Data 607 - Final Project

Joseph Simone

12/1/2019

```
library(jsonlite)
library(tidyverse)
library(tidyr)
library(dplyr)
library(tibble)
library(knitr)
library(ggplot2)
library(mapproj)
library(fiftystater)
library(kableExtra)
```

Data Management and Acquisition Final Project

Mass and School Shootings in America Analysis over a 40 year Span

Background

Not only is the topic of Gun Violence in America considered to be a "*hot*" "*topic*" for debate, the School and Mass Shooting epidemic in this country is one that can " hit close to home" for most of us. During the research phase of this project, I could not decide whether I was going to focus just on "*School*" related shooting or stick with "$Mass$" shootings as a whole. Therefore, first I am going to explore School shooting data as it pertains to Mass Shooting Data.

Data Sources

Open Data Soft - Mass Shooting in America

Mass Shooting in America

Center for Homeland Defense and Security - K-12 School Shooting Database

K-12 School Shooting Database

Research Question(s) ?

1. "According to the Mass Schooting Data-Set, which percentage of the data involves School related shootings?"
2. "According to the School Shooting Data-Set, which State has the most occurences of gun related school instances?"

Mass Shooting Data

Importation of Data-Set

```
json_mass_file <-"https://raw.githubusercontent.com/josephsimone/Data_607_Final_Project/master/mass-sho
```

```
mass <- fromJSON(json_mass_file)
```

```
names(mass)
```

```
## [1] "datasetid"        "recordid"         "fields"
## [4] "geometry"         "record_timestamp"
```

Mass Shooting Data Extrations

All the information for this Data-Set are stored in a Nested Json Category labeled "Fields"

```
mass_fields <- as.data.frame(mass$fields)

mass_fields <- mass_fields %>%
    select(date, everything())
```

```
names(mass_fields)
```

```
##  [1] "date"
##  [2] "history_of_mental_illness_detailed"
##  [3] "school_related"
##  [4] "shooter_name"
##  [5] "type_of_gun_detailed"
##  [6] "number_of_victims_injured"
##  [7] "longitude"
##  [8] "date_detailed"
##  [9] "place_type"
## [10] "city"
## [11] "shooter_age_s"
## [12] "total_number_of_fatalities"
## [13] "targeted_victim_s_general"
## [14] "number_of_handguns"
## [15] "state"
## [16] "number_of_semi_automatic_guns"
## [17] "number_of_rifles"
## [18] "targeted_victim_s_detailed"
## [19] "location"
## [20] "possible_motive_general"
## [21] "average_shooter_age"
## [22] "number_of_automatic_guns"
## [23] "history_of_mental_illness_general"
## [24] "relationship_to_incident_location"
## [25] "data_source_3"
## [26] "description"
## [27] "type_of_gun_general"
## [28] "data_source_1"
## [29] "data_source_2"
## [30] "military_experience"
## [31] "data_source_4"
```

```
## [32] "fate_of_shooter_at_the_scene"
## [33] "number_of_shotguns"
## [34] "total_number_of_victims"
## [35] "shooter_race"
## [36] "class"
## [37] "shooter_s_cause_of_death"
## [38] "latitude"
## [39] "geopoint"
## [40] "caseid"
## [41] "day_of_week"
## [42] "total_number_of_guns"
## [43] "number_of_victim_fatalities"
## [44] "possible_motive_detailed"
## [45] "title"
## [46] "shooter_sex"
## [47] "data_source_5"
## [48] "data_source_6"
## [49] "data_source_7"
```

