One friend of th...  
## $ weapons\_obtained\_legally <chr> "-", "-", "-", "Yes", "Yes", ...  
## $ where\_obtained <chr> "-", "-", "-", "-", "-", "Nev...  
## $ weapon\_type <chr> "-", "semiautomatic handgun",...  
## $ weapon\_details <chr> "-", "-", "-", "AR-15-style r...  
## $ race <chr> "Black", "-", "White", "White...  
## $ gender <chr> "-", "M", "M", "M", "M", "M",...  
## $ sources <chr> "https://www.nytimes.com/2019...  
## $ mental\_health\_sources <chr> "-", "-", "https://www.nytime...  
## $ sources\_additional\_age <chr> "-", "-", "-", "-", "-", "-",...  
## $ latitude <dbl> 40.70736, 30.36471, 31.92597,...  
## $ longitude <dbl> -74.08361, -87.28857, -102.27...  
## $ type <chr> "Spree", "Mass", "Spree", "Ma...  
## $ year <int> 2019, 2019, 2019, 2019, 2019,...

summary(mass\_shootings) #displays summary statistics for numeric variables

## case location date   
## Length:117 Length:117 Length:117   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## summary fatalities injured total\_victims   
## Length:117 Min. : 3.000 Min. : 0.00 Min. : 3.00   
## Class :character 1st Qu.: 5.000 1st Qu.: 1.00 1st Qu.: 7.00   
## Mode :character Median : 6.000 Median : 4.00 Median : 11.00   
## Mean : 8.103 Mean : 12.33 Mean : 20.44   
## 3rd Qu.: 9.000 3rd Qu.: 11.00 3rd Qu.: 18.00   
## Max. :58.000 Max. :546.00 Max. :604.00   
## location.1 age\_of\_shooter prior\_signs\_mental\_health\_issues  
## Length:117 Length:117 Length:117   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## mental\_health\_details weapons\_obtained\_legally where\_obtained   
## Length:117 Length:117 Length:117   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## weapon\_type weapon\_details race   
## Length:117 Length:117 Length:117   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## gender sources mental\_health\_sources  
## Length:117 Length:117 Length:117   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## sources\_additional\_age latitude longitude   
## Length:117 Min. :21.32 Min. :-157.88   
## Class :character 1st Qu.:33.56 1st Qu.:-117.28   
## Mode :character Median :38.25 Median : -90.67   
## Mean :37.30 Mean : -96.58   
## 3rd Qu.:41.49 3rd Qu.: -81.29   
## Max. :48.46 Max. : -71.08   
## type year   
## Length:117 Min. :1982   
## Class :character 1st Qu.:1999   
## Mode :character Median :2012   
## Mean :2008   
## 3rd Qu.:2017   
## Max. :2019

### *Obervations:*

* Date, age\_of\_shooter variable’s class is to be transformed to correct data type
* Total victims: There is a strange distribution - Max is huge compared to the 3rd quartile. Mean is 2 times higher than median which implies that the distribution is skewed towards the upper limit.
* Fatalities and Injured variables are the subsets of Total victims variable and have the same issue.

## CLEANING DATA

### *Untidy data:*

* The dataset has a column named “location” which contains the “City” and “State” together separated by “,” where the shootings have taken place. This indicates an untidy data since cities and states should have different columns in order to see the distribution of mass shootings over cities or states.
* To convert this untidy data to tidy data, we need to separate “Location” column into two different columns named “City” and “State”

### *Separating Location to “City” and “State”:*

head(mass\_shootings$location, n=20)

