DO RIGHT-TO-CARRY LAWS ACTUALLY REDUCE CRIME? : A DATA ANALYSIS OF THE "GUNS" DATASET IN R

Adel Broussard

ULID: C00162321

INFX 502

Contents

Dataset

Description

The dataset, "More Guns, Less Crime?", was obtained by John Lott and David Mustard in an attempt to answer the question: "Do right-to-carry laws result in *reduced* crime?". Lott's initial publication, *More Guns, Less Crime* (Lott 1998) sought to prove that states that enacted right-to-carry concealed firearm laws actually saw a *reduction* in crime (Ayres & Donohue 2003). In their paper, *Shooting Down the More Guns, Less Crime Hypothesis* (2003), Ayres and Donohue explain that John Lott believes arming citizens has a "protective effect for the community." The original dataset is composed of 1,173 observations on 13 variables across the years 1977 to 1999. The dataset includes information for all 50 US states, plus the District of Columbia. In the original dataset are two categorical variables and ten numerical variables. A description of all variables can be found below in *Variable Descriptions.*

Cleaning the Data

I obtained the original dataset by first installing the AER package using the "install.packages" command. Next, I loaded the AER package by using the "library" function. Finally, I loaded the dataset into R by utilizing the "data" function and used the "head" command to make sure it was properly loaded. I exported this data into a CSV file by using the "write.table" command. I did this so that I could visualize the data tabularly and also prepare to clean the data. I also used the "save" function to save the dataset in R for future use.

Installing and Loading the AER Package:

```
> install.packages(pkgs="AER")
> library(package="AER")
```

Loading the "Guns" Dataset:

```
> data(Guns)
> head(Guns)
  year violent murder robbery prisoners     afam     cauc     male population   income   density   state law
1 1977   414.4   14.2    96.8        83 8.384873 55.12291 18.17441   3.780403 9563.148 0.0745524 Alabama  no
2 1978   419.1   13.3    99.1        94 8.352101 55.14367 17.99408   3.831838 9932.000 0.0755667 Alabama  no
3 1979   413.3   13.2   109.5       144 8.329575 55.13586 17.83934   3.866248 9877.028 0.0762453 Alabama  no
4 1980   448.5   13.2   132.1       141 8.408386 54.91259 17.73420   3.900368 9541.428 0.0768288 Alabama  no
5 1981   470.5   11.9   126.5       149 8.483435 54.92513 17.67372   3.918531 9548.351 0.0771866 Alabama  no
6 1982   447.7   10.6   112.0       183 8.514000 54.89621 17.51052   3.925229 9478.919 0.0773185 Alabama  no
```

Exporting the Data:

```
> write.csv(Guns, file="~/Desktop/Guns.csv", FALSE)
```

Saving the Dataset in R:

```
save(Guns, file = "Guns.Rdata")
```

My goal in cleaning the original dataset was threefold. First, I wanted to make the data types consistent for easier handling and analysis. Second, I wanted to make sure there were no missing ("NA") values. Finally, I wanted to remove outliers for better analysis. Provided below is the original dataset structure.

```
> str(Guns)
'data.frame':   1173 obs. of  13 variables:
 $ year      : Factor w/ 23 levels "1977","1978",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ violent   : num  414 419 413 448 470 ...
 $ murder    : num  14.2 13.3 13.2 13.2 11.9 10.6 9.2 9.4 9.8 10.1 ...
 $ robbery   : num  96.8 99.1 109.5 132.1 126.5 ...
 $ prisoners : num  83 94 144 141 149 183 215 243 256 267 ...
 $ afam      : num  8.38 8.35 8.33 8.41 8.48 ...
 $ cauc      : num  55.1 55.1 55.1 54.9 54.9 ...
 $ male      : num  18.2 18 17.8 17.7 17.7 ...
 $ population: num  3.78 3.83 3.87 3.9 3.92 ...
 $ income    : num  9563 9932 9877 9541 9548 ...
 $ density   : num  0.0746 0.0756 0.0762 0.0768 0.0772 ...
 $ state     : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ law       : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Changing Data Types:

As is evident in the original data structure, there are both integer and numeric data types. I wanted to change the one integer type to numeric for better handling and analysis. To do this, I used the "as.numeric" command. I used the "is.numeric" command to check my work. I also wanted to change the factor data types (for the state and law variables) to character types for simplicity and better understanding on my part. To do this, I used the "as.character" command and the "is.character" command to check my work. I used the "str" function to view the changes.

```
> Guns$prisoners <- as.numeric(Guns$prisoners)
> is.numeric(Guns$prisoners)
[1] TRUE
> Guns$state <- as.character(Guns$state)
> is.character(Guns$state)
[1] TRUE
> Guns$law <- as.character(Guns$law)
> is.character(Guns$law)
[1] TRUE
> str(Guns)
'data.frame':   1173 obs. of  13 variables:
 $ year      : Factor w/ 23 levels "1977","1978",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ violent   : num  414 419 413 448 470 ...
 $ murder    : num  14.2 13.3 13.2 13.2 11.9 10.6 9.2 9.4 9.8 10.1 ...
 $ robbery   : num  96.8 99.1 109.5 132.1 126.5 ...
 $ prisoners : num  83 94 144 141 149 183 215 243 256 267 ...
 $ afam      : num  8.38 8.35 8.33 8.41 8.48 ...
 $ cauc      : num  55.1 55.1 55.1 54.9 54.9 ...
 $ male      : num  18.2 18 17.8 17.7 17.7 ...
 $ population: num  3.78 3.83 3.87 3.9 3.92 ...
 $ income    : num  9563 9932 9877 9541 9548 ...
 $ density   : num  0.0746 0.0756 0.0762 0.0768 0.0772 ...
 $ state     : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ law       : chr  "no" "no" "no" "no" ...
```

Checking for NA Values:

```
> any(is.na(Guns))
[1] FALSE
```

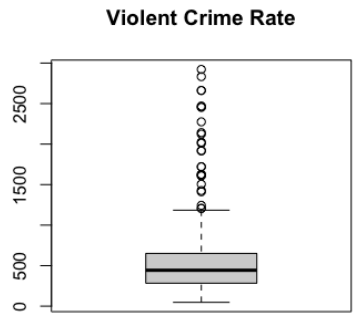
Outliers:

In order to identify the outliers, I needed to separate the dataset by year. Since crime will naturally increase or decrease over time, it made more sense to observe outliers per year than over a twenty-two-year period. To do this, I simply created separate CSV files for each year and imported those files into R using the "read.csv" command. Below is an example for the 1977 dataset.

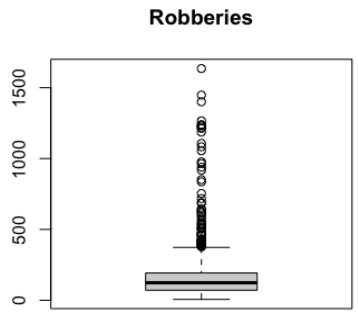
```
> Guns1977 <- read.csv("~/Desktop/Guns1977.csv")
```

To visualize and identify the outliers, I used the "Boxplot" function for each relevant numeric variable. An added benefit of using the "Boxplot" function is that it assigns a value to the outlier that corresponds with the line of data in the csv file, which makes it much simpler to find. Below are examples of the boxplot function for 1977.

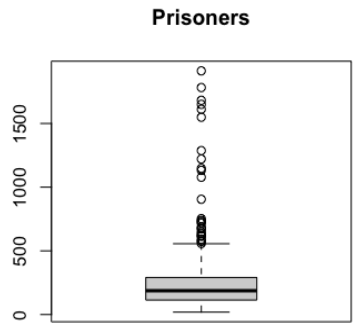
```
> Boxplot(Guns1977$violent, main =
"Violent Crime Rate")
```




```
> Boxplot(Guns1977$robbery, main =
"Robberies")
```




```
> Boxplot(Guns1977$prisoners, main =
"Prisoners")
```



I assessed the outliers for only the above three variables since I am analyzing the frequency of *crime*. It did not make logical sense to assess the prisoners, afam, cauc, or male outliers. Once I identified the outliers for these values in each year, I made sure to cross-analyze them with density, since it is generally accepted that crime rate has a positive correlation with density. For example, the outlier value for the murder rate in 1977 (9) correlates to the line of data for the District of Columbia. However, when I went into the dataset for that year, I saw that the population density was 11.10212 while the density for all other states was under 1.0. In fact, I saw that across all twenty-two years, the District of Columbia was a consistent outlier in all the data. Because of this, I made the decision to exclude the District of Columbia entirely from my analysis. I went through each year and deleted the 9th row (the row that corresponded with DC) from the datasets. Below is an example from the 1977 dataset.

