# Assignments

## Assignment 1

### Collaborators: Sara Whitelaw, Halle Wasser

### Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
library(datasets)
```

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

```
dat<-USArrests
```

It is useful to rename the dataset because it makes the file easier to access and use in later coding functions.

### Problem 2

Use this command to make the state names into a new variable called State.

```
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

```
names(dat)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

The variables are Murder, Assault, UrbanPop, and Rape.

### Problem 3

What type of variable (from the DVB chapter) is `Murder`?

Answer: Quantitative Variable

What R Type of variable is it?

Answer: Numeric in a list form

**Problem 4**

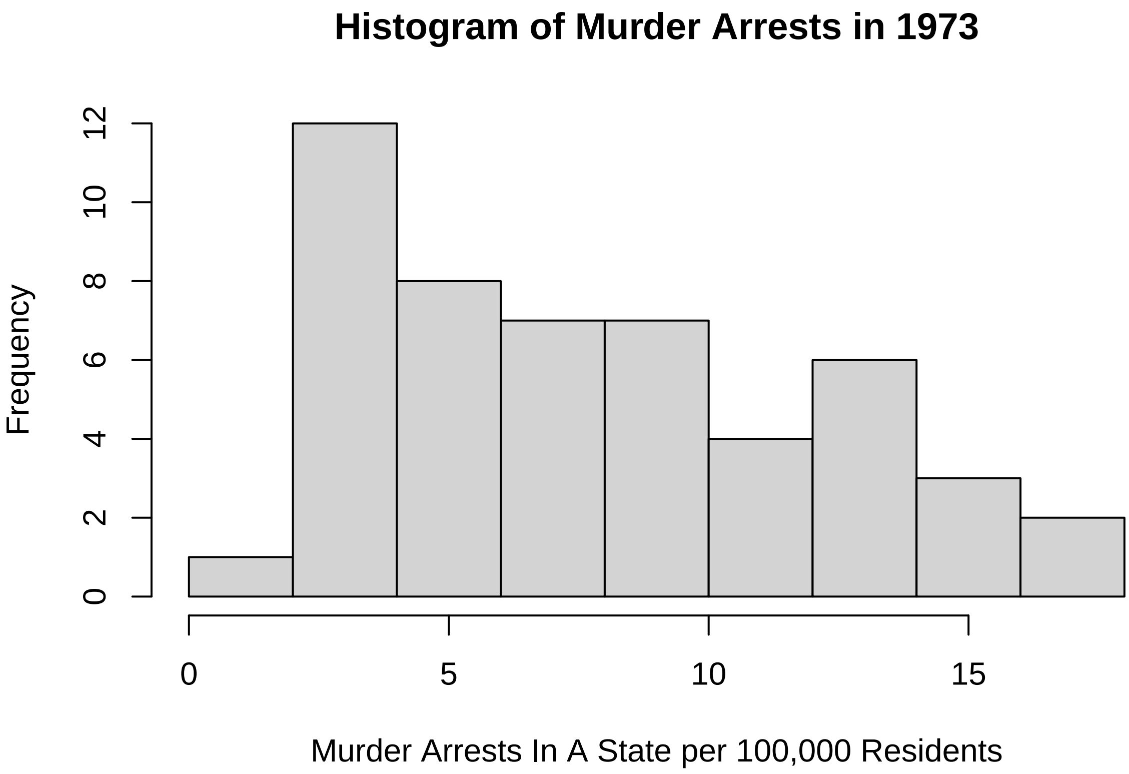
What information is contained in this dataset, in general? What do the numbers mean?

Answer: This data set gives us information on the frequency of murder, assault, and rape arrests per 100,000 residents in each state of the US in 1973. It also gives information on the percent of the population living in urban areas of each state.

**Problem 5**

Draw a histogram of `Murder` with proper labels and title.

```
hist(dat$Murder,main="Histogram of Murder Arrests in 1973", xlab="Murder Arrests In A State per 100,000
```



**Problem 6**

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

