Violents Crimes all over the United States

Samuel Dummer

10/5/2021

Introduction

The United States seems to have much higher crime rates than other countries, but this is only because we have a large population. In general, our crime level is below average in proportion to our population. In this specific data set, we are covering violent crime data throughout the various states. This includes violent crimes such as murder, rape, robbery, assault, property, burglary, larceny, and auto crimes. Within this data, there are 12 variables that were observed. First, there was "state" which was just a list of all of the separate states. Next, there is population which provides the population of each state which helps us to calculate the proportion of the crimes since without including the proportion, the states with higher population would have more crime. Next, there is the annual number of crimes which helps to provide the amount of crimes committed every year. There is also "Per10K" which gives us the number of crimes per 10 thousand people in each state. Next there are the number of each violent crimes listed above. Additionally, I added a new variable called "total" which provided the total number of crimes for each state. This variable helps to find the proportions of each crime in each state, such as, the number of murders in proportion to the number of crimes. This helps to show the most frequent crime for each state. Further, it provides and outline of where the majority of the crime in the United States happens.

In this script, we will be creating different graphical representations of this data to help describe and analyze what the data represents. This will be done mainly through maps of the United States, but also through other graphs such as bar plots, histograms, or scatter plots.

Cleaning Enviroment, Setting up Directory, and Loading In Libraries (and Functions)

Before we are able to load in the data, we must set everything up. We do so by clearing the environment of anything we had been previously doing, then we set the directory. Lastly we load in any necessary libraries, so in this case we load in "tidyverse", "maps", "socviz", and "gridExtra". Additionally we load in the functions "myfunctions.R" and "theme_map.R" to provide us with any other function that might be necessary when graphing.

```
# directory and environment
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
# libraries and functions
source("myfunctions.R")
source("theme_map.R")
library(tidyverse)
library(maps)
library(socviz)
library(gridExtra)
```

Loading in the Data Set

Now that everything is set up, we are now able to read in the data on violent crimes in the US. This is the data that we will be analyzing and mapping later on. We read the data in as the variable "crime".

```
crime <- read.csv("newcrimes.csv", header=T)
head(crime)
```

```
##        State Population Annual Per10K Murder  Rape Robbery Assault Property
## 1    Alabama    4858979  22952     47    276  2005    4701   13745   154094
## 2     Alaska     738432   5392     73     41   771     629    3243    20334
## 3    Arizona    6828065  28012     41    319  3378    6249   16970   215240
## 4   Arkansas    2978204  15526     52    165  1763    2050   10265    99018
## 5 California   39144818 166883     43   1699 11527   48680   91803   947192
## 6   Colorado    5456574  17515     32    151  3039    3039   10325   135510
##   Burglary Larceny   Auto
## 1    39715  104238  10141
## 2     3150   15445   1739
## 3    43562  154091  17587
## 4    24790   68627   5601
## 5   202670  592670 151852
## 6    23472   99464  12574
```