```
> Guns1977 <- Guns1977[-9,]
```

Alternatively, if analysis called for use of the entire dataset (not just yearly), I excluded the District of Columbia using the following:

```
> Guns <-
Guns[-207,][-206,][-205,][-204,][-203,][-202,][-201,][-200,][-199,][-198,][-197,][-196,][-195,][-194,][-193,][-192,][-191,][-190,][-189,][-188,][-187,][-186,][-185,]
```

Variable Description

I obtained descriptions of the variables in this dataset from R-Project. The exact site can be found at https://search.r-project.org/CRAN/refmans/AER/html/Guns.html. I created a table of the variables and their descriptions, which can be found below. Please note that this table reflects the original variable descriptions, and the mode for "prisoners", "state", and "law' have been changed to numeric, character, and character, respectively.

Variable	Description	Mode
state	US state being referenced	factor
year	Year being referenced, from 1977-1999	factor
violent	Rate of violent crime (incidents per 100,000 members of the population	numeric
murder	Rate of murder (incidents per 100,000)	numeric
robbery	Rate of robbery (incidents per 100,000)	numeric
prisoners	Incarcerated prisoners per 100,000 residents for the previous year	integer
afam	Percentage of the state that is African-American, ages 10 to 64	numeric
cauc	Percentage of the state that is Caucasian, ages 10 to 64	numeric
male	Percentage of the state that is male, ages 10 to 29	numeric
population	State population, in millions of people	numeric
income	Per capita personal income in the state (US dollars)	numeric
density	Population per square mile of land area, divided by 1,000	numeric
law	Indicates whether a state has a shall-carry law in effect for that year	factor

Purpose and Expectations

Gun violence is arguably one of the most controversial topics within the current zeitgeist. It's the hot-button topic of every political debate and even many family dinners. It seems as though everyone has an opinion, but no one has a resolution. A popular solution seems to be stricter firearm regulations. However, others believe arming citizens is the solution. They seem to think that if all citizens are armed, fewer people will be willing to commit violent crime against others out of fear that they will retaliate with their firearm. These opinions are hard to definitively prove true or false, but we can become better educated on possible solutions by looking at relevant data.

I will attempt to prove my own hypothesis that stricter gun laws lead to fewer gun-related crimes. Using the data in this set, that means that I expect to see more violent crime rate, murder, and robbery in states that have right-to-carry laws than in states that do not (i.e. states that have stricter gun laws). In states that change their legislation from having no right-to-carry laws to enacting said law, I expect to see an increase in gun-related crime. I do not plan on analyzing in-depth the variables of race, gender, or socioeconomic status, as that is beyond the scope of this paper. It must be noted that these are my educated predictions and, while I do have my own personal opinions on the matter, I plan to let the data speak for itself and will not manipulate the data to conform to my own beliefs.

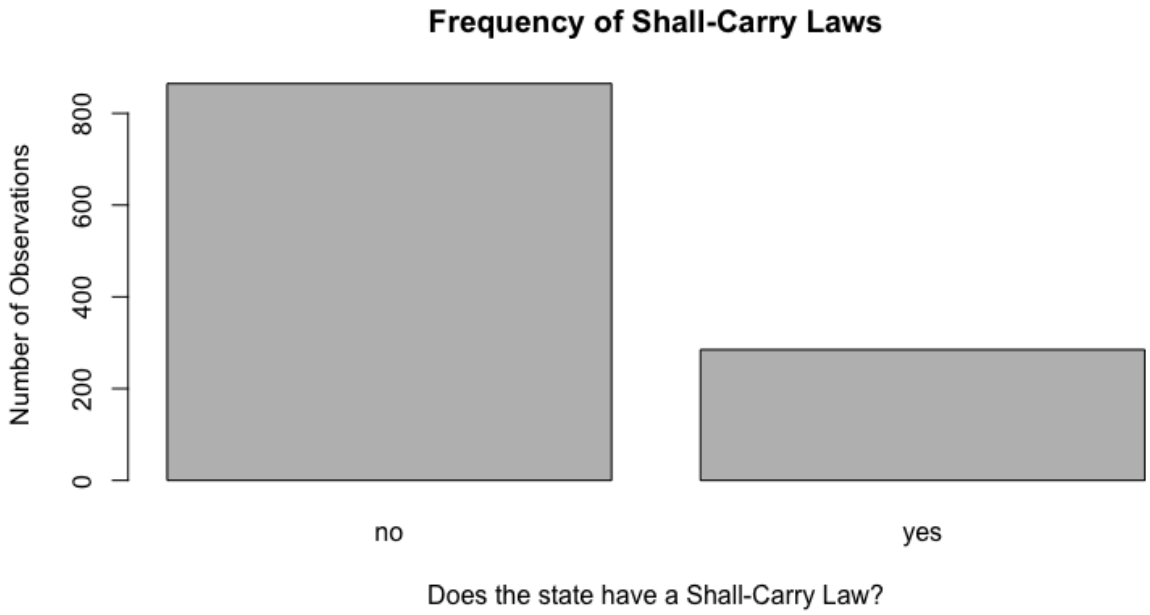
Data Analysis

Categorical Variable Analysis

Barplots

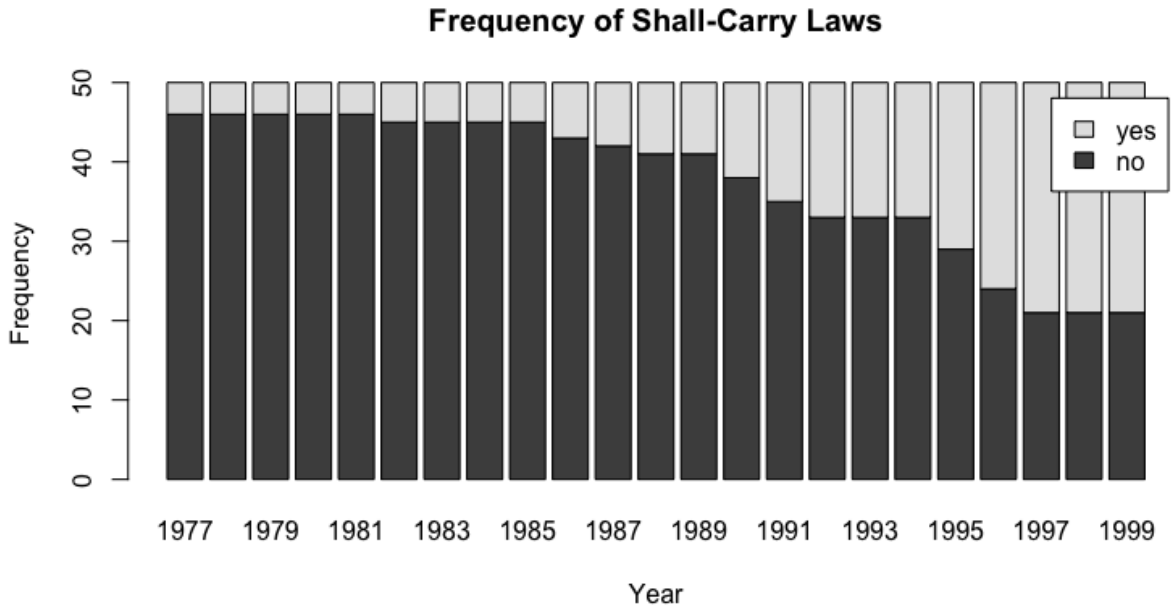
In a perfect dataset, the frequency of shall-carry laws would be evenly split. However, that is not reality. Before analyzing any of the data, I wanted to be aware of, and keep in mind, which is more popular: to have shall-carry laws or not. To visualize this, I create the barplot below.

```
> barplot(table(Guns$law), main = "Frequency of Shall-Carry Laws", xlab = "Does the state have a Shall-Carry Law?", ylab = "Number of Observations")
```



We can see clearly that, within this dataset, it is far more common for a state to not have right-to-carry laws than to have them. While this is important information to have, I also wanted to visualize the popularity of shall-carry laws over time.

```
> barplot(table(Guns$law, Guns$year), legend.text=TRUE, xlab="Year", ylab="Frequency",
main="Frequency of Shall-Carry Laws")
```



As is evident in the above graph, shall-carry laws became increasingly popular over time. This is important information to keep in mind as we further analyze the data. Instead of looking at raw numbers, it will be imperative that we look at the trends across time and how those compare to the increase in shall-carry laws.

Contingency Tables

To further corroborate the above graphs, I created a contingency table for the *law* and *year* variables using the table() function. We can see numerically that shall-carry laws grew in popularity across time.

