615 Final Report

Mass Shootings Analysis in USA

Jinfei Xue

Dec 17, 2018

1 Data Information

1.1 Introduction

The definition of mass shooting used for this database is 3 or more shooting victims (not necessarily fatalities), not including the shooter. The shooting must not be identifiably gang, drug, or organized crime related. The souce of the Stanford Mass Shootings of America (MSA) data is https://library.stanford.edu/projects/mass-shootings-america. The data was collected in a project began in 2012, in reaction to the mass shooting in Sandy Hook, CT. In their initial attempts to map this phenomena it was determined that no comprehensive collection of these incidents existed online. The Stanford Geospatial Center set out to create a single point repository for as many mass shooting events as could be collected via online media. The result was the Stanford MSA. The data can be downloaded from https://github.com/StanfordGeospatialCenter/MSA.

```
MSD_S <- read.csv("Stanford_MSA_Database.csv")
```

1.2 Data Cleaning

```
# Generate a new column: Total.Number.of.Injured
MSD_S$Total.Number.of.Injured <- MSD_S$Number.of.Civilian.Injured +
  MSD_S$Number.of.Enforcement.Injured

# Select interesting variables into new dataset: MSD
MSD <- MSD_S %>%
  dplyr::select(Title, Location, City, State, Latitude, Longitude,
         Total.Number.of.Injured, Total.Number.of.Fatalities,
         Total.Number.of.Victims, Date, Day.of.Week,
         Number.of.shooters, Average.Shooter.Age, Shooter.Sex, Shooter.
Race,
         Type.of.Gun...General, Total.Number.of.Guns,
         Fate.of.Shooter.at.the.scene, Fate.of.Shooter, Shooter.s.Cause.
of.Death,
         School.Related, Place.Type, Targeted.Victim.s...General,
         History.of.Mental.Illness...General, Military.Experience)
```

```r
# Split date into Year and Month
MSD$Year <- format(as.Date(MSD$Date, format="%m/%d/%Y"),"%Y")
MSD$Month <- format(as.Date(MSD$Date, format="%m/%d/%Y"),"%m")

# Transfer catergories of School.Related
MSD$School.Related[MSD$School.Related=="Killed"]<- "Unknown"
MSD$School.Related[MSD$School.Related=="no"]<- "No"

# Combine categories of shooter.Race
MSD$Shooter.Race[MSD$Shooter.Race=="Asian American/Some other race"]<-
"Asian American"
MSD$Shooter.Race[MSD$Shooter.Race=="Black American or African American/
Unknown"]<-
  "Black American or African American"
MSD$Shooter.Race[MSD$Shooter.Race=="Some other race"]<- "Some Other Rac
e"
MSD$Shooter.Race[MSD$Shooter.Race=="White American or European American
/Some other Race"]<-
  "White American or European American"

# Check NA
sapply(MSD, function(x) sum(is.na(x)))
```

```
##                                Title                                Locat
ion
##                                    0
  0
##                                 City                                   St
ate
##                                    0
  0
##                             Latitude                               Longit
ude
##                                    0
  0
##             Total.Number.of.Injured          Total.Number.of.Fatalit
ies
##                                    0
  0
##             Total.Number.of.Victims                                    D
ate
##                                    0
  0
##                          Day.of.Week                  Number.of.shoot
ers
##                                    0
  0
##                 Average.Shooter.Age                          Shooter.
Sex
##                                    0
```

```
  0
##                         Shooter.Race                Type.of.Gun...Gene
ral
##                                    0
  0
##                 Total.Number.of.Guns         Fate.of.Shooter.at.the.sc
ene
##                                    0
  0
##                      Fate.of.Shooter             Shooter.s.Cause.of.De
ath
##                                    0
  0
##                       School.Related                           Place.T
ype
##                                    0
  0
##         Targeted.Victim.s...General History.of.Mental.Illness...Gene
ral
##                                    0
  0
##                   Military.Experience                                Y
ear
##                                    0
  0
##                                Month
##                                    0

# Structure of the data
str(MSD)

## 'data.frame':    335 obs. of  27 variables:
##  $ Title                              : Factor w/ 334 levels "49th S
treet Elementary School",..: 307 254 197 57 211 158 32 118 306 309 ...
##  $ Location                           : Factor w/ 257 levels "Aiken,
 South Carolina",..: 11 148 165 39 178 136 81 211 45 130 ...
##  $ City                               : Factor w/ 251 levels "Aiken",
"Albuquerque",..: 11 147 164 39 175 136 80 208 44 130 ...
##  $ State                              : Factor w/ 48 levels "Alabama
","Alaska",..: 41 3 19 14 32 5 5 5 38 29 ...
##  $ Latitude                           : num  30.2 33.4 30.1 41.8 42.
1 ...
##  $ Longitude                          : num  -97.8 -111.8 -89.9 -87.
7 -78.4 ...
##  $ Total.Number.of.Injured            : int  32 1 13 3 7 6 2 9 5 2
 ...
##  $ Total.Number.of.Fatalities         : int  16 5 10 1 3 1 7 2 2 1
 ...
##  $ Total.Number.of.Victims            : int  48 6 22 4 10 8 9 11 7 3
 ...
```

```
##  $ Date                                : Factor w/ 296 levels "1/1/20
15","1/10/2015",..: 260 49 84 5 81 100 247 16 42 143 ...
##  $ Day.of.Week                         : Factor w/ 7 levels "Friday",
"Monday",..: 2 3 4 5 2 5 2 2 3 1 ...
##  $ Number.of.shooters                  : Factor w/ 6 levels "","1","2
","3",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Average.Shooter.Age                 : Factor w/ 63 levels "12","13
","14",..: 22 11 19 3 8 11 38 6 12 8 ...
##  $ Shooter.Sex                         : Factor w/ 4 levels "Female",
"Male",..: 2 2 2 2 2 2 2 1 2 2 ...
##  $ Shooter.Race                        : Factor w/ 11 levels "Asian A
merican",..: 10 10 3 9 10 10 10 10 3 10 ...
##  $ Type.of.Gun...General               : Factor w/ 11 levels "\nMulti
ple guns",..: 5 4 5 4 5 10 8 8 4 4 ...
##  $ Total.Number.of.Guns                : Factor w/ 10 levels "0","1",
"10","2",..: 9 2 4 4 4 2 2 2 2 2 ...
##  $ Fate.of.Shooter.at.the.scene        : Factor w/ 6 levels "Arrested
","Custody",..: 4 2 4 2 2 2 2 2 2 2 ...
##  $ Fate.of.Shooter                     : Factor w/ 8 levels "","Arres
ted",..: 6 3 6 3 3 3 3 3 3 3 ...
##  $ Shooter.s.Cause.of.Death            : Factor w/ 6 levels "Killed",
"Killed/Suicide",..: 1 4 1 4 4 4 4 4 4 4 ...
##  $ School.Related                      : Factor w/ 5 levels "Killed",
"no",..: 5 5 3 5 5 5 5 5 5 5 ...
##  $ Place.Type                          : Factor w/ 30 levels "College
/University/Adult education",..: 1 1 5 11 27 1 1 11 1 27 ...
##  $ Targeted.Victim.s...General         : Factor w/ 30 levels "","A so
cial altercation led to the shooting.",..: 12 28 15 28 12 28 3 28 28 28
 ...
##  $ History.of.Mental.Illness...General: Factor w/ 3 levels "No","Unk
nown",..: 3 3 3 3 1 2 3 3 2 3 ...
##  $ Military.Experience                 : Factor w/ 3 levels "No","Unk
nown",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ Year                                : chr  "1966" "1966" "1972" "1
974" ...
##  $ Month                               : chr  "08" "11" "12" "01" ...
```