Looking over Structure and Summary of the Data

Once the data is read, it is always important that the first thing you do is you observe the structure and summary of the data. This can be done by using the glimpse(), str(), and summary() command. This helps to give us a good overview of the data and find any missing data or changes needed to be done.

```
glimpse(crime)
```

```
## Rows: 50
## Columns: 12
## $ State      <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California", "~
## $ Population <int> 4858979, 738432, 6828065, 2978204, 39144818, 5456574, 35908~
## $ Annual     <int> 22952, 5392, 28012, 15526, 166883, 17515, 7845, 4720, 93626~
## $ Per10K     <int> 47, 73, 41, 52, 43, 32, 22, 50, 46, 38, 30, 22, 38, 39, 29,~
## $ Murder     <int> 276, 41, 319, 165, 1699, 151, 86, 54, 1149, 580, 26, 32, 68~
## $ Rape       <int> 2005, 771, 3378, 1763, 11527, 3039, 782, 386, 8563, 3048, 4~
## $ Robbery    <int> 4701, 629, 6249, 2050, 48680, 3039, 3159, 1269, 24914, 1241~
## $ Assault    <int> 13745, 3243, 16970, 10265, 91803, 10325, 4495, 2867, 72895,~
## $ Property   <int> 154094, 20334, 215240, 99018, 947192, 135510, 69070, 27900,~
## $ Burglary   <int> 39715, 3150, 43562, 24790, 202670, 23472, 11955, 5768, 1432~
## $ Larceny    <int> 104238, 15445, 154091, 68627, 592670, 99464, 51005, 20865, ~
## $ Auto       <int> 10141, 1739, 17587, 5601, 151852, 12574, 6110, 1267, 42579,~
```

```
str(crime)
```

```
## 'data.frame':    50 obs. of  12 variables:
##  $ State     : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ Population: int  4858979 738432 6828065 2978204 39144818 5456574 3590886 945934 20271272 10214860
```

```
##  $ Annual    : int  22952 5392 28012 15526 166883 17515 7845 4720 93626 38643 ...
##  $ Per10K    : int  47 73 41 52 43 32 22 50 46 38 ...
##  $ Murder    : int  276 41 319 165 1699 151 86 54 1149 580 ...
##  $ Rape      : int  2005 771 3378 1763 11527 3039 782 386 8563 3048 ...
##  $ Robbery   : int  4701 629 6249 2050 48680 3039 3159 1269 24914 12417 ...
##  $ Assault   : int  13745 3243 16970 10265 91803 10325 4495 2867 72895 22052 ...
##  $ Property  : int  154094 20334 215240 99018 947192 135510 69070 27900 679446 331316 ...
##  $ Burglary  : int  39715 3150 43562 24790 202670 23472 11955 5768 143220 76428 ...
##  $ Larceny   : int  104238 15445 154091 68627 592670 99464 51005 20865 493647 228034 ...
##  $ Auto      : int  10141 1739 17587 5601 151852 12574 6110 1267 42579 26854 ...
```

```
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
```

Fixing Column Names

When reading in the data we realized that all the variables were capitalized, so to make is more comprehensive, we used the "str_to_lower" command to make all the variable names lower case.

```
names(crime) <- str_to_lower(names(crime))
names(crime)
```

```
##  [1] "state"      "population" "annual"     "per10k"     "murder"
##  [6] "rape"       "robbery"    "assault"    "property"   "burglary"
## [11] "larceny"    "auto"
```

Mutating the Data

The first transformation to perform was to create a new variable that was the total of all the crimes committed. This will help us is observing which state have the highest number of crime. Additionally, we can use it to see which state has the highest specific crime in proportion to the total number of crimes.

```
crime <- mutate(crime, total = murder + rape + robbery + assault + property + burglary + larceny + auto)
head(crime)
```

```
##        state population annual per10k murder  rape robbery assault property
## 1    Alabama    4858979  22952     47    276  2005    4701   13745   154094
## 2     Alaska     738432   5392     73     41   771     629    3243    20334
## 3    Arizona    6828065  28012     41    319  3378    6249   16970   215240
## 4   Arkansas    2978204  15526     52    165  1763    2050   10265    99018
## 5 California   39144818 166883     43   1699 11527   48680   91803   947192
## 6   Colorado    5456574  17515     32    151  3039    3039   10325   135510
##   burglary larceny   auto   total
## 1    39715  104238  10141  328915
## 2     3150   15445   1739   45352
## 3    43562  154091  17587  457396
## 4    24790   68627   5601  212279
## 5   202670  592670 151852 2048093
## 6    23472   99464  12574  287574
```

Reading in Map Data

Before we are able to create any maps, we must do a few things. First we must read in the map data itself which provides the latitudes and longitudes for the mapping program. Then we must create a new variable "region" that allows us to merge the two data sets together. Lastly, we are able to use the "left_join" command to merge the data sets.

