Assignments

Contents

This page will contain all the assignments you submit for the class.

Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ```` ```{r} ``` ```` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

Assignment 1

Collaborators: Lorem Ipsum.

This assignment is due on Canvas on Monday 9/20 before class, at 10:15 am. Include the name of anyone with whom you collaborated at the top of the assignment.

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?

It is useful to rename datasets because it gives us a shorthand to work with. So in this case, instead of referring to the data with "USArrests" we can ref to it with dat.

```
dat <- USArrests
```

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`.

```
summary(dat)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
names(dat)
```

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

The four variables are "Murder", "Assault", "UrbanPop", and "Rape" (and the state variable which we created).

Problem 3

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

Answer: It is a quantitative variable because this variable is representing some numerical value in relation to a state.

What R Type of variable is it?

Answer: This variable is a character because the word murder itself is represented in a string format.

Problem 4

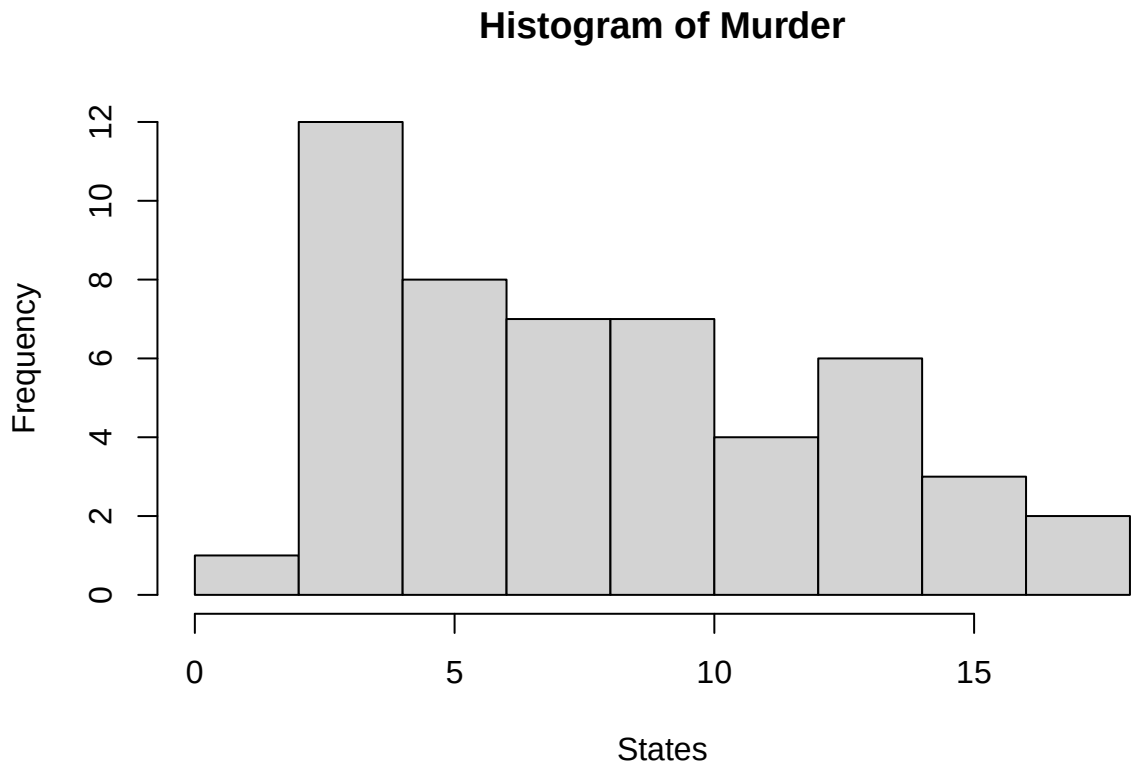
What information is contained in this dataset, in general? What do the numbers mean?

Answer: The dataset contains information about murder, assault, and rape. Additionally, it seems to give us some numbers for a states urban population to help see the relation aswell. These numbers show us the relationship with often they are occuring) of these different variables in different states. For example a number for murder is telling it there was some amount of murders within this state (and we can compare this to other states by seeing how much more or less these crimes occur in other states).

