Violent Crime Analysis

Harper Schwab

10/20/2020

Introduction

According to the National Institute of Justice, " In a violent crime, a victim is harmed by or threatened with violence. Violent crimes include rape and sexual assault, robbery, assault and murder" (DOJ, Violent Crime). The data-set used in this analysis measures on a state-by-state basis, the numbers of violent crimes. The crimes included are murder, rape, robbery, assault, property, burglary, larceny, and automotive crimes, as well as a total amount of violent crime, violent crime per ten-thousand individuals, and the population of each state. Before analyzing the data, I believe states with higher populations will have the highest amount of violent crimes because crime should be somewhat proportional to the size of the population. However, when looking at more of a per-capita basis those states may not be in the lead. Two states are unable to be shown in the data maps, Hawaii and Alaska. Alaska has the highest per capita violent crime rate at 541.1 per 100,000 people (Stebbins, Dangerous states: Which states have the highest rates of violent crime and most murders? 2020). One thing to think about for a new iteration of this analysis is to have those two states appear on the data maps. It does not seem that poverty rate is a factor in the violent crime level when comparing a chart form Wikipedia and the data provided, however for states like California, however, it is unknown if higher percentages of those below the poverty line commit violent crimes in certain states, which would be an interesting topic to research further. Better, per capita measurements can help show where the populations of states are truly 'more violent', or have worse policing. This project, mainly confirmed my suspicions about violent crime, however, I was surprised that they went this much in line with the population size correlating positively to the amount of crime.

Data Cleaning

The data, while clean, was unusable off the start. The columns were classified as non-numeric which made it unable to compute. This was because the numbers had comma separators for the larger numbers. Below is an example of how the numbers looked before the change.

```
sample_n(crime, 3)
```

```
##        State Population Annual Per10K Murder  Rape Robbery Assault Property
## 1     Maine  1,329,328  1,729     13     21   485     304     890   26,421
## 2      Iowa  3,123,899  8,936     29     60 1,128   1,045   6,264   65,056
## 3 Nebraska  1,896,190  5,212     27     54   862   1,043   3,316   47,479
##   Burglary Larceny   Auto
## 1    5,030  20,592    799
## 2   14,428  46,477  4,151
## 3    7,950  35,074  4,455
```

The following is what I did to change the numbers into a workable format.

```
crime$Population <- as.numeric(gsub(',','',crime$Population))
crime$Annual <- as.numeric(gsub(',','',crime$Annual))
crime$Murder <- as.numeric(gsub(',','',crime$Murder))
crime$Rape <- as.numeric(gsub(',','',crime$Rape))
crime$Robbery <- as.numeric(gsub(',','',crime$Robbery))
crime$Assault <- as.numeric(gsub(',','',crime$Assault))
crime$Property <- as.numeric(gsub(',','',crime$Property))
crime$Burglary <- as.numeric(gsub(',','',crime$Burglary))
crime$Larceny <- as.numeric(gsub(',','',crime$Larceny))
crime$Auto <- as.numeric(gsub(',','',crime$Auto))
```

Now take a look at that same data sample. See the difference? It's much better.

```
sample_n(crime, 3)
```

```
##          State Population Annual Per10K Murder Rape Robbery Assault Property
## 1         Utah    2995919   7071     24     67 1454    1312    3513    84711
## 2     Missouri    6083672  30261     50    403 2379    5592   18482   176237
## 3  Connecticut    3590886   7845     22     86  782    3159    4495    69070
##    Burglary Larceny  Auto
## 1     11518   65895  7298
## 2     35258  124622 16357
## 3     11955   51005  6110
```

I then looked at the demographiocs for the data using the str(), dim(), and summary() commands. I was able to see that all of the numeric data, was classified as numeric and usable, while the state column was non-numeric. The dim() comman showed me how many variables I had to work with and the Summary() command showed basic statistical information, which I will show below.

```
summary(crime)
```

```
##      State             Population            Annual            Per10K
##  Length:50          Min.   :  586107    Min.   :    739    Min.   :12.00
##  Class :character   1st Qu.: 1857144    1st Qu.:   5602    1st Qu.:26.00
##  Mode  :character   Median : 4547908    Median :  15962    Median :35.00
##                     Mean   : 6417926    Mean   :  24364    Mean   :36.30
##                     3rd Qu.: 7084780    3rd Qu.:  27875    3rd Qu.:42.75
##                     Max.   :39144818    Max.   : 166883    Max.   :73.00
##      Murder           Rape           Robbery          Assault
##  Min.   :  10.0   Min.   :  110.0   Min.   :   53   Min.   :  432
##  1st Qu.:  55.5   1st Qu.:  773.8   1st Qu.: 1060   1st Qu.: 3365
##  Median : 165.0   Median : 1547.0   Median : 3248   Median :10037
##  Mean   : 282.9   Mean   : 2323.5   Mean   : 6401   Mean   :14682
##  3rd Qu.: 369.5   3rd Qu.: 2503.2   3rd Qu.: 7173   3rd Qu.:17500
##  Max.   :1699.0   Max.   :11527.0   Max.   :48680   Max.   :91803
##     Property         Burglary          Larceny            Auto
##  Min.   :  6729   Min.   :  1689   Min.   :  7273   Min.   :   244
##  1st Qu.: 39060   1st Qu.:  8206   1st Qu.: 27994   1st Qu.:  3947
##  Median :115144   Median : 23912   Median : 81709   Median :  9720
##  Mean   :162501   Mean   : 36619   Mean   :115123   Mean   : 15472
##  3rd Qu.:194395   3rd Qu.: 44029   3rd Qu.:135388   3rd Qu.: 13803
##  Max.   :947192   Max.   :202670   Max.   :592670   Max.   :151852
```