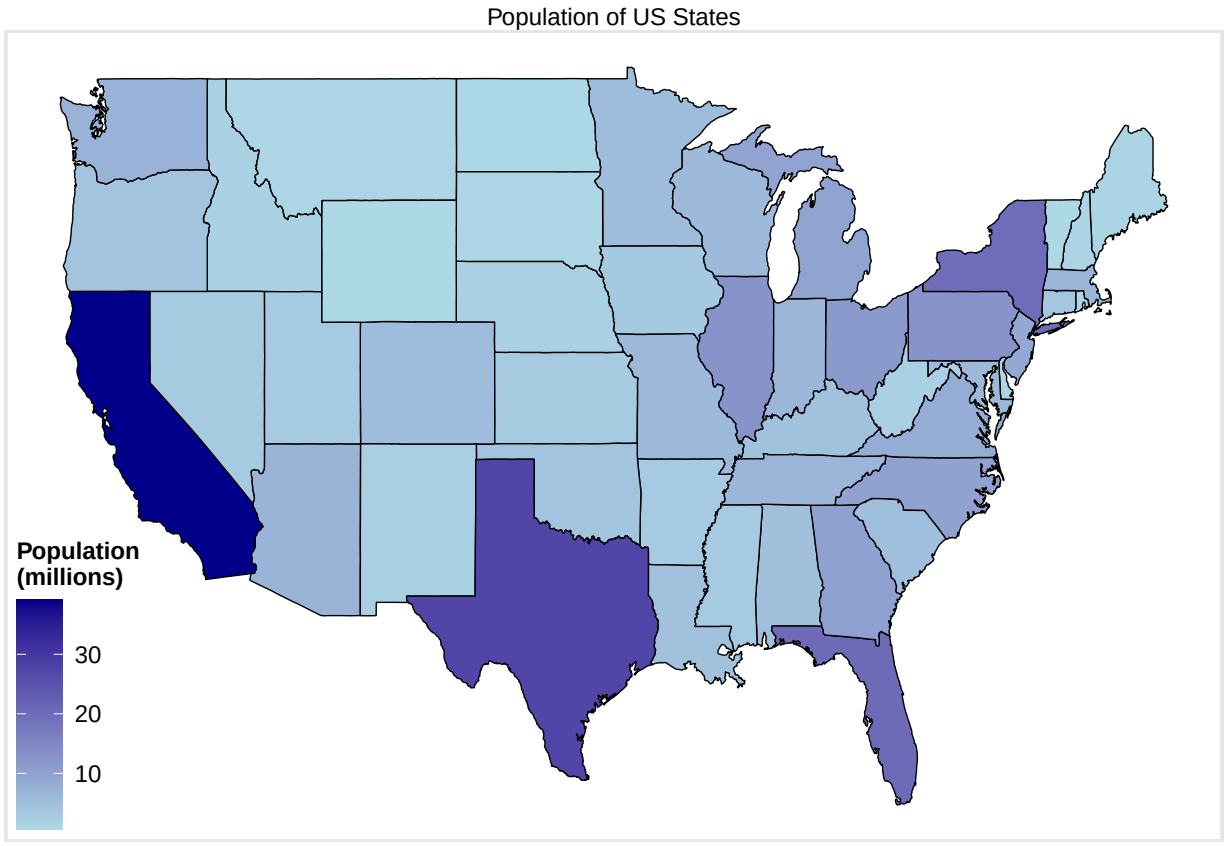
```
us_states <- map_data("state")
crime$region <- tolower(crime$state)
us_states_crime <- left_join(us_states, crime)
head(us_states_crime)
```

```
##        long      lat group order  region subregion   state population annual
## 1 -87.46201 30.38968     1     1 alabama      <NA> Alabama    4858979  22952
## 2 -87.48493 30.37249     1     2 alabama      <NA> Alabama    4858979  22952
## 3 -87.52503 30.37249     1     3 alabama      <NA> Alabama    4858979  22952
## 4 -87.53076 30.33239     1     4 alabama      <NA> Alabama    4858979  22952
## 5 -87.57087 30.32665     1     5 alabama      <NA> Alabama    4858979  22952
## 6 -87.58806 30.32665     1     6 alabama      <NA> Alabama    4858979  22952
##   per10k murder rape robbery assault property burglary larceny  auto  total
## 1     47    276 2005    4701   13745   154094    39715  104238 10141 328915
## 2     47    276 2005    4701   13745   154094    39715  104238 10141 328915
## 3     47    276 2005    4701   13745   154094    39715  104238 10141 328915
## 4     47    276 2005    4701   13745   154094    39715  104238 10141 328915
## 5     47    276 2005    4701   13745   154094    39715  104238 10141 328915
## 6     47    276 2005    4701   13745   154094    39715  104238 10141 328915
```

