Final Project

William Saltzburg

12/09/2020

Abstract

Every day around 100 people die from a gun in the United States, and little has been done on the federal level to prevent this from occuring. Policymakers have made substantive research on gun violence prevention difficult with provisions such as the Dickey Amendment. My purpose in this final project is to use reliable data from the Center for Disease Control and methods learned in GOVT-470 to extrapolate trends in gun violence deaths from 2012 to 2018.

Introduction and Goals

A salient issue in American politics is gun violence. On average more than 37,000 americans lose their lives to gun violence each year.[1] There has been no federal action to mitigate the prevalence of gun violence in 27 years, and a large reason for this is partisan politics. An important fixture in gun violnce research is the Dickey Amendment to the annual appropriations bill. The amendment states that no federal funding can be directed to advocate for "gun control." In 2018, it was clarified that the amendment does not prohibit research, a step in the right direction for better understanding gun violence as an issue of public health.[2] Because of research prohibitions, like the Dickey Amendment, there is a lack of robust statistical analysis on the issue of gun violence. The purpose of my research is to better understand patterns in gun violence that could inform public policy in a meaningful way. I will use methods and skills I learned in Intro to Applied Political Data Science (American Univeristy, GOVT-470) to investigate data from the CDC and cleaned by a function developed by data scientists at fivethirtyeight.[3]

Project Mechanics

For this project, I have utilized a github repository to house the data, large functions, and files relevant to the final project. The github page is public and can be found at this link: https://github.com/willsaltzburg/finalprojectWMS

The github page houses the following documents:

- **CDC_parser_WMS.R**
- **guns.csv**
- **Final_Project.Rmd**
- **Final_Project.pdf**
- **README.md**

[1] "Everytown Research - EveryStat," EverytownResearch.org/everystat, 2020, https://maps.everytownresearch.org/everystat.

[2] Rostron A. (2018). The Dickey Amendment on Federal Funding for Research on Gun Violence: A Legal Dissection.

[3] Ben Casselman, Matthew Conlen, and Reuben Fischer-Baum, "Gun Deaths In America," FiveThirtyEight (FiveThirtyEight, July 13, 2016), https://fivethirtyeight.com/features/gun-deaths/.

Inspiration and Motivation

FiveThirtyEight is an organization that focuses on data journalism, most notably writing on poltics and sports. In 2016, the publication wrote an article on gun violence deaths in the United States (not including territories). The article included a data visualization tool that users could interact with to better understand the data. Inspired by this article, I wanted to replicate their findings, extraploate their functions to analyze more years of data, and create different tools and visualizations to better comprehend the gun violence epidemic in the United States.[4]

About Data

I used data published by the Center for Disease Control's (CDC) Multiple Cause of Death Database[5], which FiveThirtyEight reports is, "[T]he most comprehensive estimate of firearm deaths [in the United States]"[6].

The Mortality Multiple Cause files from the CDC are highly complicated and difficult to parse. Along with the dataset publication each year, the CDC provides a codebook. This codebook allows data scientists to connect alpha-numeric codes attached to different types of death, levels of education, and other discrete variables. Since many people die in the United States each year, the annual dataset is understandably very large. Attempting to parse and clean this massive dataset was far beyond my capabilities. Fortunately, in fivethirtyeight's "Gund Deaths in America", mentioned in a prior section, included the code they used to parse, clean, and merge data from the CDC. The end result yielded a clean data set with a handful of insightful variables.

Once cleaned, there are 11 different variables; I will describe a few of the variables of interest:

- **year**: the year in which the gun death occured (2012-2014)
- **month**: the month in which the gun death occured, where 1 is January and 12 is December (1-12)
- **intent**: record if the gun death was accidnetal, homicide, suicide, or undetermined.
- **police**: a binary indicator if the death was from the firearm of a police officer.
- **sex**: the sex of the gun violence vicitim
- **age**: the age, in years, at the time of death
- **race**: the race of the individaul killed in categories: Asian Pacific Islander, Black, Hispanic, Native American/Native Alaskan, White
- **education**: the level of education at the time of death: Less than a High School Diploma, High School Diploma or GED, Some college, Bachelors degree or beyond.

Data Cleaning: CDC_parser_WMS

I have used a slightly augmented function written by FivetThirtyEight to parse the data so that just deaths from guns are included and it is clean.[7] I have slightly edited the function to be able to process more years of data. The FiveThirtyEight function was only able to process for years 2012 through 2014. The edited function can process data from 2012 to 2018. In addition I changed the nomenclature of the final dataset so that the final data set would be titled "guns.csv." I renamed the function since there are minor differences between fivethirtyeight's function and my own. I renamed the function `CDC_parser_WMS`; WMS are my initials. The entire function can be found in the github repository.

[4] Ben Casselman, Matthew Conlen, and Reuben Fischer-Baum, "Gun Deaths In America," FiveThirtyEight (FiveThirtyEight, July 13, 2016), https://fivethirtyeight.com/features/gun-deaths/.

[5] "Data Access - Vital Statistics Online," Centers for Disease Control and Prevention (Centers for Disease Control and Prevention, October 9, 2020), https://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm.

[6] Ben Casselman, Matthew Conlen, and Reuben Fischer-Baum, "Gun Deaths In America," FiveThirtyEight (FiveThirtyEight, July 13, 2016), https://fivethirtyeight.com/features/gun-deaths/.

[7] Ben Casselman, "Fivethirtyeight/Guns-Data," GitHub (FiveThirtyEight, June 2, 2017), https://github.com/fivethirtyeight/guns-data/blob/master/CDC_parser.R.