ggplot2 Graphs

I wanted to start out using some of possible commands with ggplot2. I decided to create a scatter-plot that compared the murder with the rape ammount for each state. Size measured the population for each state, while the color lets us know which state it is.

```
RapeMurder <- ggplot(crime, aes(x=Murder, y=Rape))+ geom_point(aes(col=State, size=Population)) +
    geom_smooth(method='loess',se=F)+
    labs(title="Murder vs Rape", y='Rape', x = 'Murder')
RapeMurder
```

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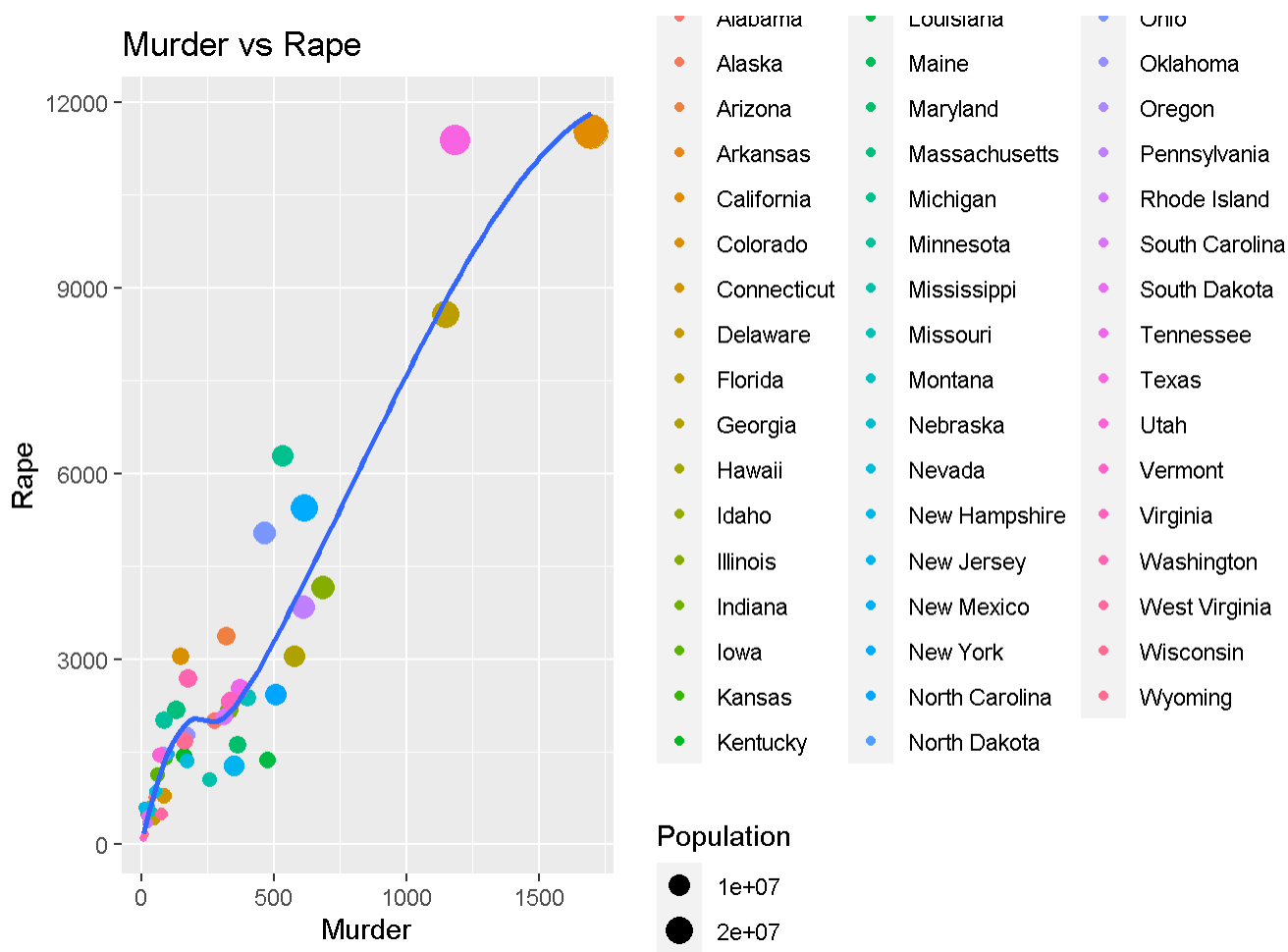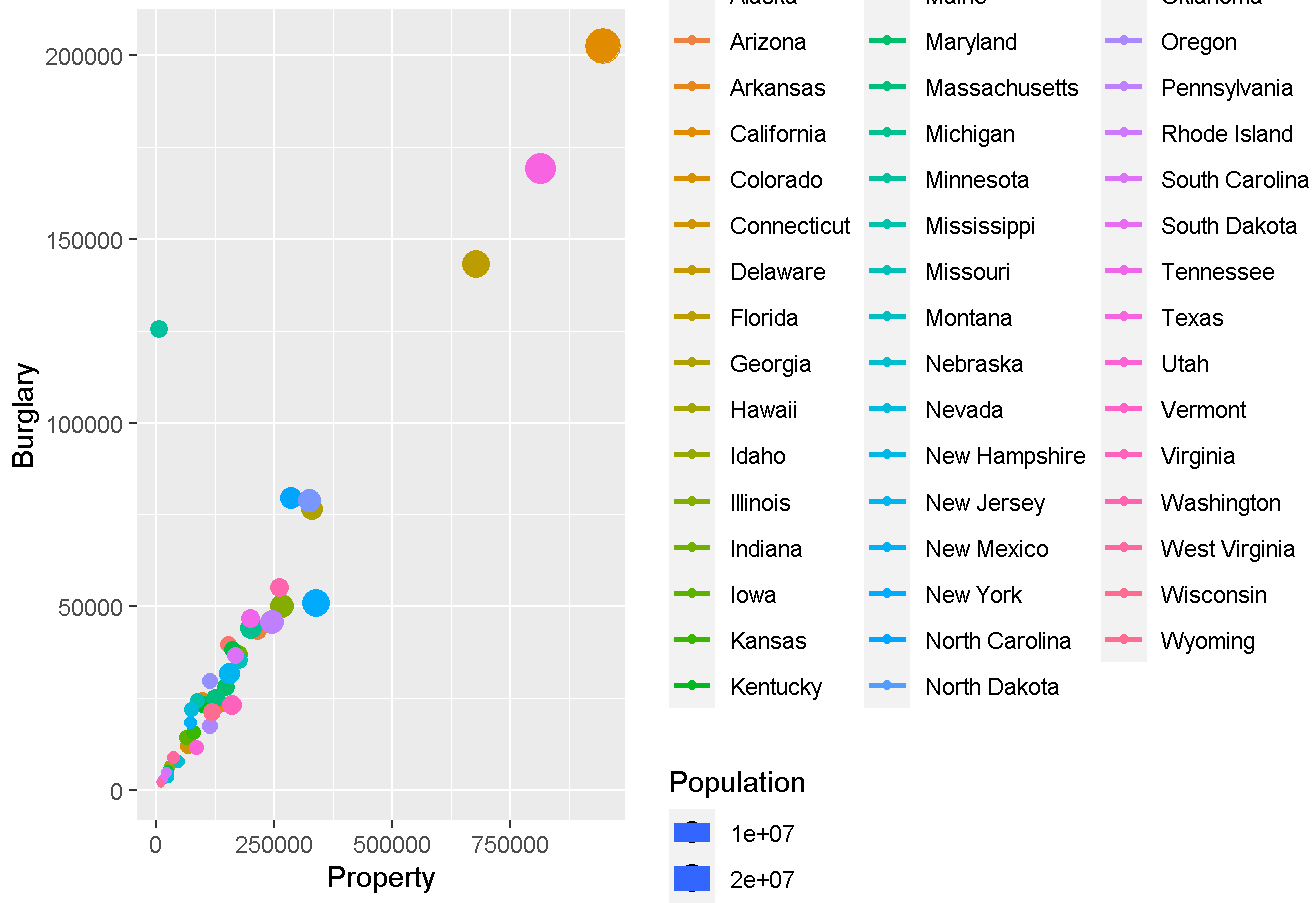

