Assignment 1 Final Output

Contents

This a summary of the descriptive statistics from Diego and Isabelle's first assignment for the Collaborative Social Science Data Analysis class.

Data Set 1: Average Yearly Temperatures in New Haven 1912-1971

Introduction to the DataSet

The core R data set nhtemp, titled "Average Yearly Temperatures in New Haven" is a time series of 60 observations recording the mean annual temperature in degrees Fahrenheit in New Haven, Connecticut, between 1912 and 1971.

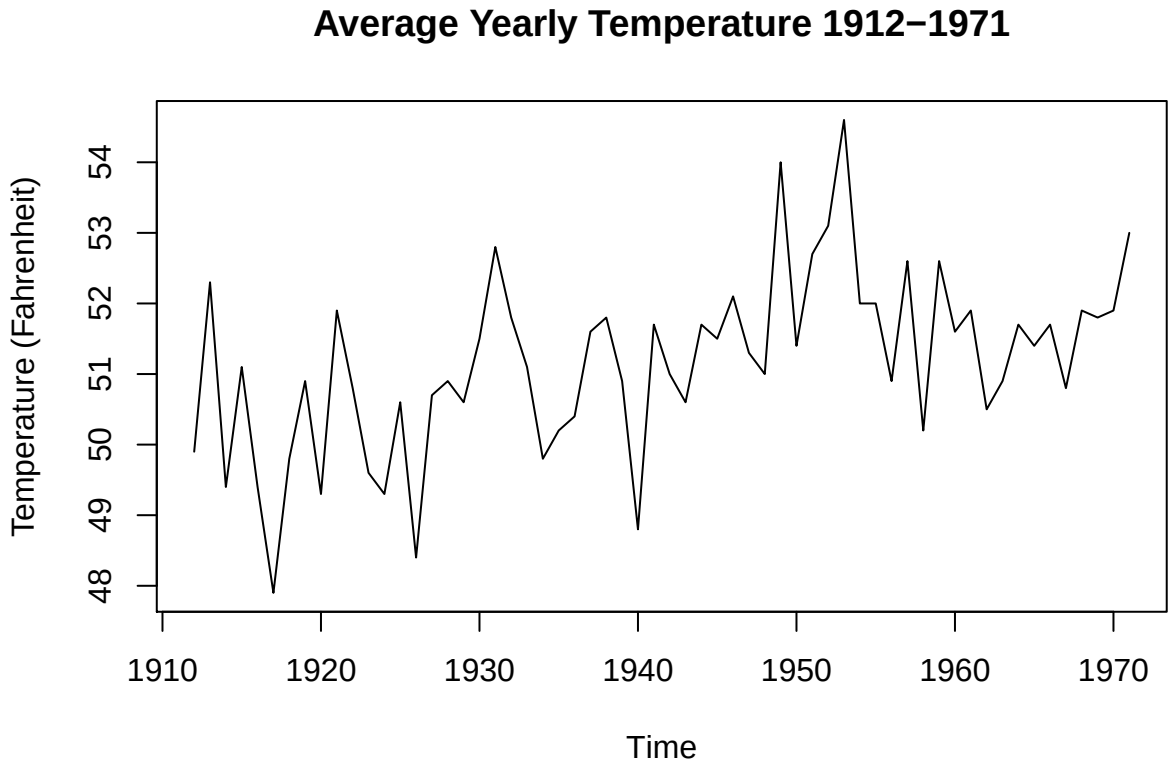
Descriptive Statistics

If we look at the summary statistics, we learn that the average temperature has fallen within a small range for the duration of the 60 years. Otherwise, it tells us very little about the actual climate or any change in climate over time, as the temperatures are aggregated by year and not by month or season.