The method of cleaning the data from fivethirtyeight includes packages from the tidyverse; they are `readr`, `dplyr`, `tidyr`, `magrittr`, `ggplot2`.[8]

Installation

For my project, I will use functions from packages within the tidyverse to support my work.

```r
library(tidyverse)
```

After cleaning the data using `CDC_parser_WMS()`, described above and adapted from FiveThirtyEight, I save the data in a .csv file that has all recorded gun deaths from 2012 to 2018. Below you can see the code necessary to load the data into RStudio.

```r
guns <- read_csv("guns.csv")
```

Visualizations and Statistical Analysis

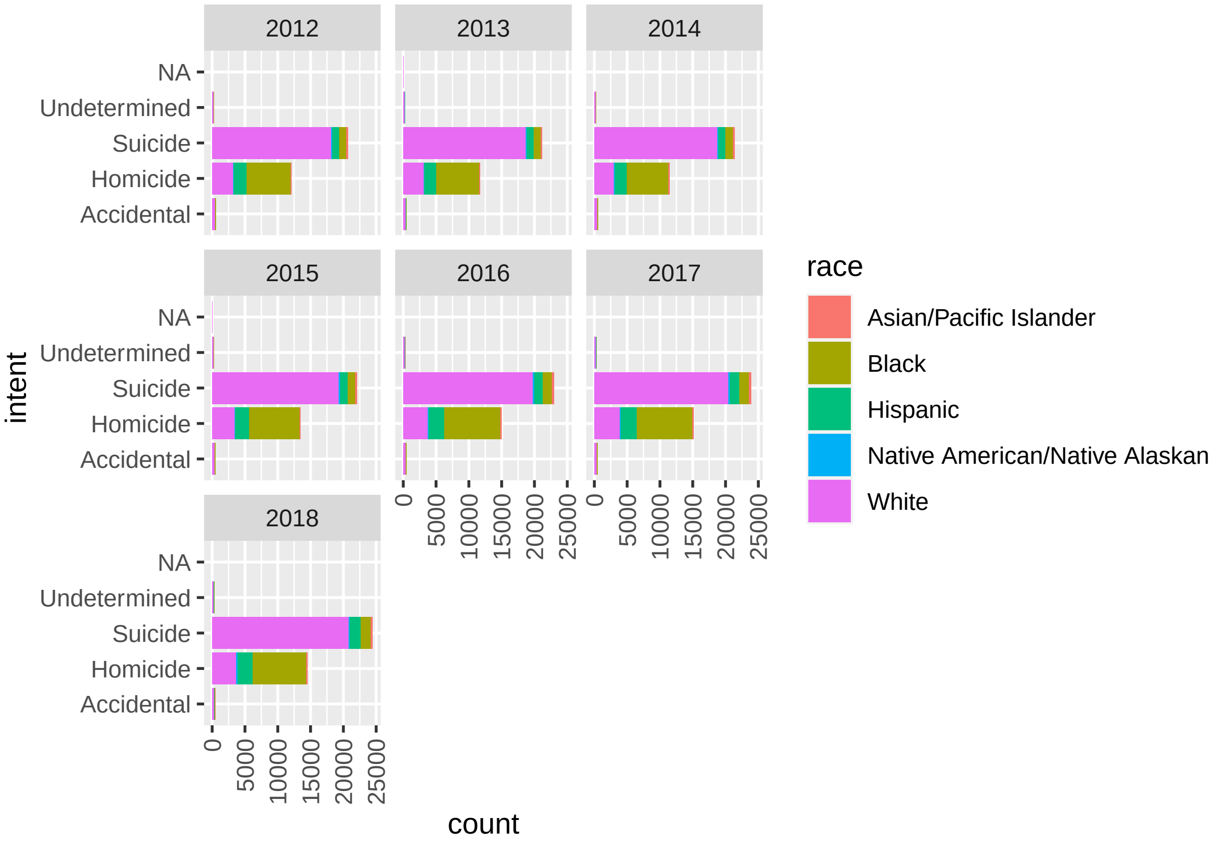
I will create visualizations that interact key variables. These visualizations will inspire specific and pointed questions that can be answered by specific data analysis tools, like modeling and linear regressions.

Intent, Race and Frequency over time

Something I will explore in this project is the evolution of how people are dying from gun violence over time. My hope is that trends will highlight a potential opportunity to curb gun violence. Particularly, I will focus on homicides. The reason I will focus on homicides, rather than suicides (which clearly accounts for the largest proportion of gun deaths each year, seen below) is that it appears as though people of color (black and hispanic people) are dying at rates that are disporportionate to their size in the population for homicides. I will test this hypothesis later with a test of statistical significance. What is crucial about this hypothesis is that if people of color are dying at rates that are statistically significant and different from their proportion in society (according to the census bureau) then it is reasonable to assume there is some system that is perpetrating the death of people of color.

```r
ggplot(data = guns) +
  geom_bar(mapping = aes(x = intent, fill = race)) +
  facet_wrap(~ year) +
  coord_flip() +
  guides(x = guide_axis(angle = 90))
```

[8] Ben Casselman, "Fivethirtyeight/Guns-Data," GitHub (FiveThirtyEight, June 2, 2017), https://github.com/fivethirtyeight/guns-data/blob/master/CDC_parser.R.



Police Shootings

An important question to consider is what role do the police play in perpetrating gun violence in the United States. This is an issue of great importance and relevance as earlier this year the death of people of color at the hands of the police was the cause of great social upheaval. There have been calls from advocacy groups to strip police officers of the ability to use lethal force, since people argue that they have abused their power. This policy action should not be taken without robust empiricial evidence that police officers disproportionately kill people of color.

Police Shootings - All Races

To statistically evaluate the role of police shootings, I created a dataset from `guns.csv` that contains only instances of gun deaths where police were involved. Below is the code I wrote to make this data set. In addition, I want to explore how police related gun deaths have evolved over time. For this reason, I have created strings of data of the number of police involved shootings from 2012 to 2018. The code used to make these strings are in the block below. Finally, I used a linear regression to model the evolution of police related gun deaths.

```
# Creating the data set police_shootings
police_shootings <- guns %>%
  filter(police == 1)

# Creating a string for all of the years for which data is captured
gun_years <- (2012:2018)

# By creating datasets for each year of gun deaths, I can easily capture
```

```r
# how many police related gun deaths there are in a year.
police_2012 <-police_shootings %>% filter(year == 2012)
police_2013 <-police_shootings %>% filter(year == 2013)
police_2014 <-police_shootings %>% filter(year == 2014)
police_2015 <-police_shootings %>% filter(year == 2015)
police_2016 <-police_shootings %>% filter(year == 2016)
police_2017 <-police_shootings %>% filter(year == 2017)
police_2018 <-police_shootings %>% filter(year == 2018)

# Creating a string for all the number of total gun deaths each year
gun_police_years <- c(472, 469, 467, 492, 513, 555, 543)

# This linear regression will capture the relationship between
# the number of police gun deaths over time.
lm_police_all <- lm(gun_police_years ~ gun_years)

lm_police_all
```

```
##
## Call:
## lm(formula = gun_police_years ~ gun_years)
##
## Coefficients:
## (Intercept)    gun_years
##   -30515.04        15.39
```