There seems to be a positive correlation between these two violent crimes. Where, generally, the higher the murder Rate, the higher the frequency of rape. The interesting part of the plot is between 0-500 murders and 0-3000 rapes the slope decreases for a short range where there are higher amount of murders but less rape. Another thing to note, is the massive difference between the scales of the x qand y axis. Murder ranges from 0-2000 while rape ranges from 0-12000. This goes to show how much more rape is committed than murder. This shocking difference is astounding to me. It helps illustrate just how common the violent crime of rape is in the United states, even when compared to murder.

The next graph I created was a jitter plot of property and burglary crime per state. Much like the last one the size was affected by population, while the color showed what state each data point was represented.

```
jp <- ggplot(crime, aes(Property,Burglary, color = State, size = Population))
jp <- jp+ geom_point()+geom_smooth(method='lm', se=F)+
    labs(title="Jitterplot of Property and Burglary", y='Burglary', x='Property')
jp
```

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


Jitterplot of Property and Burglary

Much like the relationship between rape and murder, there is a positive correlation between property crime and burglaries. The Relationship between the two are more linear than the rape/murder graph. With this graph it is also easy to see the affect population has on these violent crimes. THe more crime, the higher the crime rate. This makes sense, because population should have a somewhat proportional relationship between number of crimes.

