Week 12 Assignment

Ruben Brionez Jr

March 2nd 2024

Final Term Project Part 1

Introduction

I always have a problem when it comes to selecting a research topic, there are so many different routes to take, so many questions to be answered that it is hard to settle on one. I originally intended my topic to be about firearms and gun violence, but I've modified my topic a bit and have settled on another.

My updated topic will be:

Firearm Sales and Gun Deaths in Florida

Reasearch Questions?

1. Have Background checks increased in Florida over time?
2. Who is purchasing the most firearms in Florida?
    - Demographic Questions
3. Is there monetary data available for purchases of firearms?
4. What has been the trend in Gun Deaths in Florida over time?
5. Is there a positive correlation between Gun Deaths and Firearm purchases in Florida?

Approach?

My approach will be to research and compare data sets of gun deaths, population, and firearm sales based on background checks. This will require filtering each data set to Florida related variables. I will also be looking specifically at 2017 as the population data set is specific to this single year.

How the approach addresses (fully or partially) the problem?

My approach will attempt to find a correlation or lack of correlation between the number of firearm purchases and gun deaths in Florida to determine if more firearms cause an increase in gun deaths or has no obvious effect/correlation.

Data

The data sets have been pulled from Kaggle.com. The data includes FBI background checks for firearm purchases, census data for population, and gun violence data.

Firearms Background Check Data: *https://www.kaggle.com/datasets/saurabhshahane/fbi-criminal-background*

The data in this set includes background checks for various types of firearm purchases.

Variables include:

- "handgun"- a firearm commonly thought of as a pistol.
- "long-gun" - a firearm more commonly known as a rifle or shotgun.
- "other" - this would include any other type of item that would require a background check, this may include ammunition in some states.

Population Data: *https://www.kaggle.com/datasets/peretzcohen/2019-census-us-population-data-by-state*

The data set includes population by state, and county. It also includes various demographic variables such as, race, income, man or woman.

Gun Death Data: *https://www.kaggle.com/datasets/ahmedeltom/us-gun-deaths-by-county-19992019*

The data set includes variables such as dates, county, and states, as well as number of deaths.

When reviewing the data I use the str() function to review the structure of each variable and get an idea for the types of data I will be working with.

Required Packages

The required packages I know I will need are "dplyr" and "ggplot2". These two packages seem to be the most used packages in class.

Another that I have seen quite often is "coefplot", which aids in the plotting of the coefficient values from fitted models, such as a linear regression model.

There is a good possibility that as I begin working with the data frames I will find additional packages that will be required.

Plots and Tables Needed

Plots and tables area an essential part of data analysis. I constantly find myself thinking "Let me see how this looks" when reviewing data frames. Some of the plots I know that I will be using are below:

- Histograms
- Scatter plots
- Linear Regression Model Plots
- Coefficient Plots

Questions for Future Steps

It is hard to know what skills I may be missing until I run into an issue I cannot solve. However, it does seem that I would need additional information on comparing data frames or joining data frames for comparison.

I would also like to have some additional information on how best to present the results of the analysis.

Final Term Project Part 2

How did you import and clean your data?

I imported my data by using the readcsv() function. The data from Kaggle.com was already in a csv format which made it very easy to import right away.

```r
library(ggplot2)
library(dplyr)
background_df <- read.csv("nics-firearm-background-checks.csv")
population_df <- read.csv("acs2017_census_tract_data.csv")
deaths_df <- read.csv("gun_deaths_us_1999_2019.csv")
```

I began cleaning the data by reviewing the structure of the data frames using the str() function in R. This allowed me to see the types of data I would be working with for each data frame.

I then began checking the columns for the variables that I would want to use in my research for each data frame. In all cases I decided to leave the columns since I could later filter the columns I needed.

I ultimately created new data frames of each original data frame. The original data frames were filtered by State, Florida, to create the new data frames as well as Year of 2017.

```r
#Reviewing data from original Data frame filtering and saving to new data frame
flBackChecks <- background_df %>% group_by(state) %>%
  filter(state == "Florida", grepl('2017', month))

# Filter by all years
all_flBackChecks <- background_df %>% group_by(state) %>%
  filter(state == "Florida")

#Reviewing data from original Data frame filtering and saving to new data frame
flPop <- population_df %>% group_by(State) %>% filter(State == "Florida")

#Reviewing data from original Data frame filtering and saving to new data frame
flDeaths <- deaths_df %>% group_by(State_Name) %>%
  filter(State_Name == "Florida")
```

The reason for the year being 2017 is that the population data is only from the year 2017. In order to compare all data frames I though it would be good to check against the same years.