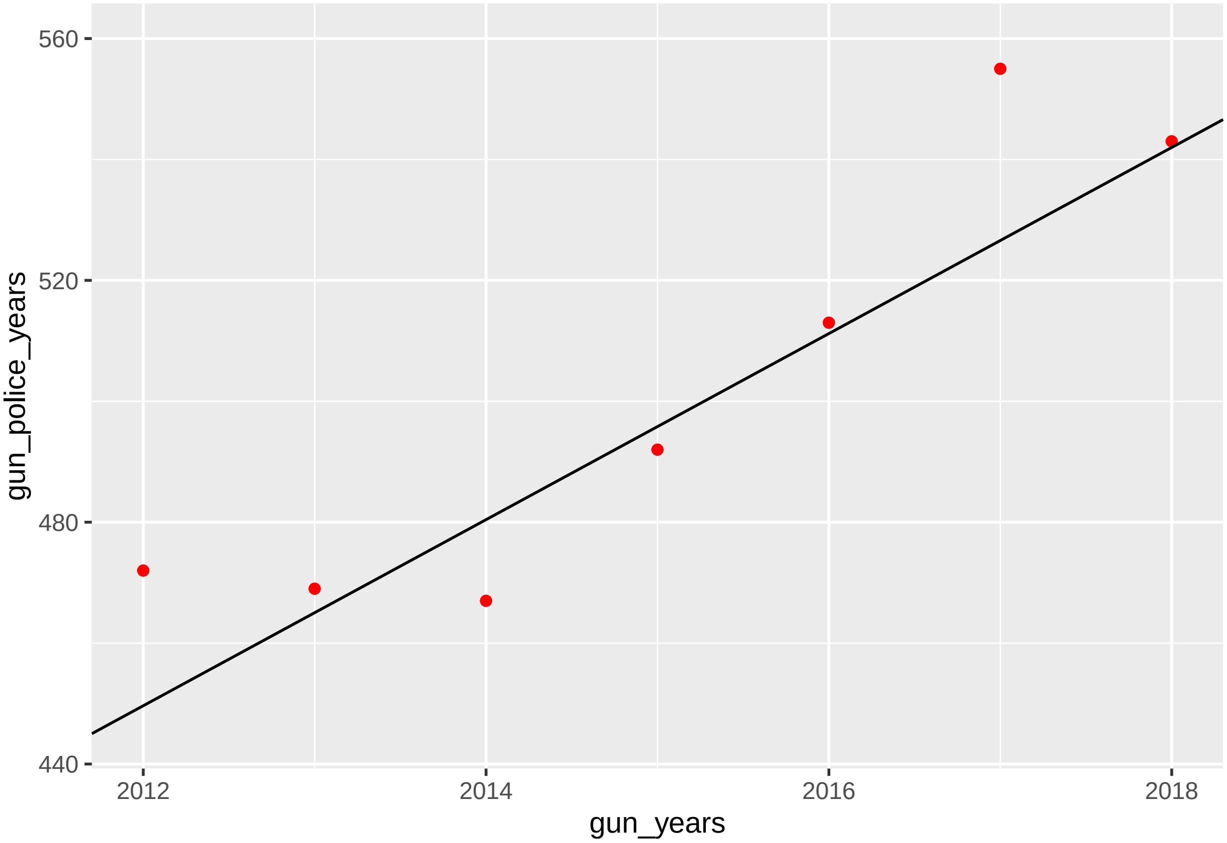
There is an important interpreation of the linear regression model. We can see from the output that the intercept is -30515.04. The number alone is not relevant to our findings, especially becaues it would be impossible to have negative deaths. What this number represents is, based on the model, how many police involved gun deaths there would be if our data stretched all the way back to the year 0. This information is irrelevant to my research. To capture a more relevant intercept, we can employ simple linear algebra and calcuate the following:

```r
# Intercept + (slope * first year of observation)

adjusted_intercept <- -30515.04 + (15.39 * 2012)
adjusted_intercept
```

```
## [1] 449.64
```

The plot below shows the evolution of all gun deaths with police involvement from 2012 to 2018. We can notice a slight positive trend in the data. I have overlayed the linear regression model to illustrate that police-involved gun deaths are increasing over time.

```r
ggplot() +
  geom_point(mapping = aes(x = gun_years, y = gun_police_years), color = "red") +
  geom_abline(intercept = -30515.04, slope = 15.39) +
  ylim(445, 560)
```



Police Shootings - People of Color (Black and Hispanic)

Following this graphical analysis and modeling, I want to repeat the same process for just people of color. I will look specifically at deaths of people who are either black or hispanic. In this analysis, I am searching to uncover if police deaths of people of color are increasing at or above the rate of the whole group of police deaths.

