Project Part 1

Your name - Blank for Part 1

Introduction

This project aims to better understand shootings in the United States and the effectiveness of policies that have been or could be implemented to improve the tragic problem. This analysis will focus on one primary dataset that features shooting incidents and a number of variables associated with them, as well as two supplementary datasets that will aid in the contextualization of the information and the accuracy of the presentation of facts and visuals. This initial report specifically describes the data, including its structure and contents, and begins an exploratory analysis to begin to understand the data beyond the superficial level.

Data Characteristics

Gun Incidents

The first dataset on gun incidents comes from Kaggle.com (Reference 1). The original data came from gunviolencearchive.org, where the information is verified by the non-profit organization and published publically for anyone to have access to this accurate information. The publisher of the data on Kaggle web scraped the data from the site using python, as this allowed him or her to capture all the information, rather than being forced to select only a few categories when downloading directly. The final compiled dataset was then sorted by date and published to the site. Since these data are a running account of shootings in the U.S., it is a dataset that comes close to representing the entire population. However, since it is almost certain that not all shootings are reported, it might be beneficial to treat the data as a sample and to try to interpret signficant findings that might be applied to the whole population. Looking at the original website, gunviolencearchive.org, there is a section specifically on methodology. It describes a thorough procedure for maintaining their site. Although intially it may seem that the organization might be partisan, they do cite efforts to include all relevant information to a shooting. For instance, they include defensive uses of weapons, whether it is home defense, rape defense or retail store defense. According to the website, they “utilize automated queries, manual research through over 2,000 media sources, aggregates, police blotters, police media outlets and other sources daily. Each incident is verified by both initial researchers and secondary validation processes." This seems like a reputable source for information, as they include links to the articles that they are using to record the incident, allowing researchers to investigate incidents for themselves and add any factors to their analysis.

Gun Laws

Noticing that the first dataset used has a record of location the incidents occurred, the decision was made to include a supplementary dataset (Reference 2). This dataset is also from Kaggle.com and provides policy information by state for each year between 1991 through 2017. The purpose of the inclusion of this dataset was to facilitate analysis of the interaction of gun deaths and policies enacted by governments. The data presented was aggregated for Kaggle by the author, Jacob Boysen. 100 of the provisions included in the set came from "Michael Siegel, MD, MPH, Boston University School of Public Health, with funding from the Robert Wood Johnson Foundation, Evidence for Action: Investigator-Initiated Research to Build a Culture of Health program (grant #73337), using data derived from the Thomson Reuters Westlaw legislative database." The information on the remaining 33 policies was derived from Everytown for Gun Safety and Legal Science, LLC. While the latter organization is known for being partisan to some degree, such as when it improperly reported the number of school shootings in 2018 (Reference 4), the information in this dataset is whether or not a state has a specific policy in place. There is no reason to doubt the accuracy of this national organization about whether a state has a firearm policy or not, as the public could very easily fact check this information. Since the rest of information came from a Thompson Reuters Westlaw database, the report is published in the American Journal of Public Health, and features multiple peer-reviewed sources, it is reasonable to say that this source is legitimate.

Population

The final dataset used come from the U.S. Census (Reference 3). Since it is from the U.S. Government itself, it would be considered a reliable source. The only information in this dataset is the predicted population totals by year between 2010 and the present, which will be used to get the number of shooting incidents per resident of a state. This is a summary of sample data that is being predicted on. This could cause some discrepancy in the predicted population totals and the actual population total, but it is going to be one of the best estimates for population, since the Census Bureau collects population data. Since these data are so straightforward, there will be no section on its structure and understanding.

Data Structure and Understanding

Gun Incidents

The structure of the Gun incidents dataset is that each observation represents an incident involving a gun in the United States. There are 29 conditions that are reported for the incidents, some of which will be removed. These conditions include the date, time, and location of the incident, the type and number of weapons involved, number killed and injured, the congressional district, notes on the incident, the age, sex, health status after the incident, and role of the people involved, the state house and senate district in which the event occurred, and the sources for how the event was discovered to have occurred and other news reports on it. There are only three quantitative variables in this dataset, the number of people killed and injured and the number of firearms used. This may prove to be difficult to work with as a result, but that does not mean that this dataset is weak by any means. With fewer numeric variables, a standard linear regression or multiple linear regression might be difficult to interpret, as there will be quite a few dummy variables due to the number of categorical variables. This just means that other statistical procedures should be used, such as t-tests for the difference in means.