In regards to things that I do not know how to do, I would like to better understand how to find missing values and a best practice method for removing them or replacing them. I feel that this step wasn't given a lot of time and I need a better understanding of the process.

What does the final data set look like?

The new data frames were filtered by State. Two of the data frames were also filter by year (2017). By using filter() and group-by() the observations of the data frames were decreased. In the case of the original "background_df", which was the background check data frame, the observation went from 16,445 to 12 observations.

The new data frames can be seen below:

```r
# Florida Background Check Data
head(flBackChecks)
```

```
## # A tibble: 6 x 27
## # Groups:   state [1]
##   month   state   permit permit_recheck handgun long_gun other multiple admin
##   <chr>   <chr>    <int>          <int>   <int>    <int> <int>    <int> <int>
## 1 2017-12 Florida  20978              0   68075    37127  4067     2495     0
## 2 2017-11 Florida  19099              0   61694    31631  3585     2622     0
## 3 2017-10 Florida  18090              0   49568    24580  3958     2089     9
## 4 2017-09 Florida  10784              0   39199    17949  2319     1721     1
## 5 2017-08 Florida  18362              0   46682    21717  2847     2048     0
## 6 2017-07 Florida  16978              0   47430    18941  2975     1844    20
## # i 18 more variables: prepawn_handgun <int>, prepawn_long_gun <int>,
## #   prepawn_other <int>, redemption_handgun <int>, redemption_long_gun <int>,
## #   redemption_other <int>, returned_handgun <int>, returned_long_gun <int>,
## #   returned_other <int>, rentals_handgun <int>, rentals_long_gun <int>,
## #   private_sale_handgun <int>, private_sale_long_gun <int>,
## #   private_sale_other <int>, return_to_seller_handgun <int>,
## #   return_to_seller_long_gun <int>, return_to_seller_other <int>, ...
```

```r
# Florida Population Data
head(flPop)
```

```
## # A tibble: 6 x 37
## # Groups:   State [1]
##      TractId State County TotalPop   Men Women Hispanic White Black Native Asian
##        <dbl> <chr> <chr>     <int> <int> <int>    <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1    1.20e10 Flor~ Alach~     6834  3096  3738      9    67.1  18.6    0      3.3
## 2    1.20e10 Flor~ Alach~     3849  1806  2043      6.6  56.4  31      0.5    3.3
## 3    1.20e10 Flor~ Alach~     2374  1151  1223     14.2  57    18.1    0.7    7.5
## 4    1.20e10 Flor~ Alach~     5996  2617  3379      8.7  28    54.2    0      1.2
## 5    1.20e10 Flor~ Alach~     5202  2617  2585      9.3  68    17.8    0      2.3
## 6    1.20e10 Flor~ Alach~     4689  1779  2910      1.2   7.1  85.1    0      1.2
## # i 26 more variables: Pacific <dbl>, VotingAgeCitizen <int>, Income <dbl>,
## #   IncomeErr <dbl>, IncomePerCap <dbl>, IncomePerCapErr <dbl>, Poverty <dbl>,
## #   ChildPoverty <dbl>, Professional <dbl>, Service <dbl>, Office <dbl>,
## #   Construction <dbl>, Production <dbl>, Drive <dbl>, Carpool <dbl>,
## #   Transit <dbl>, Walk <dbl>, OtherTransp <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <int>, PrivateWork <dbl>, PublicWork <dbl>,
## #   SelfEmployed <dbl>, FamilyWork <dbl>, Unemployment <dbl>
```

```r
#Florida Gun Deaths
head(flDeaths)
```

```
## # A tibble: 6 x 15
## # Groups:   State_Name [1]
##       X  Year County   County.Code State State_Name State.Code Deaths Population
##   <int> <int> <chr>          <int> <chr> <chr>           <int>  <int>      <int>
## 1  2249  1999 Alachua~       12001 FL    Florida            12     17     215847
## 2  2250  1999 Bay Cou~       12005 FL    Florida            12     17     148149
## 3  2251  1999 Brevard~       12009 FL    Florida            12     45     472138
## 4  2252  1999 Broward~       12011 FL    Florida            12    136    1594130
## 5  2253  1999 Charlot~       12015 FL    Florida            12     13     140240
## 6  2254  1999 Citrus ~       12017 FL    Florida            12     20     116634
## # i 6 more variables: Crude.Rate <dbl>,
## #   Crude.Rate.Lower.95..Confidence.Interval <dbl>,
## #   Crude.Rate.Upper.95..Confidence.Interval <dbl>, Age.Adjusted.Rate <dbl>,
## #   Age.Adjusted.Rate.Lower.95..Confidence.Interval <dbl>,
## #   Age.Adjusted.Rate.Upper.95..Confidence.Interval <dbl>
```