```
summary(dat$Murder)
```

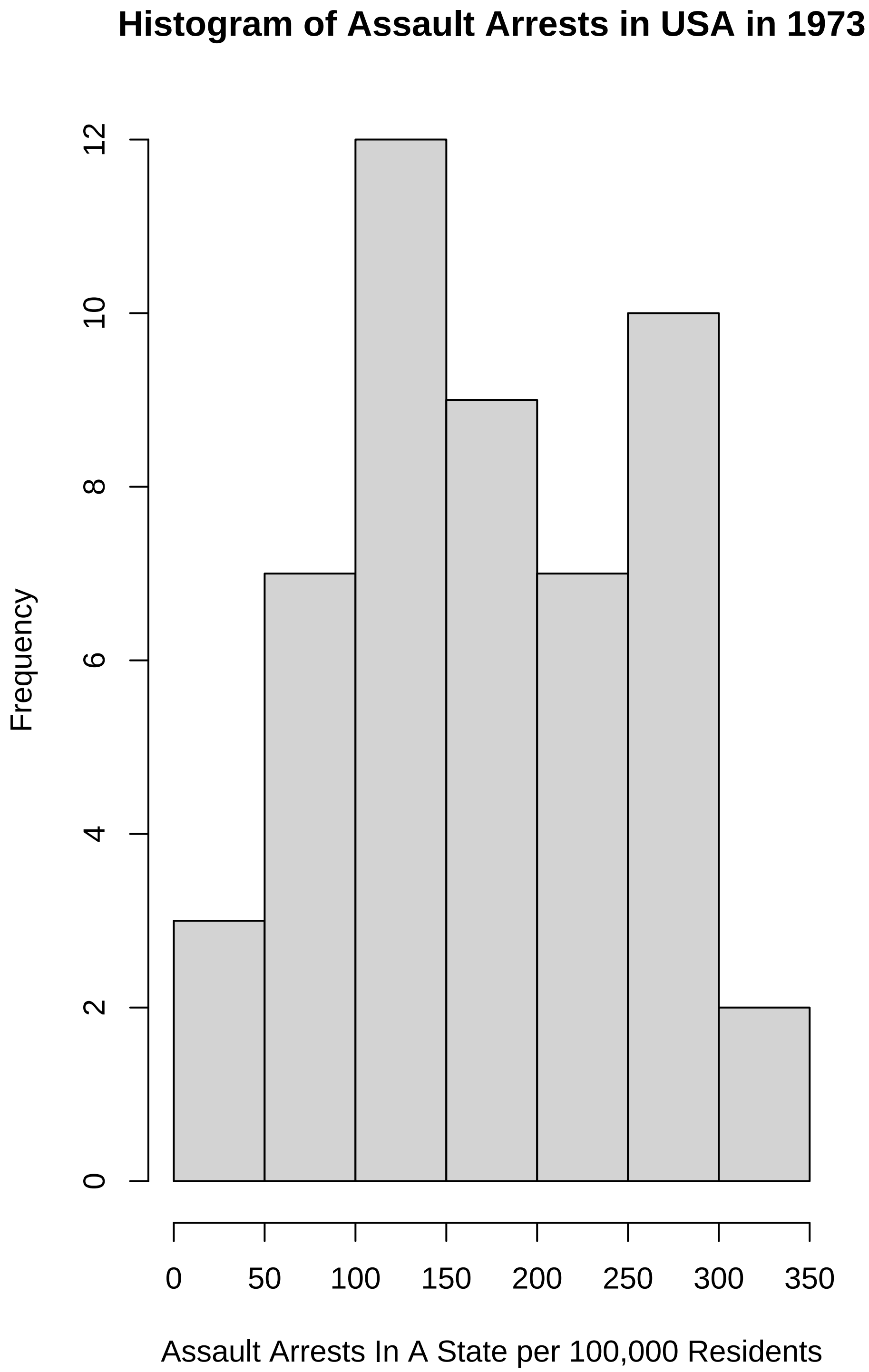
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

The mean of the Murder data is 7.788 and the median is 7.25. The mean is the average value of the data set. In other words, the mean is all of the numbers in a data set added together and then divided by the total amount of numbers present in the set. In contrast, the median is the middle value of the data when the numerical