```
> table(Guns$law, Guns$year)
```

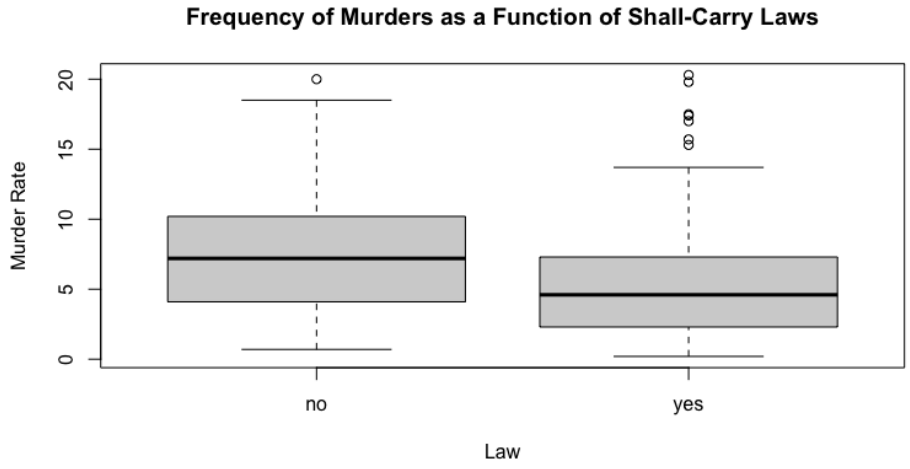
	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
no	46	46	46	46	46	45	45	45	45	43	42	41	41	38	35
yes	4	4	4	4	4	5	5	5	5	7	8	9	9	12	15

	1992	1993	1994	1995	1996	1997	1998	1999
no	33	33	33	29	24	21	21	21
yes	17	17	17	21	26	29	29	29

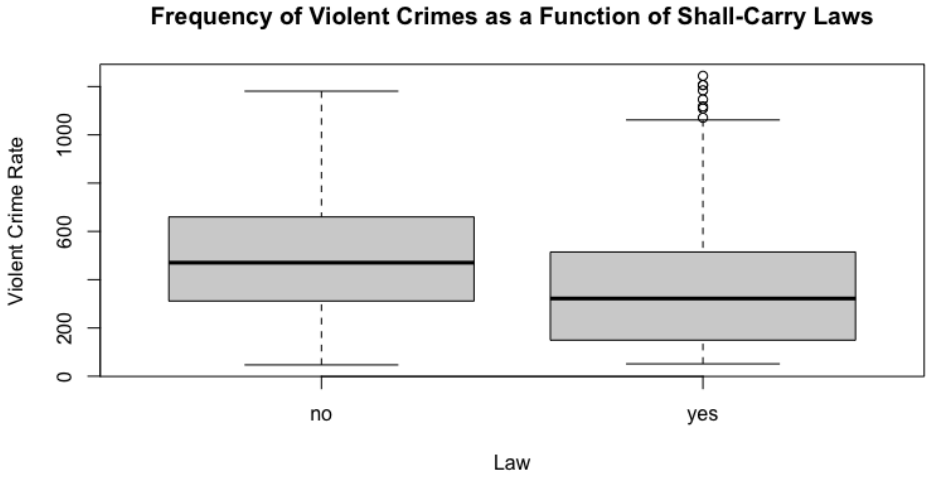
Boxplots

The most logical first step in testing my hypothesis was to analyze the frequency of murders, robberies, and violent crimes against the *law* variable. To do this, I used the plot() function to create boxplots for the *murder*, *robbery*, and *violent* variables and compare their frequencies against the *yes* or *no* observations of the *law* variable. This would show me if more crimes occur when there are, or are not, shall-carry laws.

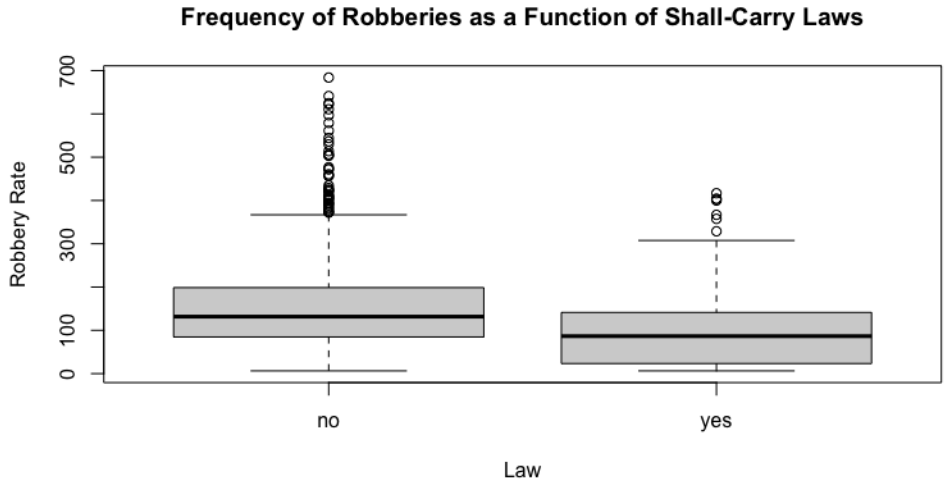
```
> plot(Guns$law, Guns$murder, main="Frequency of Murders as a Function of Shall-Carry Laws", xlab="Law", ylab="Murder Rate")
```




```
> plot(Guns$law, Guns$violent, main="Frequency of Violent Crimes as a Function of Shall-Carry Laws", xlab="Law", ylab="Violent Crime Rate")
```



