# Project 1: Wrangling, Exploration, Visualization

SDS322E

## Data Wrangling, Exploration, Visualization

**Alice Liu AL47697**

**Introduction** Paragraph or two introducing your datasets and variables, why they are interesting to you, etc.

```
library(tidyverse)
library(readr)
setwd("~/project1-main")
mw_data = read.csv("Minimum Wage Data.csv")
crime = read_csv("US_violent_crime.csv")
```

*I found both datasets off of kaggle. My first dataset is the minimum wage data (mw_data) by state in the US from 1969 to 2020. My second dataset is violent crime arrest rate (crime) by state in 1973. The variables in mw_data are year, state, actual state's minimum wage, state minimum wage in 2020 dollars, actual federal minimum wage, federal minimum wage in 2020 dollars, effective minimum wage, because if the state minimum wage is lower, it will assume the federal minimum wage. The remaining 6 columns in the mw_data dataset are average CPI (consumer price index), Department of Labor's unclean, scraped data, Department of Labor's lowest minimum wage (the one not enforced), the Department of Labor's lowest minimum wage in 2020 dollars, the Department of Labor's higher, enforced minimum wage, and lastly, the Department of Labor's higher, enforced minimum wage in 2020 dollars. My second dataset (crime) only has four variables: state, murder rate (per 100,000), assault rate (per 100,000), percent of urban population, and rape rate (per 100,000). I am interested in if there is a relationship between effective minimum wage and crime rate. Since many claim that the cost of living correlates to minimum wage, I would predict that minimum wage and crime rate would have an inverse relationship. In this project, I will want to use the effective minimum wage (in 2020 dollars) for each state in 1973, since my 'crime' dataset only has data from 1973.*

**Tidying: Reshaping** If your datasets are tidy already, demonstrate that you can reshape data with pivot wider/longer here (e.g., untidy and then retidy). Alternatively, it may be easier to wait until the wrangling section so you can reshape your summary statistics. Note here if you are going to do this.

```
# My datasets are tidy already, so I will untidy, then retidy
# & used it later in the project
mw_wider <- mw_data %>% select(1:2, Effective.Minimum.Wage) %>%
    pivot_wider(names_from = Year, values_from = Effective.Minimum.Wage)
mw_wider
```

```
## # A tibble: 54 x 54
##    State `1968` `1969` `1970` `1971` `1972` `1973` `1974` `1975` `1976` `1977`
##    <fct>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Alab~   1.15   1.15   1.3    1.3    1.6    1.6    1.6    1.6    2.2    2.2
##  2 Alas~   2.1    2.1    2.1    2.1    2.1    2.1    2.1    2.1    2.8    2.8
##  3 Ariz~   1.15   1.15   1.3    1.3    1.6    1.6    1.6    1.6    2.2    2.2
##  4 Arka~   1.15   1.15   1.3    1.3    1.6    1.6    1.6    1.6    2.2    2.2
```

```
##  5 Cali~   1.65   1.65   1.65   1.65   1.65   1.65   1.65   1.65   2.2    2.2
##  6 Colo~   1.15   1.15   1.3    1.3    1.6    1.6    1.6    1.6    2.2    2.2
##  7 Conn~   1.4    1.4    1.6    1.6    1.85   1.85   1.85   1.85   2.21   2.21
##  8 Dela~   1.25   1.25   1.3    1.3    1.6    1.6    1.6    1.6    2.2    2.2
##  9 Dist~   1.25   1.25   1.6    1.6    1.6    1.6    1.6    1.6    2.25   2.25
## 10 Flor~   1.15   1.15   1.3    1.3    1.6    1.6    1.6    1.6    2.2    2.2
## # ... with 44 more rows, and 43 more variables: `1978` <dbl>, `1979` <dbl>,
## #   `1980` <dbl>, `1981` <dbl>, `1982` <dbl>, `1983` <dbl>, `1984` <dbl>,
## #   `1985` <dbl>, `1986` <dbl>, `1987` <dbl>, `1988` <dbl>, `1989` <dbl>,
## #   `1990` <dbl>, `1991` <dbl>, `1992` <dbl>, `1993` <dbl>, `1994` <dbl>,
## #   `1995` <dbl>, `1996` <dbl>, `1997` <dbl>, `1998` <dbl>, `1999` <dbl>,
## #   `2000` <dbl>, `2001` <dbl>, `2002` <dbl>, `2003` <dbl>, `2004` <dbl>,
## #   `2005` <dbl>, `2006` <dbl>, `2007` <dbl>, `2008` <dbl>, `2009` <dbl>,
## #   `2010` <dbl>, `2011` <dbl>, `2012` <dbl>, `2013` <dbl>, `2014` <dbl>,
## #   `2015` <dbl>, `2016` <dbl>, `2017` <dbl>, `2018` <dbl>, `2019` <dbl>,
## #   `2020` <dbl>
```

```
mw_wider %>% pivot_longer(2:54, names_to = "Year", values_to = "Effective.Minimum.Wage")
```

```
## # A tibble: 2,862 x 3
##    State   Year  Effective.Minimum.Wage
##    <fct>   <chr>                  <dbl>
##  1 Alabama 1968                    1.15
##  2 Alabama 1969                    1.15
##  3 Alabama 1970                    1.3
##  4 Alabama 1971                    1.3
##  5 Alabama 1972                    1.6
##  6 Alabama 1973                    1.6
##  7 Alabama 1974                    1.6
##  8 Alabama 1975                    1.6
##  9 Alabama 1976                    2.2
## 10 Alabama 1977                    2.2
## # ... with 2,852 more rows
```