## [1] "Jersey City, New Jersey" "Pensacola, Florida"   
## [3] "Odessa, Texas" "Dayton, Ohio"   
## [5] "El Paso, Texas" "Gilroy, California"   
## [7] "Virginia Beach, Virginia" "Aurora, Illinois"   
## [9] "State College, Pennsylvania" "Sebring, Florida"   
## [11] "Chicago, Illinois" "Thousand Oaks, California"   
## [13] "Pittsburgh, Pennsylvania" "Perryman, Maryland"   
## [15] "Bakersfield, California" "Cincinnati, Ohio"   
## [17] "Annapolis, Maryland" "Santa Fe, Texas"   
## [19] "Nashville, Tennessee" "Yountville, California"

mass\_shootings <- separate(mass\_shootings, location, c("City", "State"), sep = ",", remove=FALSE)  
  
#erase blank spaces from the beginning and end of each state name  
mass\_shootings$State<-str\_trim(mass\_shootings$State)

### *Tidy Data:*

my\_data <- as\_tibble(mass\_shootings)  
#my\_data %>% select(location:State)  
my\_data %>% dplyr::select(location:State)

## # A tibble: 117 x 3  
## location City State   
## <chr> <chr> <chr>   
## 1 Jersey City, New Jersey Jersey City New Jersey   
## 2 Pensacola, Florida Pensacola Florida   
## 3 Odessa, Texas Odessa Texas   
## 4 Dayton, Ohio Dayton Ohio   
## 5 El Paso, Texas El Paso Texas   
## 6 Gilroy, California Gilroy California   
## 7 Virginia Beach, Virginia Virginia Beach Virginia   
## 8 Aurora, Illinois Aurora Illinois   
## 9 State College, Pennsylvania State College Pennsylvania  
## 10 Sebring, Florida Sebring Florida   
## # ... with 107 more rows

### *Changing multiple column names*

colnames(mass\_shootings)[colnames(mass\_shootings) %in% c("age\_of\_shooter","location.1","prior\_signs\_mental\_health\_issues")] <- c("age", "type\_of\_location","mental\_health\_issues")  
  
#summary(mass\_shootings)

### *Changing Date variable to date type*

mass\_shootings$date<-mdy(mass\_shootings$date)

### *Changing Age variable to numeric type*

mass\_shootings$age<-as.numeric(mass\_shootings$age,message=FALSE, warning=FALSE)  
#summary(mass\_shootings)

### *Changing “mental\_health\_issues”, “race”, “gender” columns’ classes to factors*

cols\_to\_factors<-c("mental\_health\_issues", "race", "gender")  
mass\_shootings[,cols\_to\_factors]<-lapply(mass\_shootings[,cols\_to\_factors], as.factor)

### *Standardize levels for “gender”, “race” and “mental\_health\_issues”:*

### *Gender:*

levels(mass\_shootings$gender)

## [1] "-" "F" "Female" "M"   
## [5] "Male" "Male & Female"

levels(mass\_shootings$gender)[levels(mass\_shootings$gender)=="M"] <- "Male"  
levels(mass\_shootings$gender)[levels(mass\_shootings$gender)=="F"] <- "Female"  
levels(mass\_shootings$gender)[levels(mass\_shootings$gender)=="Male & Female"] <- "Male/Female"  
levels(mass\_shootings$gender)[levels(mass\_shootings$gender)=="-"] <- "Unknown"  
  
levels(mass\_shootings$gender)

## [1] "Unknown" "Female" "Male" "Male/Female"

table(mass\_shootings$gender)

##   
## Unknown Female Male Male/Female   
## 1 3 112 1

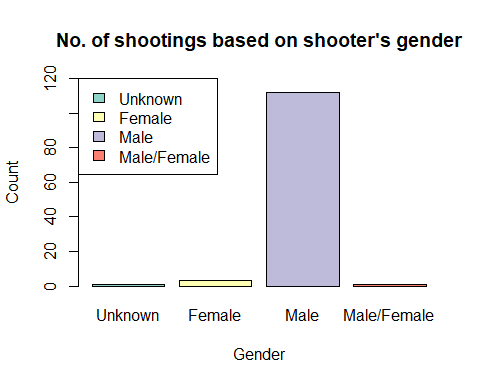
nlevels(mass\_shootings$gender)

## [1] 4

sum(table(mass\_shootings$gender))==nrow(mass\_shootings)