```
> plot(Guns$law, Guns$robbery, main="Frequency of Robberies as a Function of Shall-Carry Laws", xlab="Law", ylab="Robbery Rate")
```

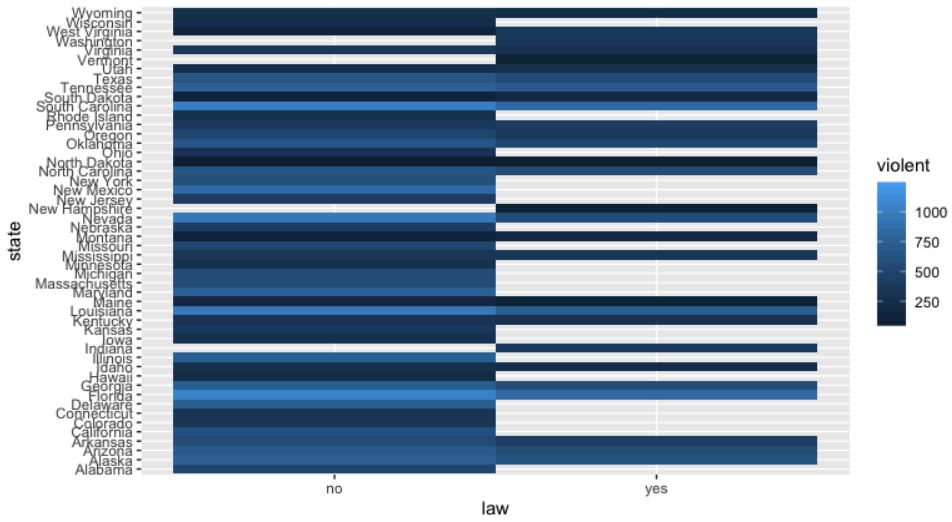


From the plots above, we can see that, for all three variables, the median rate of the crime is higher when there is not a shall-carry law in effect. For all of the variables, the IQR's of the plots are short, meaning there's not much variability in the actual rate of the crimes within the dataset when looking at the median. The whiskers of the plots are relatively long, with the upper whisker being longer than the lower whisker for all variables. This suggests that, while the majority of data points are centered around the median, there are quite a few that are on the higher end of the range. There only seem to be a few outliers with the exception of the *robbery* plot. We can see that there is a significant amount of outliers on the upper end. This suggests that the data is likely not normally distributed. Just by looking at these plots, one would gather that enacting shall-carry laws reduces crime. However, these plots do not take into account that there is less data available for the years when shall-carry laws went into effect, as they did not gain popularity until the tail end of the dataset. So, while the above plots tell us a lot about the distribution of our data, we cannot solely rely on them to make a conclusion.

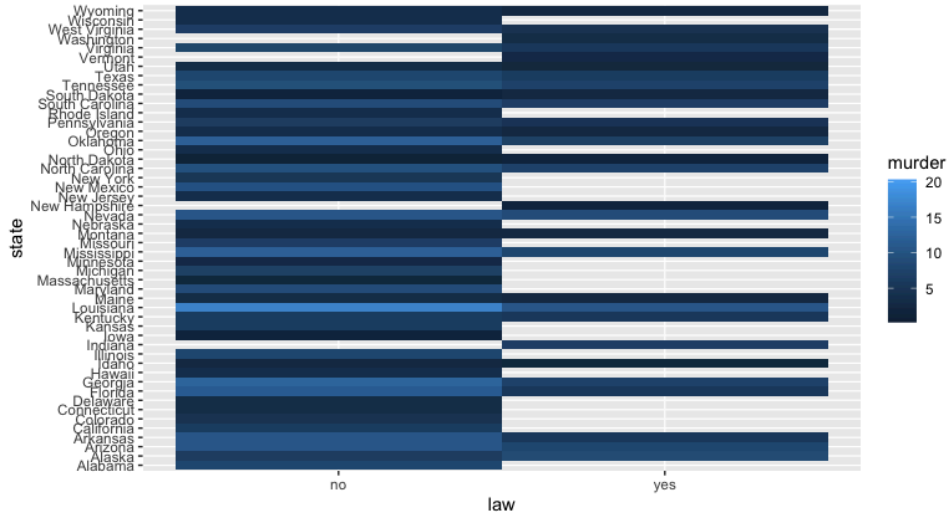
Heatmaps

The dataset does not include robust categorical data. However, I wanted to visualize the relationship between the *law* variable and the *murder*, *violent*, and *robbery* variables. To do this, I used the ggplot() function in the ggplot2 package to create a heatmap in order to further visualize the categorical data available.

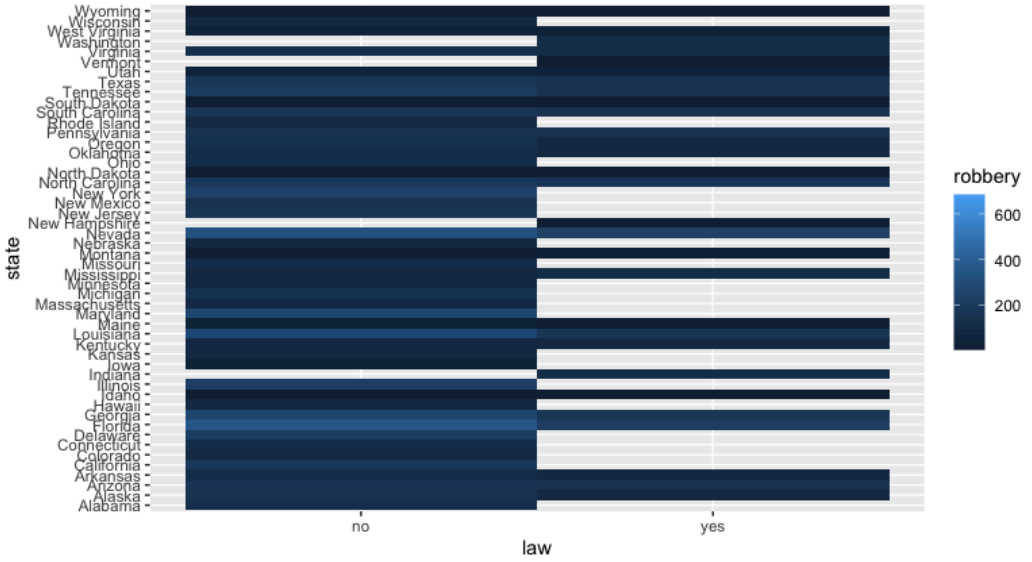
```
> ggplot(Guns, aes(x=law, y=state, fill=violent))+geom_tile()
```




```
> ggplot(Guns, aes(x=law, y=state, fill=murder))+geom_tile()
```



```
> ggplot(Guns, aes(x=law, y=state, fill=robbery))+geom_tile()
```



From the above heatmaps, we can again see that more states do *not* have shall-carry laws, for the years that the data was collected. What we can see with these graphs that we could not with the others, though, is the increase or decrease in crime as states enact shall-carry laws. For example, in the first heatmap, we can see that violent crime actually went down once shall-carry laws were enacted in the following states: Texas, Tennessee, South Carolina, Oregon, Oklahoma, North Carolina, Nevada, Maine, Louisiana, Georgia, Florida, Arkansas, Arizona, and Alaska. In fact, that seems to be the general trend for all three variables.

Numerical Variable Analysis

Scatter Plots and Correlation Matrices

I began my analysis of numerical variables by running a correlation matrix of all numeric variables. I first had to remove all non-numeric data.

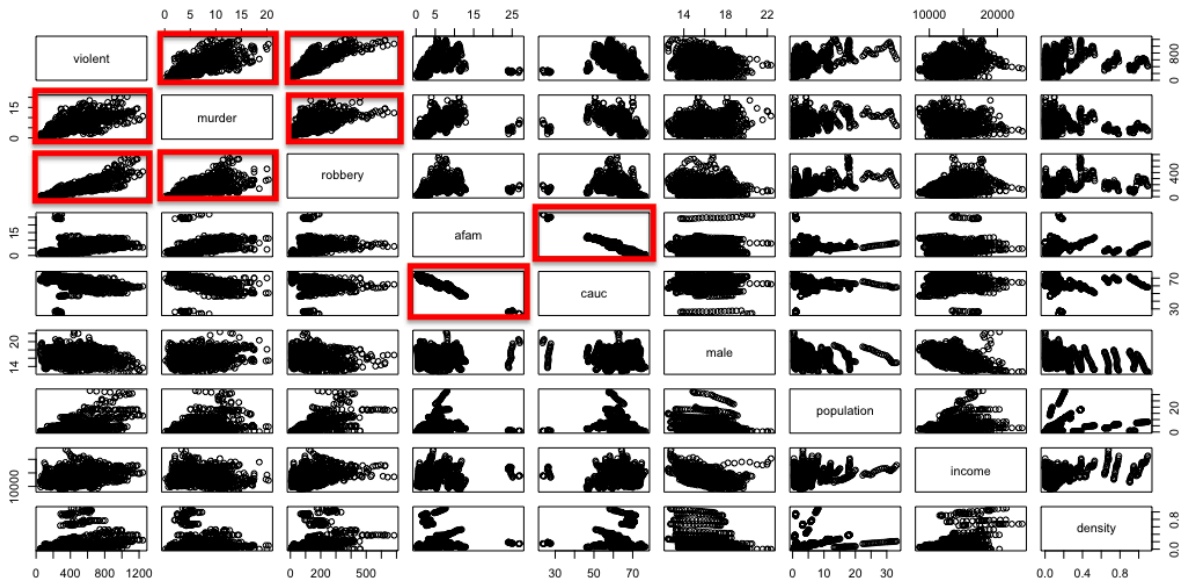
```
> Guns <- Guns[,-13][,-12][,-5][,-1]
> cor(Guns)
```

	violent	murder	robbery	afam	cauc	male
violent	1.0000000	0.70744229	0.8528373	0.35971991	-0.35043842	-0.17822233
murder	0.7074423	1.00000000	0.6264604	0.46200479	-0.46250376	0.18166542
robbery	0.8528373	0.62646045	1.0000000	0.33052951	-0.31878078	-0.10371373
afam	0.3597199	0.46200479	0.3305295	1.00000000	-0.97466958	0.03865876
cauc	-0.3504384	-0.46250376	-0.3187808	-0.97466958	1.00000000	-0.03766789
male	-0.1782223	0.18166542	-0.1037137	0.03865876	-0.03766789	1.00000000
population	0.5410316	0.38495083	0.6493391	0.14213137	-0.15460723	-0.10447151
income	0.3010535	-0.10121140	0.3592404	0.14714383	-0.05046700	-0.52686140
density	0.2215428	-0.07574882	0.3673820	0.08747197	-0.07505950	-0.20381114

	population	income	density
violent	0.5410316	0.3010535	0.22154283
murder	0.3849508	-0.1012114	-0.07574882
robbery	0.6493391	0.3592404	0.36738200
afam	0.1421314	0.1471438	0.08747197
cauc	-0.1546072	-0.0504670	-0.07505950
male	-0.1044715	-0.5268614	-0.20381114
population	1.0000000	0.2632227	0.20146801
income	0.2632227	1.0000000	0.47645385
density	0.2014680	0.4764539	1.00000000

I then used the plot() function to visualize these correlations.

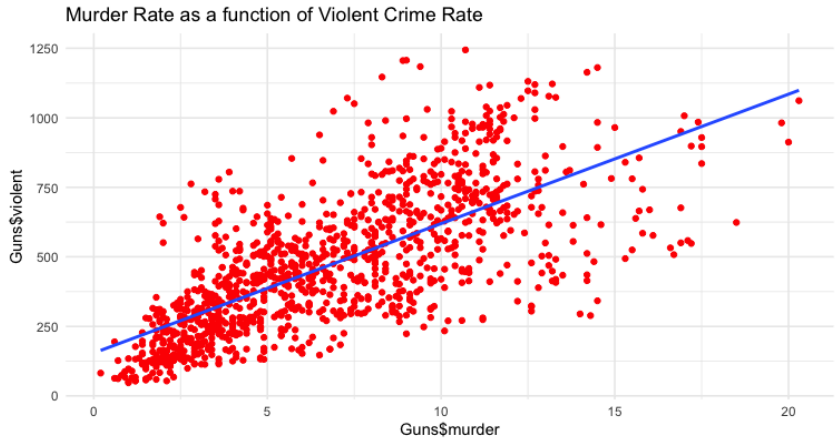
```
> plot(Guns)
```



The two plots in the middle show a seemingly significant relationship, however they simply represent the negative correlation between races. As is evident in the correlation matrix and plot above, the *murder*, *violent*, and *robbery* variables are all highly correlated with one another. This makes sense, as one would expect crimes of one type to increase as crimes of other types increase. To visualize this, I plotted the correlations between these three variables.

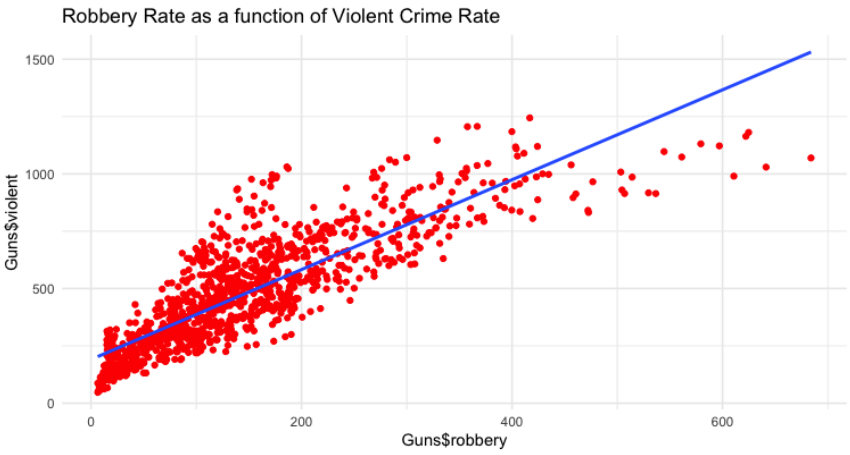
Murder & Violent Variables