As mentioned, this dataset includes the weapon type(s) for each shooting and the sex, age, and health status for each person involved. This adds great depth the dataset, though its structure provides a problem. The values in each of these four columns are in a structure like this: “0::Male||1::Female”, where the number is associated with a specific individual and the value is the column information for that individual. This allows the statistician to track each individual’s sex, health, and age, but provides difficulty in its list-like representation in the dataframe. This will require some cleaning of the data before beginning to conduct tests or analysis. To fix this problem, new columns will be created. This analysis does not focus on individuals, so creating variables like the number of males injured, the number of females arrested, and the number of victims who are minors should prove to be useful for these purposes. Besides the issues with the data structure, there is definitely potential areas of missing data. Based on the way these data were collected, mainly through the internet, there is a strong chance that incidents of gun use were overlooked in reporting or simply missed in the search by GunViolenceArchive. This could lead to problems of bias in the final model, though the methods for collecting data appear to be robust.

Gun Laws

The data for gun laws are much more straight forward. Each column is a provision that a state either has (1) for that year or does not (0). The data are in a long format, where each observation represents one of the fifty states for a year between 1991 and 2017. The final column is a sum of the number of gun laws that a state has for that year. Ultimately though a subset will most likely be used, so that there can be a focus on the policies that are most often advocated for in the media. As noted above, there are over 133 provisions included in the data, so it would not be proper to included an explanation of all of these variables. Rather, when introducing analyses including particular ones, the relevant information will supplement the results and conclusions.

Exploratory Data Analysis

To get an overall understanding of the relationship between policy and gun related deaths and injuries, it is crucial to understand if there is a relationship between the number of laws and the number of these violent incidents. Do states with more laws have less frequency of incidents? In the table below, the data for the states with the highest and lowest frequencies of gun deaths are presented, along with the number of injuries for these states. This, however, will bring in another factor to be considered though. For this to be appropriately represented, it will be desired to look at the number of incidents per person in that state. To do this, it is best to use the 2017 Census estimate to divide the number of incidents by the state population.

```
lawtotal_2016 <- gun_laws_states[gun_laws_states$year == 2016,c("state", "lawtotal")]
gun_data1$date <- as.character(gun_data1$date)
year_vect <- sapply(gun_data1$date, function(x) strsplit(x,"/")[[1]][3])
gun_data1$year <-year_vect
subset_2016_data <- gun_data1[gun_data1$year == "16",]
killed_2016 <- tapply(subset_2016_data$n_killed,subset_2016_data$state, sum)
injured_2016 <- tapply(subset_2016_data$n_injured,subset_2016_data$state, sum)
state_2016_data <- cbind(killed_2016, injured_2016)

head(state_2016_data[order(as.data.frame(state_2016_data)[,"killed_2016"]),])
```

```
##                killed_2016 injured_2016
## Vermont                 13           10
## Rhode Island           14           84
## Maine                  18           37
## North Dakota           18           33
## Wyoming                18           12
## New Hampshire          20           40
```

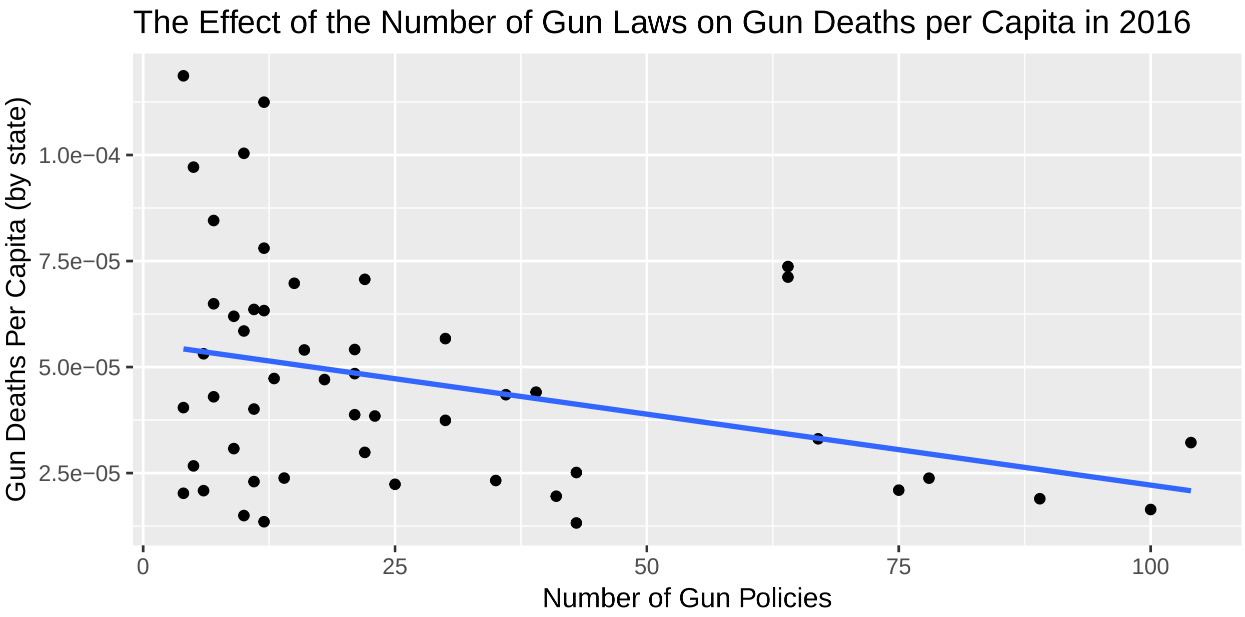
```
tail(state_2016_data[order(as.data.frame(state_2016_data)[,"killed_2016"]),])
```

```
##                killed_2016 injured_2016
## North Carolina         576         1205
## Ohio                   628         1455
## Illinois               946         4137
## Florida               1001         1846
## California            1265         1896
## Texas                 1313         1614
```