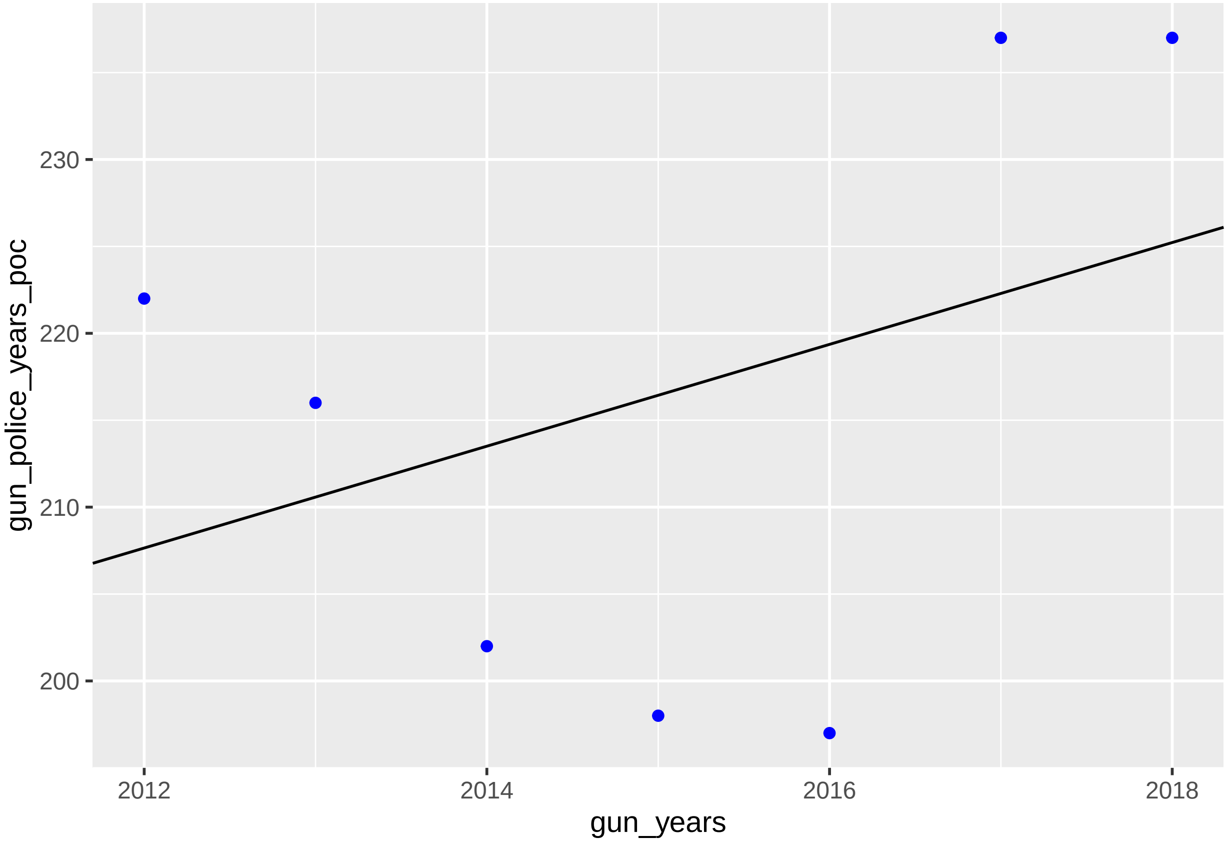
```
# By creating datasets for each year of gun deaths, I can easily capture
# how many police related gun deaths of people of color there are in a year.
police_2012_poc <-police_shootings %>% filter(year == 2012 &
                                              (race == "Black" | race == "Hispanic"))
police_2013_poc <-police_shootings %>% filter(year == 2013 &
                                              (race == "Black" | race == "Hispanic"))
police_2014_poc <-police_shootings %>% filter(year == 2014 &
                                              (race == "Black" | race == "Hispanic"))
police_2015_poc <-police_shootings %>% filter(year == 2015 &
                                              (race == "Black" | race == "Hispanic"))
police_2016_poc <-police_shootings %>% filter(year == 2016 &
                                              (race == "Black" | race == "Hispanic"))
police_2017_poc <-police_shootings %>% filter(year == 2017 &
                                              (race == "Black" | race == "Hispanic"))
police_2018_poc <-police_shootings %>% filter(year == 2018 &
                                              (race == "Black" | race == "Hispanic"))

# Creating a string for all the number of total gun deaths each year
gun_police_years_poc <- c(222, 216, 202, 198, 197, 237, 237)
```

```r
# This linear regression will capture the relationship between
# the number of police gun deaths over time.
lm_police_all_poc <- lm(gun_police_years_poc ~ gun_years)

lm_police_all_poc
```

```
##
## Call:
## lm(formula = gun_police_years_poc ~ gun_years)
##
## Coefficients:
## (Intercept)    gun_years
##   -5685.500        2.929
```

Again the intercept of this linear regression requires interpretation. The substnative intercept that is helpful for our data is not `-5685.500` it is 207.648. This is the fitted value of the linear model at 2012, the first year of our data collection. In addition, it is crucial that we look at the slope of these data points. The slope is `2.929` and considering we are considering hundereds of deaths per year, this slope seems very insignificant and contains high residuals. Below is the plot of people of color gun deaths from police over time with the linear regression overlayed.

```r
ggplot() +
  geom_point(mapping = aes(x = gun_years, y = gun_police_years_poc), color = "blue") +
  geom_abline(intercept = -5685.500, slope = 2.929)
```



We can see that there are high residuals since many of the data lie far away from the linear regression model. An interpretation of this phenomenon is that the linear model does not best explain the trends we see in the raw data. It is difficult to draw a substantive conclusion from this data since the coefficient is so small.

Police Shootings - Proportions

The final analysis that I would like to perform in the graphical analysis of police-involved shootings is investgating the proportion of police gun deaths that are people of color. Again I will follow the same procedure to produce graphs and linear models.

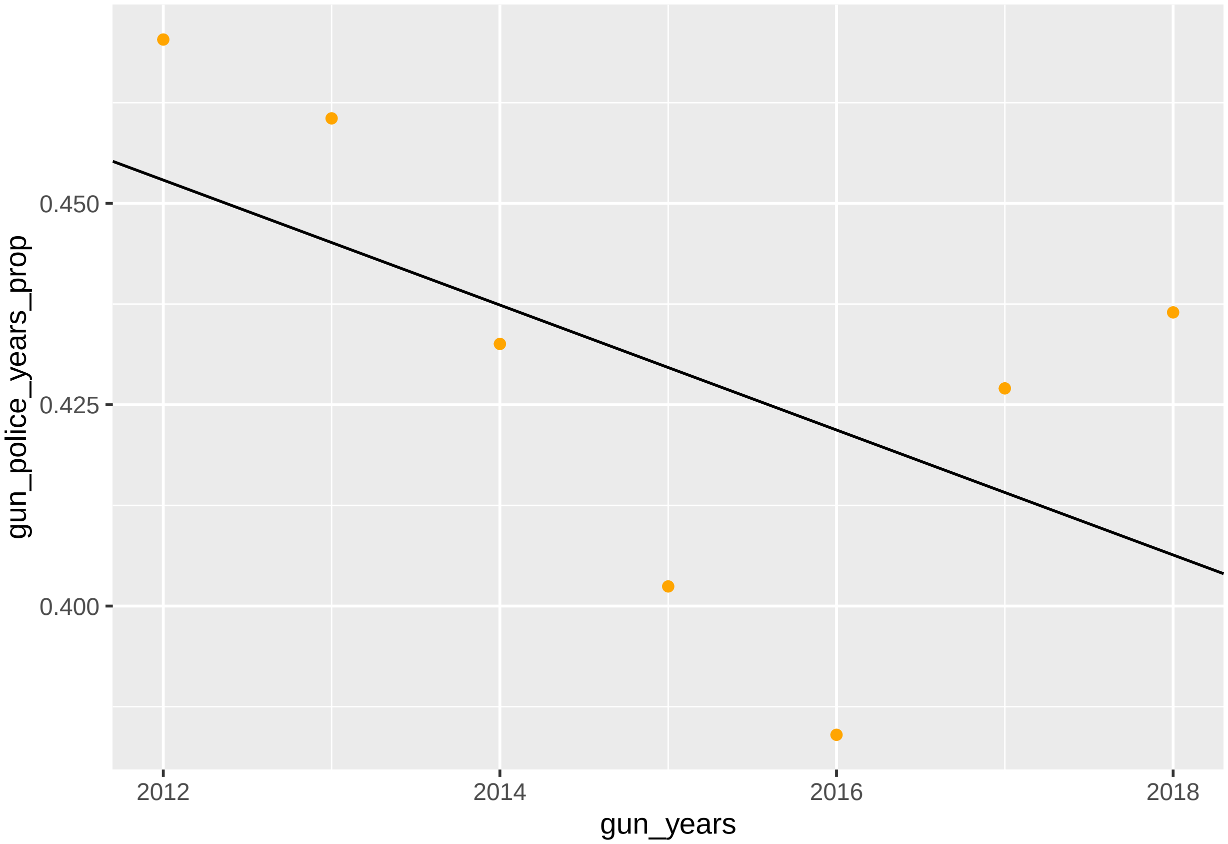
```
# This creates a string of values between 0 and 1 that represent the proportion of
# police-involved gun deaths where a person of color died
gun_police_years_prop <- gun_police_years_poc / gun_police_years

lm_police_prop <- lm(gun_police_years_prop ~ gun_years)
lm_police_prop
```

```
##
## Call:
## lm(formula = gun_police_years_prop ~ gun_years)
##
## Coefficients:
## (Intercept)    gun_years
##   16.061987    -0.007758
```