```
> ggplot(Guns, aes(Guns$murder, Guns$violent)) + geom_point(color="red") +
theme_minimal() + geom_smooth(method="lm", se=FALSE) + ggtitle("Murder Rate as a
function of Violent Crime Rate")
```



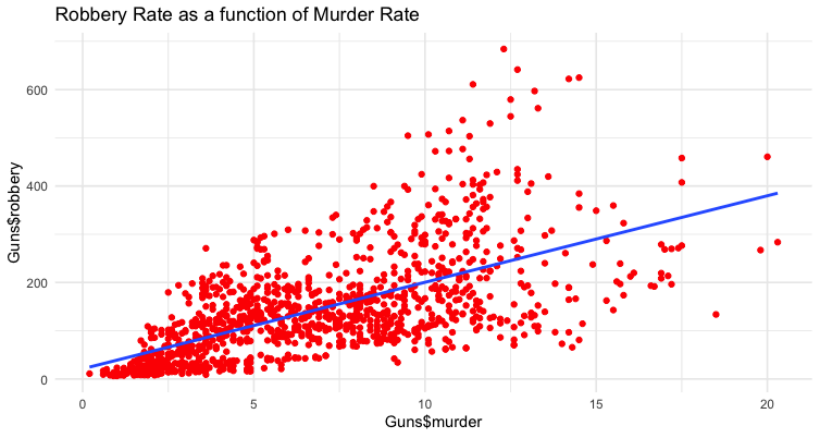
Robbery & Violent Variables

```
> ggplot(Guns, aes(Guns$robbery, Guns$violent)) + geom_point(color="red") +
theme_minimal() + geom_smooth(method="lm", se=FALSE) + ggtitle("Robbery Rate as a
function of Violent Crime Rate")
```



Murder & Robbery Variables

```
> ggplot(Guns, aes(Guns$murder, Guns$robbery)) + geom_point(color="red") +
theme_minimal() + geom_smooth(method="lm", se=FALSE) + ggtitle("Robbery Rate as a
function of Murder Rate")
```

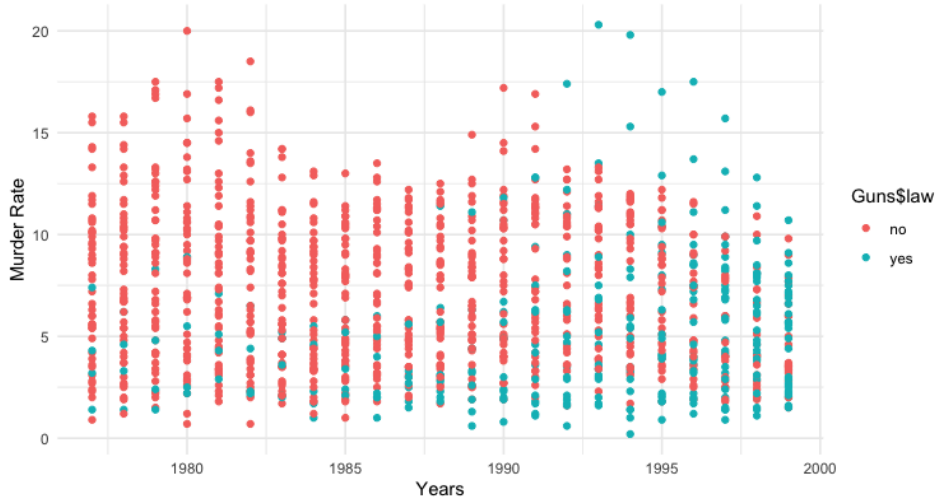


While all three variables have a high correlation with one another, *robbery* and *violent* have the highest (0.85). The data can be corroborated with the following fact from the FBI's Uniform Crime Reporting Program: "Over 60% of assaults, including the heinous crime of rape, happens during home invasions" (Burglary Statistics).

Returning to my hypothesis which argues that enacting right-to-carry laws will *increase* gun violence, I wanted to create a plot that directly addressed this. To do this, I again used the three crime variables from above (*robbery*, *violent*, and *murder*) and plotted them across years. I then color-coded each coordinate to identify whether or not that state, at that time, had right-to-carry laws or not. I found the following:

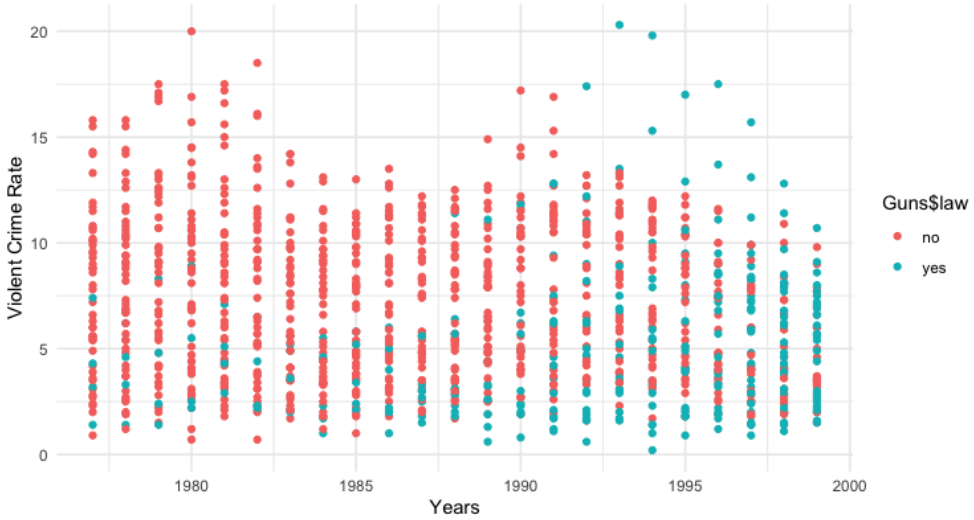
Murder Rate Across Time