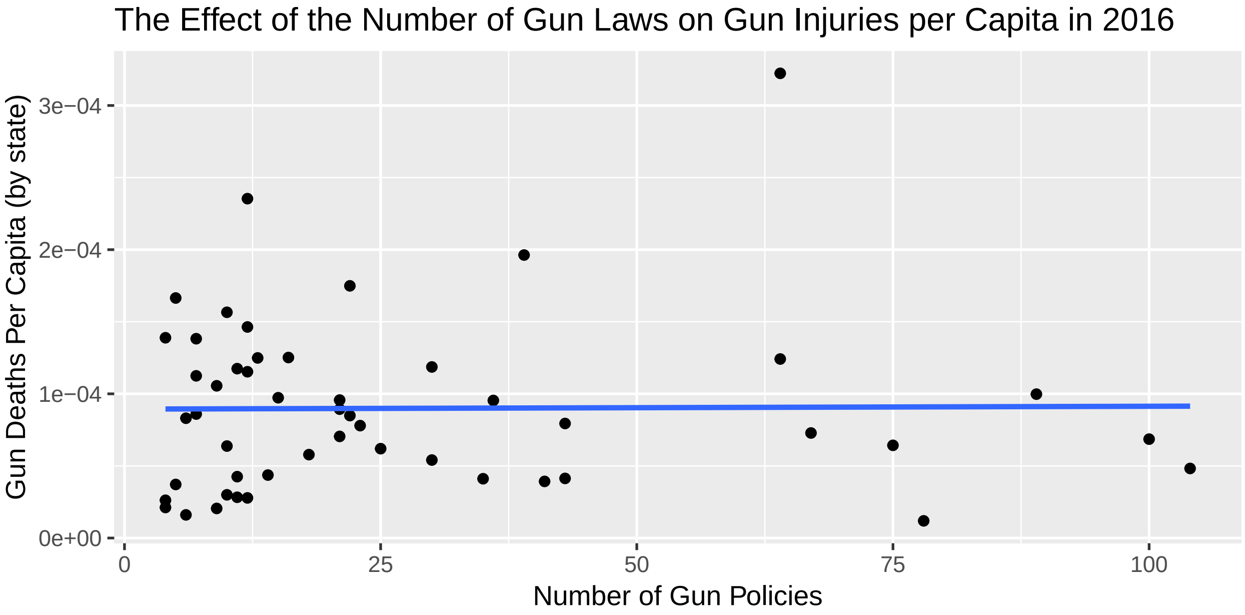
```
state_2016_all <- cbind(state_2016_data[-9,],lawtotal_2016)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
state_population <- population[-9,"POPESTIMATE2016"]
summary_df <- cbind(state_2016_all, state_population)
ggplot(summary_df, aes(y = killed_2016/state_population, x = lawtotal)) +
geom_point() +
stat_smooth(method = "lm", se = F) +
labs(title = "The Effect of the Number of Gun Laws on Gun Deaths per Capita in 2016",
       y = "Gun Deaths Per Capita (by state)", x = "Number of Gun Policies" )
```