*To tidy the the minimum wage dataset, I selected 3 columns: State, Year, and Effective Minimum Wage. I then used pivot wider to widen the dataset by having 'Year' as my variables and minimum wage as my values. I chose 'Year' because it is easier to visualize how minimum wage changes each year per state. Afterwards, I pivoted longer to retidy my data.*

```
statedata <- left_join(crime, mw_data, by = c(X1 = "State"))
joined <- left_join(crime, mw_data, by = c(X1 = "State"))
nrow(mw_data)
```

### Joining/Merging

```
## [1] 2862
```

```
nrow(crime)
```

```
## [1] 50
```

```
nrow(joined)
```

```
## [1] 2650
```

*I decided to do a left join with crime being my first dataframe, and then adding rows with matches from my*

*minimum wage dataframe. I did this because my minimum wage dataframe includes the 50 states, DC, Guam, Pureto Rico, and the US Virgin Islands, whereas the 'crime' dataframe only has the United States' 50 states. Therefore, the US territories and DC would have NA for murder, assault, urban population, and rape if I were to do a full join. Therefore, the observations dropped were the 3 territories and DC. I do not see many problems with this, since they have smaller populations compared to the 50 states, their datapoints/values may be outliers of my data. Total of observations after I joined is 2650, which makes sense because there are 50 states for the 53 years from my minimum wage dataset. My crime dataset has 50 rows; minimum wage dataset has 2862 rows.*

```
# Dplyr functions:
statedata <- statedata %>% filter(Year == 1973)  #Since my crime dataset only has stats from 1973, I wa

statedata <- statedata %>% select(1:5, Effective.Minimum.Wage.2020.Dollars)  #Only using Effective Min.

statedata <- statedata %>% mutate(totalcrime = Murder + Assault +
    Rape)  #New variable is total violent crime rate (per 100,000)

statedata %>% arrange(desc(totalcrime))  #Visualize the states with the highest crime rates per 100,000
```