```
> ggplot(Guns, aes(Guns$year, Guns$murder, color = Guns$law)) + geom_point() +
xlab("Years") + ylab("Murder Rate") + theme_minimal()
```



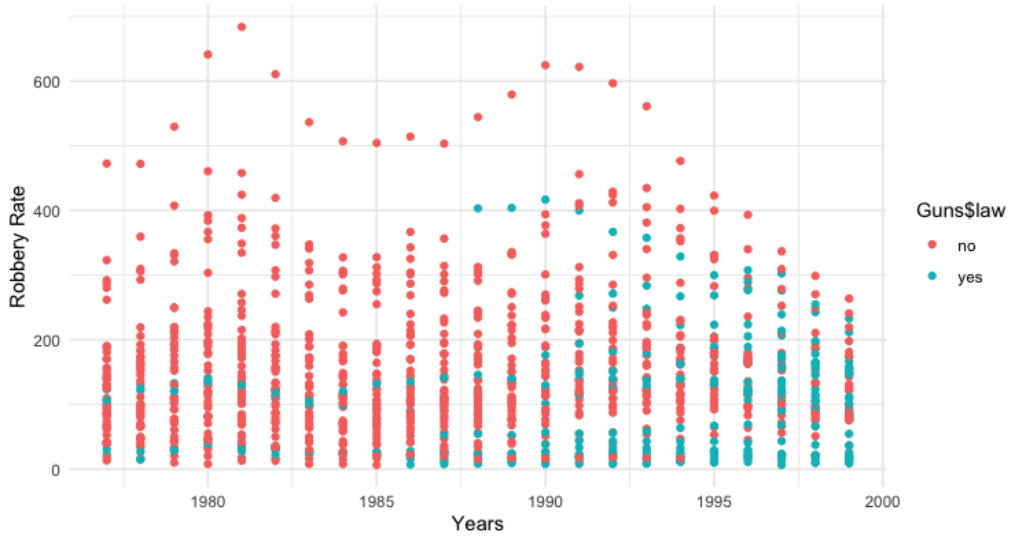
Violent Crime Rate Across Time

```
> ggplot(Guns, aes(Guns$year, Guns$murder, color = Guns$law)) + geom_point() +
xlab("Years") + ylab("Violent Crime Rate") + theme_minimal()
```



Robbery Rate Across Time

```
> ggplot(Guns, aes(Guns$year, Guns$robbery, color = Guns$law)) + geom_point()
+xlab("Years") + ylab("Robbery Rate") + theme_minimal()
```



When looking at the first graph, *Murder Rate Across Time*, we can infer a few things. First, we can see that right-carry-laws became increasingly popular after 1990. Second, the crime rate is bimodal, with peaks around 1980 and 1995. These have historically been attributed to the crack-cocaine epidemic of the 1980's (Section I: Gun Violence in the United States). Third, there is no clear trend to suggest that right-to-carry laws increase or decrease crime, at least in these charts. In fact, these three inferences are true for all three graphs. Further analysis is required.

Summary Statistics

While at this point I feel familiar with the variables in this dataset, I still wanted to get a better statistical understanding of them. To do this, I used the summary() function for all variables except for *year*, *state*, and *law*.

```
> summary(Guns[ , c(-1, -12, -13)]
```

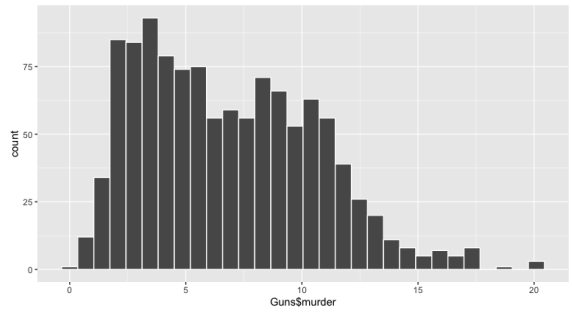
```
    violent           murder          robbery          prisoners          afam              cauc       
 Min.   :  47.0   Min.   : 0.200   Min.   :  6.4   Min.   : 19.0   Min.   : 0.2482   Min.   :23.52  
 1st Qu.: 281.5   1st Qu.: 3.600   1st Qu.: 70.2   1st Qu.:113.0   1st Qu.: 2.1603   1st Qu.:60.29  
 Median : 437.5   Median : 6.300   Median :122.4   Median :183.0   Median : 3.9651   Median :65.34  
 Mean   : 472.2   Mean   : 6.833   Mean   :143.7   Mean   :211.5   Mean   : 4.9652   Mean   :63.71  
 3rd Qu.: 635.2   3rd Qu.: 9.600   3rd Qu.:188.5   3rd Qu.:285.0   3rd Qu.: 6.4426   3rd Qu.:69.29  
 Max.   :1244.3   Max.   :20.300   Max.   :684.0   Max.   :736.0   Max.   :26.9796   Max.   :76.53  
      male         population          income         density          
 Min.   :12.51   Min.   : 0.4027   Min.   : 8555   Min.   :0.0007071  
 1st Qu.:14.66   1st Qu.: 1.2424   1st Qu.:11897   1st Qu.:0.0315201  
 Median :15.89   Median : 3.2889   Median :13316   Median :0.0795802  
 Mean   :16.09   Mean   : 4.9006   Mean   :13623   Mean   :0.1636141  
 3rd Qu.:17.53   3rd Qu.: 5.7378   3rd Qu.:15162   3rd Qu.:0.1682232  
 Max.   :22.35   Max.   :33.1451   Max.   :23647   Max.   :1.0976430  
```

What I can gather just from looking at the output of this function is that violent crimes make up the majority of reported crime in this dataset. *Robbery* follows, then *murder*. We can also see that the population within this dataset is majority Caucasian. What is extremely interesting, and important to note, is the *income*. The average income in the United States in 1977 was around $13,000 (Money Income in 1977 of Households in the United States) and the average in 1999 was $42,000 (Census 2000 Brief: Household Income: 1999) . Within this dataset, the max income (across all years) is $23,647. This means that the population is majority low socioeconomic status.

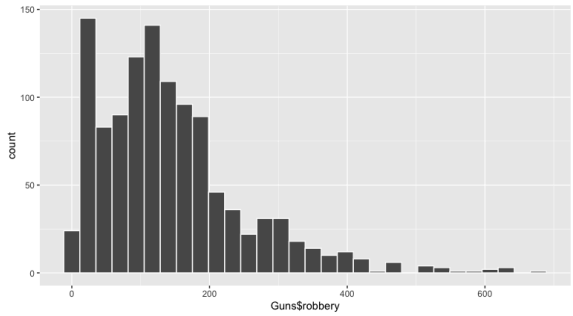
Histograms and Density Plots

To determine the distribution of the variables, I created the following histograms for each of them. As we can see, they are all positively skewed.

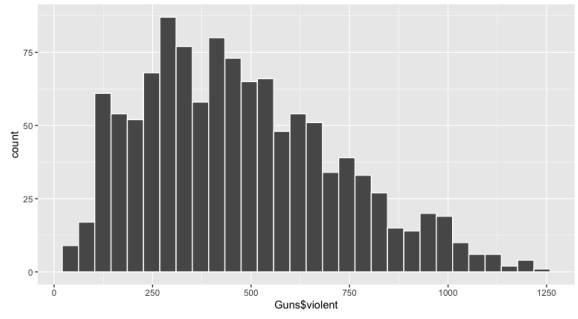
```
ggplot(Guns, aes(Guns$murder))+geom_histogram(color="white")
```



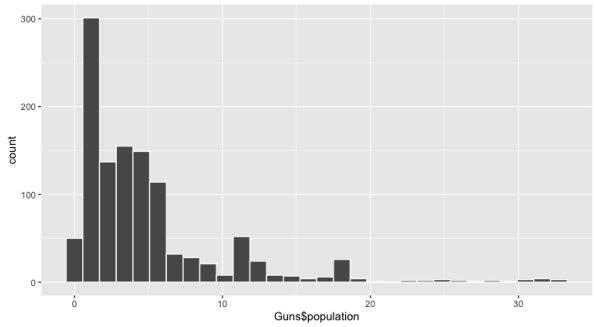

```
ggplot(Guns, aes(Guns$robbery))+geom_histogram(color="white")
```



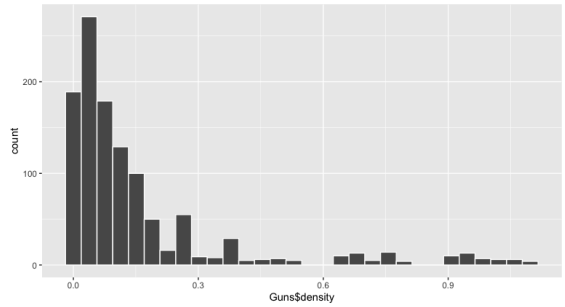