Data Maps

started to visualize the crime data on a map of the continental United States. First, I had to create a base map of the U.S. without any data measurements. I took map data of of the position of the different states from map_data and created a dataset. I then used a ggplot map where the longitude and latitude data-points were used to create a picture of each state. I decided to use the albers projection because it looks the nicest for what I am doing, and used the theme_map() function to get rid of the background lines. I also Used the merge command so I measure data from the crime dataset in my maps.

```
## Loading required package: grid
```
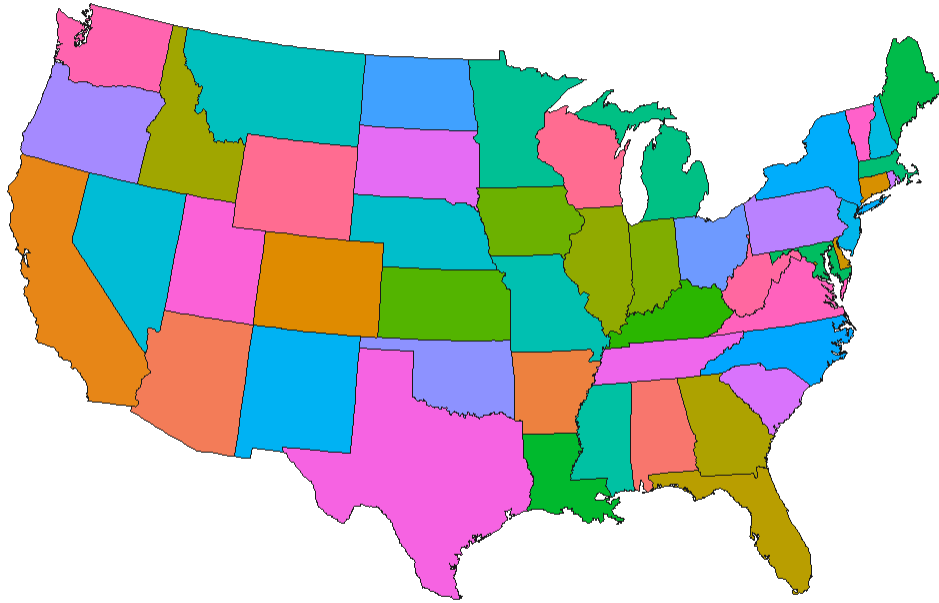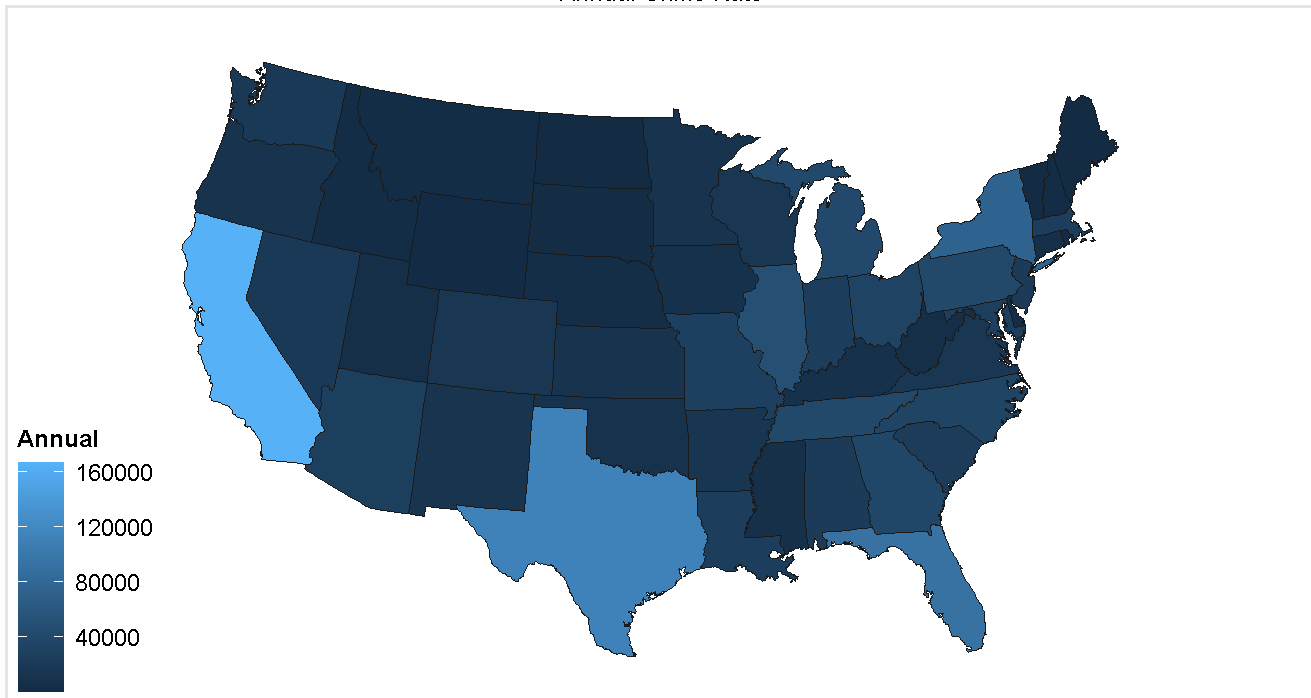
```
## Joining, by = "region"
```

Thee first map I created was a Map of total annual crime.

```
#Map Total Annual Crime
#
totalCrime <- ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Annual))+
  geom_polygon(color='gray10',size=0.1)+
  coord_map(projection='albers', lat0=25,lat1=50)+
  labs(title='Annual Crime Rate', full=NULL)  +theme_map()
totalCrime
```

California has the highest annual crime rate followed by Texas, Florida and New York. These states have high populations, so it makes sense that the crime number is higher for those states. The mid-west, for instance, has very low crime amount in comparison to the states mentioned earlier. To account for population differences I will map out the crime per 10 thopusand individuals column which should show a better representation of crime. That plot is embedded below.

```
crime10K <- ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Per10K))+
    geom_polygon(color='gray10',size=0.1)+
    coord_map(projection='albers', lat0=25,lat1=50)+
    labs(title='Annual Crime Per 10K Rate', full=NULL)  +theme_map()
crime10K
```