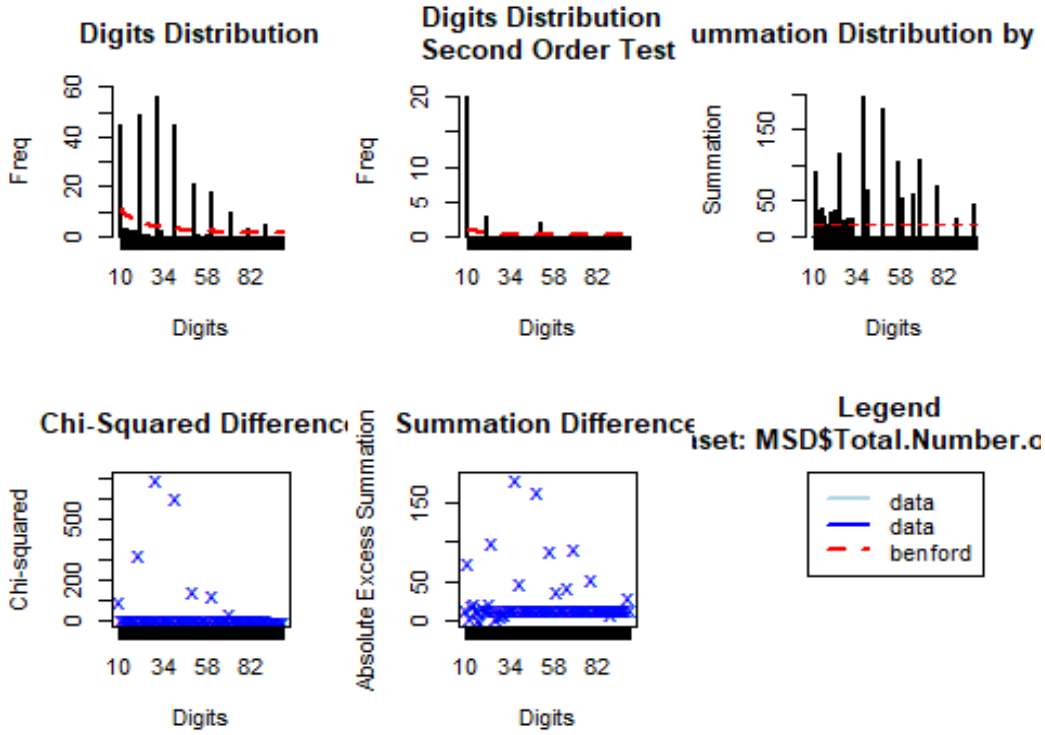
2 Data Validity Analysis

In this part, we can juestify whether total number of victims in each mass shooting accident follow the benford distribution to find out the potential fraud published data.

2.1 Total number of injured

```
bfd.injured <- benford(MSD$Total.Number.of.Injured)
bfd.injured
```

```
##
## Benford object:
##
## Data: MSD$Total.Number.of.Injured
## Number of observations used = 275
## Number of obs. for second order = 26
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.434
##          Var  0.071
##  Ex.Kurtosis -0.868
##     Skewness -0.212
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     30         52.08
## 2     20         43.17
## 3     40         42.05
## 4     10         33.62
## 5     50         18.63
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  MSD$Total.Number.of.Injured
## X-squared = 2243.9, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  MSD$Total.Number.of.Injured
## L2 = 0.033577, df = 2, p-value = 9.771e-05
##
## Mean Absolute Deviation: 0.0177041
## Distortion Factor: -50.13374
##
## Remember: Real data will never conform perfectly to Benford's Law. Y
ou should not focus on p-values!

plot(bfd.injured)
```

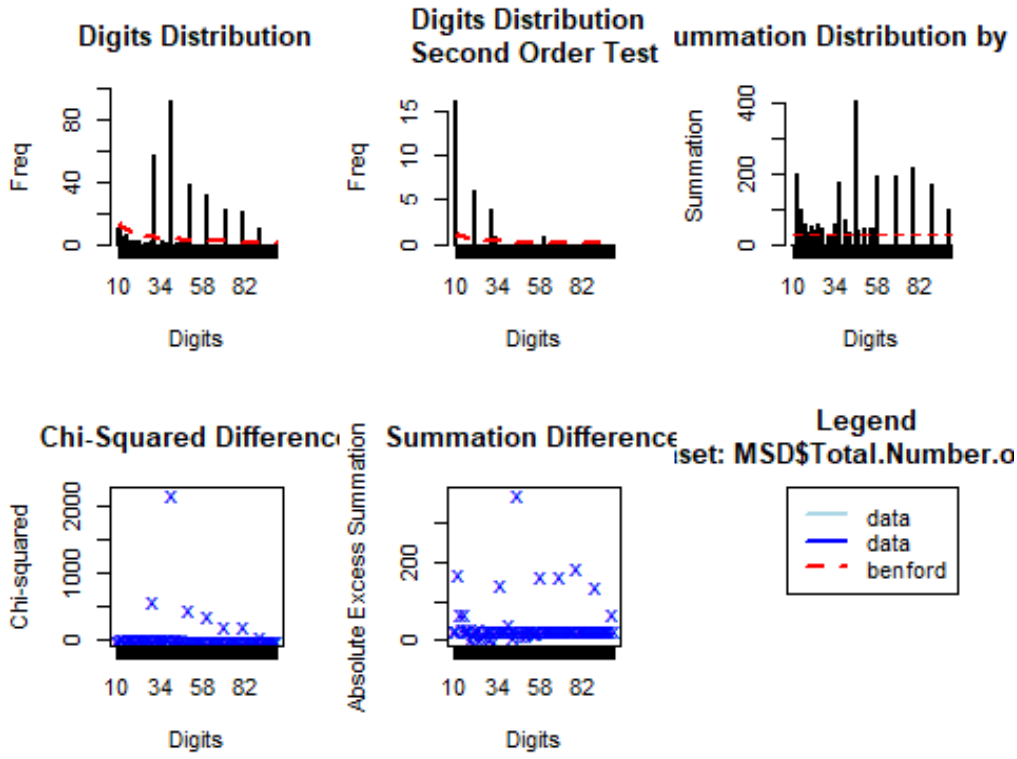


The result of chi-squared test shows p-value is smaller than 0.05. Therefore, we should reject the null hypothesis, which means it is probable that there may exist some fake data in the total number of injured. Also, from the plots, we can also find out the total number of injured does not follow benford law very well. But we should notice real data will never conform perfectly to Benford's Law.