```
ggplot(Guns, aes(Guns$violent))+geom_histogram(color="white")
```



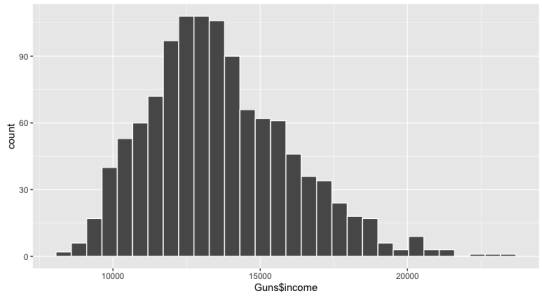

```
ggplot(Guns, aes(Guns$population))+geom_histogram(color="white")
```




```
ggplot(Guns, aes(Guns$density))+geom_histogram(color="white")
```




```
ggplot(Guns, aes(Guns$income))+geom_histogram(color="white")
```

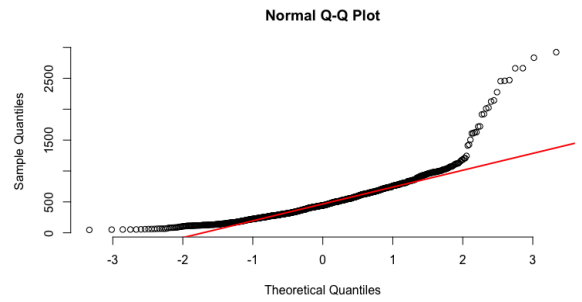


I then performed the Shapiro-Wilk test on the above variables.

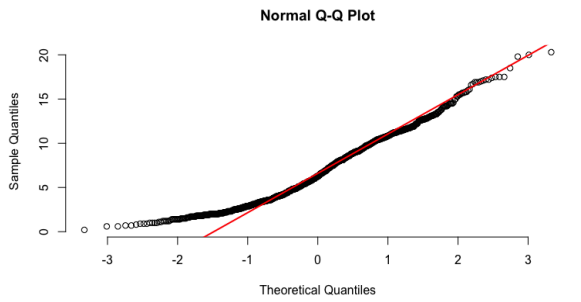
> shapiro.test(Guns$murder) Shapiro-Wilk normality test data: Guns$murder W = 0.96112, p-value < 2.2e-16	> shapiro.test(Guns$robbery) Shapiro-Wilk normality test data: Guns$robbery W = 0.89462, p-value < 2.2e-16
> shapiro.test(Guns$violent) Shapiro-Wilk normality test data: Guns$violent W = 0.96775, p-value = 2.647e-15	> shapiro.test(Guns$population) Shapiro-Wilk normality test data: Guns$population W = 0.73846, p-value < 2.2e-16
> shapiro.test(Guns$density) Shapiro-Wilk normality test data: Guns$density W = 0.65615, p-value < 2.2e-16	> shapiro.test(Guns$income) Shapiro-Wilk normality test data: Guns$income W = 0.97589, p-value = 6.457e-13

The Shapiro-Wilk Test corroborates the histograms and confirms that none are distributed normally. Because of this, using a Quantile-Quantile Plot is useful.

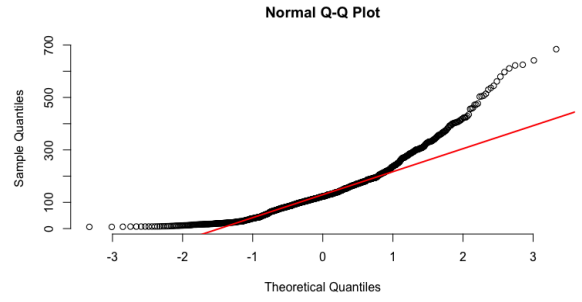
```
> qqnorm(Guns$violent, pch=1, frame=FALSE)
> qqline(Guns$violent, col="red", lwd=2)
```



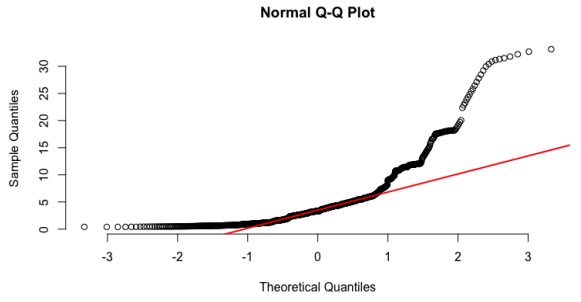

```
> qqnorm(Guns$murder, pch=1, frame=FALSE)
> qqline(Guns$murder, col="red", lwd=2)
```



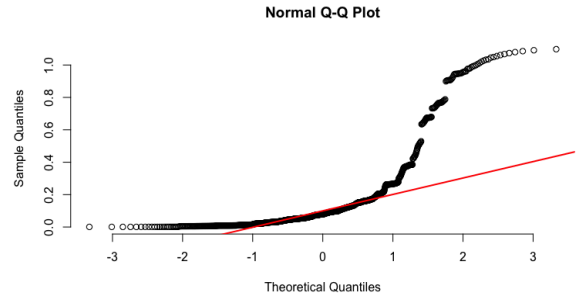

```
> qqnorm(Guns$robbery, pch=1, frame=FALSE)
> qqline(Guns$robbery, col="red", lwd=2)
```



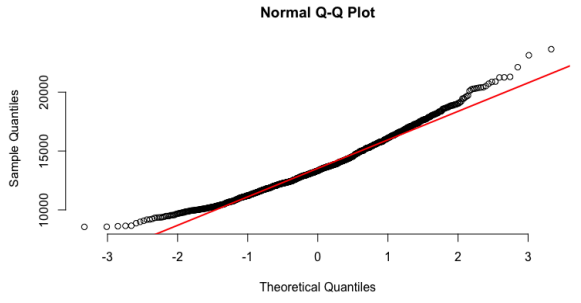

```
> qqnorm(Guns$population, pch=1, frame=FALSE)
> qqline(Guns$population, col="red", lwd=2)
```




```
> qqnorm(Guns$density, pch=1, frame=FALSE)
> qqline(Guns$density, col="red", lwd=2)
```




```
> qqnorm(Guns$income, pch=1, frame=FALSE)
> qqline(Guns$income, col="red", lwd=2)
```



These QQ Plots further confirm what the histograms and Shapiro-Wilk Tests have already shown. The most normally distributed variables are the *income* and *murder* variables, but even those are skewed. The *violent*, *density*, *population*, and *violent* variables are heavily influenced by outliers. The *density* and *population* outliers are especially prevalent, which may explain why we did not see a strong correlation between them. This could also influence their correlation with other variables, which may explain why there was not as strong of a correlation between density and *violent*, *murder*, or *robbery*, as I had predicted.

Multiple Linear Regression

So far I have not seen a clear or definite relationship between *law* and the three crime variables. In order to test if this is due to my own error or if there truly is no relationship in this dataset, I performed multiple linear regression using the lm() and summary() functions. My first model is very straightforward, as I investigate the relationship between the crime variables and the *law* variable. I converted the *law* variable to numeric in order to run the analysis. The equation for the first regression can be found below.

$$Y = \beta_0 + \beta_1 numericlaw$$

Multiple Regression #1

```
> numericlaw <- as.numeric(Guns$law)
> mr1 <- lm(violent~numericlaw, data=Guns)
> summary(mr1)

Call:
lm(formula = violent ~ numericlaw, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-455.17 -202.00  -41.62  153.01  863.25

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   623.30      21.78  28.617  < 2e-16 ***
numericlaw   -121.12      16.50  -7.343 3.96e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 241.5 on 1148 degrees of freedom
Multiple R-squared:  0.04486,   Adjusted R-squared:  0.04403
F-statistic: 53.92 on 1 and 1148 DF,  p-value: 3.955e-13


> mr2 <- lm(murder~numericlaw, data=Guns)
> summary(mr2)
Call:
lm(formula = murder ~ numericlaw, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6434 -3.1434 -0.4136  2.6465 15.0161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4028     0.3291  28.575  < 2e-16 ***
numericlaw   -2.0595     0.2492  -8.264 3.85e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.649 on 1148 degrees of freedom
Multiple R-squared:  0.05615,   Adjusted R-squared:  0.05533
F-statistic: 68.29 on 1 and 1148 DF,  p-value: 3.849e-16
```

```
> mr3 <- lm(robbery~numericlaw, data=Guns)
> summary(mr3)
Call:
lm(formula = robbery ~ numericlaw, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-152.34  -74.57  -25.79   41.89  525.26