Annual Crime Per 10K Rate
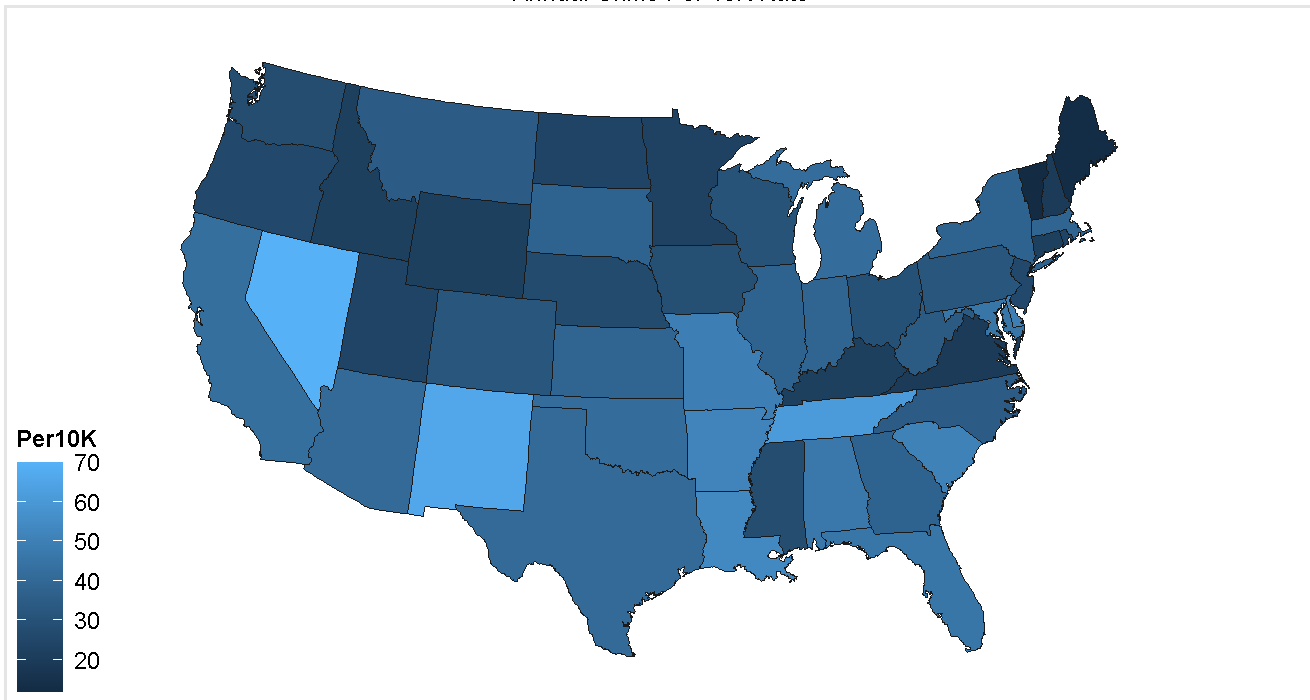
This map is better because it accounts for population differences which can skew numbers heavily in states such as New York, Florida, and California. The lighter colored states have a higher crime rate per ten-thousand residents. With this map it become apparent that states such as Nevada, Arizona, and Tennessee have the highest rate pof crime when we account for population differences.

Next I did a series of maps which analyzed different types of crime per state. The first one I did was Murder.

```
#Next map will be a Murder Map
#
murder <- ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Murder))+
  geom_polygon(color='gray10',size=0.1)+
  coord_map(projection='albers', lat0=25,lat1=50)+
  labs(title='Annual Murder Rate by State', full=NULL) +theme_map()
murder
```
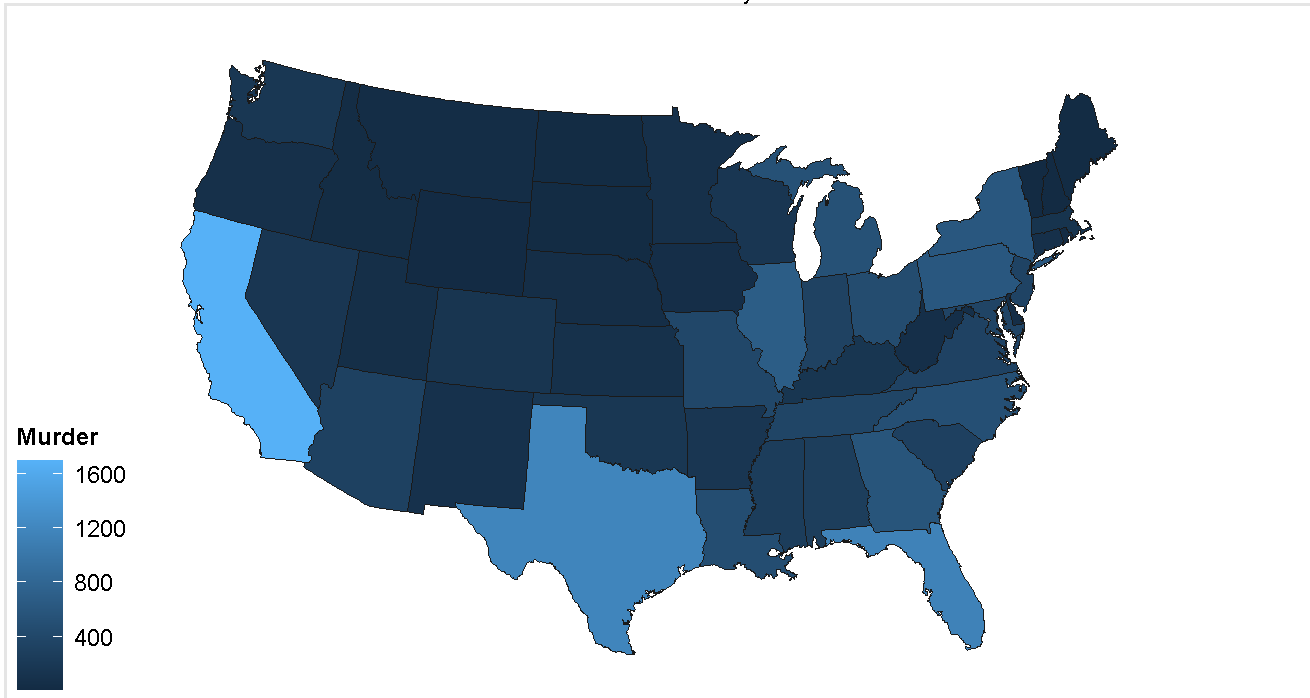
Again California has the highest rate the crime out of all of the other states. Texas and Florida follow, which match up with the total crime map from before. States such as Illinois, North Carolina, New York, Georgia, and Pennsylvania are in a group of higher murder rates.