Mass Shooting Data-Set Tidying

Date Format to Match the Second Data-Set

```
mass_fields$date <- format(as.Date(mass_fields$date , format = "%Y-%m-%d"), "%m/%d/%Y")
head(mass_fields,1)
```

```
##         date
## 1 12/30/1999
##
## 1 The shooter did not have any history of treatment for mental illness, and his family, friends and
##   school_related           shooter_name
## 1             No Silvio Izquierdo-Leyva
##                                                        type_of_gun_detailed
## 1 9mm Lorcin semi-automatic pistol, .38 caliber Charter Arms semi-automatic pistol
##   number_of_victims_injured longitude               date_detailed
## 1                         3 -82.44504 Thursday, December 30, 1999
##                           place_type  city shooter_age_s
## 1 Retail/Wholesale/Services facility Tampa            36
##   total_number_of_fatalities               targeted_victim_s_general
## 1                          5 Colleague/Workmate/Business acquaintance
##   number_of_handguns   state number_of_semi_automatic_guns
## 1                  2 Florida                             2
##   number_of_rifles targeted_victim_s_detailed       location
## 1                0                 Coworkers Tampa, Florida
##   possible_motive_general average_shooter_age number_of_automatic_guns
## 1                 Unknown                  36                        0
##   history_of_mental_illness_general relationship_to_incident_location
## 1                                No      Place of business/employment
##                                        data_source_3
## 1 http://murderpedia.org/male.I/i/izquierdo-leyva.htm
##
## 1 On December 30, 1999, a 36-year-old employee of the Radisson Bay Harbor Inn in Tampa, Florida arri
##   type_of_gun_general                             data_source_1
```

```
## 1             Handgun http://www.vpc.org/studies/wgun991230.htm
##                                                                                                    data_source_2
## 1 http://www.sptimes.com/News/123000/news_pf/TampaBay/A_year_later__the_str.shtml
##   military_experience
## 1             Unknown
##
## 1 http://news.google.com/newspapers?id=CSUdAAAAIBAJ&sjid=ZaYEAAAAIBAJ&pg=2043,2634069&dq=silvio+leyva
##   fate_of_shooter_at_the_scene number_of_shotguns total_number_of_victims
## 1                     Arrested                 0                       8
##      shooter_race class shooter_s_cause_of_death latitude
## 1 Some other race    MS           Not applicable 27.99602
##              geopoint caseid day_of_week total_number_of_guns
## 1 27.99602, -82.44504     77    Thursday                    2
##   number_of_victim_fatalities possible_motive_detailed
## 1                           5       No motive is known
##                      title shooter_sex data_source_5 data_source_6
## 1 Radisson Bay Harbor Inn        Male          <NA>          <NA>
##   data_source_7
## 1          <NA>
```

Creation of a subsetted Mass Shooting Data-Frame

This newly created Data_Frame contains a count of whether or not an occurence in this Data-Set was a School Related Shooting or not.

```
mass_school_related <-as.data.frame(mass_fields %>% count(school_related))
mass_school_related
```

```
##   school_related   n
## 1         Killed   1
## 2             no   1
## 3             No 220
## 4        Unknown  12
## 5            Yes  73
```

```
NewRow2 <- mass_school_related$n[2] + mass_school_related$n[3]

NewRow <- mass_school_related$n[1] + mass_school_related$n[4]
mass_school_related <- rbind(mass_school_related, NewRow, NewRow2)
mass_school_related <- mass_school_related[-c(1,2,3,4), ]
mass_school_related$school_related[3] = "No"
mass_school_related$school_related[2] = "Unknown"


kable(mass_school_related)
```