2.2 Total Number of Victims

```
bfd.victims <- benford(MSD$Total.Number.of.Victims)
bfd.victims

##
## Benford object:
##
## Data: MSD$Total.Number.of.Victims
## Number of observations used = 335
## Number of obs. for second order = 29
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.583
##          Var  0.056
##  Ex.Kurtosis  0.342
##     Skewness -0.832
##
```

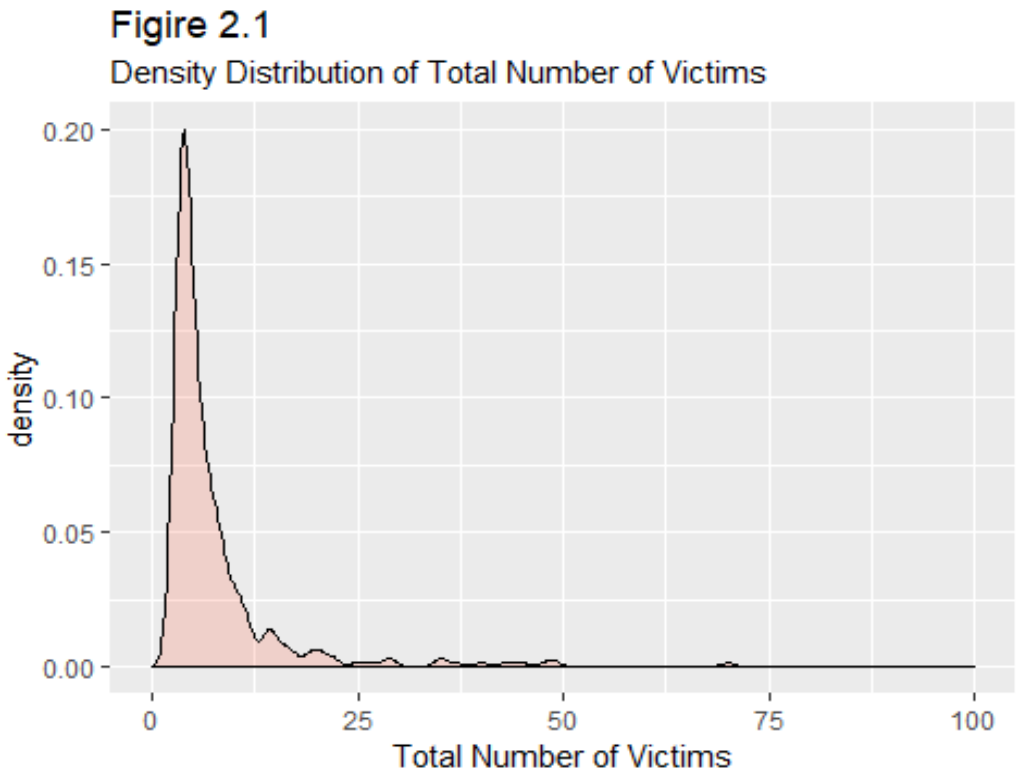
```
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     40         88.41
## 2     30         53.23
## 3     50         36.12
## 4     60         29.60
## 5     70         19.94
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  MSD$Total.Number.of.Victims
## X-squared = 4265.4, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  MSD$Total.Number.of.Victims
## L2 = 0.19392, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.01697329
## Distortion Factor: -51.1163
##
## Remember: Real data will never conform perfectly to Benford's Law. Y
ou should not focus on p-values!

plot(bfd.victims)
```



The result of chi-squared test shows p-value is smaller than 0.05. Therefore, we should reject the null hypothesis, which means it is probable that there may exist some fake data in the total number of victims. Also, from the plots, we can also find out the total number of victims does not follow benford law very well. But we should notice real data will never conform perfectly to Benford's Law.

2.3 Distribution Plot

```
ggplot(MSD, aes(x=Total.Number.of.Victims)) +
    geom_histogram(aes(y=..density..),binwidth=100,
                   colour="white", fill="tomato") +
    geom_density(alpha=.2, fill="tomato")+
    labs(title="Figire 2.1",
         subtitle="Density Distribution of Total Number of Victims",
         x="Total Number of Victims")+
    xlim(c(0,100))
```

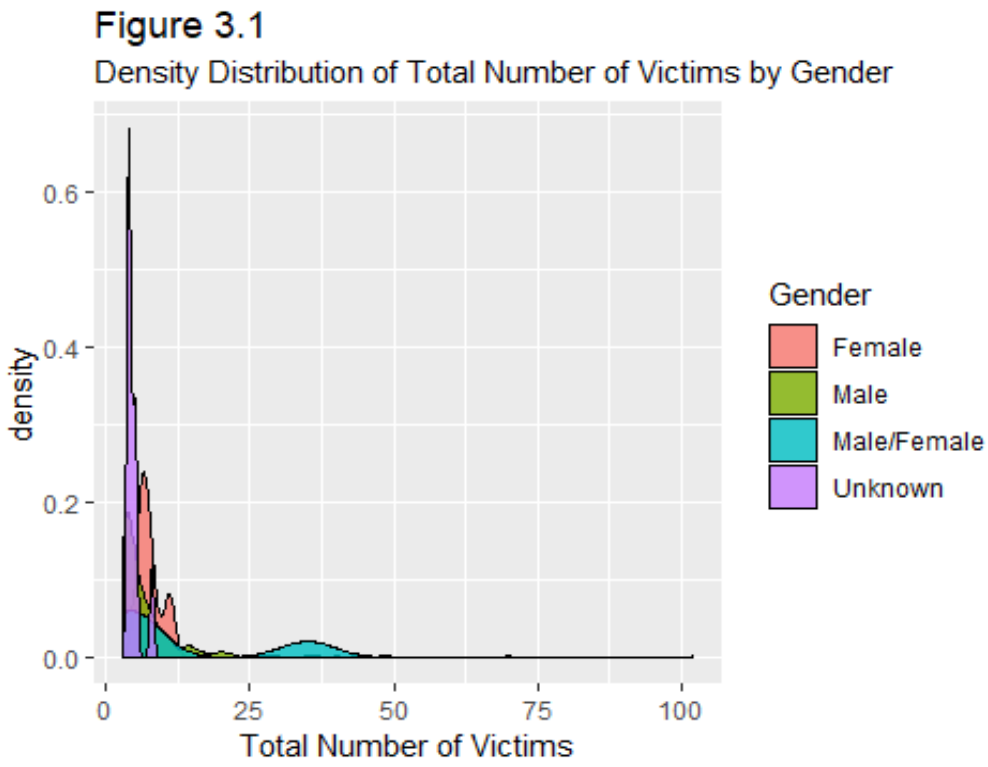


From the plot, we can see large total number of victims happens less frequent than small total number of victims. The distribution is obviously right-skewed.