The next map is a map of automotive violent crimes.

```
#Next map will be a automotive a Map
#
auto <- ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Auto))+
  geom_polygon(color='gray10',size=0.1)+
  coord_map(projection='albers', lat0=25,lat1=50)+
  labs(title='Annual Automotive Crime by State', full=NULL) +theme_map()
auto
```
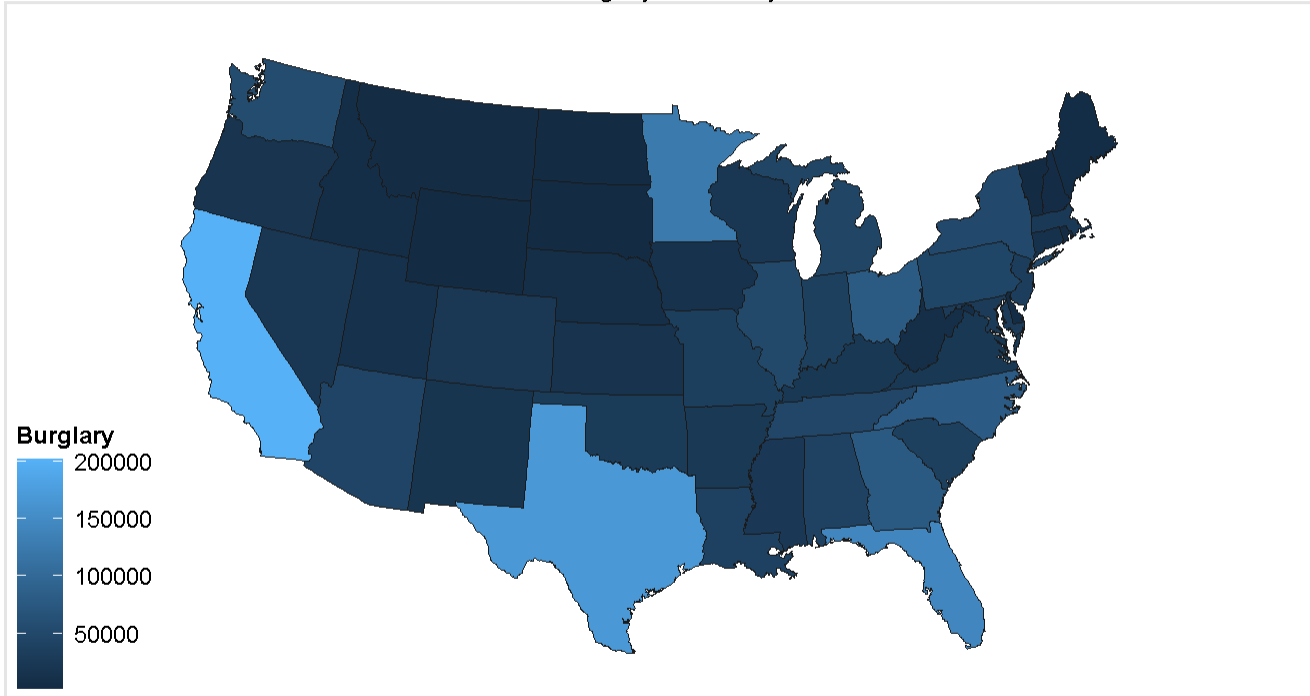
Annual Automotive Crime by State
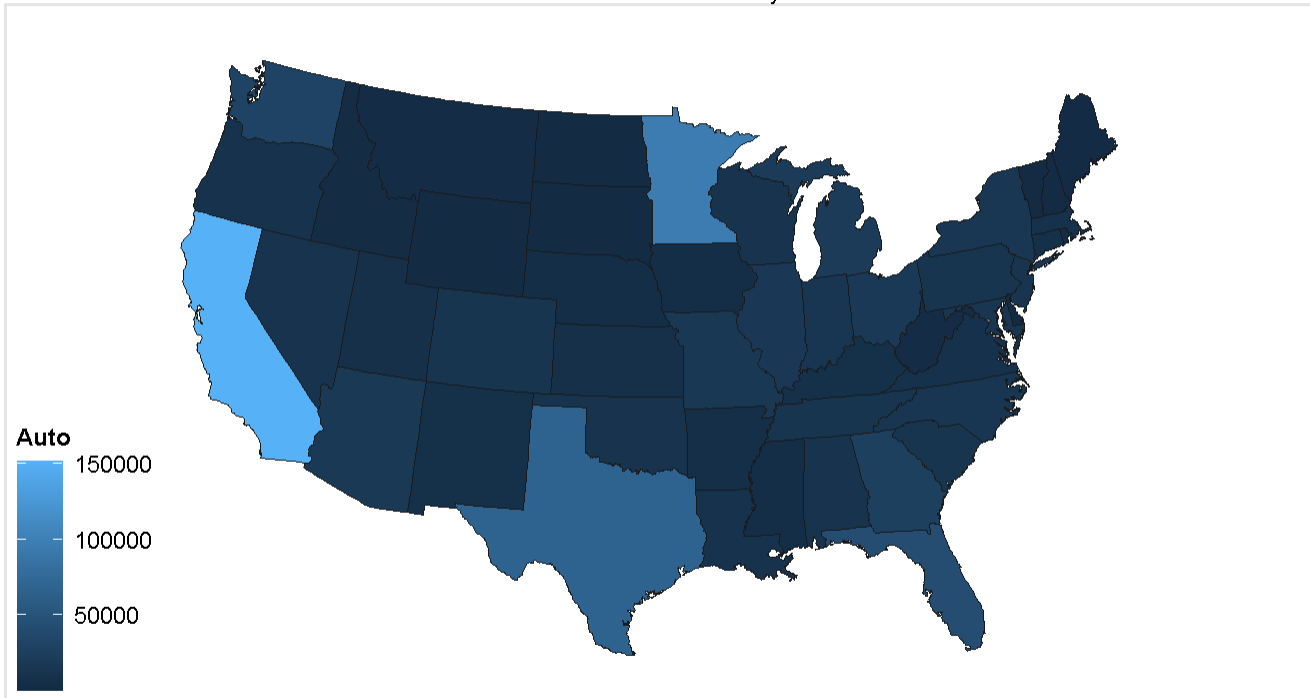
The most