Problem 5

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

```
hist(dat$Murder, main = "Histogram of Murder", xlab = "States")
```



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
```

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

The mean is 7.788 and the median and 7.250. The mean is the average of the dataset while the median gives us a central value of our dataset. A quartile tells us the variability around the median. So the 1st and 3rd quartiles show us the variability before the median is reached and after the median is reached. R gives us this data to show us where it might be more or less skewed.

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.

Answer (for data on the other two variables) : For assaults, the mean is 170.8 and the median and 159.0.

For rapes, the mean is 21.23 and the median and 20.10.

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

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

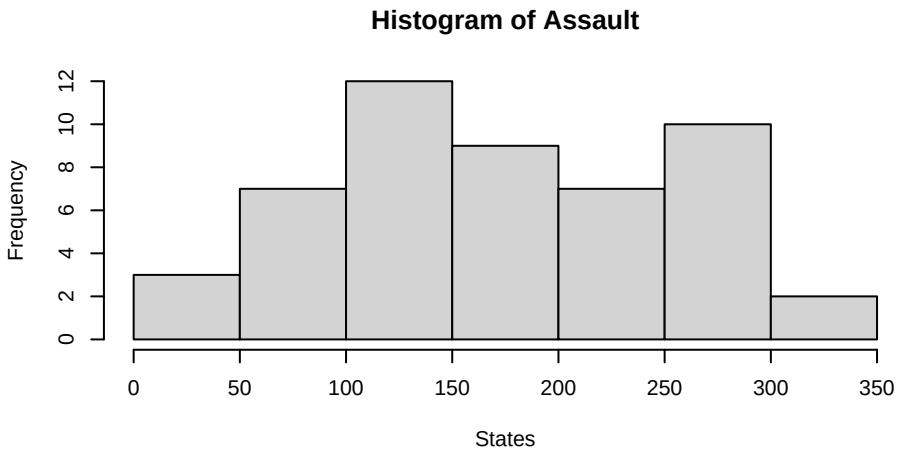
```
summary(dat$Assault)
```

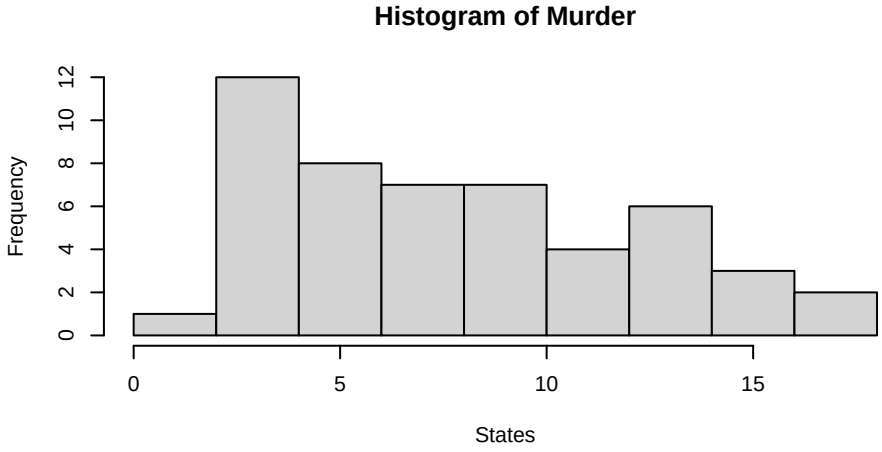
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

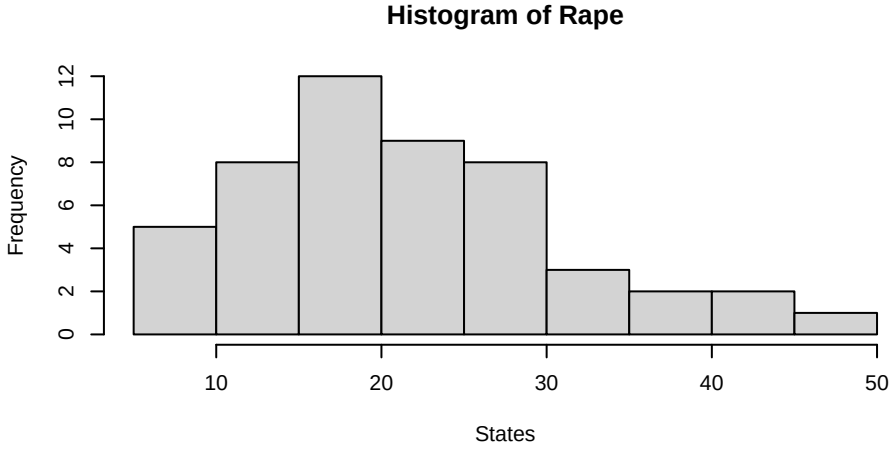
```
summary(dat$Rape)
```

```
##    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$Assault, main = "Histogram of Assault", xlab = "States")
hist(dat$Murder, main = "Histogram of Murder", xlab = "States")
hist(dat$Rape, main = "Histogram of Rape", xlab = "States")
```







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

Answer: It helps us combine multiple plots that we created into one big vertical plot.

What can you learn from plotting the histograms together?

Answer: We can see the correlation between the data. If there are some points where there is a peak at the

same time then we can generalize and say that state could be more dangerous than others. This could work likewise for the converse situation.

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:

```
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 first two lines import the libraries map and ggplot2. Line 154 imports dat wants to define the map id of states with murder. The second line is filling out map with state data.The third line maps it out in a x and y axis so we can see the data in states.