What information is not self-evident?

When reviewing the data frames a few things are not self evident. In regards to the topic at hand, Firearm and Gun Deaths in Florida, it is not clear at all how the two are related, if at all.

Just by viewing the data it is not evident if background checks and firearm sales has any trend or pattern for the year of 2017.

It is also not evident whether there has been a population increase or decrease for the year of 2017.

It is not clear what percentage of the Florida population has purchased or attempted to purchase a firearm for the year of 2017.

What are different ways you could look at this data?

There are many different ways to look at this data. Instead of filtering by State, I could have left all states to get a broader look at the data across multiple states.

There could also be an analysis done of the types of firearms sold since this data is included in the background check data. However, my intent is to simpley look at all firearms.

In the population data frame, flpop, there are a total of 37 variables that could be explored. Among those are, County, Men, Women, and various race identifiers. Since the topic is more broad those variables do not need to be explored for now but could be used in a regression model.

How do you plan to slice and dice the data?

My plan for the data started with filtering by State and Year. This was done to narrow the topic to only include Florida and the Year of 2017 statistics. Next, my plan is to sum the columns of the background check data to get a total of all firearm purchases for Florida in 2017.

I would then like to compare that data to some of the population data as well as the "flDeaths" data to begin the analysis.

After reviewing and attempting to plot some of the data as it is, it appears there would be some benefits to adding additional columns to some of the data frames. Perhaps adding a total column of all firearms background checks to "flBackChecks" would make it easier for comparison and analysis.

An example of this can be seen below:

```
# Summing the columns of the data frame flBackChecks
totalFirearms <- flBackChecks$handgun + flBackChecks$long_gun +
  flBackChecks$other + flBackChecks$multiple
# Printing the results
totalFirearms
```

```
##  [1] 111764  99532  80195  61188  73294  71190  76766  73084  77508  90554
## [11]  87795  78411
```

```
# Summing the returned vector
sum(totalFirearms)
```

```
## [1] 981281
```

The code above may need to be modified, it appears that it returns a vector that is 1:12, while the sum() returns the sum of the vector. This would be challenging to work with the existing data frames.

How could you summarize your data to answer key questions?

There are a few different ways to summarize the data after exploring and analysis. It is my intent to summarize the data to reflect a few key questions.

There should be a summary of the percentage of the total population and the number of background checks completed for the purchases of firearms.

There should also be a summary of the percentage of firearm deaths compared to the total population as well as a percentage of firearm deaths to the percentage of firearms sold vs total population. In other words I would like to find if the total firearms sold had any effect on the total firearm deaths.

Some of the summary statistics that would be required are: * sum() of all firearm sales * sum() of population * sum() of firearm deaths * mean() of firearm sales, and firearm deaths

What types of plots and tables will help you to illustrate the findings to your questions?