Again, to interpret the intercept we perform simple algebra to find that the fitted value of gun violence deaths in 2012 is 0.452891. The significance of this is that our fitted value intercept is 0.452891; this means that our model predicts that in 2012 45.2891% of police involved gun deaths were people of color. This is alarming because the proportions of Black and Hispanic people combined only equal 31.9% of the total population, according to the Census Bureau[9]. Next I will dig deeper into this phenomenon with a test of statistical significance. Below is a graph of how the proportion of police-involved gun deaths of people of color have evolved overtime.

```
ggplot() +
  geom_point(mapping = aes(x = gun_years, y = gun_police_years_prop), color = "orange") +
  geom_abline(intercept = 16.061987, slope = -0.007758)
```

[9] “U.S. Census Bureau QuickFacts: United States,” Census Bureau QuickFacts, accessed December 1, 2020, https://www.census.gov/quickfacts/fact/table/US/PST045219.



Significance Tests

Referenced earlier, it is important that beyond identifying trends, that a robust statistical analysis include a testing of statistical significance. Particularly in the last section where the proportion of people of color was modeled, this is a good opportunity to test if the average proportion of police-involved gun deaths of people of color is statistically different from their proportion in society. For this test, I have written a function that will compute a significance test. The function takes two inputs: x, which is a string of numbers, and y, the value of the null hypothesis. The function is below:

```
ttest <- function(x, y){
  xbar <- mean(x, na.rm = TRUE)
  sqrt_n <- (length(x))^(1/2)
  st_dev <- sd(x)
  tstat <- (xbar - y) / (st_dev / sqrt_n)
  print(tstat)
}
```

The output of this function is a t-statistic. The absolute value of this number will inform whether our hypothesis is statistically different from the mean of the data. According to the Census Bureau, the proportion of Black people in society is 13.4% and the proportion of Hispanic people is 18.5%.[10] Their combined proportion in the population is 31.9%

Below we run the ttest to see if the proportion of gun deaths are statistically different from their proportion in society.

[10] https://www.census.gov/quickfacts/fact/table/US/PST045219

```r
ttest(gun_police_years_prop, 0.319)
```

```
## [1] 9.75063
```

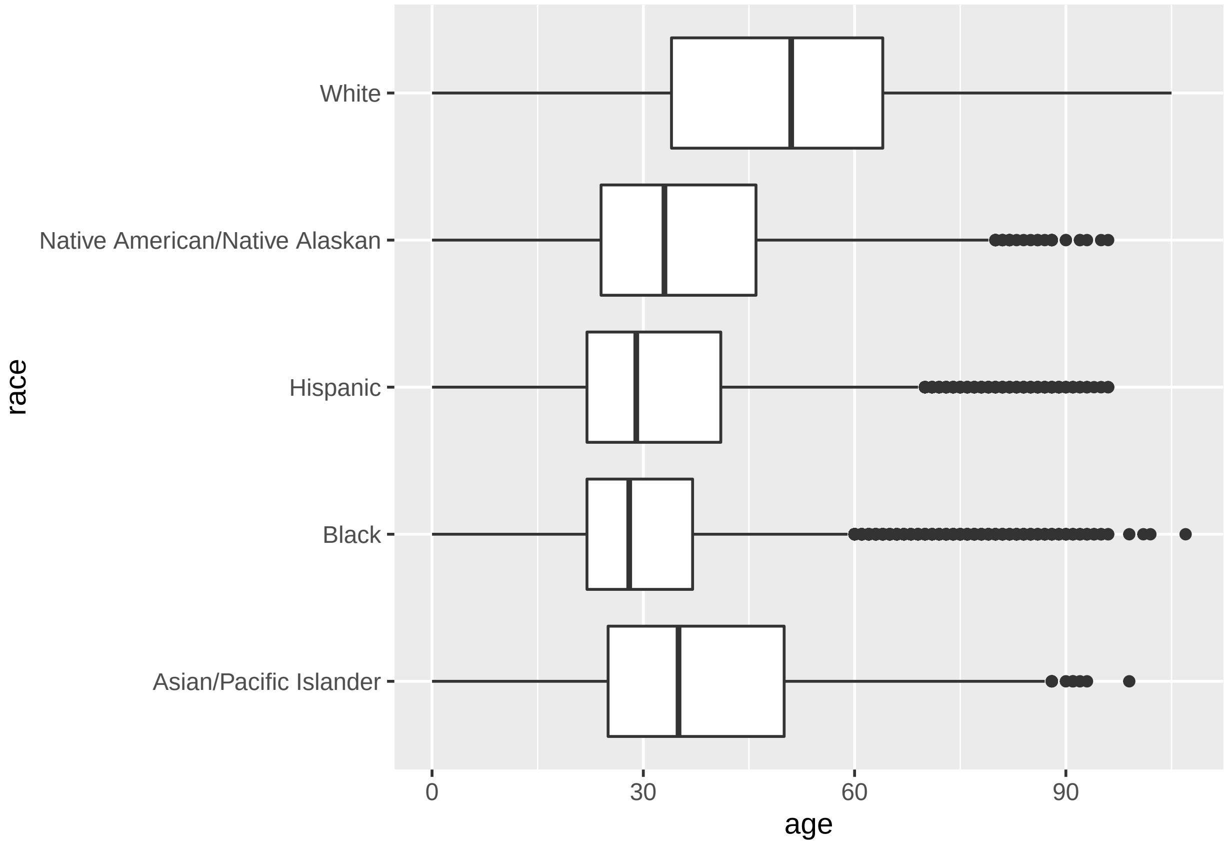
The value t-statistic vlaue of 9.75063 warrants two conclusions. First, that the proportion of people of color in society is almost certainly less than the proportion of police-involved deaths of people of color. Second, there must be a reason (other than difference between the sample mean and hypothesis) that accounts for the large size of the t-statistic. The absolute value of the t-statistic is influenced negatively by the standard deviation. This means when the standard deviation is small, the t-statistic is large. The standard devation includes a summation in the numerator. What happens is the following: as the sample size grows, the standard deviation grows and the t-statistic shrinks. Since we are looking at 7 years of data, and each entry of data is one year, we have a very small sample of data. For this reason, the standard deviation is small, which leads to a large t-statistic. This statistical intuition explains the size of the t-statistic. Despite a small sample size, I am confident that the proportion of police-involved gun deaths of people of color is statistically different from the proportion in society. I defend using anual data rather than monthly data for this analysis because there are only a few hundred police-involved gun deaths each year. To look month-by-month would subject the interpretation to larger ebbs and flows in the data, rather than a smooth linear trend. For example, some months may have more or less deaths than others for a myraid of reasons. February may habitually see less deaths because it has the least amount of days. The winter months may see more deaths because people are inside and intances of domestic violence are up, as opposed to the summer months. From just this data, there is no way to control for the month to month factors that could influence the number of gun deaths. Year-to-year data is more uniform and easier to compare, which is why I defend using it over monthly data, even thought it produced a large t-statistic.