surprising piece of data here, is Minnesota which has the second highest rate of all the states. I would think maybe the temperatures could affect this but because none of the other northern states are that high, that theory is improbable.

The last Map I did was the burglary rate which follows:

```
burglar <- ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Burglary))+
    geom_polygon(color='gray10',size=0.1)+
    coord_map(projection='albers', lat0=25,lat1=50)+
    labs(title='Annual Burglary Number by State', full=NULL) +theme_map()
burglar
```

California again has the highest rate, followed by Texas, Florida, and Minnesota. One thing that is interesting, is that Washington state is higher than the surrounding states, excluding California.

```
burglar <- ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Burglary))+
  geom_polygon(color='gray10',size=0.1)+
  coord_map(projection='albers', lat0=25,lat1=50)+
  labs(title='Annual Burglary Number by State', full=NULL) +theme_map()
burglar
```

Annual Burglary Number by State

Burglary
200000
150000
100000
50000
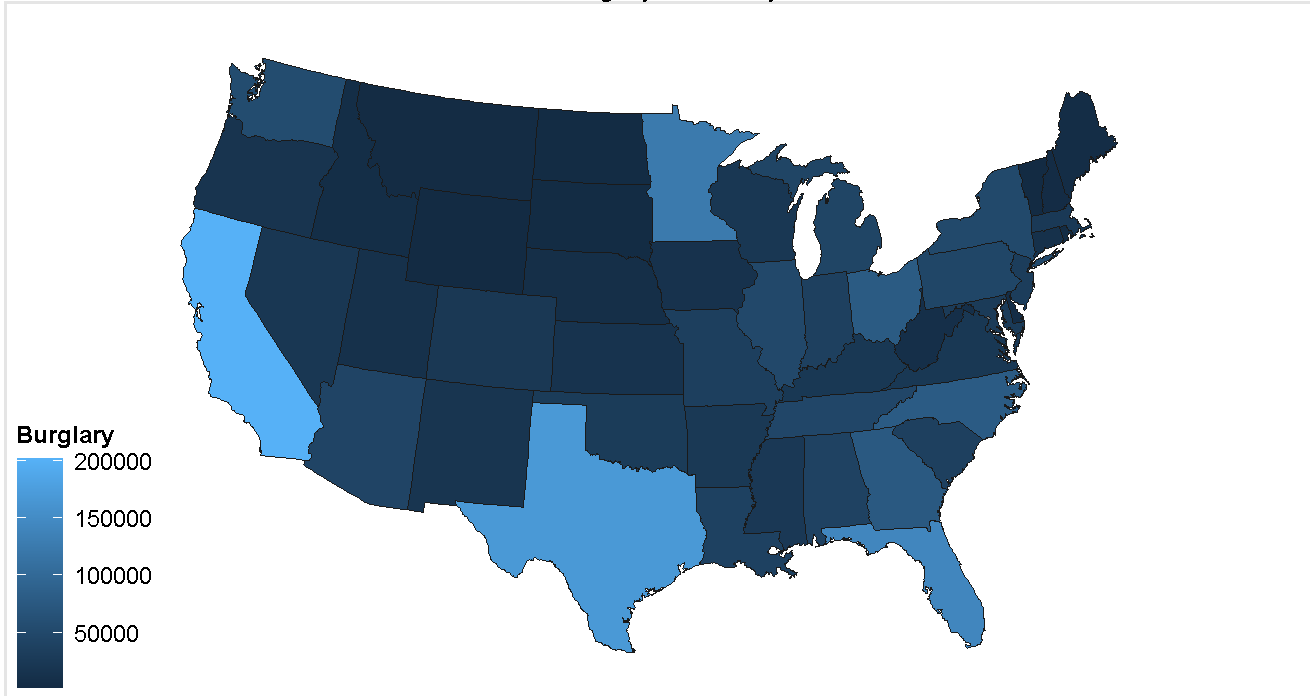
The finalmap I decided to do all had to do with material crimes: larcenyu, burglary, and robbery.

```
meterial <-ggplot(data=usCrime, mapping=aes(x=long, y=lat, group =group, fill=Larceny+Burglary+Rob
bery))+
   geom_polygon(color='gray10',size=0.1)+
   coord_map(projection='albers', lat0=25,lat1=50)+
   labs(title='Annual Burglary Number by State', full=NULL) +theme_map()
meterial
```