data values are ordered from least to greatest. Quartiles of a data set are the 3 values that divide the observed data into even fourths. R gives the 1st and 3rd quartiles to give us insight into the spread of our data.

### Problem 7

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

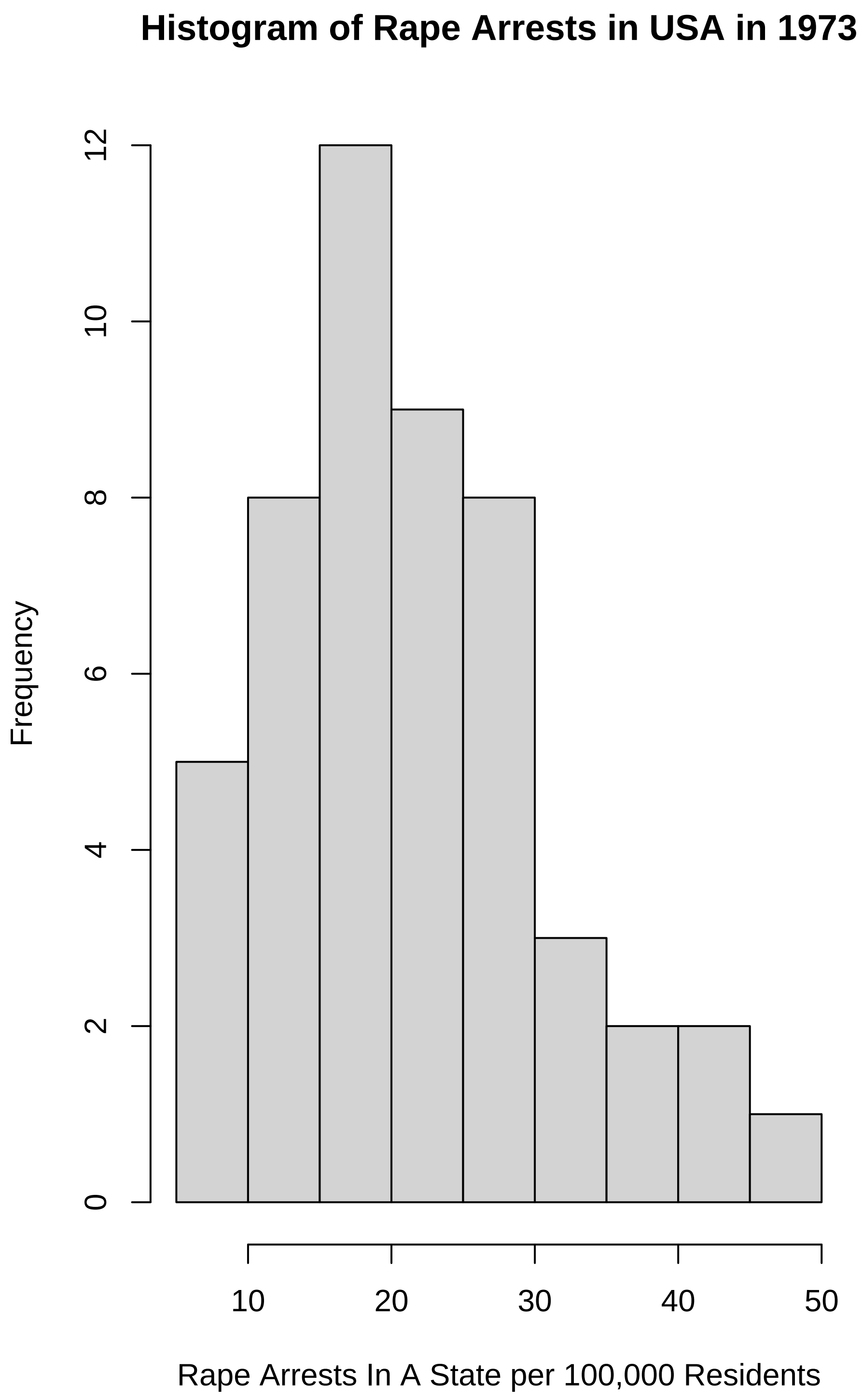
```
hist(dat$Assault,main="Histogram of Assault Arrests in USA in 1973", xlab="Assault Arrests In A State pe
```