## [1] TRUE

#plot(mass\_shootings$gender,ylim = c(0,140))  
coul <- brewer.pal(4, "Set3")  
barplot(table(mass\_shootings$gender),  
main="No. of shootings based on shooter's gender",  
xlab="Gender",  
ylab="Count",ylim = c(0,120),  
col = coul  
)  
legend("topleft",  
c("Unknown","Female","Male","Male/Female"),  
fill = coul  
)



### *Observations-*

* The above graph shows that male shooters were responsible for majority of the mass shootings(i.e. 95% of the shootings)

### *Race:*

levels(mass\_shootings$race)

## [1] "-" "Asian" "black"   
## [4] "Black" "Latino" "Native American"  
## [7] "Other" "unclear" "white"   
## [10] "White" "White "

levels(mass\_shootings$race)[levels(mass\_shootings$race)=="-"]<-"Unknown"  
levels(mass\_shootings$race)[levels(mass\_shootings$race)=="black"  
 |levels(mass\_shootings$race)=="Black"]<-"Black American"  
levels(mass\_shootings$race)[levels(mass\_shootings$race)=="white"  
 |levels(mass\_shootings$race)=="white"  
 |levels(mass\_shootings$race)=="White"  
 |levels(mass\_shootings$race)=="White "]<-"White American"  
levels(mass\_shootings$race)[levels(mass\_shootings$race)=="unclear"]<-"Unknown"   
levels(mass\_shootings$race)[levels(mass\_shootings$race)=="Native American"]<-"Native American"  
  
table(mass\_shootings$race)

##   
## Unknown Asian Black American Latino   
## 6 8 20 10   
## Native American Other White American   
## 3 5 65

### *Mental Health Issues:*

levels(mass\_shootings$mental\_health\_issues)

## [1] "-" "No" "TBD" "Unclear" "Unclear " "Unknown"   
## [7] "yes" "Yes"

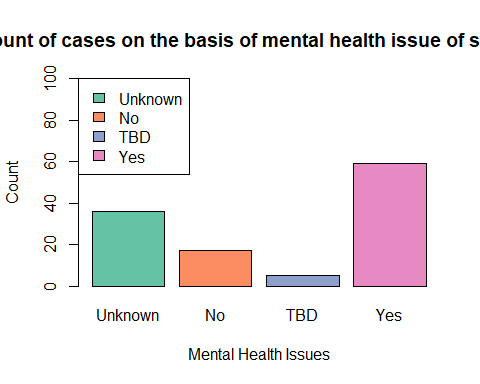
levels(mass\_shootings$mental\_health\_issues)[levels(mass\_shootings$mental\_health\_issues)=="Unclear " |levels(mass\_shootings$mental\_health\_issues)=="Unclear" |levels(mass\_shootings$mental\_health\_issues)=="unknown" | levels(mass\_shootings$mental\_health\_issues)=="-"]<-"Unknown"  
  
levels(mass\_shootings$mental\_health\_issues)[levels(mass\_shootings$mental\_health\_issues)=="yes"  
 |levels(mass\_shootings$mental\_health\_issues)=="Yes"]<-"Yes"  
   
levels(mass\_shootings$mental\_health\_issues)

## [1] "Unknown" "No" "TBD" "Yes"

table(mass\_shootings$mental\_health\_issues)

##   
## Unknown No TBD Yes   
## 36 17 5 59

#plot(mass\_shootings$mental\_health\_issues,ylim = c(0,70))  
library(RColorBrewer)  
coul <- brewer.pal(4, "Set2")  
barplot(table(mass\_shootings$mental\_health\_issues),  
main="Count of cases on the basis of mental health issue of shooter",  
xlab="Mental Health Issues",  
ylab="Count",ylim = c(0,100),  
col = coul  
)  
legend("topleft",  
c("Unknown","No","TBD","Yes"),  
fill = coul)



### *Observations-*

* The above graph shows that more than 50% of the shooters suffered through mental health issues.