3 Exploratory Data Analysis

3.1 Density Distribution of Total Number of Victims by Gender

```
g <- ggplot(MSD, aes(Total.Number.of.Victims))
g + geom_density(aes(fill=factor(Shooter.Sex)), alpha=0.8) +
    labs(title="Figure 3.1",
         subtitle="Density Distribution of Total Number of Victims by G
ender",
         x="Total Number of Victims",
         fill="Gender")
```

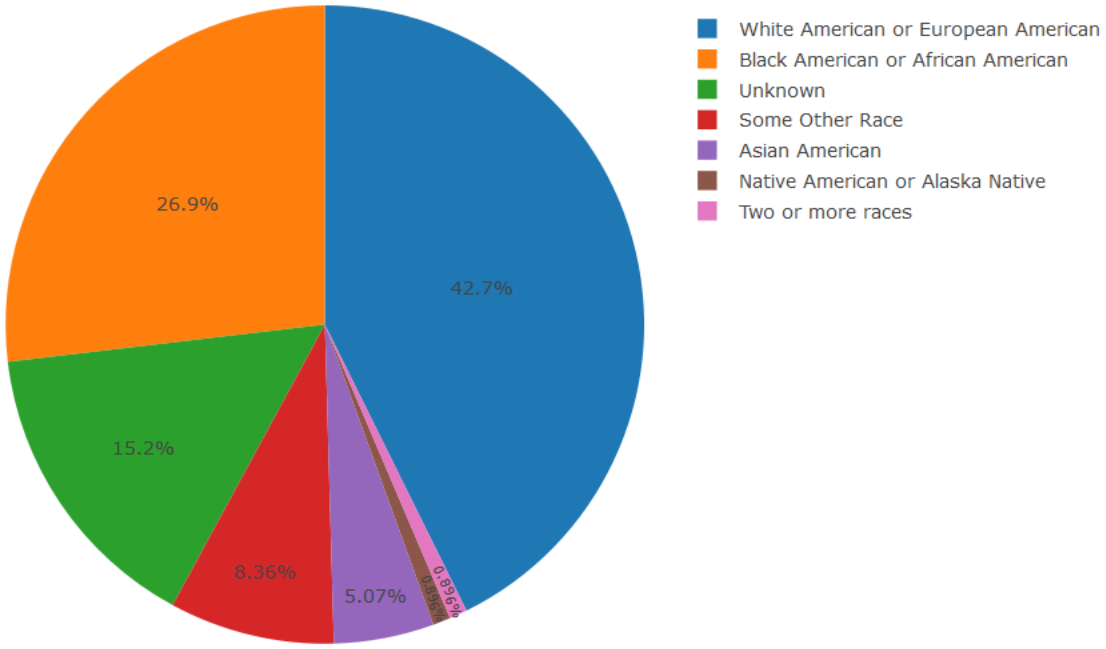


We can see distributions of total number of victims for different genders are largely different from each other.

3.2 Pie Chart of Shooter Race

```
MSD_r <- MSD %>%
  select(Shooter.Race) %>%
  group_by(Shooter.Race) %>%
  summarise(count=n())

race_pie <- plot_ly(MSD_r, labels = ~Shooter.Race, values = ~count, typ
e = 'pie',
                    textposition = 'inside',textinfo = 'percent') %>%
  layout(title="Figure 3.2 Pie Chart of Shooter Race")

ggplotly(race_pie)
```

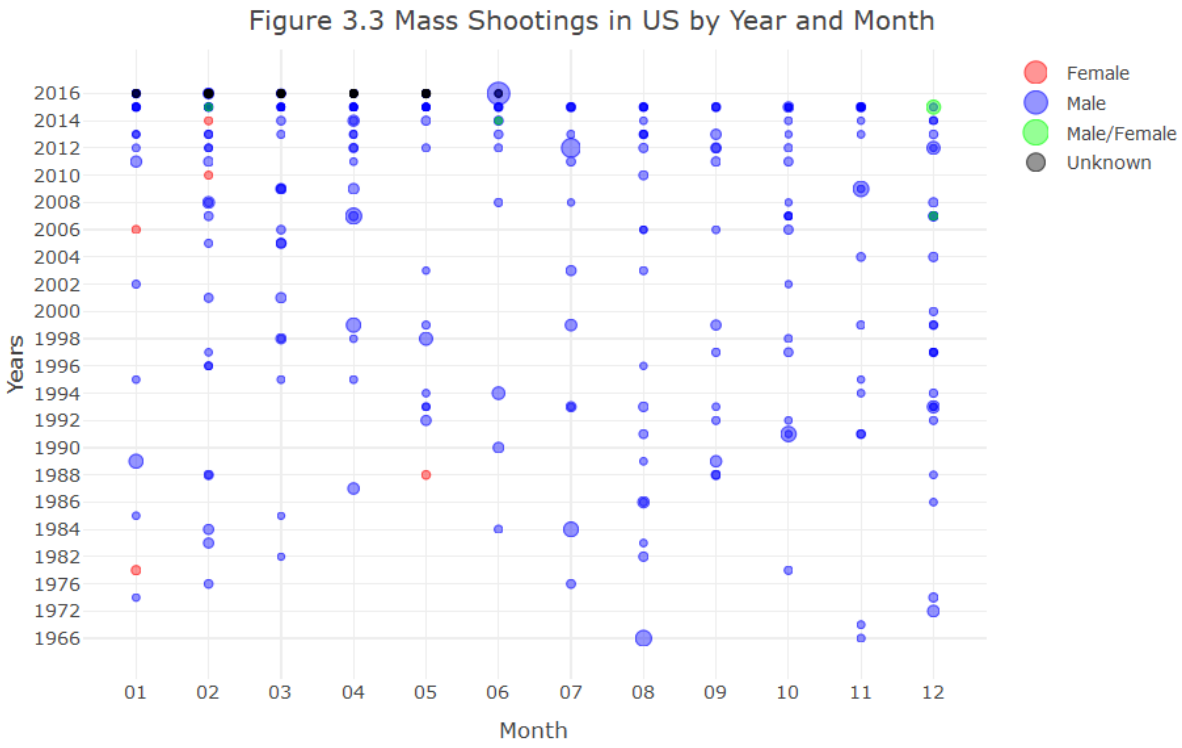
Figure 3.2 Pie Chart of Shooter Race



We can see from the pie chart that most of shooters are “White American or European American”.And only 0.896% of shooters are “Native American or Alaska Native”.