	school_related	n
5	Yes	73
6	Unknown	13
7	No	221

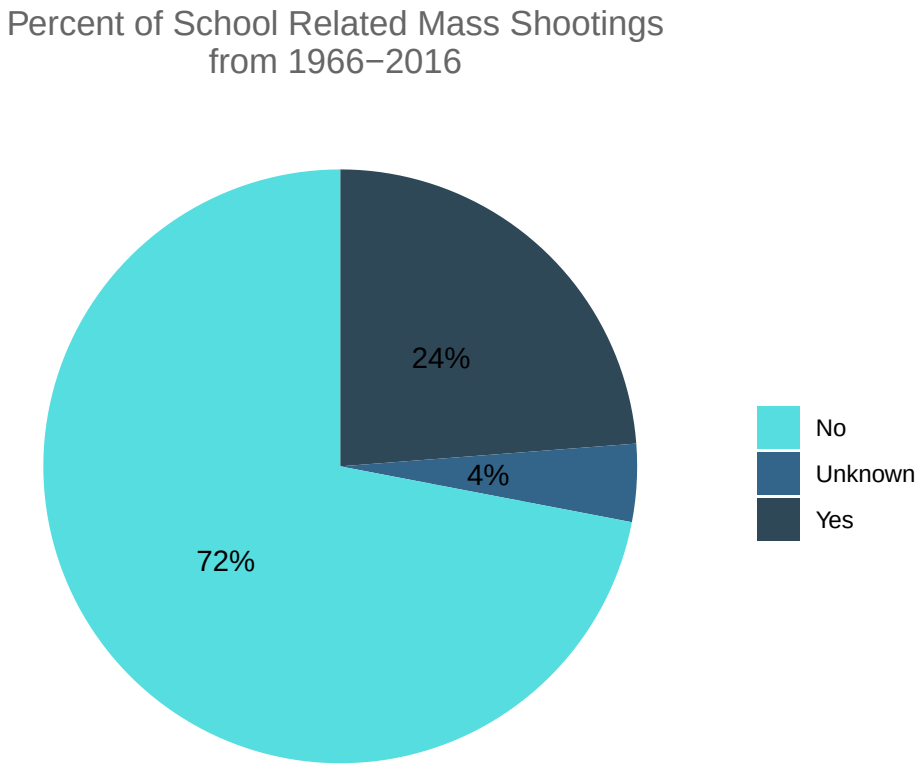
Count of each vartiable “No”, “Yes”, “Unkown”

```
total <- sum(mass_school_related$n)
```

Pie Chart and Percent Calculation

```
pie = ggplot(mass_school_related, aes(x="", y=n, fill=school_related)) + geom_bar(stat="identity", widt

pie = pie + coord_polar("y", start=0) + geom_text(aes(label = paste0(round(n / total * 100), "%")), pos

pie = pie + scale_fill_manual(values=c("#55DDE0", "#33658A", "#2F4858", "#F6AE2D", "#F26419", "#999999"

pie = pie + labs(x = NULL, y = NULL, fill = NULL, title = "Percent of School Related Mass Shootings \nf

pie = pie + theme_classic() + theme(axis.line = element_blank(),
                                    axis.text = element_blank(),
                                    axis.ticks = element_blank(),
                                    plot.title = element_text(hjust = 0.5, color = "#666666"))
```

```
pie
```




```
dim(mass_fields)
```

```
## [1] 307  49
```

School Shooting Data

Importation of Data-Set

For this section of the project, I wanted to focus on the School-Related Shootings as a sub-category of Mass Shooting, however I wanted to use a more robust Data_Set for this task. After sub-setting the First Data-Set, there was only 307 occurrences. I found this second DataSet K-12 School Shooting Database, that is updated everyday.

```
school_file <- read_csv(file ="https://raw.githubusercontent.com/josephsimone/Data_607_Final_Project/mas
```

```
dim(school_file)
```

```
## [1] 5701   47
```