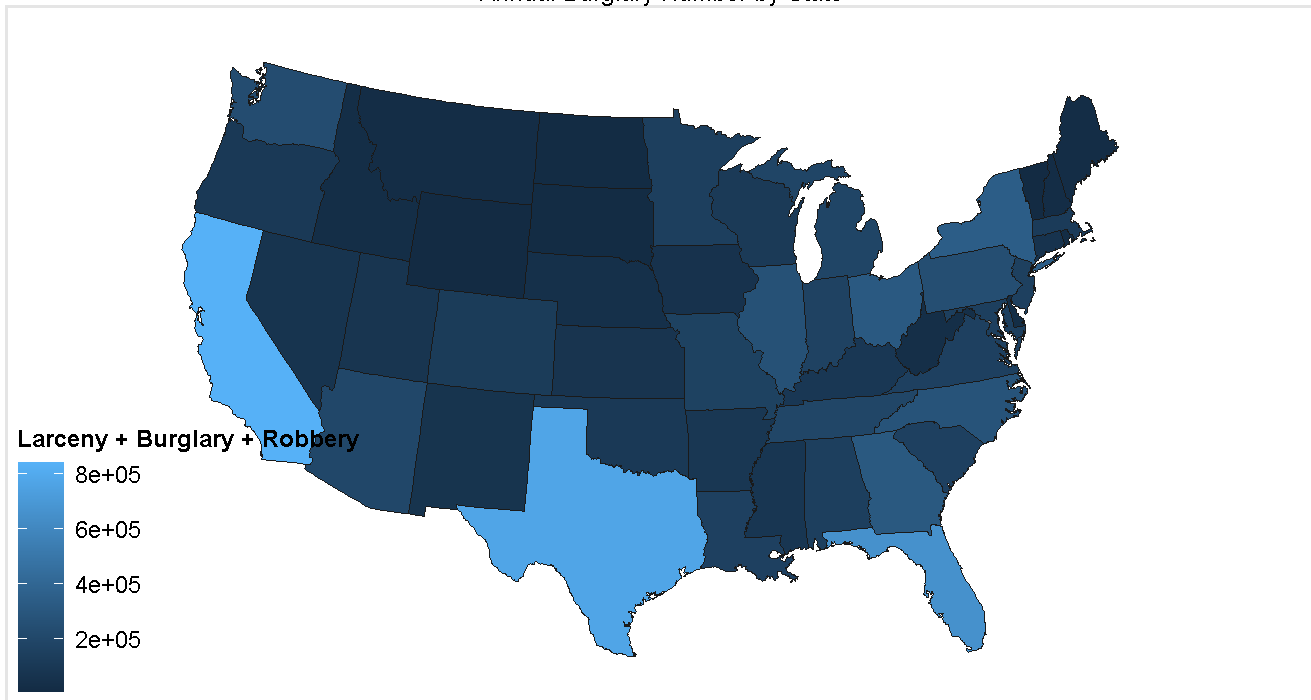
Annual Burglary Number by State

Going straight in

accordance with the population size California Texas and Florida have the highest rates. the north-eastern states of New Your Pennsylvania and Ohio all are in the relativly high range as well as North Carolina and Oregon.

Conclusion

California, by far, has the most crimes. Texas seems to have the second most amount of crimes. And while Minnesota doesn't have one of the highest overall, it has higher levels compared to many of the other states for the individual crime rates. High population is a good indicator for total higher crime rates. But as seen in the per 10K map, states other than the highest total crime rates, had the highest per 10K numbers. although, those with high crime rates were on the lighter side of the per 10K map.Overall, it seems like many violent crime rates are heavily correlated with each other, as seen by the strong positive correlations from the scatter charts.

Citations

U. (n.d.). Violent Crime. Retrieved October 21, 2020, from https://nij.ojp.gov/topics/crimes/violent-crime (https://nij.ojp.gov/topics/crimes/violent-crime)

Stebbins, S. (2020, January 13). Dangerous states: Which states have the highest rates of violent crime and most murders? Retrieved October 21, 2020, from https://www.usatoday.com/story/money/2020/01/13/most-dangerous-states-in-america-violent-crime-murder-rate/40968963/ (https://www.usatoday.com/story/money/2020/01/13/most-dangerous-states-in-america-violent-crime-murder-rate/40968963/)