3.3 Mass Shootings in US by Year and Month

```
plot_ly(data = MSD
        ,type = 'scatter'
        ,mode = 'markers'
        ,hoverinfo = 'text'
        ,x = ~Month
        ,y = ~Year
        ,size = ~Total.Number.of.Victims
        ,color = ~Shooter.Sex
        ,colors = c('Red', 'Blue', 'Green', 'Black')
        ,alpha = 0.6
        ,text = ~paste("Title: ", Title
               ,"\nLocation: ", Location
               ,'\n Date: ', Date
               ,'\n Total victims : ', Total.Number.of.Victims)) %>%
  layout(title = "Figure 3.3 Mass Shootings in US by Year and Month"
         , xaxis = list(title = "Month")
         , yaxis = list(title = "Years"))
```



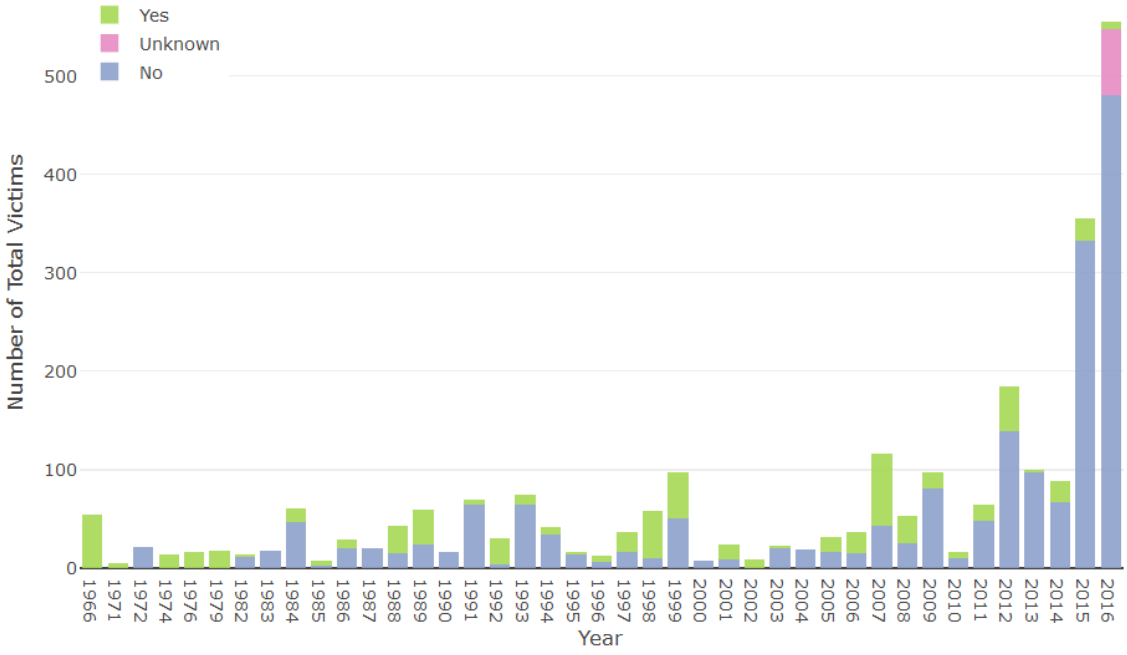
The circle size represents the number of victims in the mass shooting and the color means the gender of shooters. We can find out that most of shooters are males and the worst mass shooting happened randomly.

3.4 Total victims by Years and School Related

```
# year + race
MSD_ys <- MSD %>%
  select(Year, School.Related, Total.Number.of.Victims) %>%
  group_by(Year, School.Related) %>%
  summarise(sum=sum(Total.Number.of.Victims))

plot_ly(data = MSD_ys
        ,type = 'bar'
        ,mode = 'markers'
        ,x = ~Year
        ,y = ~sum
        ,color = ~School.Related
        ,alpha = 0.9) %>%
  layout(title = "Figure 3.4 Total victims by Years and School Related"
         , xaxis = list(title = "Year")
         , yaxis = list(title = "Number of Total Victims")
         , showlegend = T
         , barmode = 'stack'
         , position = 1
         , xaxis = list(title = "")
         , yaxis = list(title = "")
```

```
        , legend = list(x = 0, y = 1)
        , hovermode = 'compare')
```

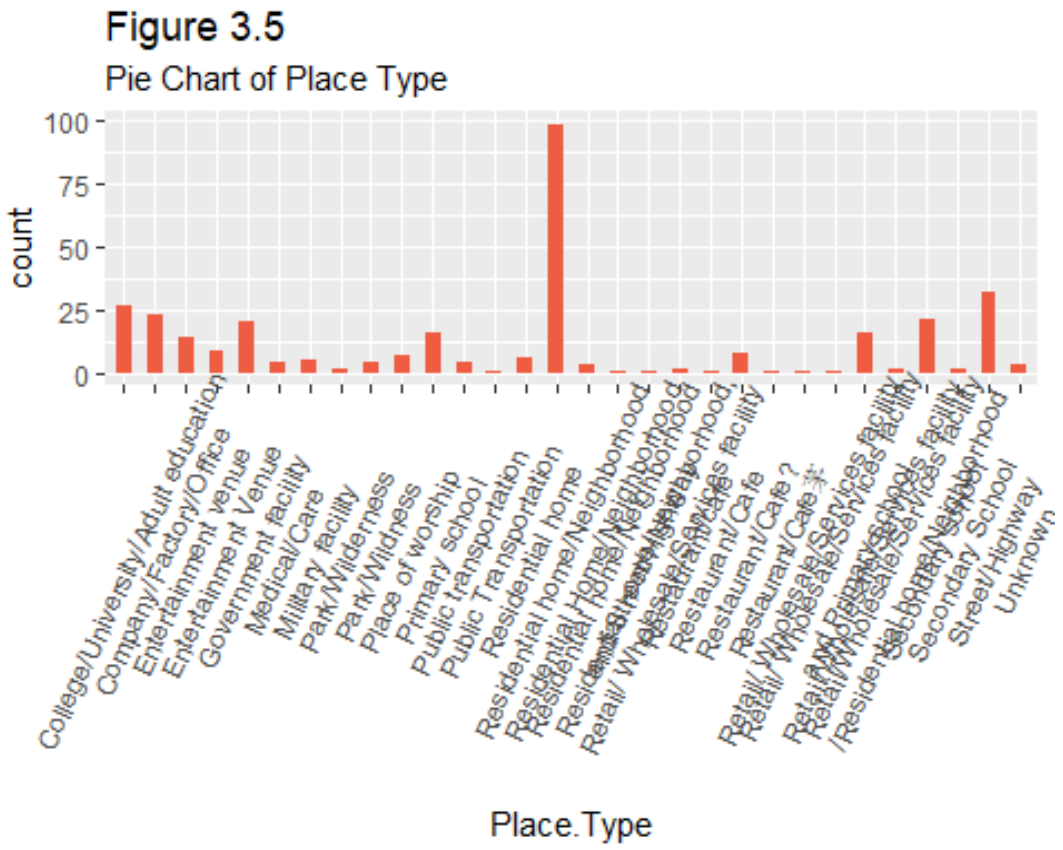
Figure 3.4 Total victims by Years and School Related



We can see the total number of victims in 2016 is the largest and there is an increasing trend over years. Besides, the large percentage of school related mass shooting cannot be ignored.