Creating the Maps and Plots

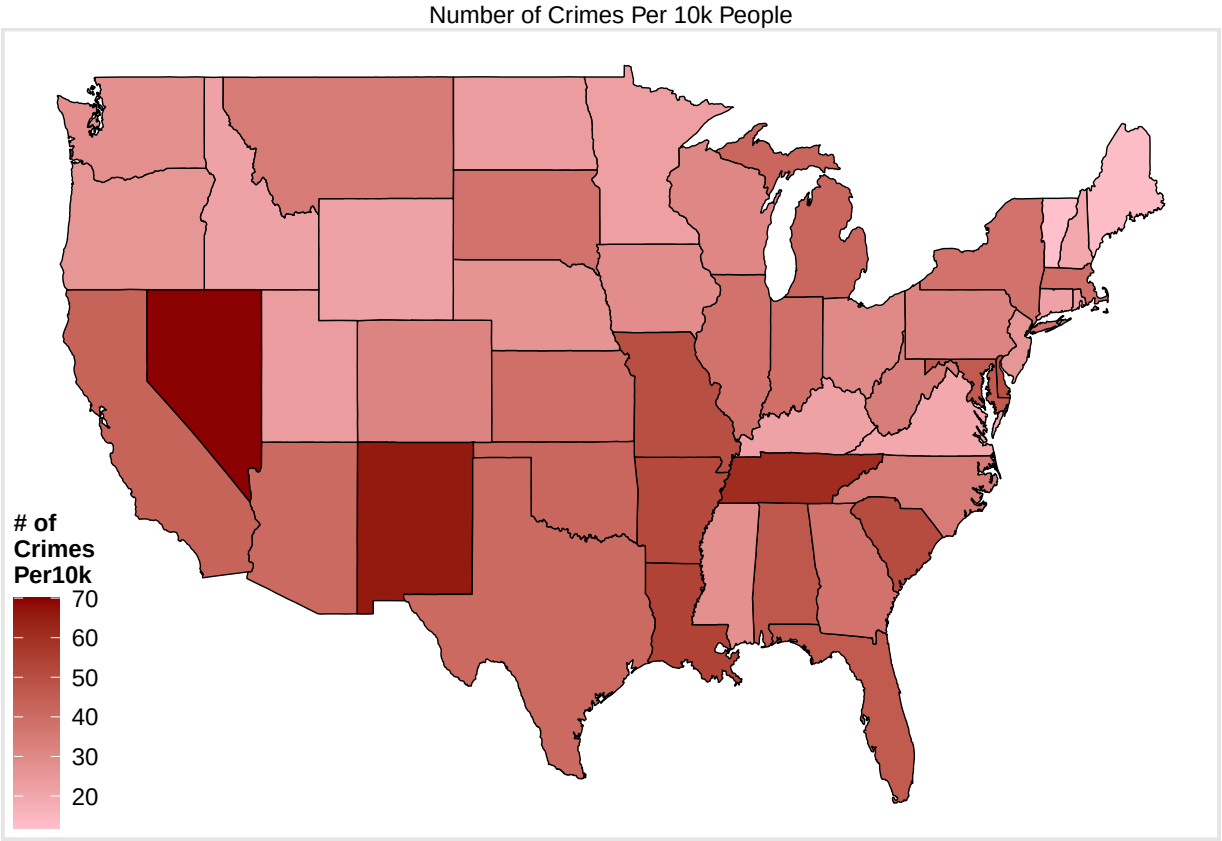
First Plot This is a plot of the overall population. This is helpful since, in later graph, there will be some skewedness to the crime in some states since they have higher populations. This shows that, currently, California has the highest population of the states meaning that it it will most likely have the highest count when it comes to violent crimes even if the crime levels there aren't proportionally high.

```
usPopulation <-
  ggplot(data = us_states_crime,
         mapping = aes(x = long, y = lat, group = group, fill = population/1000000)) +
  geom_polygon(color="black", size=0.25) +
  theme_map() +
  labs(title = "Population of US States", fill = "Population \n(millions)") +
  scale_fill_gradient(low="light blue", high="dark blue")
usPopulation
```