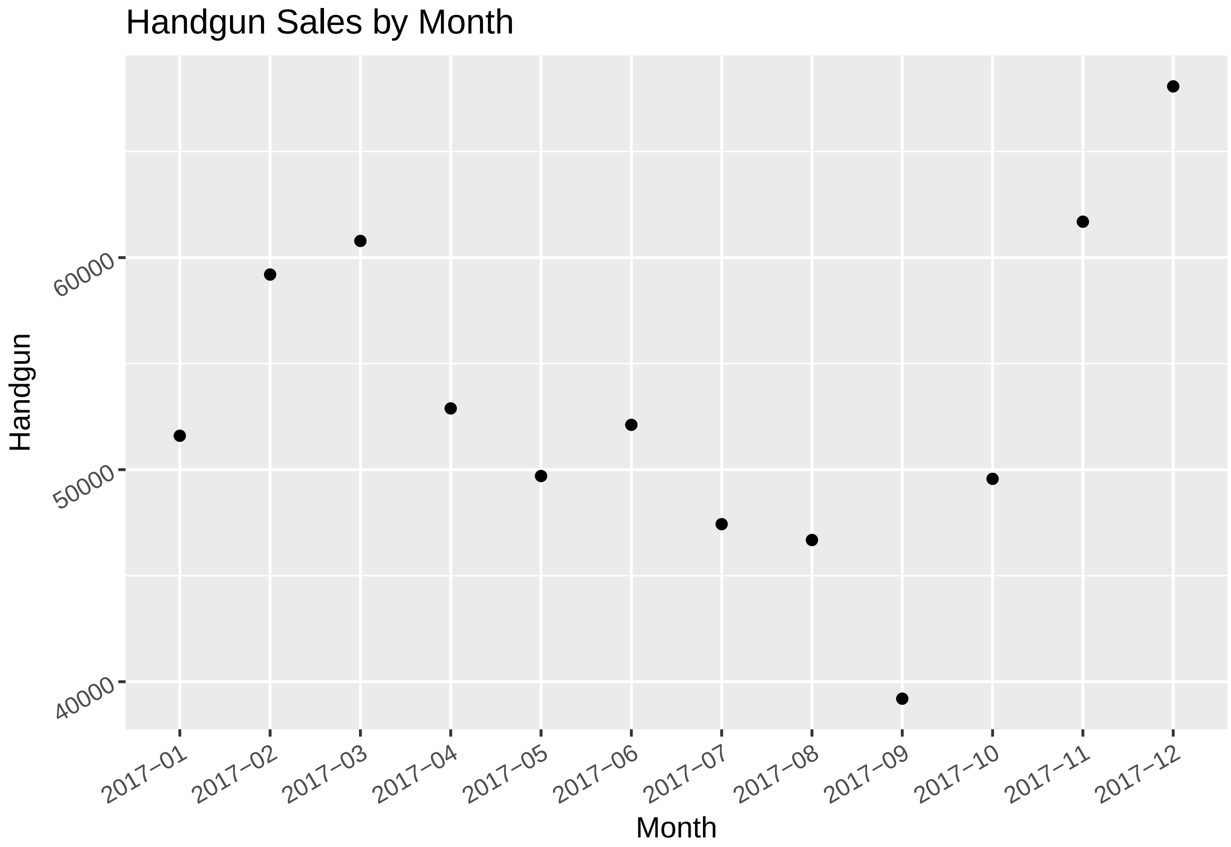
The best visualization for this topic would be a scatter plot that would show firearm purchases and firearm deaths.

This would visually show how the two variables correlate, if they do.

I would also then be able to reference the scatter plot to help create the linear or the generalized linear model for the machine learning techniques.

An example of the scatter plot can be seen below: This is not a final product.

```
library(ggplot2)
# Sample Scatter Plot from the data frame being used.
ggplot(flBackChecks, aes(x=month, y=handgun)) + geom_point() +
  labs(title="Handgun Sales by Month",x="Month", y="Handgun") +
  theme(axis.text = element_text(angle=30, hjust =1))
```



```
# A side note on the results of this plot,
# it's interesting to see that the sales increased in the months
# immediately prior to Christmas, are there that many people buying
# firearms as Christmas gifts?
```

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I would love to include machine learning techniques and I am going to do my best in doing so. Part of the reason I started this journey into Data Science was to dive deep into Machine Learning and A.I.

I will make an attempt at running a few linear regression models and multiple regression models to try and illustrate how different variables effect firearm deaths. Maybe also a model showing how race can effect the results?

I hope to also find some other applications while working with the data sets to implement the machine learning techniques that we have learned so far.

What questions do you have now, that will lead to further analysis or additional steps?

I have found a few questions that remain that are going to create further analysis.

One of those questions is, should this data have a time element to it as opposed to a single year in review? I a, currently reviewing only the year of 2017 data but it seems that it may be beneficial to have data across multiple years for better analysis.