## *Standardize State names:*

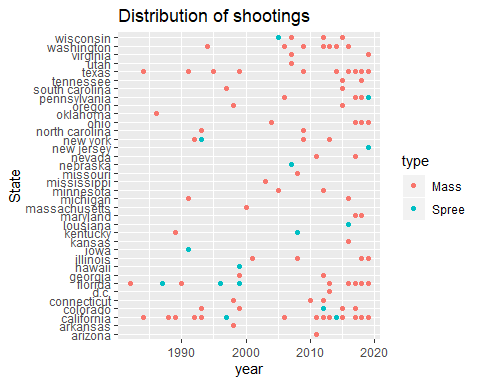
mass\_shootings$State<-as.factor(mass\_shootings$State)  
levels(mass\_shootings$State)

## [1] "Arizona" "Arkansas" "California" "Colorado"   
## [5] "Connecticut" "D.C." "Florida" "Georgia"   
## [9] "Hawaii" "Illinois" "Iowa" "Kansas"   
## [13] "Kentucky" "Lousiana" "Maryland" "Massachusetts"   
## [17] "Michigan" "Minnesota" "Mississippi" "Missouri"   
## [21] "Nebraska" "Nevada" "New Jersey" "New York"   
## [25] "North Carolina" "Ohio" "Oklahoma" "Oregon"   
## [29] "Pennsylvania" "South Carolina" "Tennessee" "Texas"   
## [33] "Utah" "Virginia" "Washington" "Wisconsin"

#Creating a data frame of all states with abbreviations:  
st.codes<-data.frame(  
 state=(c("AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL", "GA",  
 "HI", "IA", "ID", "IL", "IN", "KS", "KY", "LA", "MA", "MD", "ME",  
 "MI", "MN", "MO", "MS", "MT", "NC", "ND", "NE", "NH", "NJ", "NM",  
 "NV", "NY", "OH", "OK", "OR", "PA", "PR", "RI", "SC", "SD", "TN",  
 "TX", "UT", "VA", "VT", "WA", "WI", "WV", "WY")),  
 full=(c("alaska","alabama","arkansas","arizona","california","colorado",  
 "connecticut","district of columbia","delaware","florida","georgia",  
 "hawaii","iowa","idaho","illinois","indiana","kansas","kentucky",  
 "louisiana","massachusetts","maryland","maine","michigan","minnesota",  
 "missouri","mississippi","montana","north carolina","north dakota",  
 "nebraska","new hampshire","new jersey","new mexico","nevada",  
 "new york","ohio","oklahoma","oregon","pennsylvania","puerto rico",  
 "rhode island","south carolina","south dakota","tennessee","texas",  
 "utah","virginia","vermont","washington","wisconsin",  
 "west virginia","wyoming")), stringsAsFactors = FALSE)  
  
# Standardizing state names using the data frame "st.codes":  
matched\_states<-match(mass\_shootings$State, st.codes$state)  
  
mass\_shootings[which(!is.na(matched\_states)), "State"]<-st.codes$full[match(mass\_shootings$State[which(!is.na(matched\_states))], st.codes$state)]  
mass\_shootings$State<-tolower(mass\_shootings$State)  
  
#Adding the State Abbreviation column:  
mass\_shootings$State\_Abbr <- sapply(mass\_shootings$State, function(x) st.codes$state[match(x, st.codes$full)])  
  
mass\_shootings[,c("State", "State\_Abbr")]<-lapply(mass\_shootings[,c("State", "State\_Abbr")], as.factor)

### *Plotting the distribution of shootings over the years in various states according to shooting type*

ggplot(data = mass\_shootings) +  
geom\_point(mapping = aes(x = year, y = State, color = type))+  
 labs(title="Distribution of shootings")



### *Observations-*

* The above scatter plot shows that the number of mass shootings in USA increased over the last decade in states like California, Colorado, Florida, Texas, Washington, Ohio.