School Shooting Data Tidying

```
colnames(school_file) = school_file[1, ]
```

```
colnames(school_file) = school_file[1, ]
school_file = school_file[-1, ]
school_data = as.data.frame(school_file)
```

```
names(school_file)
```

```
##  [1] "Date"
##  [2] "School"
##  [3] "City"
##  [4] "State"
##  [5] "Reliability Score (1-5)"
##  [6] "Killed (includes shooter)"
##  [7] "Wounded"
##  [8] "Total Injured/Killed Victims"
##  [9] "Gender of Victims (M/F/Both)"
## [10] "Victim's Affiliation w/ School"
## [11] "Victim's age(s)"
## [12] "Victims Race"
## [13] "Victim Ethnicity"
## [14] "Targeted Specific Victim(s)"
## [15] "Random Victims"
## [16] "Bullied (Y/N/ N/A)"
## [17] "Domestic Violence (Y/N)"
## [18] "Suicide (Shooter was only victim) Y/N/ N/A"
## [19] "Suicide (shot self immediately following initial shootings) Y/N/ N/A"
## [20] "Suicide (e.g., shot self at end of incident - time period between first shots and suicide, dif
## [21] "Suicide (or attempted suicide) by Shooter (Y/N)"
## [22] "Shooter's actions immediately after shots fired"
## [23] "Pre-planned school attack"
## [24] "Summary"
## [25] "Category"
## [26] "School Type"
## [27] "Narrative (Detailed Summary/ Background)"
```

```
## [28] "Sources"
## [29] "Time of Occurrence (12 hour AM/PM)"
## [30] "Duration (minutes)"
## [31] "Day of week (formula)"
## [32] "During School Day (Y/N)"
## [33] "Time Period"
## [34] "Location"
## [35] "Number of Shots Fired"
## [36] "Firearm Type"
## [37] "Number of Shooters"
## [38] "Shooter Name"
## [39] "Shooter Age"
## [40] "Shooter Gender"
## [41] "Race"
## [42] "Shooter Ethnicity"
## [43] "Shooter's Affiliation with School"
## [44] "Shooter had an accomplice who did not fire gun (Y/N)"
## [45] "Hostages Taken (Y/N)"
## [46] NA
## [47] NA
```

Dropping of Duplicated Rows During Import

```
school_file <- school_file[ -c(46:47) ]
school_tbl = as.data.frame(school_file)
```

```
dim(school_tbl)
```

```
## [1] 5700   45
```

This Data-Set has significantly more instances than the Mass Shooting Data-Set, so I thought it was more effective to use this Data-Set for a second data Visualization

Creation of a subsetted School Shooting Data-Frame

This newly created Data_Frame contains a count of how many occurence in this Data-Set take place in the Same State.

School Shooting Data-Set Tidying

```
school_states <-as.data.frame(school_tbl %>% count(State))
kable(school_states)
```

State	n
AK	20
AL	168
AR	72
AZ	48
CA	660
CO	84
CT	72
DC	100
DE	32
FL	360
GA	188
HI	12
IA	48
ID	8
IL	260
IN	96
KS	40
KY	56
LA	180
MA	60
MD	196
ME	8
MI	280
MN	48
MO	156
MS	68
MT	32
NC	172
ND	4
NE	28
NH	24
NJ	48
NM	40
NV	52
NY	200
OH	220
OK	44
OR	56
PA	216
RI	16
SC	108
SD	16
St. Croix, US Virgin Islands	4
TN	184
TX	548
UT	52
VA	100
VT	8
WA	128
WI	60
WV	12
WY	8

```
names(school_states)
```

```
## [1] "State" "n"
```

Dropped St. Croix, US Virgin Islands because they are not part of the Continential United States

```
school_states <- school_states[-c(43), ]
```

Convert State Abbreviates into State Names

```
school_states$State <- tolower(state.name[match(school_states$State,  state.abb)])
names(school_states)[names(school_states) == "State"] <- "state"
kable(head(school_states, 5))
```