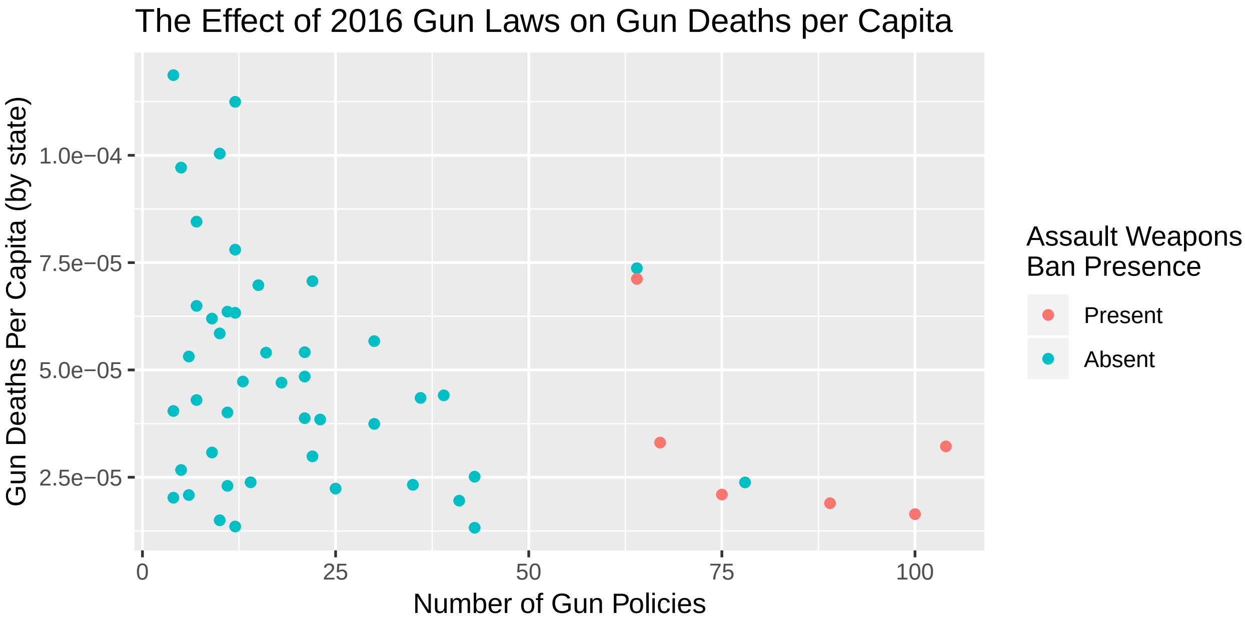


```
ggplot(summary_df, aes(y = injured_2016/state_population, x = lawtotal)) +
geom_point() + stat_smooth(method = "lm", se = F) +
labs(title = "The Effect of the Number of Gun Laws on Gun Injuries per Capita in 2016",
       y = "Gun Deaths Per Capita (by state)", x = "Number of Gun Policies" )
```



As shown, there seems to be a weak correlation overall between the number of gun deaths and injuries and the number of gun laws put in place to prevent these tragedies. The overall trend appears to be negative for deaths and almost absent for injuries, indicating that, generally, more regulation seems to lead to lower rates of gun-related deaths only. That being said, there is definitely a lot more exploration to do with the data. The analysis above includes all gun laws, not just the effective or popularly advocated provisions. Further analysis can be done to identify key provisions pushed for by the public to determine whether or not the presence of the law has a significant relationship with gun deaths or injuries. For instance, if states with assault weapon bans in place, as examined below, perhaps these have relatively lower gun deaths. The law being examined in this visual is described as a ban "on [the] sale of assault weapons beyond just assault pistols."

```
awban_2016 <- gun_laws_states[gun_laws_states$year == 2016,
                              c("state", "assault","lawtotal")]
state_2016_aw <- cbind(state_2016_data[-9,],awban_2016)
summary_df2 <- cbind(state_2016_aw, state_population)
ggplot(summary_df2, aes(y = killed_2016/state_population, x = lawtotal)) +
  geom_point(aes(color = factor(assault))) +
  labs(title = "The Effect of 2016 Gun Laws on Gun Deaths per Capita",
       y = "Gun Deaths Per Capita (by state)", x = "Number of Gun Policies",
       color = "Assault Weapons\nBan Presence") +
  scale_colour_discrete(limits = c(1, 0),labels=c("Present","Absent"))
```



From the initial examinations, it appeared that gun deaths is an area that can be explored in a greater depth. As illustrated in the first graph, there is a negative trend between the number of gun laws and gun deaths. The graph above examines this idea further, as it highlights states that have Assault Weapons Ban legislation specifically. The graphic seems to illustrate that this specific legislation could be effective at preventing gun death, but the significance of this relationship will be explored further in the next report.

References

1. https://www.kaggle.com/jameslko/gun-violence-dasta/home.
2. https://www.kaggle.com/jboysen/state-firearms/home.
3. https://www.census.gov/data/tables/2017/demo/popest/state-total.html#par_textimage_1574439295.
4. https://www.washingtonpost.com/local/no-there-havent-been-18-school-shooting-in-2018-that-numbe
   2018/02/15/65b6cf72-1264-11e8-8ea1-c1d91fcec3fe_story.html?noredirect=on&utm_term=.170e02280d3d