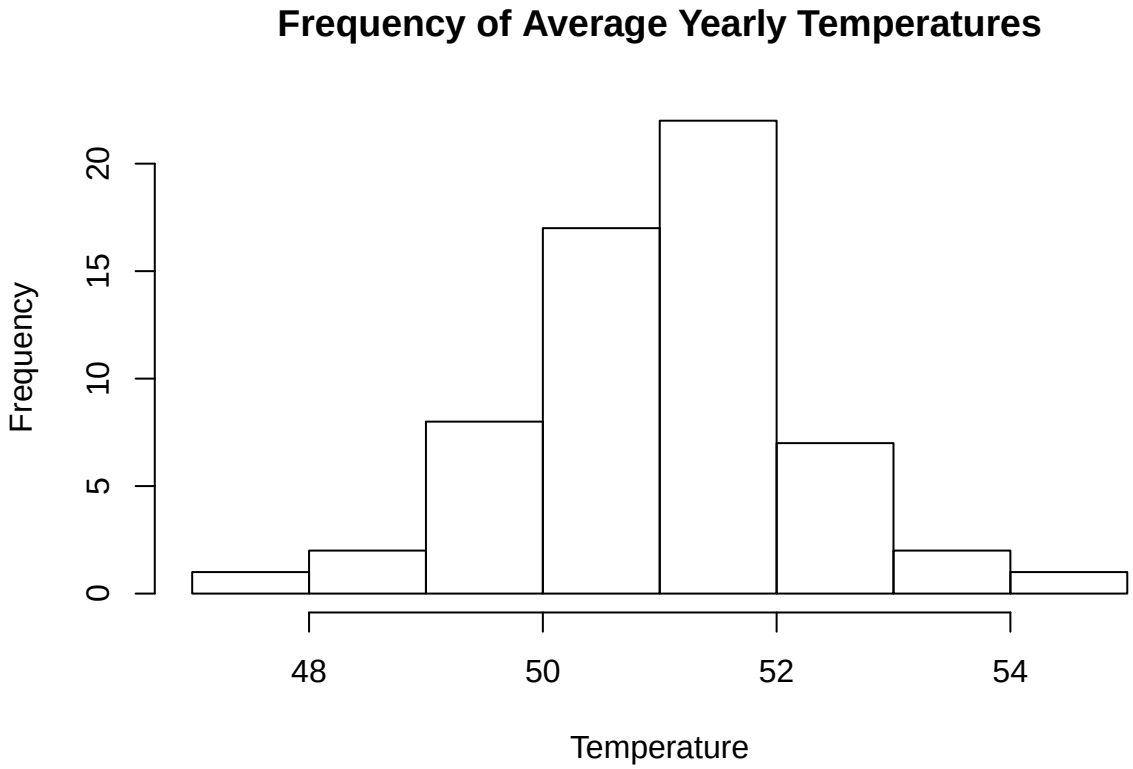
```
summary(nhtemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.90   50.58   51.20   51.16   51.90   54.60
```

However, if we look at the data over time, we can see a general upwards trend in the average temperatures, although they vary considerably from one year to the next.



The histogram shows us that most years cluster towards an average temperature of 50-52.



Data Set 2: Violent Crime Rates by US State

Introduction to the DataSet

The dataset USArrests contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. It also provides the percent of the population living in urban areas.

There are 50 observations on the 4 following variables:

-Murder (murder arrests per 100,000)

-Assault (assault arrests per 100,000)

-UrbanPop (percent urban population per state)

-Rape (rape arrests per 100,000)

Descriptive Statistics

Summary statistics

```
summary(USArrests$Murder)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

Firstly, we use the summary function to have a look at the values of the mean and the median (measures of central tendency) for each of the 4 variables. For murder arrests per 100,000 population, the mean is 7.788 and the median is 7.250.

```
summary(USArrests$Assault)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

For assault arrests per 100,000 population, the mean is 170.8 and the median is 159.

```
summary(USArrests$UrbanPop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.00   54.50   66.00   65.54   77.75   91.00
```

For percent of urban population in each state, the mean is 65.54% and the median is 66%.