state	n
alaska	20
alabama	168
arkansas	72
arizona	48
california	660

States with the Most amount of School Shootings

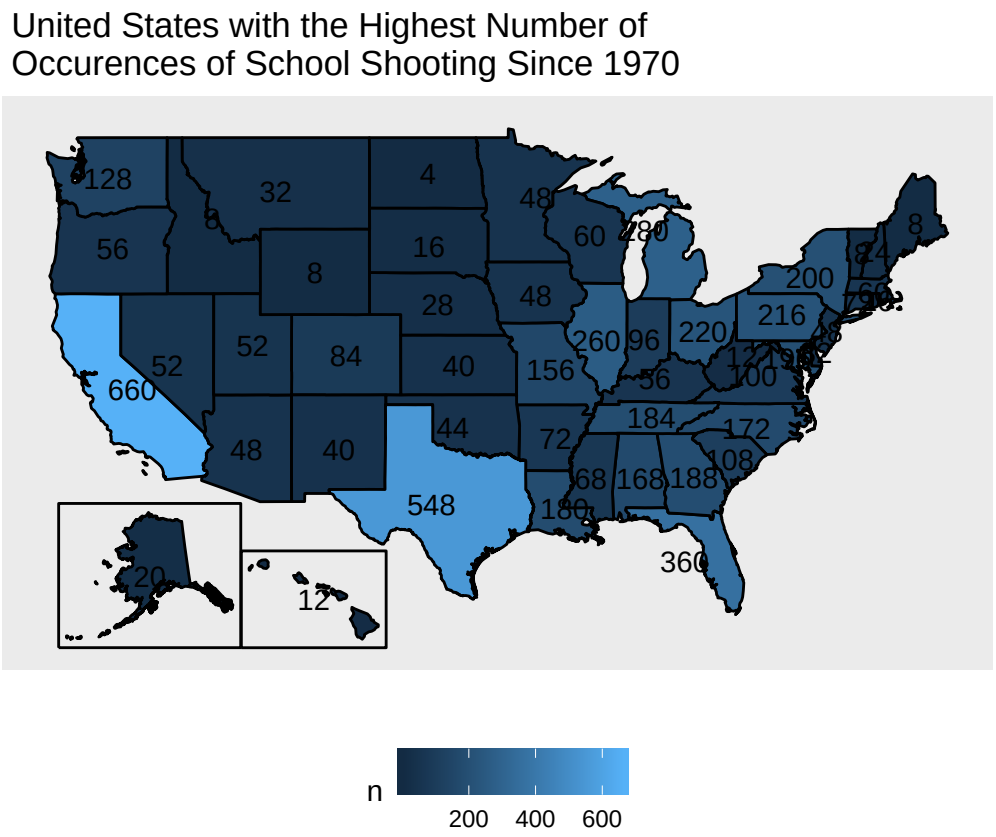
```
df_sorted_asc <- school_states[with(school_states, order(-n)), ]
kable(head(df_sorted_asc, 10))
```

	state	n
5	california	660
45	texas	548
10	florida	360
23	michigan	280
15	illinois	260
36	ohio	220
39	pennsylvania	216
35	new york	200
21	maryland	196
11	georgia	188

Us Map Plot

```
gg <- ggplot(data= school_states, aes(map_id = state)) +
  geom_map(aes(fill = n),  color= "black", map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() +
  geom_text(data = fifty_states %>%
              group_by(id) %>%
              summarise(lat = mean(c(max(lat), min(lat))),
                        long = mean(c(max(long), min(long)))) %>%
              mutate(state = id) %>%
```

```
              left_join(school_states, by = "state"), aes(x = long, y = lat, label = n ))+
  scale_x_continuous(breaks = NULL) + scale_y_continuous(breaks = NULL) +
  labs(x = "", y = "") + theme(legend.position = "bottom")

p <- gg + labs(title = " United States with the Highest Number of \n Occurences of School Shooting Sinc

p + fifty_states_inset_boxes()
```



Conclusion

Throughout my analysis, I found out some pretty interesting information from these two Data-Sets. First and formost, I found the second Data-Set to be more robust than the first. Also, I found the first Data-Set to have a lot of duplicate values.

During my analyis of the first Data-Set, that only 24% of all Mass Shooting involved a School.

Furthermore, during my analyis of the second Data-Set, that California was the state with the most occurences of School Shooting over the past 40 years.