**Wrangling**

```
## # A tibble: 50 x 7
##    X1         Murder Assault UrbanPop  Rape Effective.Minimum.Wage.2~ totalcrime
##    <chr>       <dbl>   <dbl>    <dbl> <dbl>                     <dbl>      <dbl>
##  1 Florida      15.4     335       80  31.9                      9.32       382.
##  2 North Car~   13       337       45  16.1                      9.32       366.
##  3 Maryland     11.3     300       67  27.8                      9.32       339.
##  4 Arizona       8.1     294       80  31                        9.32       333.
##  5 New Mexico   11.4     285       70  32.1                      9.32       328.
##  6 California    9       276       91  40.6                      9.61       326.
##  7 Alaska       10       263       48  44.5                     12.2        318.
##  8 South Car~   14.4     279       48  22.5                      9.32       316.
##  9 Nevada       12.2     252       81  46                        9.32       310.
## 10 Michigan     12.1     255       74  35.1                      9.32       302.
## # ... with 40 more rows
```

```
statedata <- statedata %>% rename(State = X1) %>% rename(min_wage = Effective.Minimum.Wage.2020.Dollars)

statedata %>% group_by(State) %>% summarize(minwage_crime_ratio = min_wage/totalcrime) %>%
    arrange(minwage_crime_ratio)  #Outputs minimum wage to total crime ratio in ascending order
```

```
## # A tibble: 50 x 2
##    State          minwage_crime_ratio
##    <chr>                        <dbl>
##  1 Florida                     0.0244
##  2 North Carolina              0.0255
##  3 Maryland                    0.0275
##  4 Arizona                     0.0280
##  5 New Mexico                  0.0284
##  6 South Carolina              0.0295
##  7 California                  0.0295
##  8 Nevada                      0.0300
##  9 Michigan                    0.0308
## 10 Mississippi                 0.0319
```

```
## # ... with 40 more rows
```

```
sum(str_detect(statedata$State, "^A"))  # There are 4 states in the U.S. that start with the letter 'A'
```

```
## [1] 4
```

```
# Summary Statistics

statedata %>% summarize(correlation = cor(min_wage, totalcrime))  #correlation between minimum wage and
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.162
```

```
statedata %>% summarize(correlation = cor(Murder, totalcrime))  #correlation between murder rate and to
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.819
```

```
statedata %>% summarize(correlation = cor(Assault, totalcrime))  #correlation between assault rate and
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.997
```

```
statedata %>% summarize(correlation = cor(Rape, totalcrime))  #correlation between rape rate and total
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.720
```

```
statedata %>% summarize(mean = mean(Murder))  #mean of murder rates for all the states
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  7.79
```

```
statedata %>% summarize(sd = sd(min_wage))  #standard deviation of minimum wage for all the states
```

```
## # A tibble: 1 x 1
##      sd
##   <dbl>
## 1 0.507
```

```
statedata %>% summarize(var = var(UrbanPop))  #variance of urban population percentage
```

```
## # A tibble: 1 x 1
##     var
##   <dbl>
## 1  210.
```

```
statedata %>% summarize(distinct = n_distinct(min_wage))  # number of distinct minimum wage values
```

```
## # A tibble: 1 x 1
##   distinct
```

```
##      <int>
## 1        5
```

```
statedata %>% mutate(lowcrime = totalcrime <= 185) %>% group_by(lowcrime) %>%
    summarize(mean_urbanpop = mean(UrbanPop))  #'low crime' was determined by finding the median of tot
```

```
## # A tibble: 2 x 2
##   lowcrime mean_urbanpop
##   <lgl>            <dbl>
## 1 FALSE             69.1
## 2 TRUE              61.7
```

```
statedata %>% mutate(lowcrime = totalcrime <= 185) %>% group_by(lowcrime) %>%
    summarize(sd_urbanpop = sd(UrbanPop))
```

```
## # A tibble: 2 x 2
##   lowcrime sd_urbanpop
##   <lgl>          <dbl>
## 1 FALSE           14.3
## 2 TRUE            13.9
```

```
min_median_max <- function(x) {
    array1 <- quantile(x, c(0, 0.5, 1))
    data.frame(minimum = array1[1], median = array1[2], maximum = array1[3])
}

statedata %>% summarize(min_median_max(Murder))
```

```
## # A tibble: 1 x 3
##   minimum median maximum
##     <dbl>  <dbl>   <dbl>
## 1     0.8   7.25    17.4
```

```
statedata %>% summarize(min_median_max(Assault))
```

```
## # A tibble: 1 x 3
##   minimum median maximum
##     <dbl>  <dbl>   <dbl>
## 1      45    159     337
```

```
statedata %>% summarize(min_median_max(UrbanPop))
```

```
## # A tibble: 1 x 3
##   minimum median maximum
##     <dbl>  <dbl>   <dbl>
## 1      32     66      91
```

```
statedata %>% summarize(min_median_max(Rape))
```

```
## # A tibble: 1 x 3
##   minimum median maximum
##     <dbl>  <dbl>   <dbl>
## 1     7.3   20.1      46
```

```
statedata %>% summarize(min_median_max(min_wage))
```

```
## # A tibble: 1 x 3
##   minimum median maximum
##     <dbl>  <dbl>   <dbl>
```

```
## 1    9.32   9.32    12.2
```

```
statedata %>% summarize(min_median_max(totalcrime))
```

```
## # A tibble: 1 x 3
##   minimum median maximum
##     <dbl>  <dbl>   <dbl>
## 1    53.1   185.    382.
```

```
library(kableExtra)
statedata %>% kbl(caption = "State Data in 1973") %>% kable_classic(full_width = F,
    html_font = "Cambria")
```