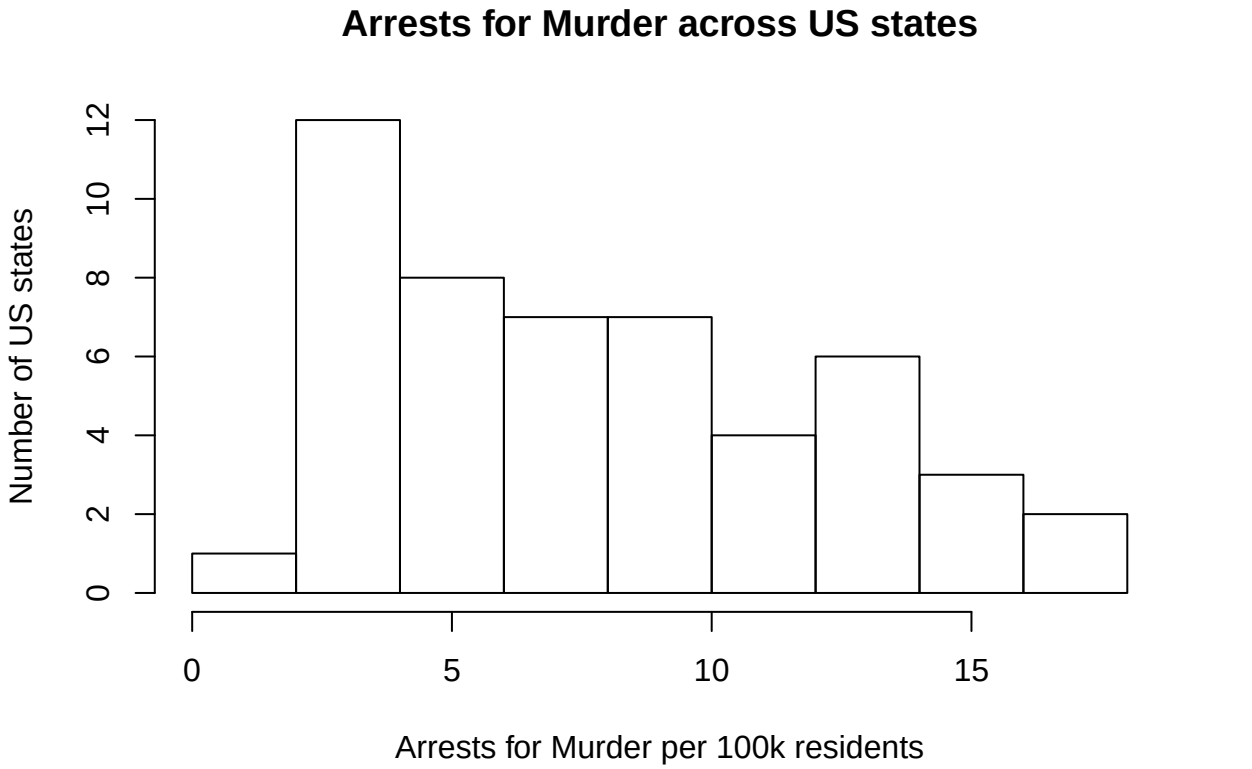
```
summary(USArrests$Rape)
```

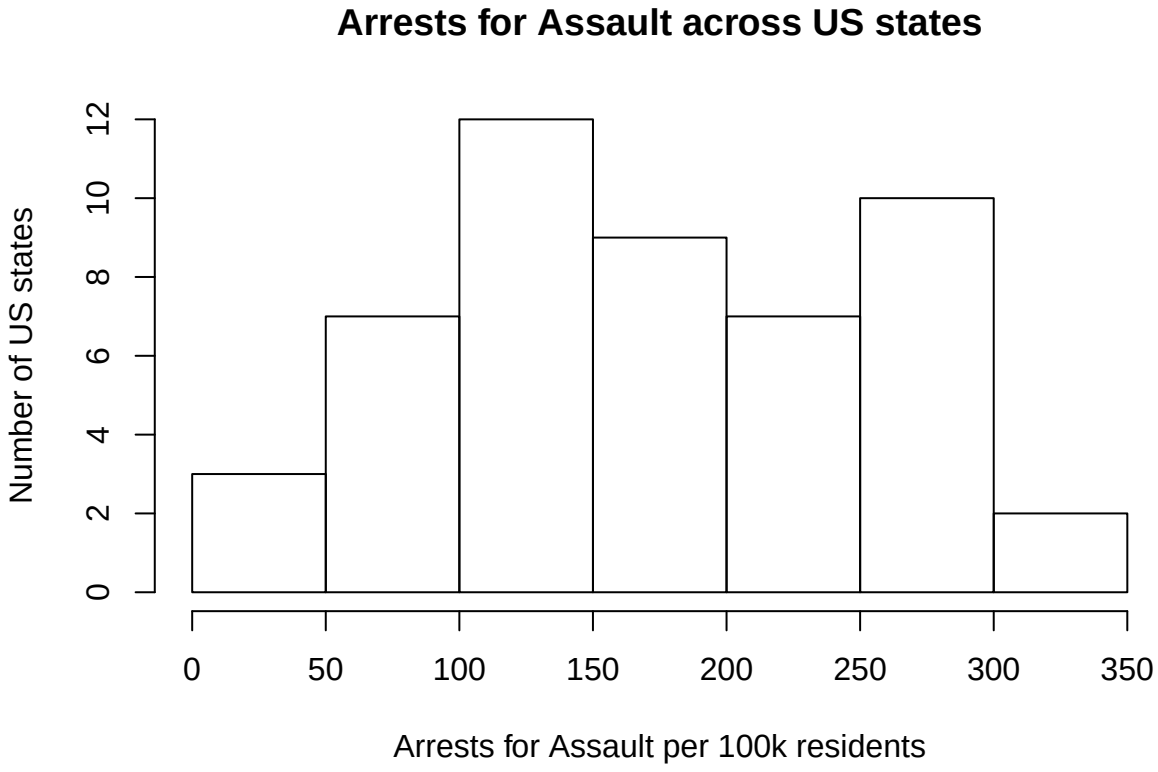
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.08   20.10   21.23   26.17   46.00
```

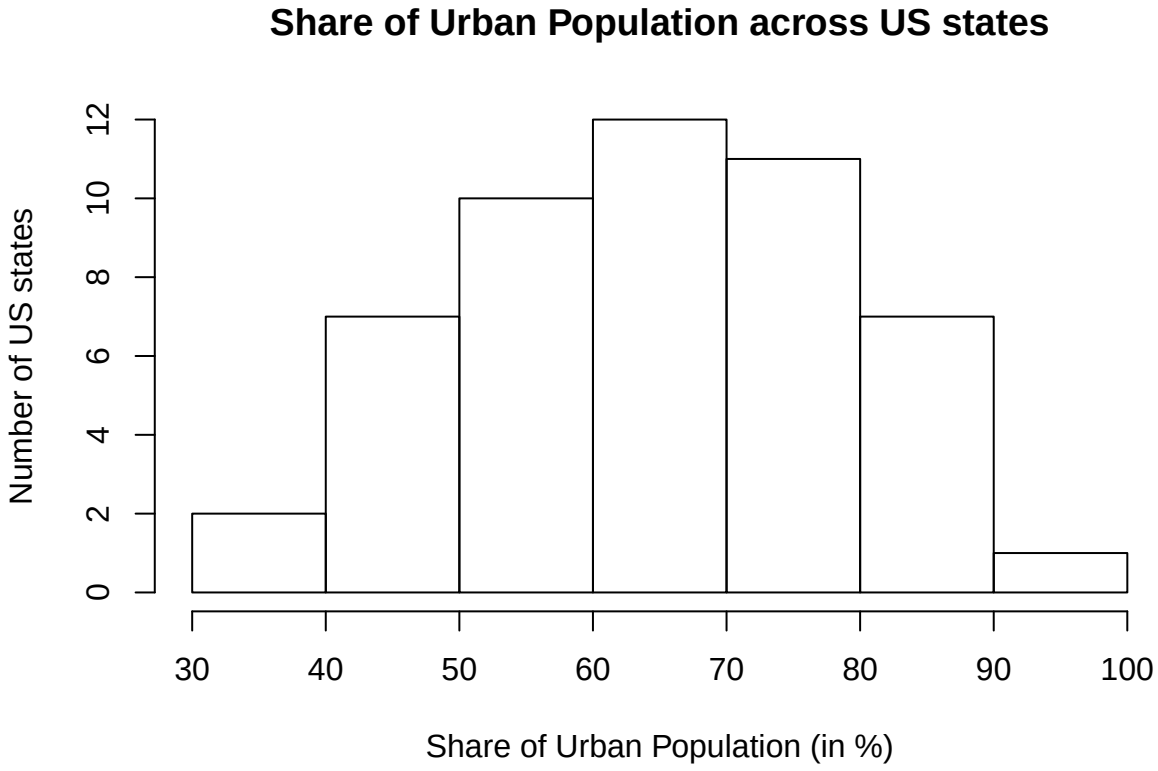
For rape arrests per 100,000 population, the mean is 21.23 and the median is 20.10.

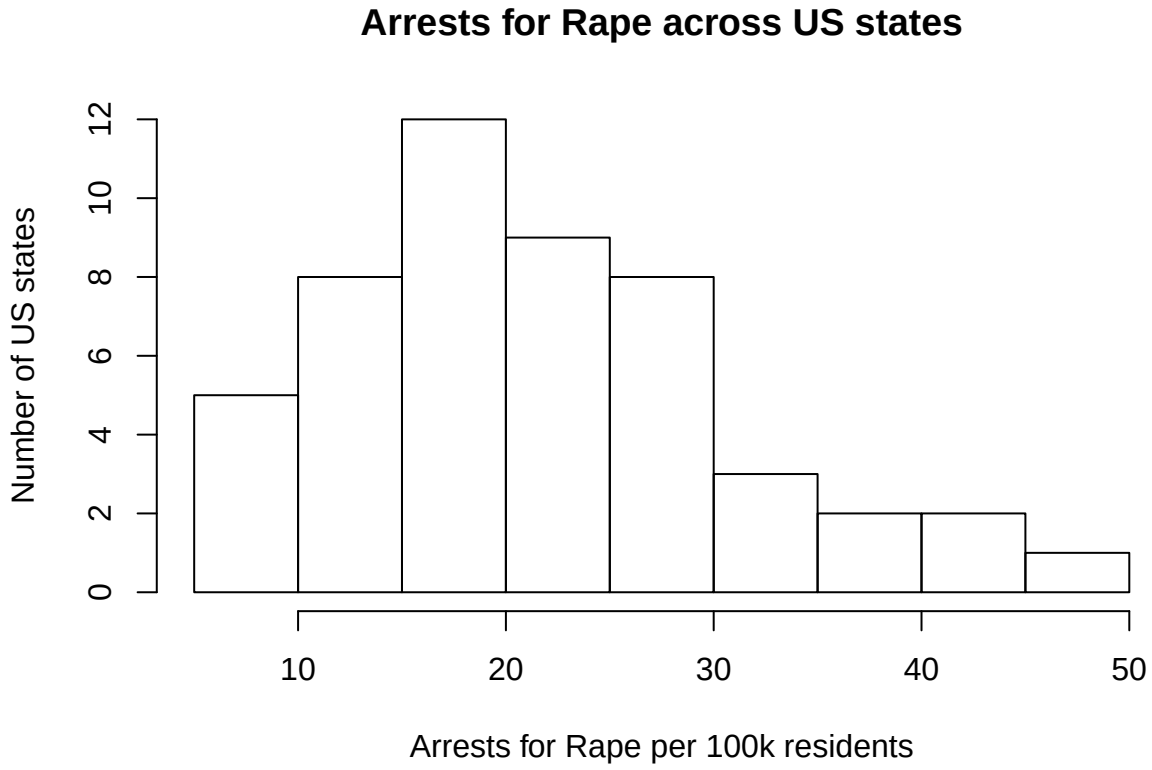
Histograms

Secondly, we generate histograms in order to have a general idea about the distribution of the 4 variables contained in our dataset. For murder arrests per 100,000 population, the histogram tells