3.5 Bar Chart of Place Type

```
MSD_place <- MSD %>%
  select(Place.Type, Total.Number.of.Victims) %>%
  group_by(Place.Type) %>%
  summarise(count=n()) %>%
  arrange(desc(count))

g <- ggplot(MSD_place, aes(Place.Type, count))
g + geom_bar(stat="identity", width = 0.5, fill="tomato2") +
      labs(title="Figure 3.5",
           subtitle="Pie Chart of Place Type",
           xlab="Place Type") +
      theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



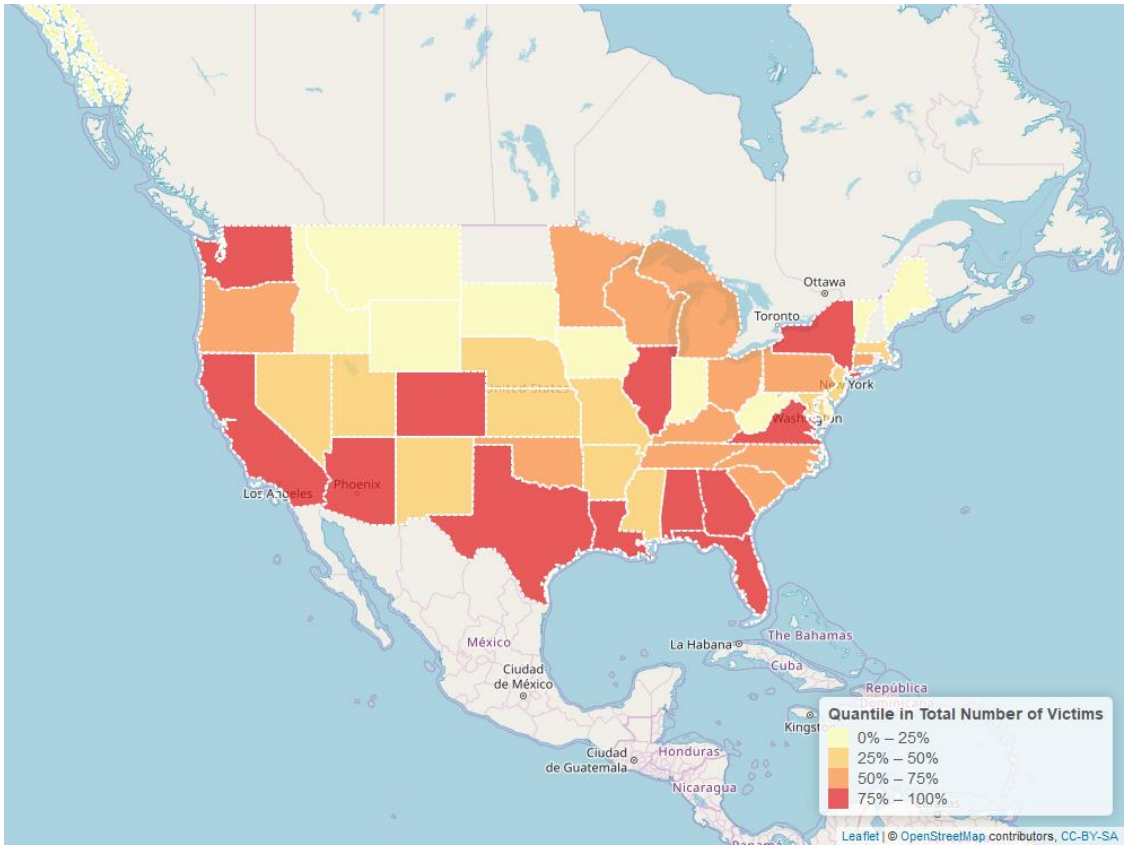
We can see the frequency of mass shootings happened in residual home is the largest, then the frequency of street/highway is the second largest.

3.6 Map for Total Number of Victims in Each State

```
MSD_state <- MSD %>%
  select(State, Total.Number.of.Injured, Total.Number.of.Victims) %>%
  group_by(State) %>%
  mutate(Sum.Injured=sum(Total.Number.of.Injured)) %>%
  mutate(Sum.Victims=sum(Total.Number.of.Victims)) %>%
  select(State, Sum.Injured, Sum.Victims) %>%
  unique()

## Download data from Natural Earth
url <- "http://www.naturalearthdata.com/http//www.naturalearthdata.com/
download/50m/cultural/ne_50m_admin_1_states_provinces.zip"
tmp <- tempdir()
file <- basename(url)
download.file(url, file)
unzip(file, exdir = tmp)
## Read the data into R
state_spatial <- readOGR(dsn=tmp,
                         layer = "ne_50m_admin_1_states_provinces",
                         encoding = "UTF-8")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:\Fiona\AppData\Local\Temp\Rtmp6xfIAV", layer: "ne_50m_adm
in_1_states_provinces"
## with 100 features
## It has 83 fields
## Integer64 fields read as strings:  ne_id

#get the states name in spatial data
a<-state_spatial@data[["gn_name"]] #because the state name in "states"
are in lower format
# state_spatial@data[["gn_name"]]<-a
data<-sp::merge(state_spatial,MSD_state,by.x="gn_name",by.y="State",sor
t=FALSE,duplicateGeoms = TRUE,all.x=FALSE)
labels <- sprintf(
  "<strong>%s</strong><br/>%s Total number of injured<br/>%g Total numb
er of victims",
  data$gn_name, data$Sum.Injured, data$Sum.Victims
) %>% lapply(htmltools::HTML)


# leaflet
m <- leaflet(data) %>%
  addTiles()%>%
  setView(-96, 37.8, 4)
#bins
pal <- colorQuantile("YlOrRd", domain = MSD_state$Sum.Victims)
#pal(states$Donations)
m_victims <- m %>% addPolygons(
  fillColor = ~pal(data$Sum.Victims),
  weight = 2,
  opacity = 1,
  color = "white",
  dashArray = "3",
  fillOpacity = 0.7,
  highlight = highlightOptions(
    weight = 5,
    color = "#666",
    dashArray = "",
    fillOpacity = 0.7,
    bringToFront = TRUE),
  label = labels,
  labelOptions = labelOptions(
    style = list("font-weight" = "normal", padding = "3px 8px"),
    textsize = "15px",
    direction = "auto")
) %>%
  addLegend(pal = pal, values = ~Sum.Victims, opacity = 0.7, title = "Q
uantile in Total Number of Victims",
            position = "bottomright")
m_victims
```

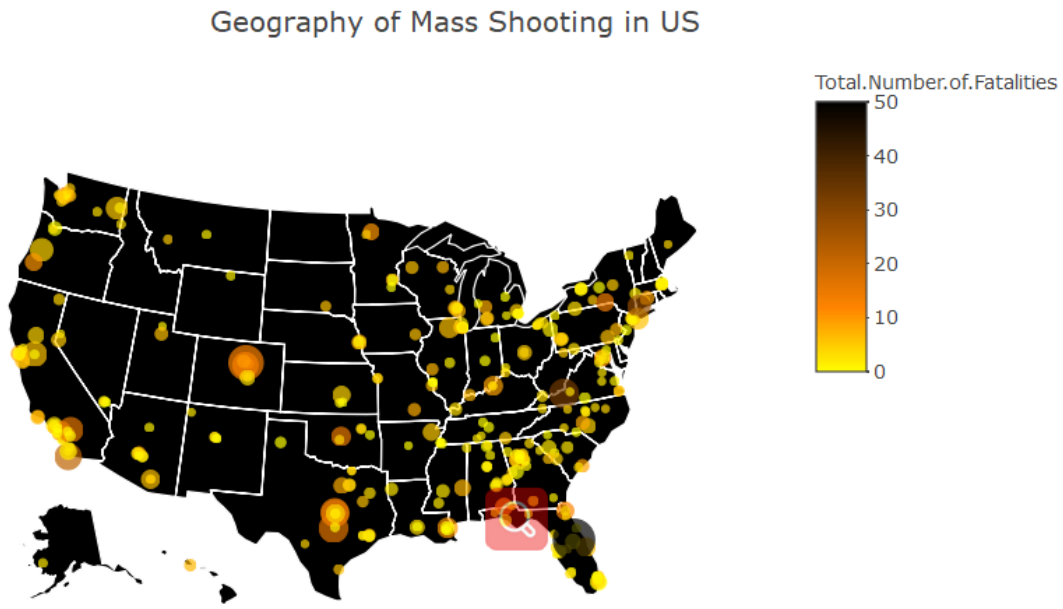
```
mapview::mapshot(m_victims, file = "mapstate.png")
knitr::include_graphics("mapstate.png")
```



3.7 Map for Mass Shooting Distribution