*I commented the procedures in the code above, so the following will only be about what I found interesting. To me, it is interesting that the correlation between minimum wage and total crime rate is very low: 0.162. In addition, the states with lower crime has a similar mean and sd to the states with higher crime. Also, the number of distinct minimum wage values was surprisingly low. Apparently, most states assume the federal minimum wage. The function that I wrote determines the minimum, median, and maximum of each variable. For my table, I used the 'kable' package.*

```
ggplot(data = statedata, aes(x = min_wage, y = totalcrime)) +
    geom_point(size = 3, aes(color = State)) + geom_smooth(method = "lm") +
    theme(legend.position = "none") + xlab("Minimum Wage(in 2020 dollars)") +
    ylab("Crime Rate(per 100,000)") + ggtitle("Minimum Wage v. Crime Rate")
```

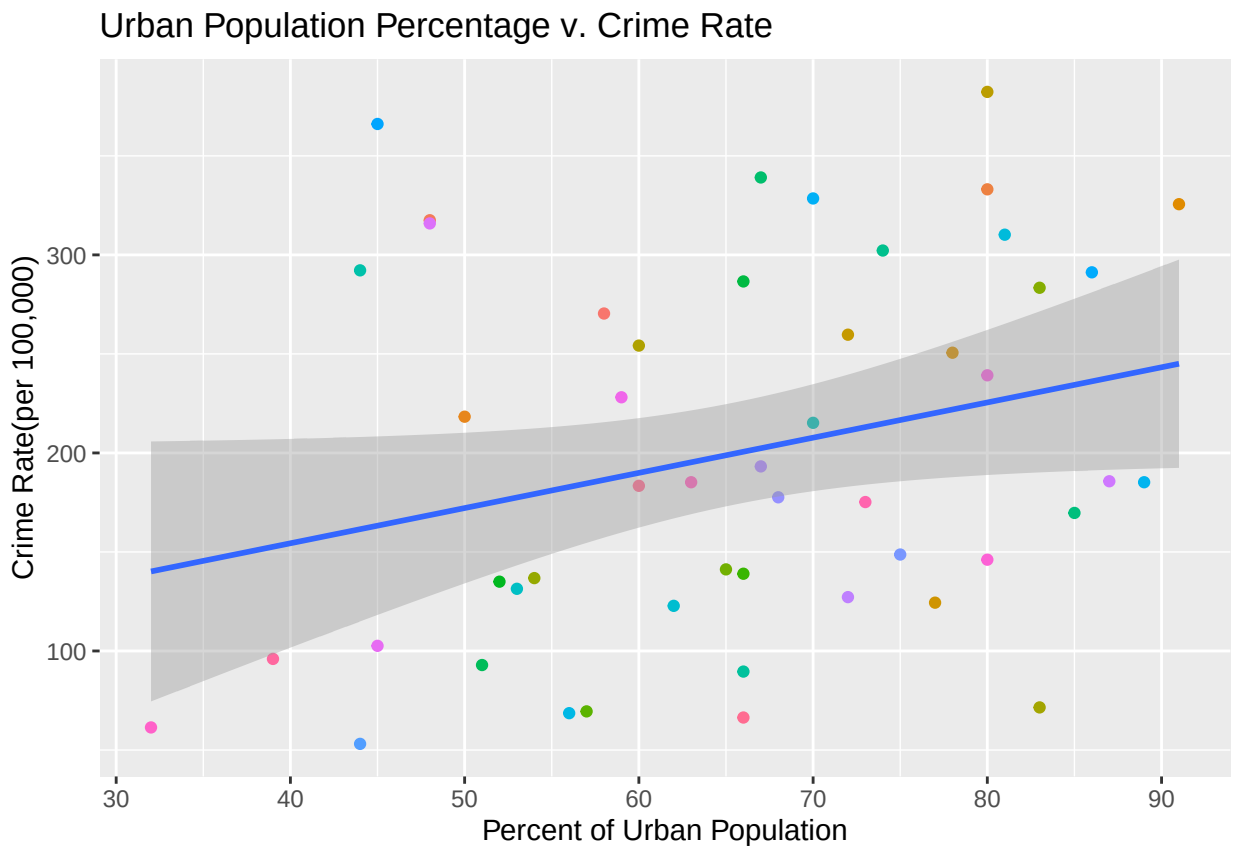
**Visualizing**

Table 1: State Data in 1973

| State | Murder | Assault | UrbanPop | Rape | min_wage | totalcrime |
|---|---|---|---|---|---|---|
| Alabama | 13.2 | 236 | 58 | 21.2 | 9.32 | 270.4 |
| Alaska | 10.0 | 263 | 48 | 44.5 | 12.23 | 317.5 |
| Arizona | 8.1 | 294 | 80 | 31.0 | 9.32 | 333.1 |
| Arkansas | 8.8 | 190 | 50 | 19.5 | 9.32 | 218.3 |
| California | 9.0 | 276 | 91 | 40.6 | 9.61 | 325.6 |
| Colorado | 7.9 | 204 | 78 | 38.7 | 9.32 | 250.6 |