us that the largest group of US states (12 states) has between 2 and 5 arrests for murder per 100,000 residents. There is only 1 state that has less than 2 arrests for murder per 100,000 residents and more than 16 states have between 6 and 10 arrests per 100,000 residents (two bars of the histogram). For assault arrests per 100,000 residents, the histogram tells us that the largest group of states (12 states) has between 100 and 150 arrests for assault per 100,000 residents. The second largest group of states (10 states) has between 250 and 300 arrests for assault per 100,000 residents. On the other hand, while 3 states only have between 0 and 50 arrests for assault per 100,000 residents, 2 states have between 300 and 350. For the share of urban population in each US state, the largest group of states (12 states) has between 60 and 70 % of their population living in urban areas. Likewise, 10 states have between 70 and 80 % of their population living in urban areas. On the other hand, 2 states have less than 40 % of their population living in urban areas, and 1 state has more than 90% of urban population. When it comes to rape arrests per 100,000 residents, the distribution appears to be right skewed since a very small number of US states have more than 30 arrests for rape per 100,000 residents. The largest number of states (12 states) have between 15 and 20 arrests for rape per 100,000 residents.



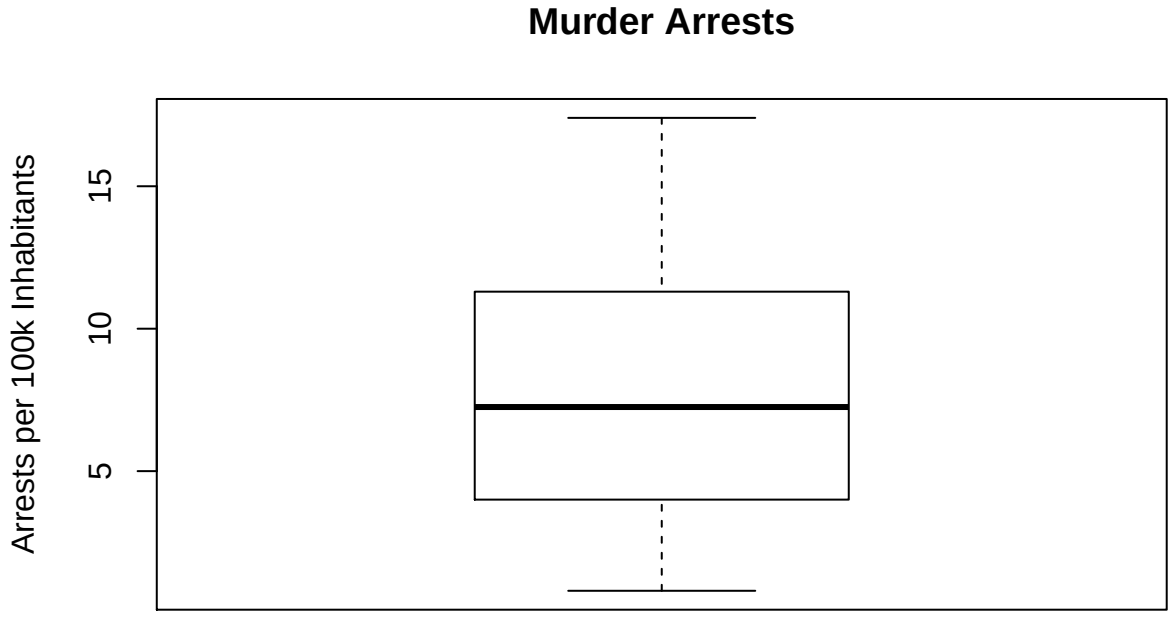

Arrests for Assault across US states
Number of US states
0
2
4
6
8
10
12
0
50
100
150
200
250
300
350
Arrests for Assault per 100k residents

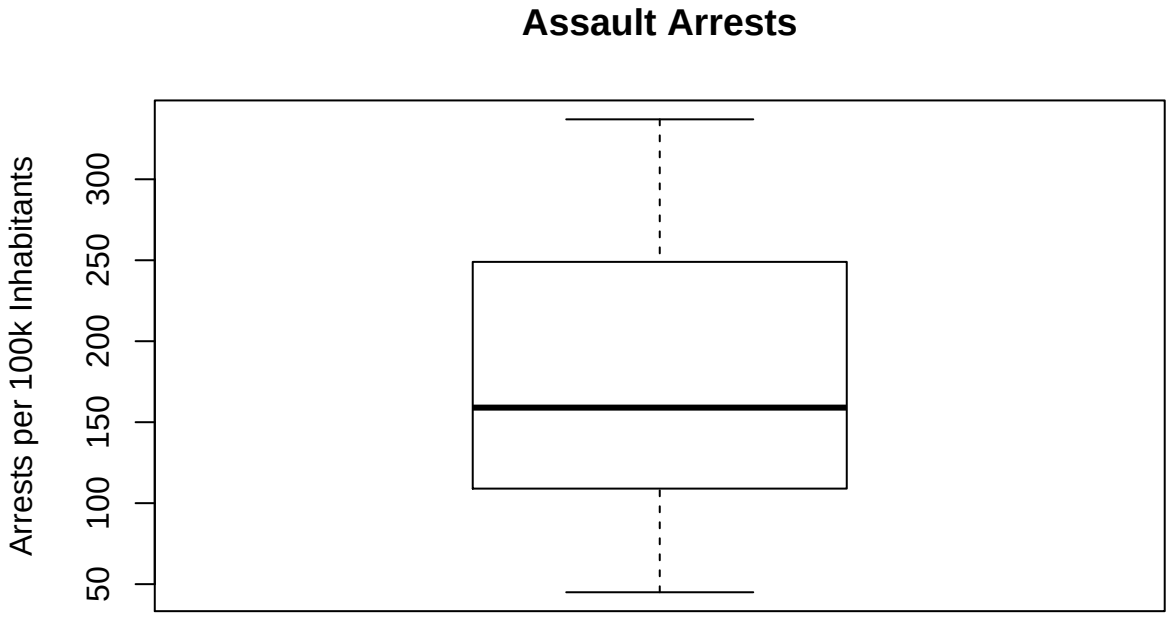

Share of Urban Population across US states
Number of US states
0
2
4
6
8
10
12
30
40
50
60
70
80
90
100