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  219.577      9.402  23.355   <2e-16 ***
numericlaw   -60.839      7.120  -8.544   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104.3 on 1148 degrees of freedom
Multiple R-squared:  0.05979,   Adjusted R-squared:  0.05897
F-statistic:    73 on 1 and 1148 DF,  p-value: < 2.2e-16
```

The p-values for all three of the above regressions were below 0.05, so I can reject the null hypothesis and can assume there is a statistically significant relationship between the variables. However, the R-squared values are relatively low at 4.5% (violent), 5.6% (murder), and 5.9% (robbery). These low values mean that the change in the response is not due to a change in the predictors (or at least only a very small percentage). Because of this, I have to accept the null hypothesis that this model is not a good fit. In other words, based on the data in this set, there is not a clear relationship between the change in law and the rate of violence.

Multiple Regression #2

Since I have discovered that shall-carry laws cannot significantly predict a change in violence rate, I am now curious if there are other variables within this dataset that can better predict an increase or decrease in violence. To find this out, I will now add in population predictors.

$$Y = \beta_0 + \beta_1 population + \beta_2 density$$

```
> mr4 <- lm(violent~population + density, data=Guns)
> summary(mr4)
```

```
Call:
lm(formula = violent ~ population + density, data = Guns)

Residuals:
   Min     1Q Median     3Q    Max
-536.1 -157.9  -42.8  127.0  632.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  332.591      8.918  37.293  < 2e-16 ***
population    24.251      1.177  20.602  < 2e-16 ***
density      126.651     27.116   4.671 3.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 206 on 1147 degrees of freedom
Multiple R-squared:  0.3059,    Adjusted R-squared:  0.3047
F-statistic: 252.8 on 2 and 1147 DF,  p-value: < 2.2e-16
```

```
> mr5 <- lm(murder~population + density, data=Guns)
> summary(mr5)
```

```
Call:
lm(formula = murder ~ population + density, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0975 -2.6669 -0.2892  2.2546 13.9728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.80577    0.14798  39.234  < 2e-16 ***
population   0.29714    0.01953  15.214  < 2e-16 ***
density     -2.62196    0.44991  -5.828  7.3e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.418 on 1147 degrees of freedom
Multiple R-squared:  0.1727,    Adjusted R-squared:  0.1712
F-statistic: 119.7 on 2 and 1147 DF,  p-value: < 2.2e-16
```

```
> mr6 <- lm(robbery~population + density, data=Guns)
> summary(mr6)
Call:
lm(formula = robbery ~ population + density, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-313.61  -53.89  -18.52   34.35  385.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.7837     3.3586   19.29   <2e-16 ***
population   12.2284     0.4433   27.59   <2e-16 ***
density     115.8235    10.2115   11.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.57 on 1147 degrees of freedom
Multiple R-squared:   0.48,     Adjusted R-squared:  0.4791
F-statistic: 529.3 on 2 and 1147 DF,  p-value: < 2.2e-16
```

As we can see above, the R-squared values here are 0.31(violent), 0.17 (murder), and 0.48 (robbery). These are significantly higher that those in Multiple Regression #1, with the R-squared value for *mr6* (robbery) being the highest yet. This tells me that this regression model is closer to predicting a relationship between the variables than the previous model, and *population* and *density* are the best predictors of robbery rates so far.

Multiple Regression #3

I am now interested to see the relationship between crime rate and demographics. There exists a plethora of information and misinformation on the relationship between violent crime and gender, socioeconomic status, and race. Using the following regression models, I aim to see if there is a connection within this dataset.

$$Y = \beta_0 + \beta_1 income + \beta_2 male + \beta_3 afam + \beta_4 cauc$$

```
> mr7 <- lm(violent~income+male+afam+cauc, data=Guns)
> summary(mr7)
Call:
lm(formula = violent ~ income + afam + cauc + male + density +
    population, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-557.05 -135.35  -22.17  129.61  576.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 333.843319 242.546216   1.376 0.168964
income        0.005209   0.003767   1.383 0.167043
afam         19.580633   7.426837   2.636 0.008491 **
cauc          1.376096   3.718560   0.370 0.711405
male        -14.576989   4.124089  -3.535 0.000425 ***
density      61.619906  28.896354   2.132 0.033183 *
population   21.822355   1.148260  19.005  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 191.7 on 1143 degrees of freedom
Multiple R-squared:  0.4007,    Adjusted R-squared:  0.3975
F-statistic: 127.4 on 6 and 1143 DF,  p-value: < 2.2e-16
```

```
> mr8 <- lm(murder~income+male+afam+cauc, data=Guns)
> summary(mr8)
Call:
lm(formula = murder ~ income + male + afam + cauc, data = Guns)

Residuals:
     Min       1Q   Median       3Q      Max
-13.0605  -2.0466  -0.2714   1.8985  13.3022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.568e+00  4.023e+00   1.136  0.25635
income      -1.744e-04  5.451e-05  -3.200  0.00141 **
male         2.253e-01  6.924e-02   3.254  0.00117 **
afam         3.994e-01  1.212e-01   3.295  0.00101 **
cauc        -1.518e-02  6.033e-02  -0.252  0.80136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.257 on 1145 degrees of freedom
Multiple R-squared:   0.25,     Adjusted R-squared:  0.2474
F-statistic: 95.42 on 4 and 1145 DF,  p-value: < 2.2e-16
```

```
> mr9 <- lm(robbery~income+male+afam+cauc, data=Guns)
> summary(mr9)

Call:
lm(formula = robbery ~ income + male + afam + cauc, data = Guns)

Residuals:
    Min      1Q  Median      3Q     Max
-249.15  -59.93  -14.68   41.44  529.52

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 463.221965 116.259169   3.984 7.19e-05 ***
income        0.020447   0.001575  12.979  < 2e-16 ***
male          8.158148   2.001212   4.077 4.89e-05 ***
afam        -13.395701   3.503200  -3.824 0.000138 ***
cauc        -10.404894   1.743629  -5.967 3.21e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.13 on 1145 degrees of freedom
Multiple R-squared:  0.2355,    Adjusted R-squared:  0.2329
F-statistic:  88.2 on 4 and 1145 DF,  p-value: < 2.2e-16
```

The above R-square value for *murder* is stronger, but not significantly. The values for *violent* and *robbery* are actually less than those in Multiple Regression #2. This suggests that demographics do not play a significant role in predicting the rate of crime, at least not within this dataset.

After performing the above regressions, I tried many other combinations of variables and found that none produced a higher R-squared value than that of Multiple Regression #2. This is troubling, though, because in order to reject the null hypothesis that my equation is not a good fit, the R-squared value would need to be higher than 0.75, and the highest that I was able to produce was 0.48. I could continue on with the linear regression using equation #2, but the results would not be statistically significant nor relevant.

Summary

Before analyzing my dataset, I was sure that I would see a clear relationship between crime rate and the enactment of shall-carry laws. I felt positive that shall-carry laws would increase crime. However, I found that most of my analyses proved my hypothesis wrong. There were more instances of crime when states did not have shall-carry laws enacted. In fact, when states that previously did not have shall-carry laws put them into effect, crime actually went down.

While the boxplots, heatmaps, and scatterplots suggested that shall-carry laws decrease crime, I did not find a clear relationship between these two variables when I ran the multiple linear regression models. I cannot say for certain, but I believe this non-conclusion is due to a lack of data. As I showed in the barplot, shall-carry laws did not gain popularity until the late 1980s. This means that the majority of the data that was collected represents when shall-carry laws were not in effect. I believe a more representative and fair sample would show different results.

Works Cited

Ayres, I. & Donohue, J. J. (2003). *Shooting down the more guns, less crime hypothesis.* 55 Stanford Law Review 1193 (2003).

Burglary Statistics. (2023, January 31). The Zebra. https://www.thezebra.com/resources/research/burglary-statistics/

Census 2000 Brief: Household Income: 1999. (2021, October 8). United States Census Bureau. https://www.census.gov/library/publications/2005/dec/c2kbr-36.html#:~:text=Median%20household%20income%20in%201999,percent%20inflation%20over%20the%20period).&text=Median%20income%20divides%20households%20into,other%20half%20having%20incomes%20below.

Lott, J. R. (1998). *More guns, less crime: Understanding crime and gun control laws.* University of Chicago Press.

Money Income in 1977 of Households in the United States. (2021, October 28). United States Census Bureau. https://www.census.gov/library/publications/1978/demo/p60-117.html#:~:text=The%20median%20money%20income%20of,was%20eroded%20by%20rising%20prices.

More Guns, Less Crime? R-Project. https://search.r-project.org/CRAN/refmans/AER/html/Guns.html

Section I: Gun Violence in the United States. https://ojjdp.ojp.gov/sites/g/files/xyckuh176/files/pubs/gun_violence/sect01.html#:~:text=The%20drug%20market%20is%20a,the%20crack%20cocaine%20drug%20trade