| Connecticut | 3.3 | 110 | 77 | 11.1 | 10.78 | 124.4 |
| Delaware | 5.9 | 238 | 72 | 15.8 | 9.32 | 259.7 |
| Florida | 15.4 | 335 | 80 | 31.9 | 9.32 | 382.3 |
| Georgia | 17.4 | 211 | 60 | 25.8 | 9.32 | 254.2 |
| Hawaii | 5.3 | 46 | 83 | 20.2 | 9.32 | 71.5 |
| Idaho | 2.6 | 120 | 54 | 14.2 | 9.32 | 136.8 |
| Illinois | 10.4 | 249 | 83 | 24.0 | 9.32 | 283.4 |
| Indiana | 7.2 | 113 | 65 | 21.0 | 9.32 | 141.2 |
| Iowa | 2.2 | 56 | 57 | 11.3 | 9.32 | 69.5 |
| Kansas | 6.0 | 115 | 66 | 18.0 | 9.32 | 139.0 |
| Kentucky | 9.7 | 109 | 52 | 16.3 | 9.32 | 135.0 |
| Louisiana | 15.4 | 249 | 66 | 22.2 | 9.32 | 286.6 |
| Maine | 2.1 | 83 | 51 | 7.8 | 9.32 | 92.9 |
| Maryland | 11.3 | 300 | 67 | 27.8 | 9.32 | 339.1 |
| Massachusetts | 4.4 | 149 | 85 | 16.3 | 10.19 | 169.7 |
| Michigan | 12.1 | 255 | 74 | 35.1 | 9.32 | 302.2 |
| Minnesota | 2.7 | 72 | 66 | 14.9 | 9.32 | 89.6 |
| Mississippi | 16.1 | 259 | 44 | 17.1 | 9.32 | 292.2 |
| Missouri | 9.0 | 178 | 70 | 28.2 | 9.32 | 215.2 |
| Montana | 6.0 | 109 | 53 | 16.4 | 9.32 | 131.4 |
| Nebraska | 4.3 | 102 | 62 | 16.5 | 9.32 | 122.8 |
| Nevada | 12.2 | 252 | 81 | 46.0 | 9.32 | 310.2 |
| New Hampshire | 2.1 | 57 | 56 | 9.5 | 9.32 | 68.6 |
| New Jersey | 7.4 | 159 | 89 | 18.8 | 9.32 | 185.2 |
| New Mexico | 11.4 | 285 | 70 | 32.1 | 9.32 | 328.5 |
| New York | 11.1 | 254 | 86 | 26.1 | 10.78 | 291.2 |