This unfortunately would add the additional steps of cleaning the new data and filtering what would be needed again.

Another question that I have found while working with the data frames is, would it be better to create a single data frame with only the variables I need instead of working across data frames?

It seems like the answer to this question will be yes, especially since I would like to incorporate machine learning techniques.

The two question that I think on the most while working with my data is, do I have the correct data and do I have enough data?

It seems that the only way to answer these questions is to continue with the analysis and implementing the functions.

Final Term Project Part 3

Introduction

In this final part of the project additional analysis of the data will be completed. Final plots and regression models will be provided for analysis to wrap up the project and provide a conclusion. A summary of the steps and process will be provided on how the topic was addressed.

Limitations of the analysis as well as implications of the analysis will be provided before concluding the project.

The Problem Statement Addressed

The problem statement to be addressed was: **Firearm Sales and Gun Deaths in Florida**

I tried to keep the topic a little open ended for this project in order to have the opportunity to explore other avenues the data provided.

How the Problem Statement was Addressed

The problem statement was addressed by analyzing different data sets. The data sets included data on background checks, populations and gun deaths. The data sets were imported, filtered using the 'dplyr' package and then plotted for analysis using the 'ggplot' package.

Linear regression models were also created to help analyze the data sets and provide predictions.

Analysis

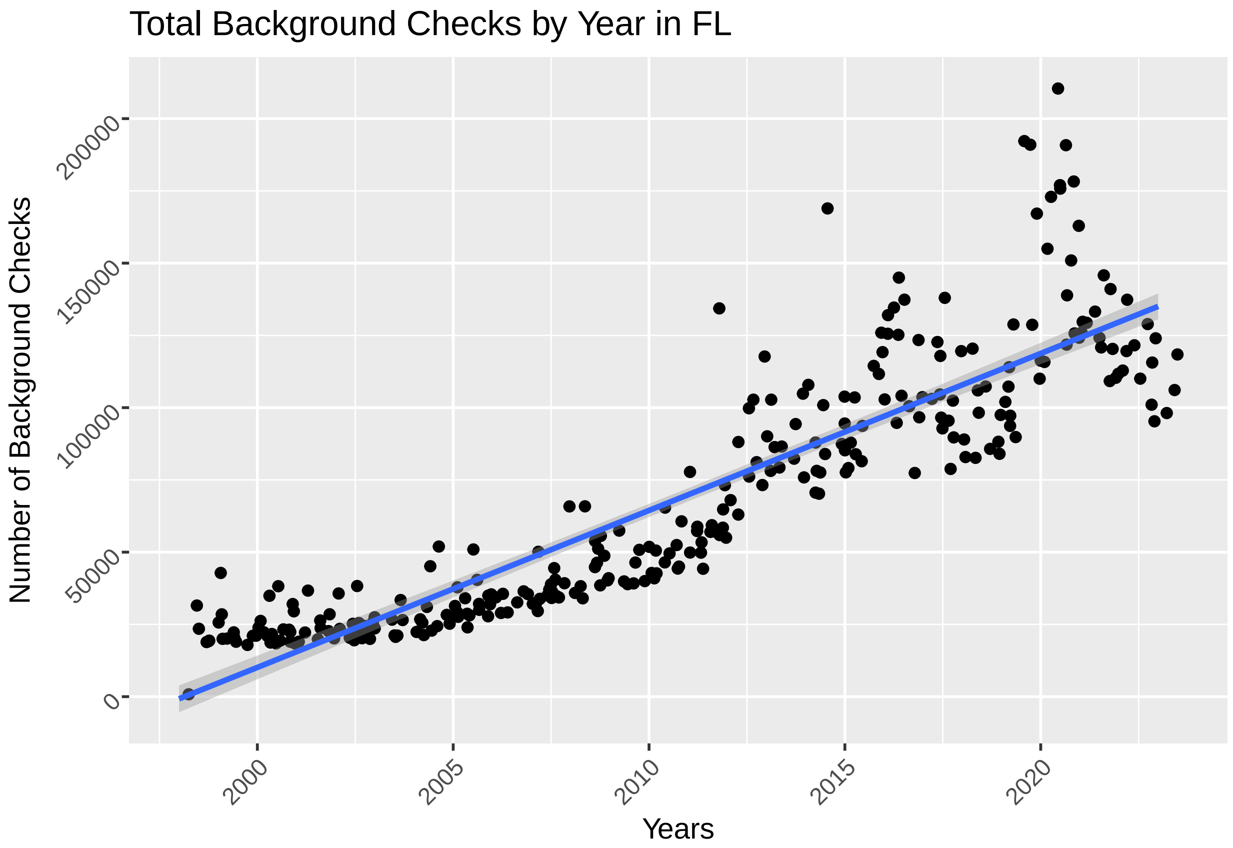
The analysis for this project included tasks such as filtering, grouping and transforming data frames. Scatter plots and linear regression models were also created for analysis.

