Assignments

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: Rachael Villari, Elizabeth Stoner, and Halle Wasser

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(knitr)
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
```

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"
```

Answer: The variables are Murder, Assault, UrbanPop, Rape, and state.

Problem 3

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

Answer: It is a quantitative/continuous variable.

What R Type of variable is it?

```
typeof("Murder")
```

```
## [1] "character"
```

Answer: It is a character data type.

Problem 4

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

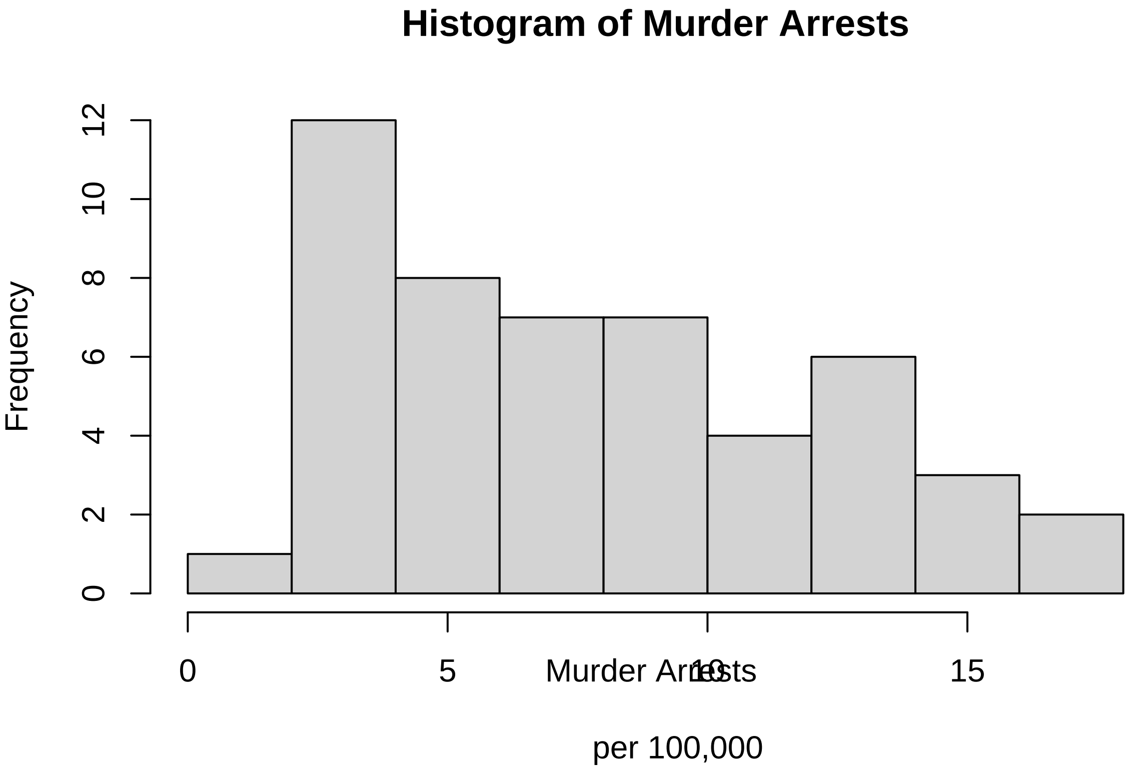
```
?USArrests
```

Answer: This dataset contains 50 observations/instances on 5 variables: Murder (murder arrests per 100,000), Assault (assault arrests per 100,000), UrbanPop (percent urban population), and Rape (rape arrests per 100,000), and state.

Problem 5

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

```
hist(dat$Murder, main="Histogram of Murder Arrests", xlab="Murder Arrests \n
     per 100,000", ylab="Frequency")
```



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
```

Answer: The mean number of murder arrests is 7.7888 per 100,000, and the median is 7.25 murder arrests per 100,000. According to HS, the mean and median are both measures of central tendency. However, the mean is described to be the value that each observation would receive if the values were redistributed among the dataset, and the mean on a histogram is the "balancing point". The median is the value that is at the center if all the values were put in an ordered list. According to HS, the interquartile range (IQR) is a measure of spread, and the "quartiles. . . are the three values which divide the distribution into even fourths. The IQR is obtained by subtracting the 1st quartile from the 3rd quartile which is why R gives us both these values.

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.