```
g <- list(
    scope = 'usa'
  , projection = list(type = 'albers usa')
  , showland = TRUE
  , landcolor = 'black'
  , subunitwidth = 1
  , countrywidth = 1
 , subunitcolor = toRGB("white")
 , countrycolor = toRGB("white")
)

plot_geo(MSD
         #, locationmode = 'USA-states'
         , sizes = c(10, 300)) %>%
  add_markers(
    x = ~Longitude
    , y = ~Latitude
    , size = ~Total.Number.of.Victims
    , color = ~Total.Number.of.Fatalities
    , colors = colorRamp(c("yellow", "red", "black"))
```

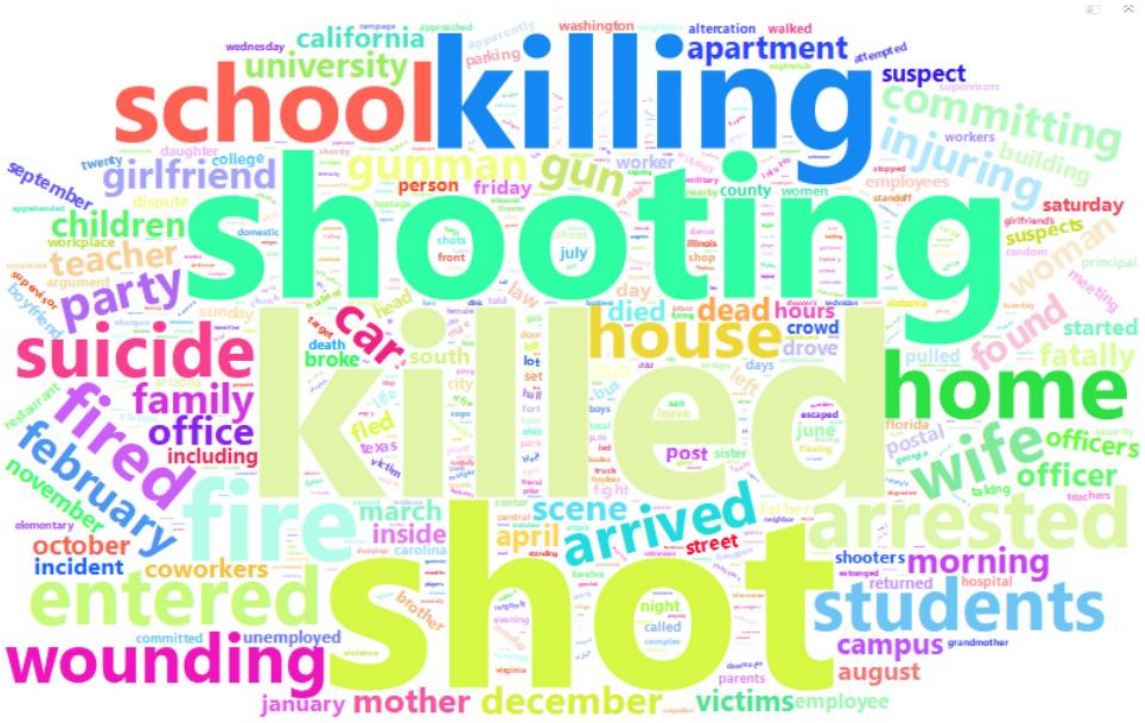
```
    , hoverinfo = "text"
    , text = ~paste(MSD$Title
                    , '\n Fatalities: ', MSD$Total.Number.of.Fatalities
                    , '\n Injured: ', MSD$Total.Number.of.Injured)
  ) %>%
  layout(title = 'Geography of Mass Shooting in US', geo = g)
```



We can see that most mass shooting happened along the east coastline.

4 Text Mining

```
# Remove numbers
text <- data.frame(text=removeNumbers(as.character(MSD_S$Description)))
text$text <- as.character(text$text)

# Sort frequency
text_n <-text %>%
  unnest_tokens(output=word, input=text) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)

# Wordcloud
wordcloud2(text_n,color="random-light",rotateRatio = 0.3)
```



We can see the most eight frequent words are “killed”, “shooting”, “school”, “killing”, “shot”, “home”, “students” and ”suicide”.

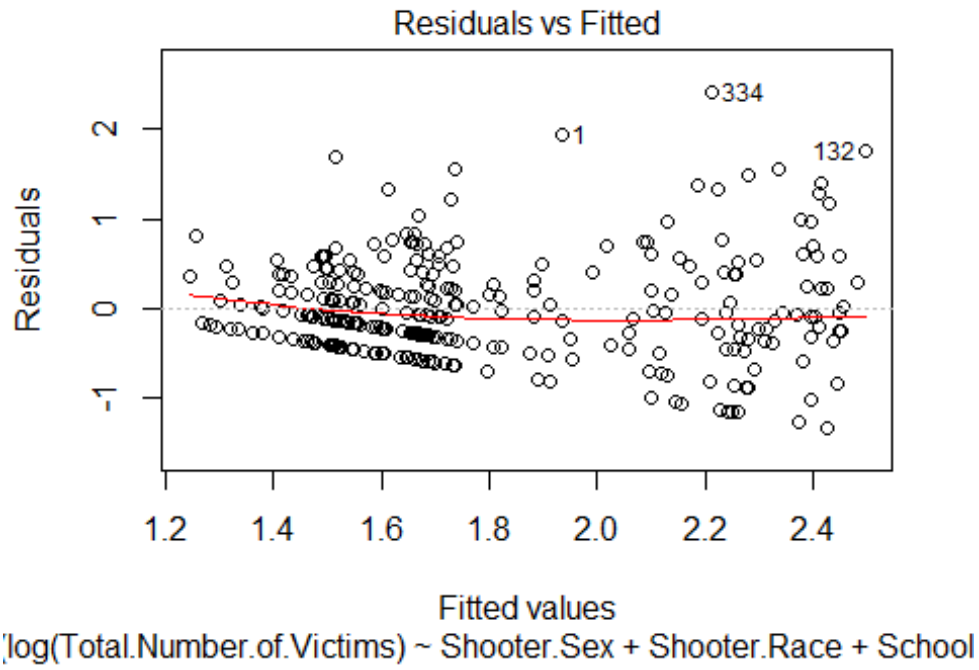
5 Modeling