| North Carolina | 13.0 | 337 | 45 | 16.1 | 9.32 | 366.1 |
| North Dakota | 0.8 | 45 | 44 | 7.3 | 9.32 | 53.1 |
| Ohio | 7.3 | 120 | 75 | 21.4 | 9.32 | 148.7 |
| Oklahoma | 6.6 | 151 | 68 | 20.0 | 9.32 | 177.6 |
| Oregon | 4.9 | 159 | 67 | 29.3 | 9.32 | 193.2 |
| Pennsylvania | 6.3 | 106 | 72 | 14.9 | 9.32 | 127.2 |
| Rhode Island | 3.4 | 174 | 87 | 8.3 | 9.32 | 185.7 |
| South Carolina | 14.4 | 279 | 48 | 22.5 | 9.32 | 315.9 |
| South Dakota | 3.8 | 86 | 45 | 12.8 | 9.32 | 102.6 |
| Tennessee | 13.2 | 188 | 59 | 26.9 | 9.32 | 228.1 |
| Texas | 12.7 | 201 | 80 | 25.5 | 9.32 | 239.2 |
| Utah | 3.2 | 120 | 80 | 22.9 | 9.32 | 146.1 |
| Vermont | 2.2 | 48 | 32 | 11.2 | 9.32 | 61.4 |
| Virginia | 8.5 | 156 | 63 | 20.7 | 9.32 | 185.2 |
| Washington | 4.0 | 145 | 73 | 26.2 | 9.32 | 175.2 |
| West Virginia | 5.7 | 81 | 39 | 9.3 | 9.32 | 96.0 |
| Wisconsin | 2.6 | 53 | 66 | 10.8 | 9.32 | 66.4 |
| Wyoming | 6.8 | 161 | 60 | 15.6 | 9.32 | 183.4 |

*My first plot demonstrates the reason why I was interested in this dataset: the relationship between minimum wage and crime rate. However, you can see that the relationship is pretty much non-existent, which is conclusive with the section above where I found that the correlation was 0.162. I predicted that the relationship would be inversely related; however, I was completely wrong, since there is a weak positive correlation between the two.*