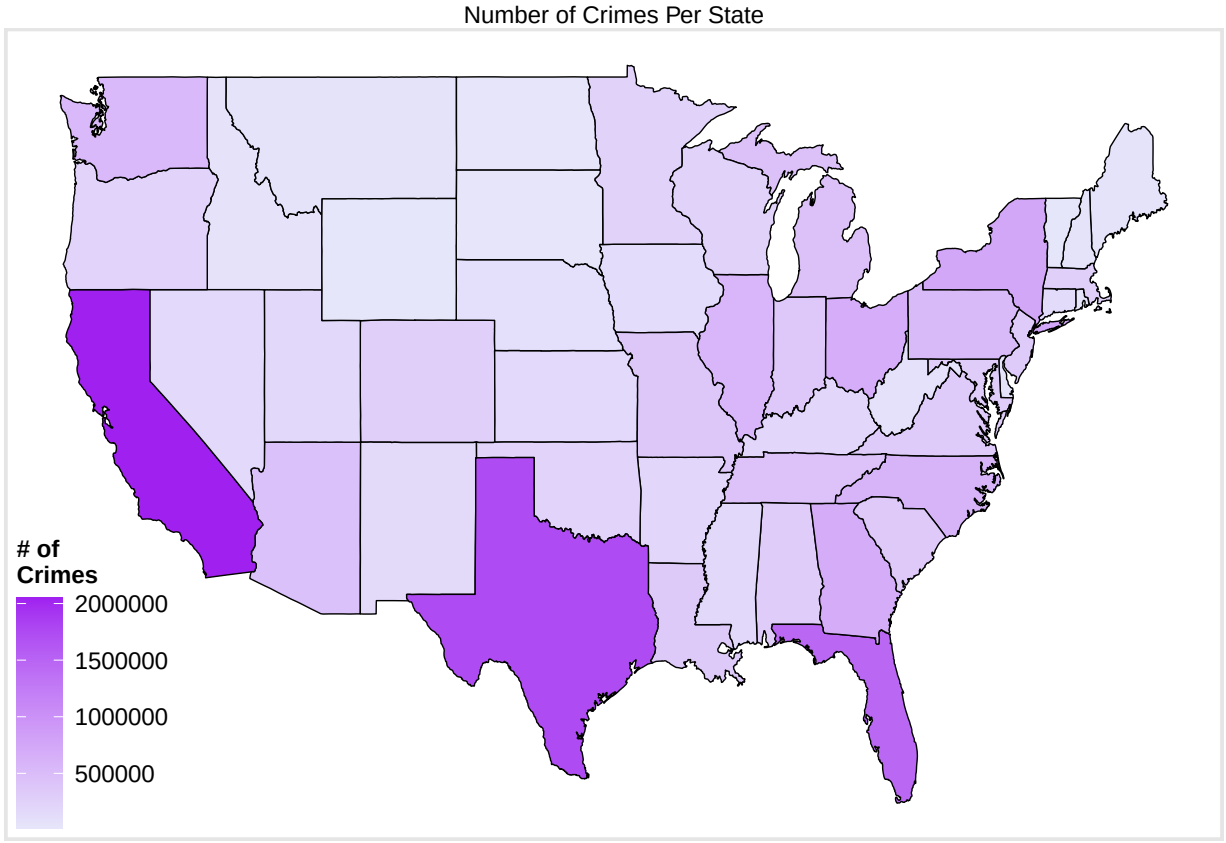
Second Plot The next plot we decided to make was a map of the US that was shaded depending on the number of crimes per 10k people. This was helpful in finding the states with the proportionally highest number of crimes. This is able to be done since the data in crimes per 10k people is proportional to the population. Here, we are able to observe that Nevada has the highest crime rate of all the states per 10k people.

```
usPer10K <-
  ggplot(data = us_states_crime,
         mapping = aes(x = long, y = lat, group = group, fill = per10k)) +
  geom_polygon(color="black", size=0.25) +
  theme_map() +
  labs(title = "Number of Crimes Per 10k People", fill = "# of \nCrimes \nPer10k") +
  scale_fill_gradient(low="pink", high="dark red")
usPer10K
```