Age

In this next section, I will produce some visualizations that show the age distrobution of gun deaths in the United States.

The histogram below shows the age distrobution of people who died from gun violence sorted by race.

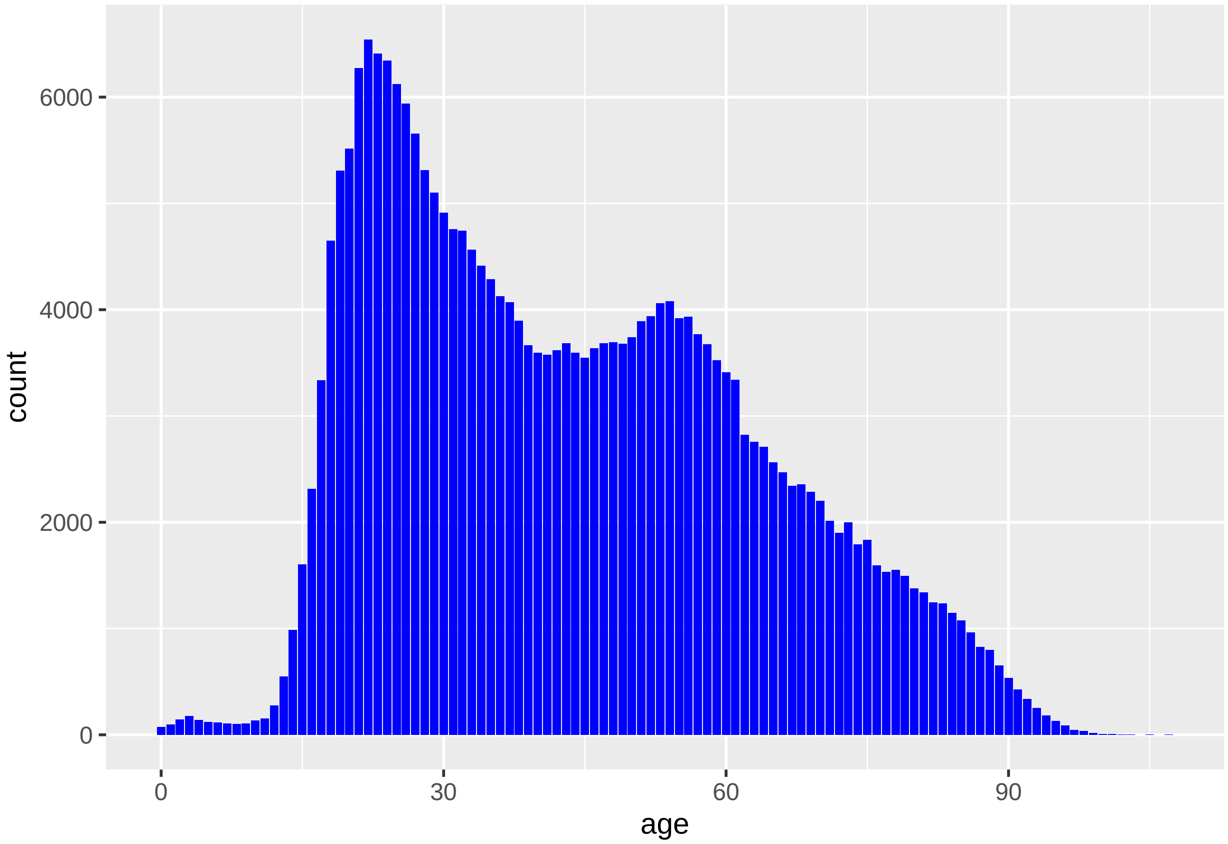
```r
ggplot(data = guns, mapping = aes(x = age, y = race)) +
  geom_boxplot()
```



From this histogram, we can see that black people have the youngest median age of death and white people have the highest median age of death.

The graph below explores the overall age distrobution in a bar format.