Share of Urban Population (in %)

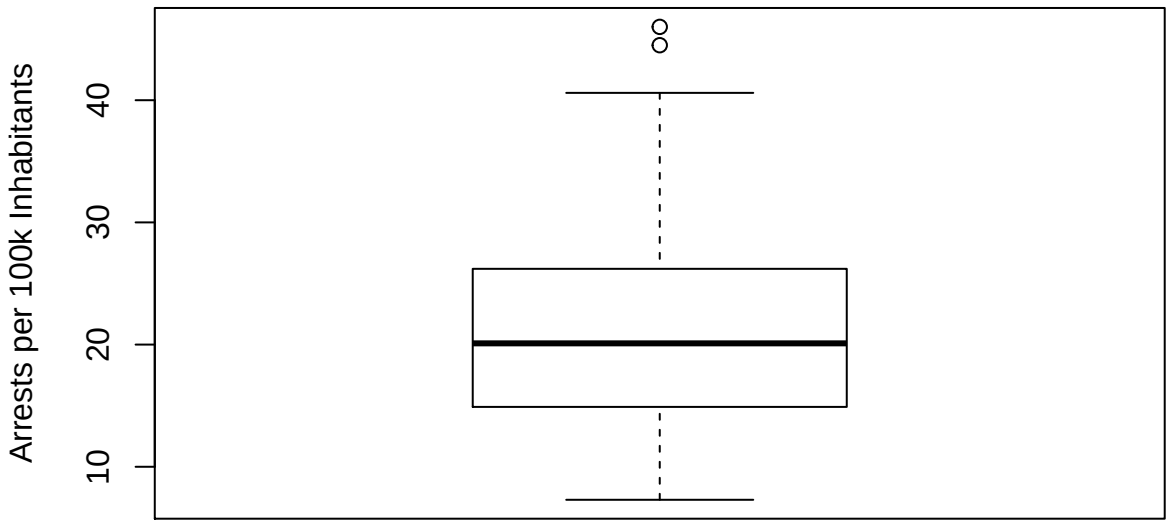


Boxplots

Now we make use of boxplots in order to provide a graphical representation of where the mean, median and interquartile range of our three crime variables are.

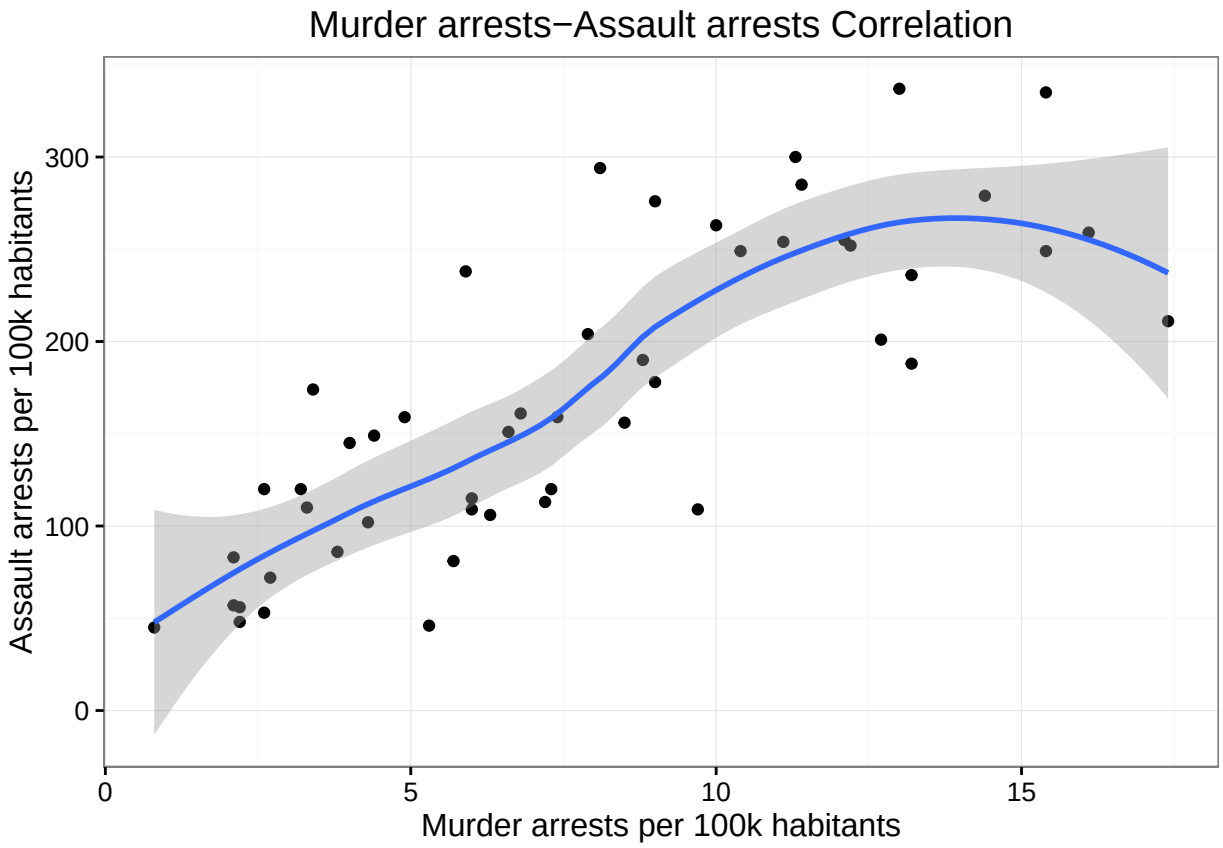



Assault Arrests
Arrests per 100k Inhabitants
50
100
150
200
250
300


Rape Arrests
Arrests per 100k Inhabitants
10
20
30
40

Correlations

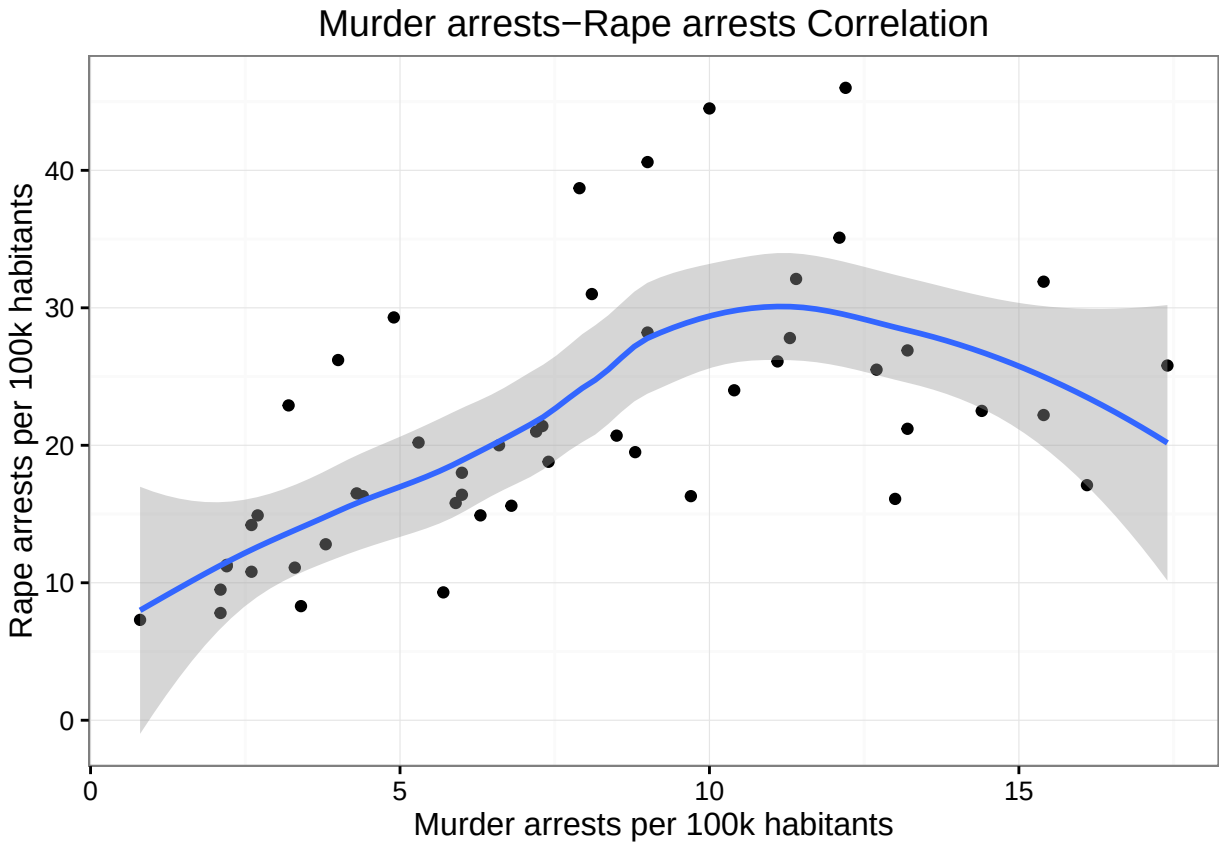
Finally, scatter plots help us assess whether there is correlation between our main variables of interest.