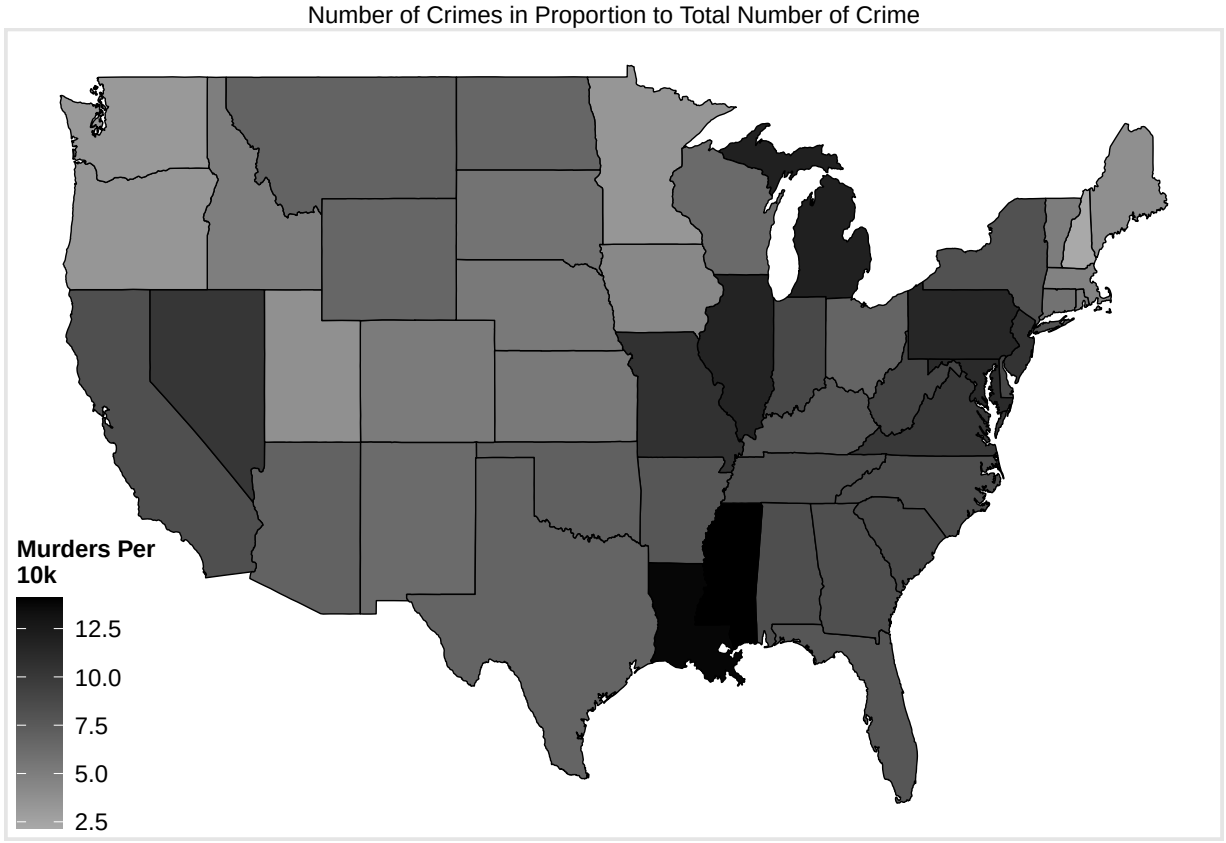
Third Plot Next we decided that it would be important to graph the total crimes in each state. This would help show where the most crime was but not the highest crime rate. This means that California has the most, not because there is the most crime, but because it has the highest population, therefore, the most crime will happen there.

```
usTotal <-
  ggplot(data = us_states_crime,
         mapping = aes(x = long, y = lat, group = group, fill = total)) +
  geom_polygon(color="black", size=0.25) +
  theme_map() +
  labs(title = "Number of Crimes Per State", fill = "# of \nCrimes") +
  scale_fill_gradient(low="lavender", high="purple")
usTotal
```