```
summary(dat$Assault)
```

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

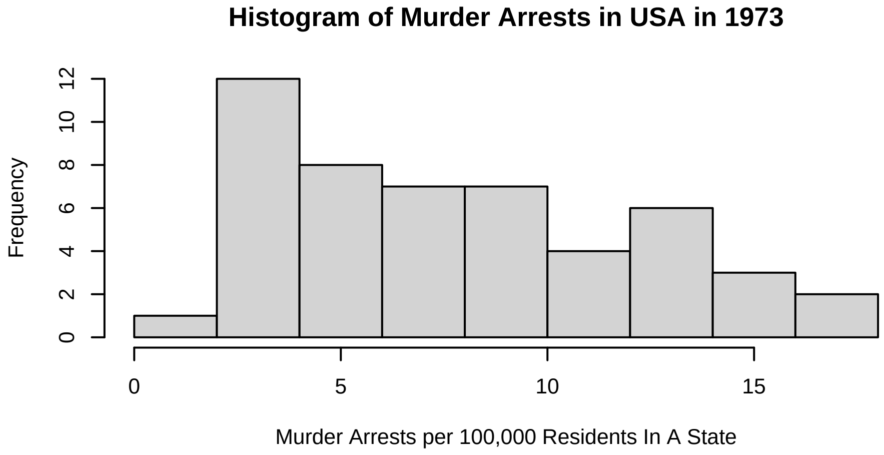
```
hist(dat$Rape,main="Histogram of Rape Arrests in USA in 1973", xlab="Rape Arrests In A State per 100,00
```

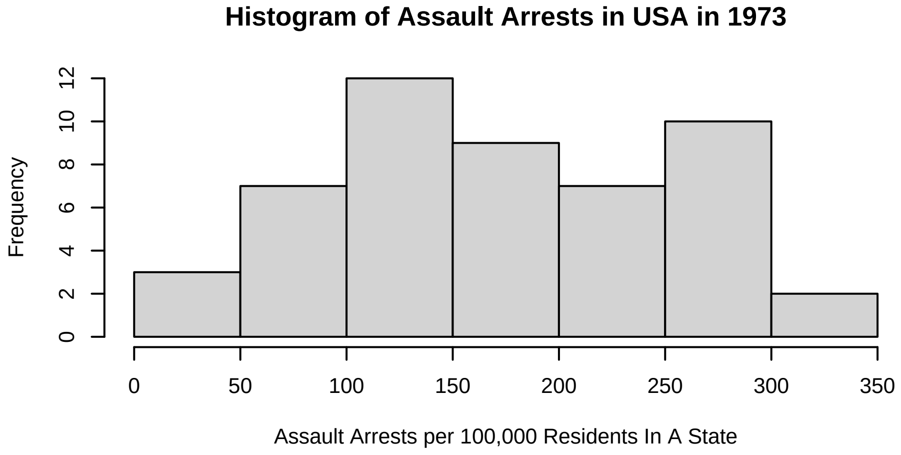


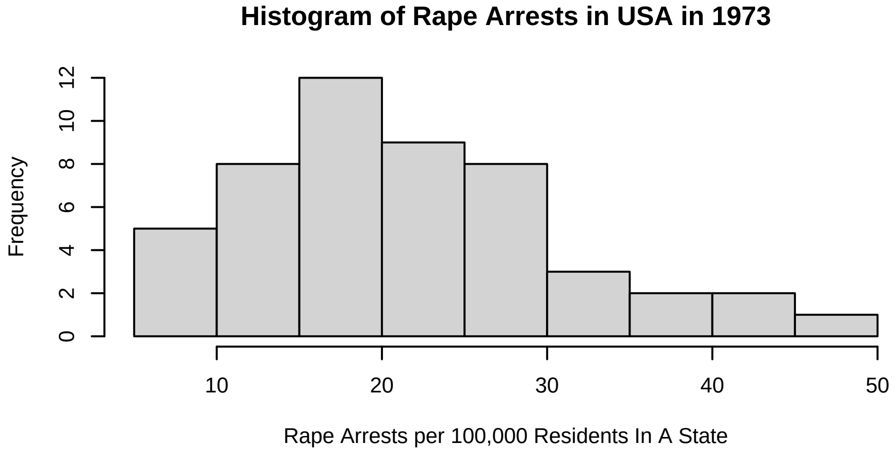
```
summary(dat$Rape)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

```
par(mfrow=c(3,1))
hist(dat$Murder,main="Histogram of Murder Arrests in USA in 1973", xlab="Murder Arrests per 100,000 Resi
hist(dat$Assault,main="Histogram of Assault Arrests in USA in 1973", xlab="Assault Arrests per 100,000 R
hist(dat$Rape,main="Histogram of Rape Arrests in USA in 1973", xlab="Rape Arrests per 100,000 Residents
```







The mean value for Assault is 170.76 and the median value is 159. The mean value for Rape is 21.232 and the median value is 20.1

What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: par is used to set or query graphical parameters.

What can you learn from plotting the histograms together?

Answer: We can see that the histograms for Murder and Rape Arrests skew to the left. We can also see that that there are more Assault arrests than Rape arrests, and more Rape arrests than Murder arrests.

### Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```r
library('maps')
library('ggplot2')

ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```

What does this code do? Explain what each line is doing.

Answer: The install.packages command allows us to download and install packages from CRAN-like repositories or from local files. The library command allows one to add the functions installed from the package mentioned above into one's library so that s/he can use those functions when coding. Lastly, the ggplot command initializes a ggplot object. It can be used to "declare the input data frame for a graphic and the specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden" (R). Overall, this set of commands creates a graph showing the frequency of murder arrests in each state, with darker blues corresponding to higher frequencies.