```
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)
```

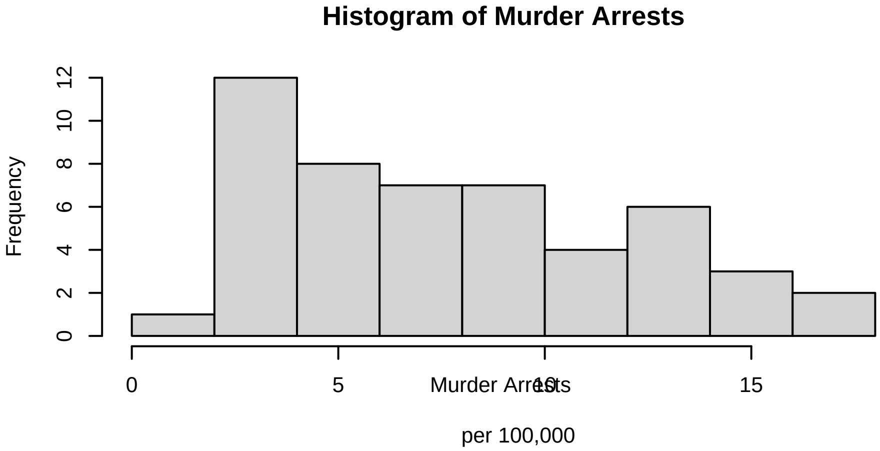
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    7.30   15.07   20.10   21.23   26.18   46.00
```

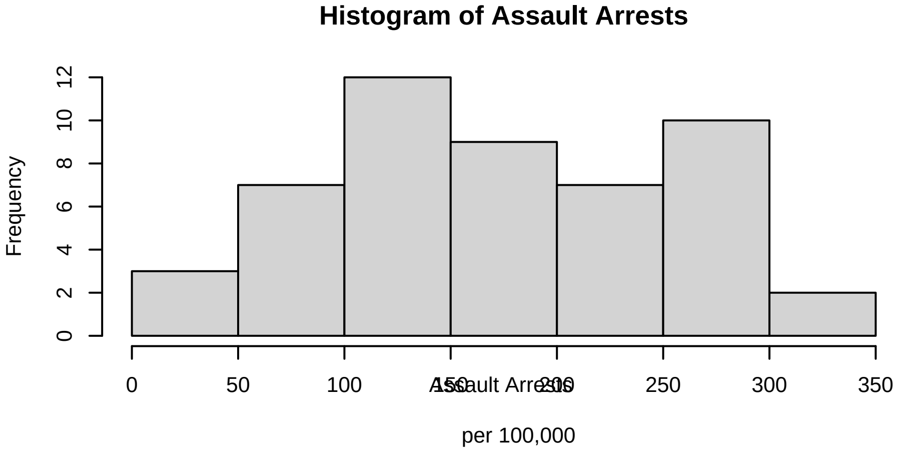
Answer: The mean number of assault arrests is 170.8 per 100,000, and the median is 159 assault arrests per 100,000. The mean number of rape arrests is 21.23 per 100,000, and the median is 20.1 rape arrests per 100,000.

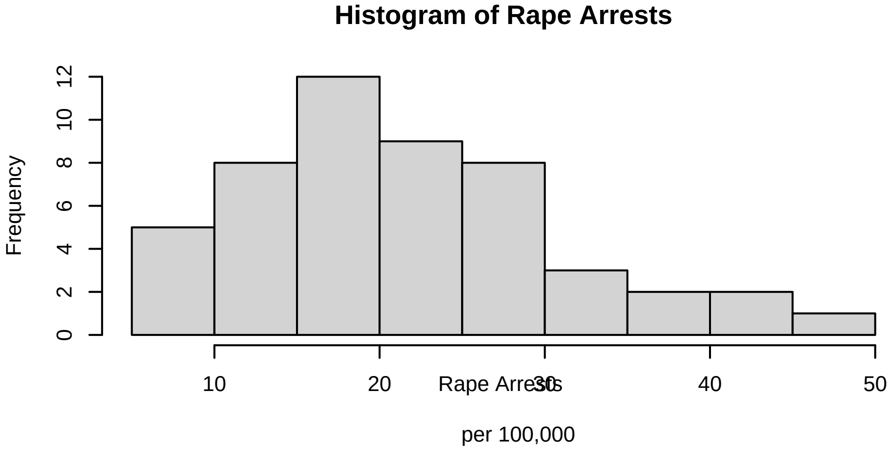
What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: There are a lot of graphical features in R. Par() allows us to search and set them.

```
par(mfrow=c(3,1))
hist(dat$Murder, main="Histogram of Murder Arrests", xlab="Murder Arrests \n
     per 100,000", ylab="Frequency")
hist(dat$Assault, main="Histogram of Assault Arrests", xlab="Assault Arrests \n
     per 100,000", ylab="Frequency")
hist(dat$Rape, main="Histogram of Rape Arrests", xlab="Rape Arrests \n
     per 100,000", ylab="Frequency")
```







What can you learn from plotting the histograms together?

Answer: Murder and rape arrests are right skewed, which is good. What I thought was interesting were how the assault arrests are pretty normally distributed. Additionally, I see the average assault arrests per 100,000 were much greater than the average murder or rape arrests per 100,000 which makes sense.

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') # load maps library into R session
library('ggplot2') # load ggplot2 into R session

ggplot(dat, aes(map_id=state, fill=Murder)) + # call ggplot function, use dat
                                              # dataframe, and set up plot aesthetics
  geom_map(map=map_data("state")) + # create map of all states
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat) # expand plot limits
```

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

Answer: This code is creating a heat map of the murder arrests per 100,000 in each state. See comments for a description of each line of code.

Assignment 2

(Coming soon)