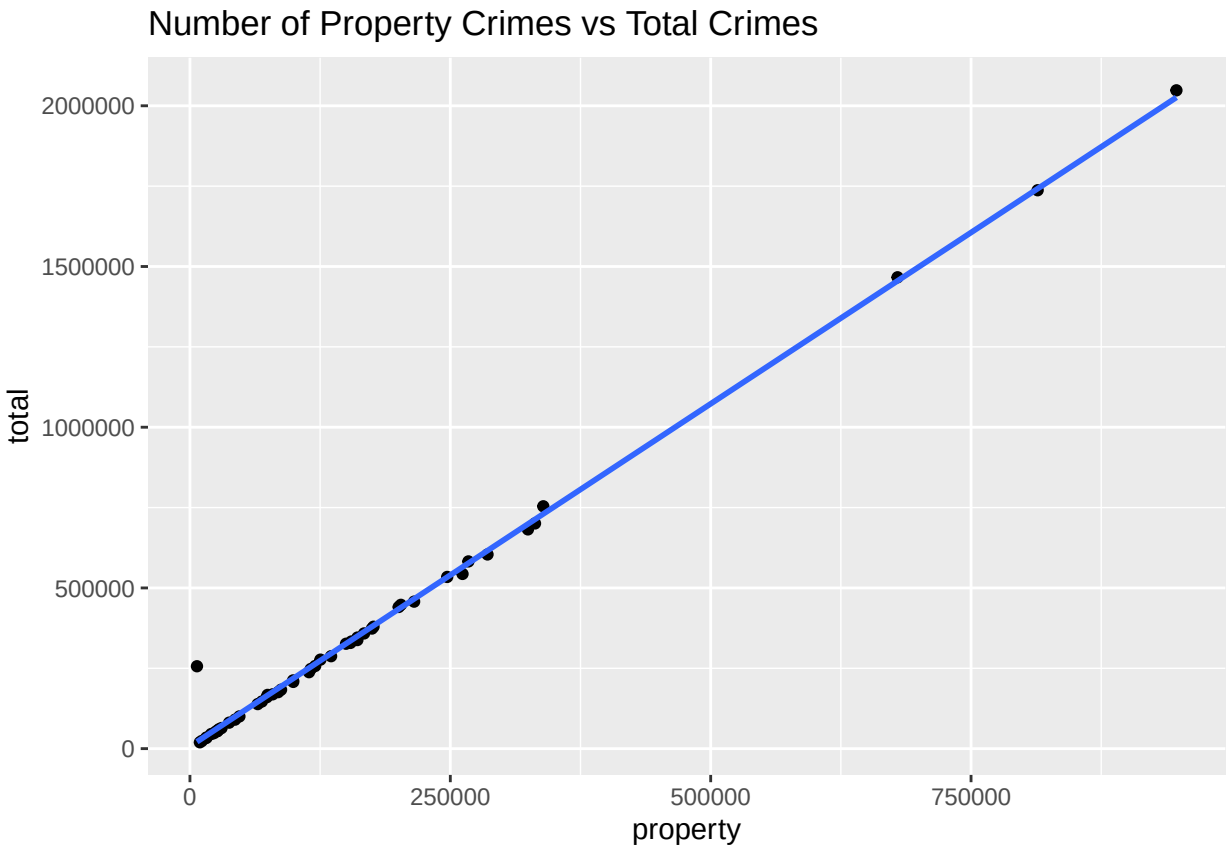
Fourth Plot This fourth and final map is a map of the murder rates in proportion to the total number of crimes. This helps to show the number of murders per 10k people. This data is very interesting since Nevada is no longer the number one state which means that most of its crimes must be other crimes such as larceny, robbery, etc. Additionally, Mississippi has an incredibly high murder rate along with Louisiana which is surprising.

```
usMurder <-
  ggplot(data = us_states_crime,
         mapping = aes(x = long, y = lat, group = group, fill = (murder/total)*10000)) +
  geom_polygon(color="black", size=0.25) +
  theme_map() +
  labs(title = "Number of Crimes in Proportion to Total Number of Crime", fill = "Murders Per \n10k") +
  scale_fill_gradient(low="dark grey", high="black")
usMurder
```