(Further analysis is below the plots)

Plots for Analysis

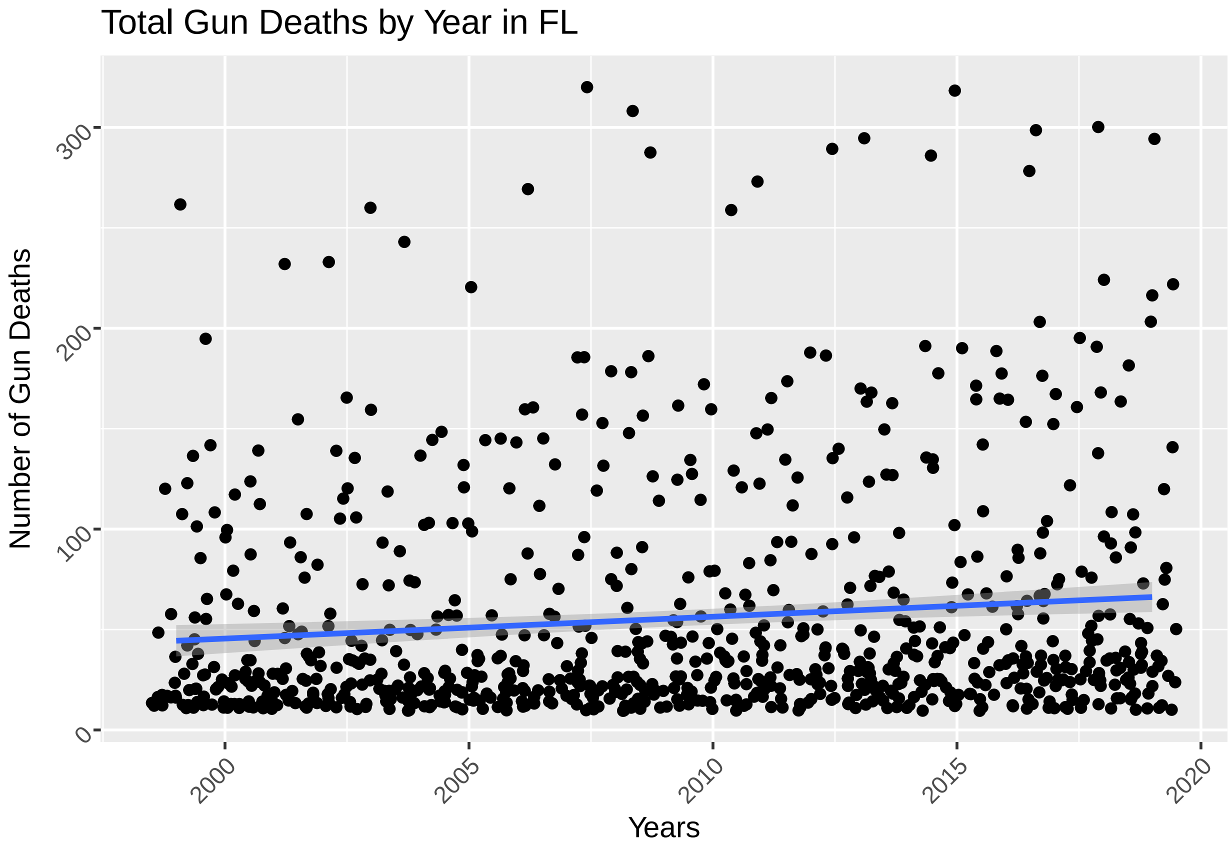
```r
# Further transformation of data
all_flBackChecks$month <- substr(all_flBackChecks$month, 1,4)

# Changing the "month" column to only years
all_flBackChecks$month <- format(as.Date(all_flBackChecks$month, format="%Y"),"%Y")

# Converted the string values to numeric in the month column
# This could be improved
all_flBackChecks$month <- as.numeric(all_flBackChecks$month)

# Plotting the Years and Total Background Checks in Florida
ggplot(all_flBackChecks, aes(x=month, y=totals)) +
  geom_jitter(width = 0.5, height = 0.5) +
  geom_smooth(method = "lm") +
  labs(title="Total Background Checks by Year in FL",x="Years",
       y="Number of Background Checks") +
  theme(axis.text = element_text(angle=45, hjust =1))
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```r
# Plotting the Years and Gun Deaths in Florida
ggplot(flDeaths, aes(x=Year, y=Deaths)) +
  geom_jitter(width = 0.5, height = 0.5) +
  geom_smooth(method = "lm") +
  labs(title="Total Gun Deaths by Year in FL", x="Years",
       y="Number of Gun Deaths") +
  theme(axis.text = element_text(angle=45, hjust =1))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Linear Regression with Values