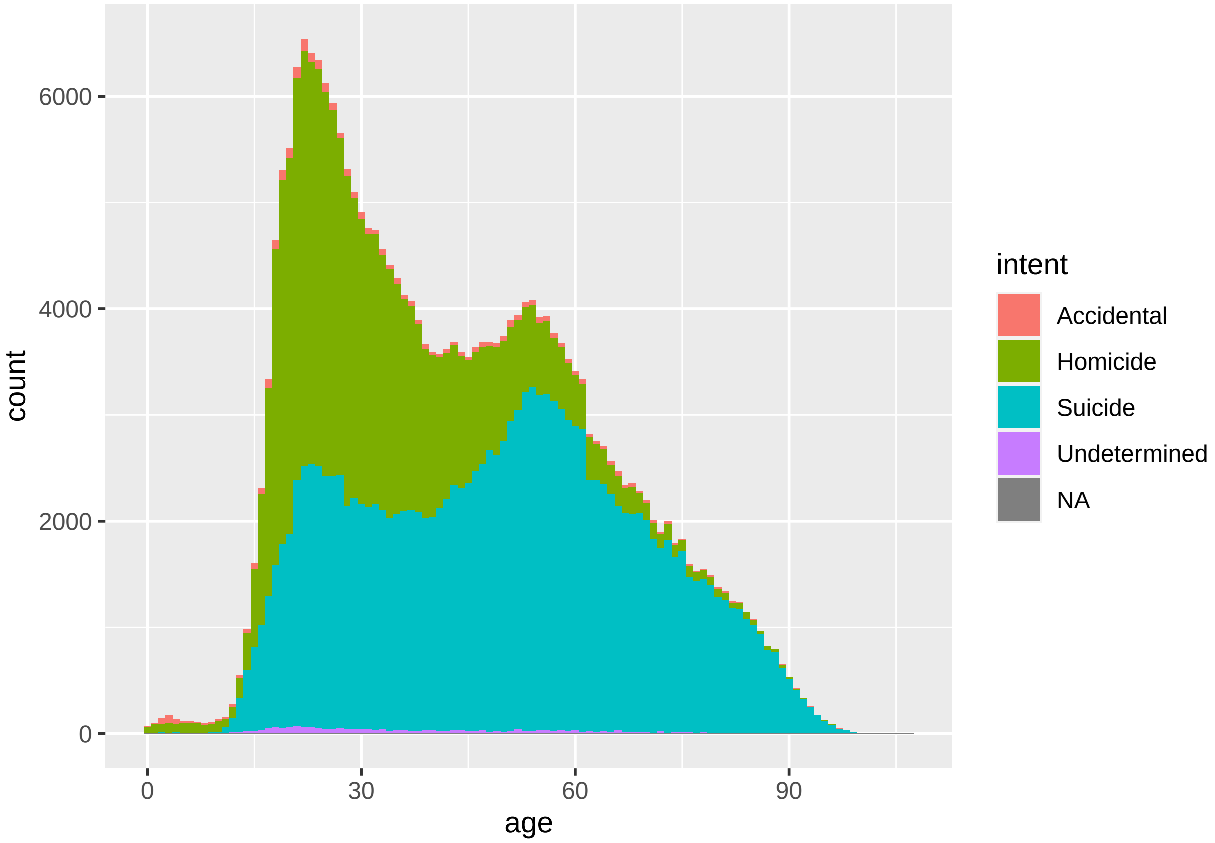
```
ggplot(data = guns, mapping = aes(x = age)) +
  geom_bar(fill = "blue")
```



From this visualization, we can see that the highest frequency of gun deaths occurs in the mid-twenties. From there, the frequency of gun deaths steadily decreases, with the exception of a short bounce in the mid-to-late 50s.

In the graph below, I fill bars by intent to see how the method of death evolves with age.

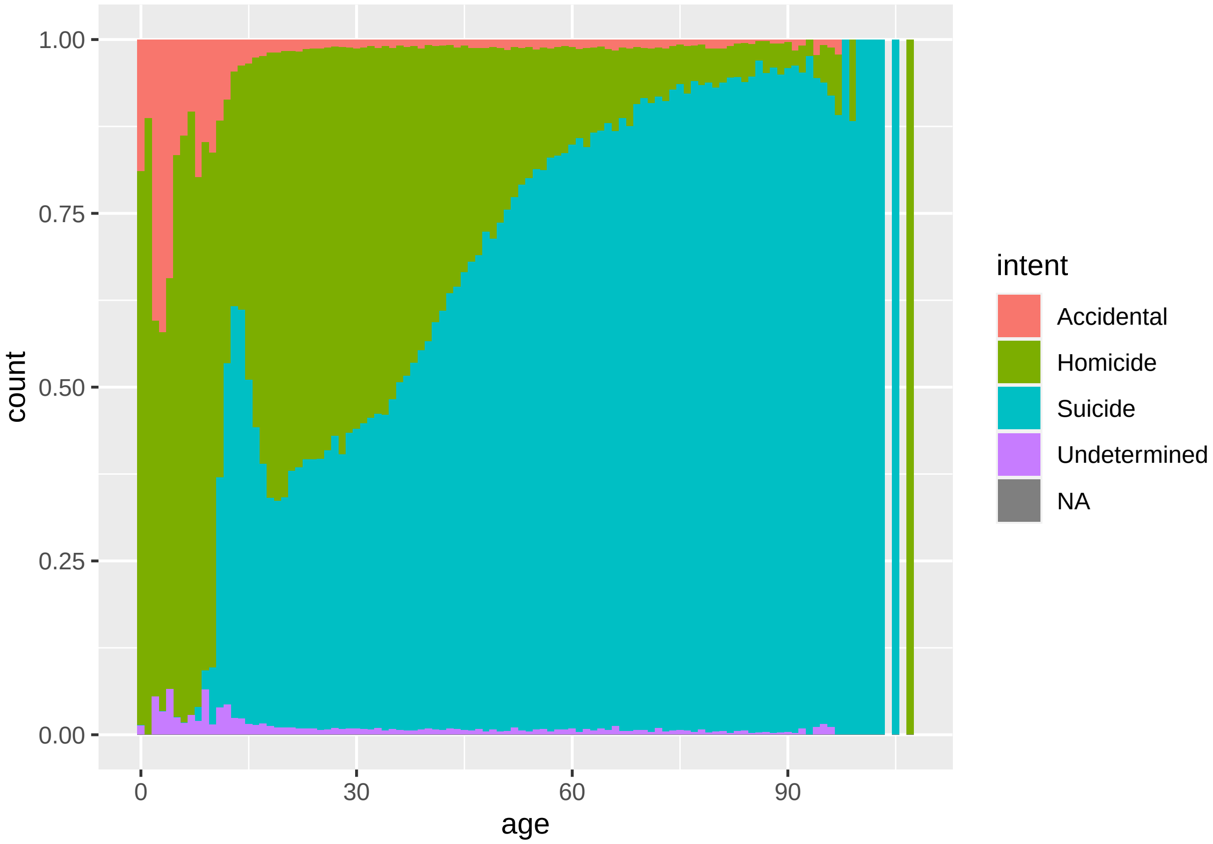
```
ggplot(data = guns, mapping = aes(x = age)) +
  geom_histogram(mapping = aes(fill = intent), binwidth = 1)
```



The extrapolation from this is quite interesting. Starting around age 20, homicide is the leading intent of gun deaths. The proportion of gun deaths that are homicides contiues to increase until around age 30 where the proportion of homicides decrease the the proportion of suicides increases until the end the age distrobution where almost all of the deaths are suicides. I have developed another graph to viualize the proportions evenly and not just with respect to the volume of deaths for each age.