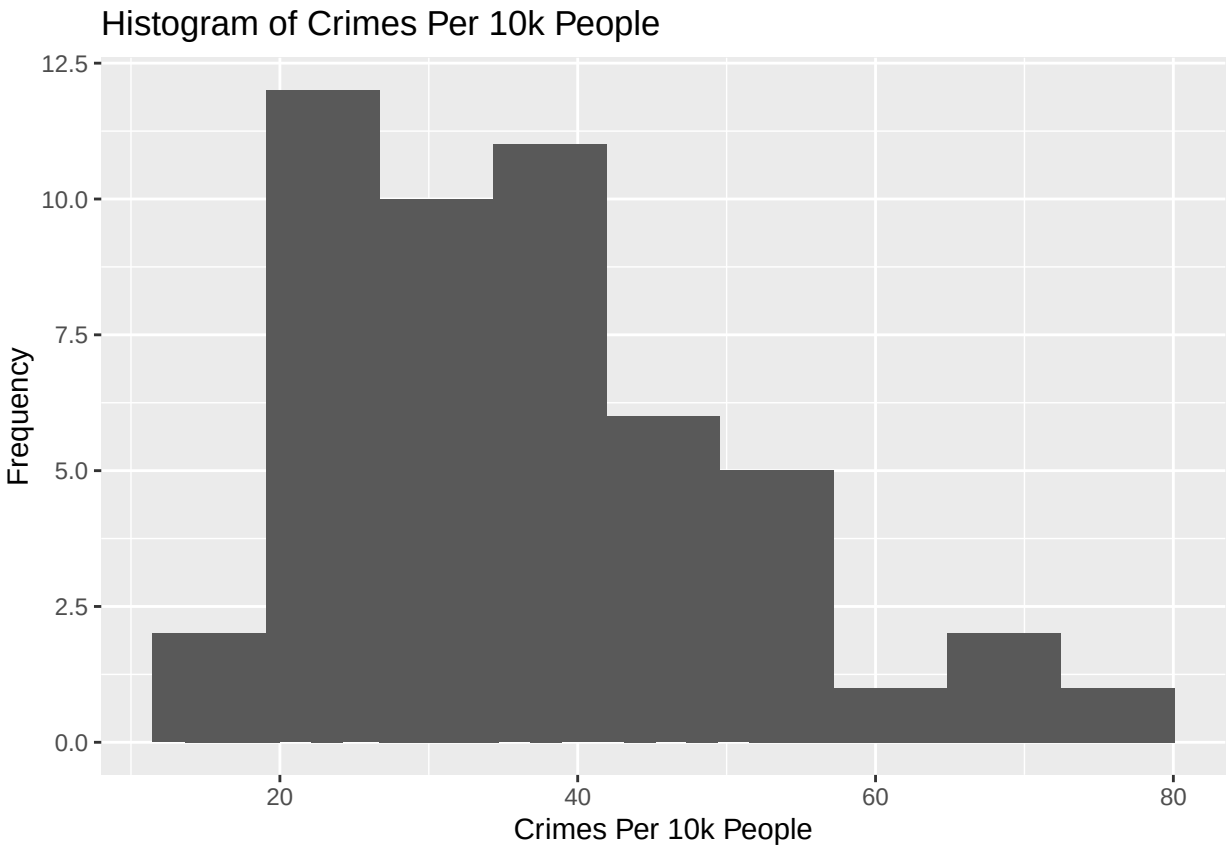
Fifth Plot This plot is actually a scatter plot with a linear regression and not a map. This shows the number of property crimes vs. the number of total crimes. Overall, this data seems to completely follow the linear regression except for there is one point that is an outlier compared to the rest. Its total crime is fairly high, but the number of property is low.

```
usScatter <-
  ggplot(crime, aes(property, total)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Number of Property Crimes vs Total Crimes")
usScatter
```



Sixth Plot This last plot is a histogram of the number of crimes per 10k people. This helps to visualize the distribution of the data. With this histogram we are able to observe that the data is generally centered around 30 crimes per 10k while there are still a few outliers with 70 and 80 crimes per 10k. Additionally, the data seems to be somewhat skewed right meaning that states tend to have lower crime rates, but there are still some with fairly high ones.

```
usHist <-
  ggplot(crime, aes(per10k)) +
  geom_histogram() +
  labs(title = "Histogram of Crimes Per 10k People", x = "Crimes Per 10k People", y = "Frequency") +
  stat_bin(bins = 9)
usHist
```



Conclusion

In conclusion, the United States seems to have very high crimes rates in Nevada, but these are generally crimes other than murder, since the highest murder rates seem to be in Louisiana and, especially, in Mississippi. Next, we also observe that the crimes per 10k people is almost normally distributed, but it is somewhat skewed right meaning that the crime rates tend to be lower with some outliers. We also observed that California and Texas had the most crime, but not the highest crime rate since they have very high populations compared to other states.