```
cor.test(USArrests$Murder, USArrests$Assault)
```

```
##
##  Pearson's product-moment correlation
##
## data:  USArrests$Murder and USArrests$Assault
## t = 9.2981, df = 48, p-value = 2.596e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6739512 0.8831110
## sample estimates:
##       cor
## 0.8018733
```

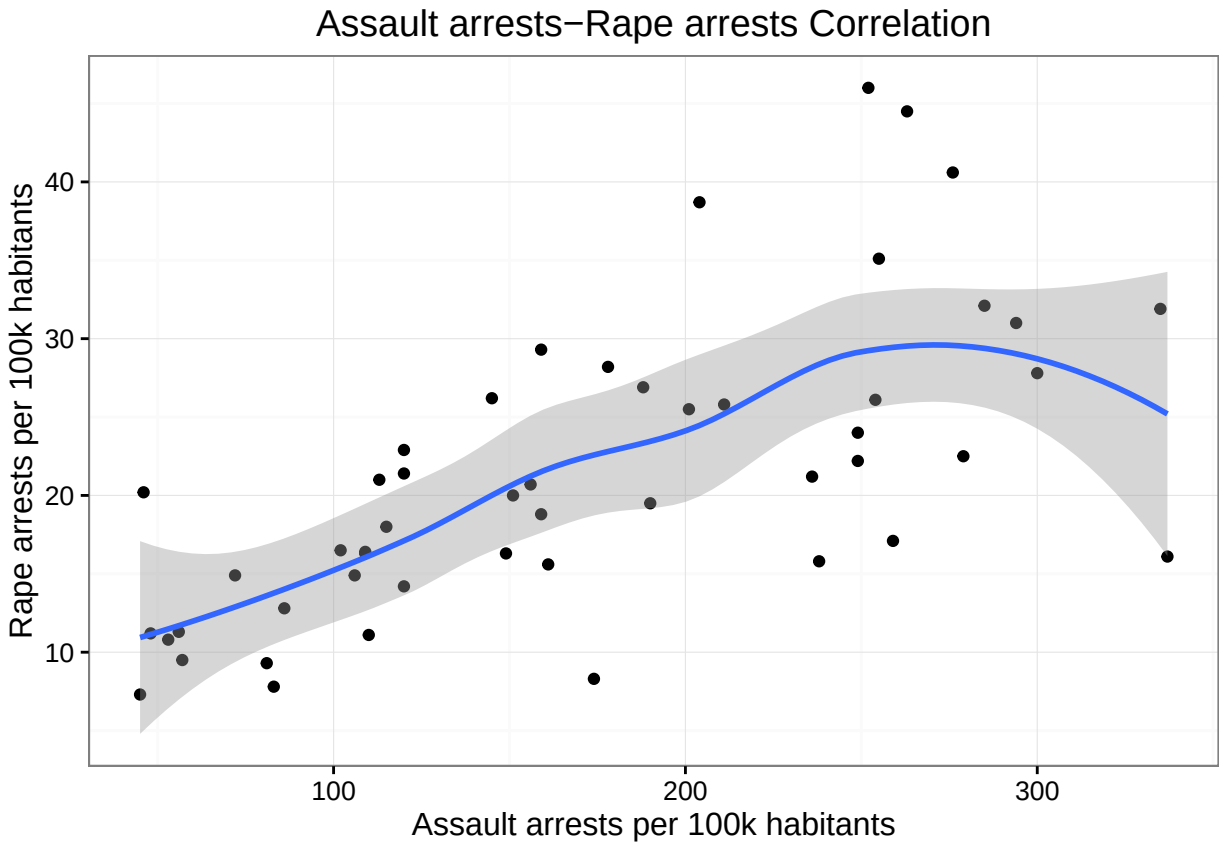
As both the scatter plot and the Pearson correlation coefficient indicate, there seems to be a strong positive correlation between murder arrests and assault arrests.




```
cor.test(USArrests$Murder, USArrests$Rape)
```

```
##
##  Pearson's product-moment correlation
##
## data:  USArrests$Murder and USArrests$Rape
## t = 4.7267, df = 48, p-value = 2.031e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3383006 0.7277619
## sample estimates:
##       cor
## 0.5635788
```

There also seems to be correlation between murder arrests and rape arrests, although slightly weaker than in the previous case.




```
cor.test(USArrests$Assault, USArrests$Rape)
```

```
##
##  Pearson's product-moment correlation
##
## data:  USArrests$Assault and USArrests$Rape
## t = 6.173, df = 48, p-value = 1.364e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4748141 0.7961645
## sample estimates:
##       cor
## 0.6652412
```

The correlation between assault arrests and rape arrests is also positive.

In conclusion, there seems to be a strong positive correlation between the three variables of interest. Nevertheless, the strongest correlation is the one between murder arrests and assault arrests.