```
# Linear regression model for the total number of background checks and years
bgchecks <- lm(totals ~ month, data = all_flBackChecks)
summary(bgchecks)
```

```
##
## Call:
## lm(formula = totals ~ month, data = all_flBackChecks)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39773 -13212  -3703  10397  91657
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.085e+07  3.224e+05  -33.67   <2e-16 ***
## month        5.432e+03  1.603e+02   33.88   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 19950 on 297 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7938
## F-statistic:  1148 on 1 and 297 DF,  p-value: < 2.2e-16
```

```
bgchecks
```

```
##
## Call:
## lm(formula = totals ~ month, data = all_flBackChecks)
##
## Coefficients:
## (Intercept)        month
##   -10853804         5432
```

```
# Linear regression model for deaths and years
deathsModel <- lm(Deaths ~ Year, data = flDeaths)
summary(deathsModel)
```

```
##
## Call:
## lm(formula = Deaths ~ Year, data = flDeaths)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56.17 -35.58 -23.76  12.17 266.89
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2131.0954   668.1358  -3.190  0.00148 **
## Year            1.0883     0.3325   3.273  0.00111 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 57.53 on 816 degrees of freedom
## Multiple R-squared:  0.01296,    Adjusted R-squared:  0.01175
## F-statistic: 10.71 on 1 and 816 DF,  p-value: 0.001109
```

```
deathsModel
```

```
##
## Call:
## lm(formula = Deaths ~ Year, data = flDeaths)
##
## Coefficients:
## (Intercept)         Year
##   -2131.095        1.088
```

Reviewing the plot titled “Total Background Checks by Year in FL” , a few observations become apparent. There is a clear visual positive correlation between “Year” and “Background Checks”; as the years increase so do the amount of background checks.

When reviewing the plot titled "Total Gun Deaths by Year in FL", there also appears to be a weak positive correlation between "Year" and "Gun Deaths".

Both P-Values for the linear regression models are close to zero, this indicates that the null hypothesis can be rejected.

Implications

The implication of the analysis is that as the years increase, both background checks for the purchase of firearms and gun deaths increase. The rate of increase differs between background checks and gun deaths but both show a positive correlation with years.

The models and plots indicate that the growth of both firearm purchases and gun deaths will continue.

These insights may help drive legislation, safety, and even more analysis of causes for the positive correlations.

Limitations

There were some limitations of the analysis. I was unable to actually join the data frames to directly compare variables. Also, providing a multiple regression model may have made the analysis more well rounded.

An analysis that included a variable of the total population would further strengthen the original analysis. We do not have an analysis of the relationship between the increase in background checks, increase in gun deaths and total population.

Concluding Remarks

The analysis has effectively shown that there is a positive correlation between "years" and "background checks" as well as "years" and "gun deaths". The linear regression models also reflect a positive linear correlation.

The P-Values calculated from the regression models are also close to zero, an indication that the null hypothesis can be rejected.

While there were some limitations of the analysis, the overall findings of the analysis addressed the topic and illustrated the relationship between variables.