The graph below will fill the entire bar relative to the proportion of intent. This will allow us to better analyze the change in proportions of intent across ages.

```r
ggplot(data = guns, mapping = aes(x = age)) +
  geom_histogram(mapping = aes(fill = intent), position = "fill", binwidth = 1)
```



An interesting take away from this graph is that a relatively large proportion of gun deaths in minors are the result of an accident. This conclusion would have been difficult to ascertain from the previous graph, because the frequency of child gun deaths is much lower than adults. The evaluations from the prior graph also are confirmed. The proportion of homicides is largest among young people and at around 30 the proportion of suicides steadily increases.

Conclusions

The data visualizations and analysis of the CDC's Multiple Mortalities database has the propensity to inform substantive change in gun violence prevention policy. In my conclusion, I will discuss the findings of my research and make suggestions rooted in my understanding of data science from this course and others I have taken previously.

The first visulaization depicted the breakdown of different intents; the methods by which people die from gun violence. From this visualization, it is evident that the most gun deaths in the United States are suicides, and that the majority of suicides have consistently been white people. While this finding doesn't necessarily inform any policy decision, it might be wise, from a public health perspective, if health care providers of white people over the age 30 have a focus on suicide prevention. The suggestion to taylor this to patients over 30 is informed by a later visualization that shows the progression of the intents of gun violnce deaths with ages. At age 30 the proportion of suicide deaths increase steadily until it makes up nearly all of the intent of gun deaths by the 90s.

Discussion of the role of the police in gun deaths is prevalent in society, especially this year. For that reason, it is cruicial that any policy action is supported by robust statistical data. Because of the long-time prohibition on gun violence prevention research, it has been difficult for researchers (especially in the public sector) ascertain reliable data. As an annecdote of this struggle, it was difficult for me to access data from the CDC to conducnt analysis for this project. If it were not for the CDC parser function from fivethirtyeight, I would not have been able to analyze the data. Despite these drawbacks, it was crucial for me to examine the trends in police-involved gun deaths, especially as it pertains to people of color.

While creating visualizations and conducting statistical tests, the top line results were that police-involved shooting are rising around 15 deaths per year, across all races. For just people of color (which I defined to be Black and Hispanic people) there was an increase of only 3 deaths where the police were involved. From these results, it is difficult to tell if there is a pattern in police-involved gun deaths, particularly for people of color.

Because the findings were largely inconclusive from those regressions, I turned to the statistical method of siginificance testing to test whether the average proportion of police-involved gun deaths was larger than the proportion of people of color in society. What I was able to conclude from this analysis is that average proportion of police-involved gun deaths of people of color is almost ceratinly higher than their proportion in society. I support this finding with a t-statistic of 9.75, which is sufficiently large to disprove that the proportion of people of color in society matches the proportion at which they die from police gun violence.

Since people of color are being killed at a proportion that is beyond random, I would suggest a policy of implicit bias training. These types of training have the propensity to better inform officers about the latent racism they may have. By training officers to combat their subconscious racism, the proportion of police-involved gun deaths of people of color will hopefully drop to a level that is consistent with their prevance in society. In other words, if the proportion of police-involved gun deaths of people of color drops to be statistically the same as their proprotion in society, it is a reasonable assumption that these deaths are as good as random.

Moving forward, I observed the interaction of age and intent. There are some interesting take aways that should inform the advocacy and policy in the future. Looking to the youngest people, the last visualization depicts that a substantial poriton of children who die from gun violence die as a result of an accient. Accidents are intrisically avoidable, that is why they are called accidents. Adovocates and policy makers should work to pass legislation that requires homes with children under 15 to properly store guns away from children; this will make it more difficult for accidents to occur. As described, around age 20 the proportion of homicides increases until around age 30. While this conclusion does not lead to any substantive policy suggestion, I think it is important to note that if a government wants to curb homicides that it should focus their tactics and policies on young people. As I previously mentioned, moving beyond age 30, the proportion of suicides increases steadily until almost all gun deaths can be attributed to suicide. There are policies that could help mitigate the frequency of suicides. A widely talked about policy are red flag laws. A red flag law is a statute that allows a partner, loved one, or civilian from pursuing a temorary restraining order against a person who poses eminent harm to themselves or the people around them. Red flag laws have the propensity to prevent people from harming themselves with guns; this could mitigate suicides.

To conclude, there is so much to be uncovered with the statistical analysis of gun violence in the United States. Because research had been stifled for so long and there is still steadfast opposition to any control on access to gun, the research on this subject is not as robust as it could be. In this projet I am proud that I can contribute to making the analysis of gun violence in the United States marginally more robust. While not all of my analysis leads to substantive policy steps, I am confident that this work will point policymakers in the right direction of how to act on gun violnce prevention.

References

Casselman, Ben. "Fivethirtyeight/Guns-Data." GitHub. FiveThirtyEight, June 2, 2017. https://github.com/fivethirtyeight/guns-data/blob/master/CDC_parser.R.

Casselman, Ben, Matthew Conlen, and Reuben Fischer-Baum. "Gun Deaths In America." FiveThirtyEight. FiveThirtyEight, July 13, 2016. https://fivethirtyeight.com/features/gun-deaths/.

"Data Access - Vital Statistics Online." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, October 9, 2020. https://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm.

“Everytown Research - EveryStat.” EverytownResearch.org/everystat, 2020. https://maps.everytownresearch.org/everystat.

Rostron A. (2018). The Dickey Amendment on Federal Funding for Research on Gun Violence: A Legal Dissection. American journal of public health, 108(7), 865–867. https://doi.org/10.2105/AJPH.2018.304450

“U.S. Census Bureau QuickFacts: United States.” Census Bureau QuickFacts. Accessed December 1, 2020. https://www.census.gov/quickfacts/fact/table/US/PST045219.