```
r1 <- lm(log(Total.Number.of.Victims)~Shooter.Sex+Shooter.Race+School.R
elated+
           as.numeric(Average.Shooter.Age)+poly(as.numeric(Total.Number.
of.Guns),degree=2), data=MSD)
summary(r1)

##
## Call:
## lm(formula = log(Total.Number.of.Victims) ~ Shooter.Sex + Shooter.Ra
ce +
##     School.Related + as.numeric(Average.Shooter.Age) + poly(as.numer
ic(Total.Number.of.Guns),
##     degree = 2), data = MSD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3269 -0.3176 -0.1056  0.2781  2.4141
##
## Coefficients:
##                                                      Estimate Std. E
rror
## (Intercept)                                          2.286846   0.32
2030
```

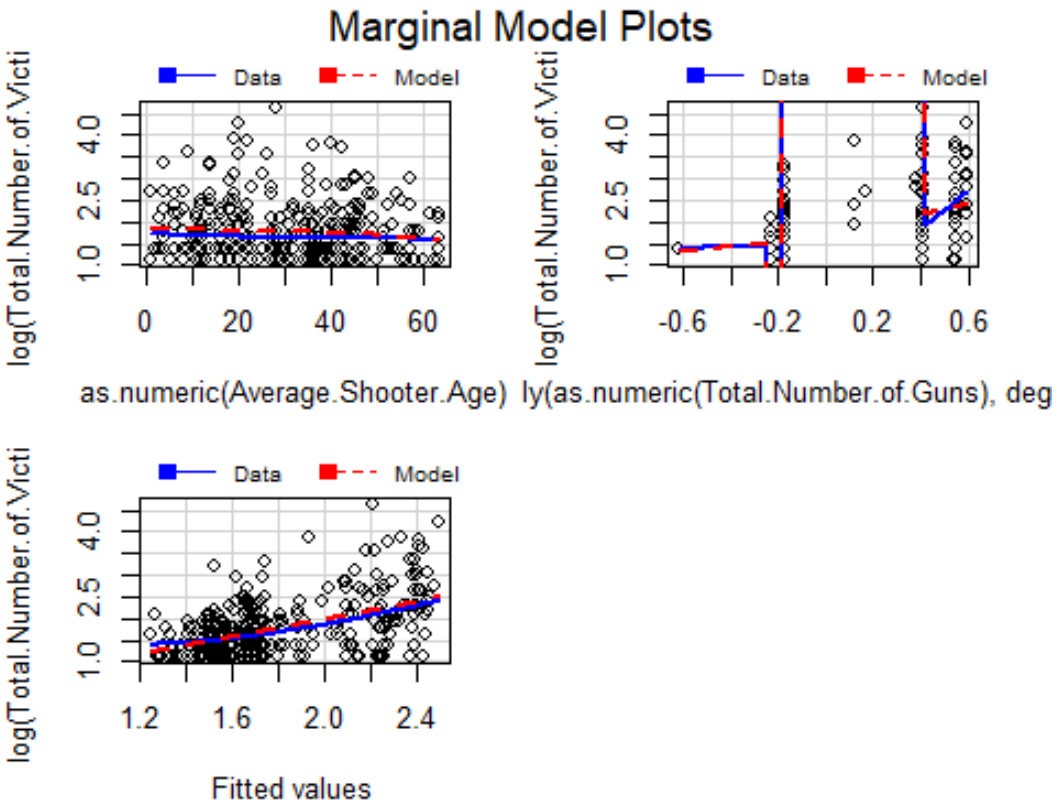
```
## Shooter.SexMale                                          -0.233329   0.26
3055
## Shooter.SexMale/Female                                   -0.645376   0.39
0731
## Shooter.SexUnknown                                        0.014205   0.31
2827
## Shooter.RaceBlack American or African American           -0.293947   0.15
1237
## Shooter.RaceNative American or Alaska Native             -0.089957   0.36
0941
## Shooter.RaceSome Other Race                              -0.196481   0.17
4111
## Shooter.RaceTwo or more races                            -0.040855   0.35
5066
## Shooter.RaceUnknown                                      -0.460007   0.18
2003
## Shooter.RaceWhite American or European American          -0.112876   0.14
4532
## School.RelatedUnknown                                     0.014221   0.16
8119
## School.RelatedYes                                        -0.088223   0.08
6544
## as.numeric(Average.Shooter.Age)                          -0.001907   0.00
2221
## poly(as.numeric(Total.Number.of.Guns), degree = 2)1       0.108433   0.65
1711
## poly(as.numeric(Total.Number.of.Guns), degree = 2)2      -5.462212   0.59
7469
##                                                        t value Pr(>|t|)

## (Intercept)                                              7.101 8.04e-12
 ***
## Shooter.SexMale                                         -0.887   0.3757

## Shooter.SexMale/Female                                  -1.652   0.0996
 .
## Shooter.SexUnknown                                       0.045   0.9638

## Shooter.RaceBlack American or African American          -1.944   0.0528
 .
## Shooter.RaceNative American or Alaska Native            -0.249   0.8033

## Shooter.RaceSome Other Race                             -1.128   0.2600

## Shooter.RaceTwo or more races                           -0.115   0.9085

## Shooter.RaceUnknown                                     -2.527   0.0120
 *
## Shooter.RaceWhite American or European American         -0.781   0.4354
```

```
## School.RelatedUnknown                                    0.085   0.9326

## School.RelatedYes                                       -1.019   0.3088

## as.numeric(Average.Shooter.Age)                         -0.859   0.3912

## poly(as.numeric(Total.Number.of.Guns), degree = 2)1      0.166   0.8680

## poly(as.numeric(Total.Number.of.Guns), degree = 2)2     -9.142  < 2e-16
 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5557 on 320 degrees of freedom
## Multiple R-squared:  0.2684, Adjusted R-squared:  0.2363
## F-statistic: 8.384 on 14 and 320 DF,  p-value: 2.134e-15
```

```
plot(r1, which=1)
```




```
car::marginalModelPlots(r1)
```



From the summary result, the p value of F statistics is small and some coefficients are significant. Besides, from the residual plot, we can see that although some points have a little decreasing trend, they relatively randomly dispersed around the horizontal line at zero (the dashed black line). And after looking at the last marginal plot, we can conclude the polynomial model can fit the data well.

So the linear model is:

$$log(Total.Number.of.Victims) = 2.286846 - 0.233329 \times Male - 0.645376 \times Male/Female + 0.014205 \times SexUnknown - 0.293947 \times BlackAmericanorAfricanAmerican - 0.089957 \times NativeAmericanorAlaskaNative - 0.196481 \times SomeOtherRace - 0.040855 \times Twoormoreraces - 0.460007 \times RaceUnknown - 0.112876 \times WhiteAmericanorEuropeanAmerican + 0.014221 \times SchoolRelatedUnknown - 0.088223 SchoolRelatedYes - 0.001907 \times Shooter.Age + 0.108433 \times guns - 5.462212 \times guns^2$$