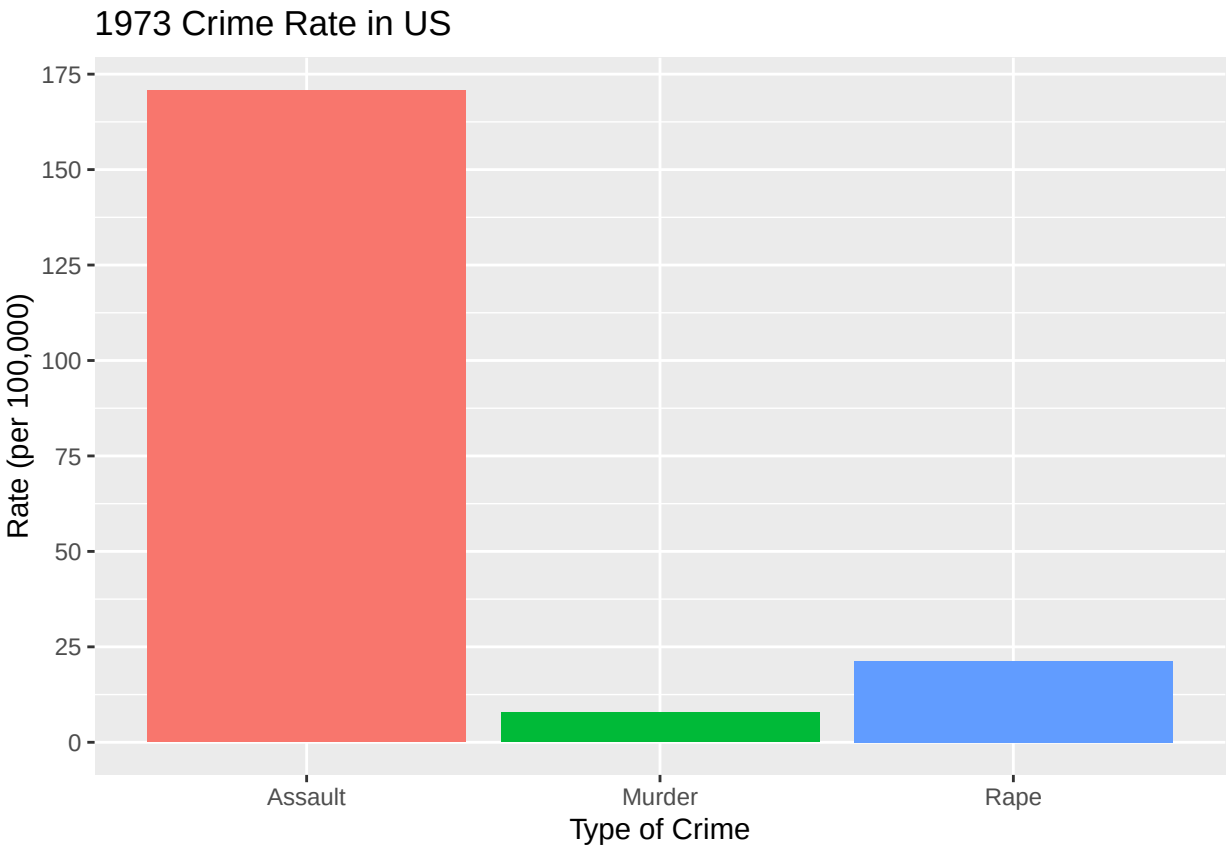
```
ggplot(data = statedata, aes(x = UrbanPop, y = totalcrime)) +
    geom_point(aes(color = State)) + geom_smooth(method = "lm") +
    theme(legend.position = "none") + xlab("Percent of Urban Population") +
    ylab("Crime Rate(per 100,000)") + ggtitle("Urban Population Percentage v. Crime Rate")
```




*I then moved on to see if there was a correlation between urban population and crime rate. I assumed that the correlation would be strong and positive, since urban populations are more dense and therefore more crime occurs. However, I was wrong once again. While the correlation is positive, it is pretty weak.*

```
statedatalonger <- statedata %>% pivot_longer(c("Murder", "Assault",
    "Rape"), names_to = "crimetype", values_to = "rate")

ggplot(data = statedatalonger, aes(x = crimetype, y = rate, fill = crimetype)) +
    geom_bar(stat = "summary") + xlab("Type of Crime") + ylab("Rate (per 100,000)") +
    ggtitle("1973 Crime Rate in US") + scale_y_continuous(breaks = seq(0,
    200, by = 25)) + theme(legend.position = "none")
```




*This plot shows the distribution of the type of crime. As you can see, the most common type of crime is assault, followed by rape, the murder. In the previous section, I used a lot of 'totalcrime' which is the sum of all three types of crime rates. I wanted to show this graph to display how it is distributed*

**Concluding Remarks** If any!