Assignments

Table of Contents

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

### Instructions for all assignments

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: Sara Whitelaw**

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)  
#This command loads the data set.

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  
#This command renames our data set "USArrests" as dat. Renaming USArrests as dat makes the name shorter and easier to use.

### Problem 2

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

dat$state <- tolower(rownames(USArrests))  
#This command renames the state names into a variable called state.

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"

This command is telling me what the different variables in the USArrests data set are.

### Problem 3

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

Answer: ‘Murder’ is a quantitative variable.

What R Type of variable is it?

Answer:In R, ‘Murder’ is a numerical variable, which means it is a quantitative variable that is expressed with decimal points.

### Problem 4

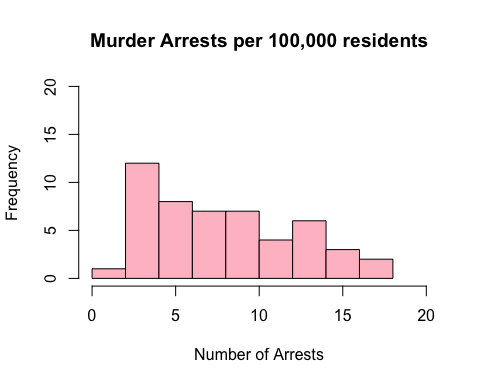
What information is contained in this dataset, in general? What do the numbers mean?

Answer: This data set contains information about different arrest rates in the United States. It contains statistics for the arrests in every state for three crimes: rape, assault, and murder. The numbers are the amount of arrests per crime per 100,000 residents in that state. The urban pop variable is the percent of residents living in an urban setting.

### Problem 5

Draw a histogram of Murder with proper labels and title.

hist(dat$Murder, main='Murder Arrests per 100,000 residents', xlab="Number of Arrests", ylab="Frequency", xlim=c(0,20), ylim=c(0,20),col="pink",breaks=10)



### 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)

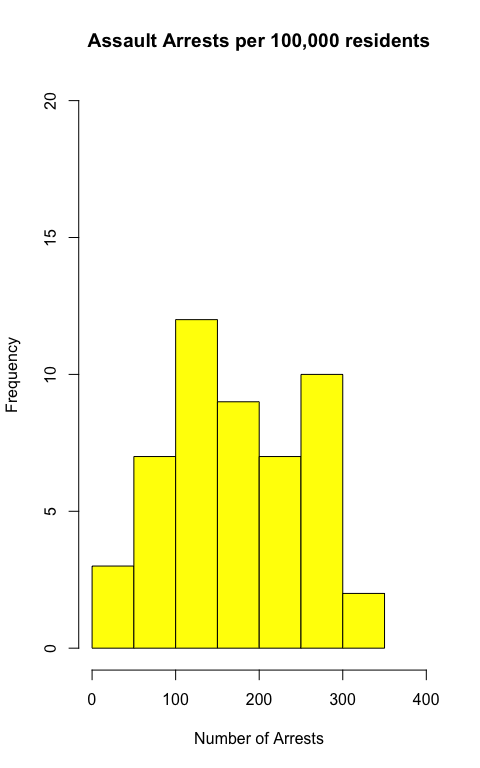
## Murder Assault UrbanPop Rape   
## Min. : 0.800 Min. : 45.0 Min. :32.00 Min. : 7.30   
## 1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50 1st Qu.:15.07   
## Median : 7.250 Median :159.0 Median :66.00 Median :20.10   
## Mean : 7.788 Mean :170.8 Mean :65.54 Mean :21.23   
## 3rd Qu.:11.250 3rd Qu.:249.0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

The average number of murder arrests per 100,000 residents is 7.788. The median for this data set is 7.250. Mean is the average value of a data set. Mean is equal to the sum of values divided by the number of values. Mean = sum of values/number of values. The Median is the middle value of a number set after all of the values have been put in an ordered list. Quartiles are three values that divide the data distribution into even fourths. In particular, Q1 and Q3 are important because we can use these two values to calculate the IQR, or interquartile range. The middle half of the data lies within the IQR.

### Problem 7

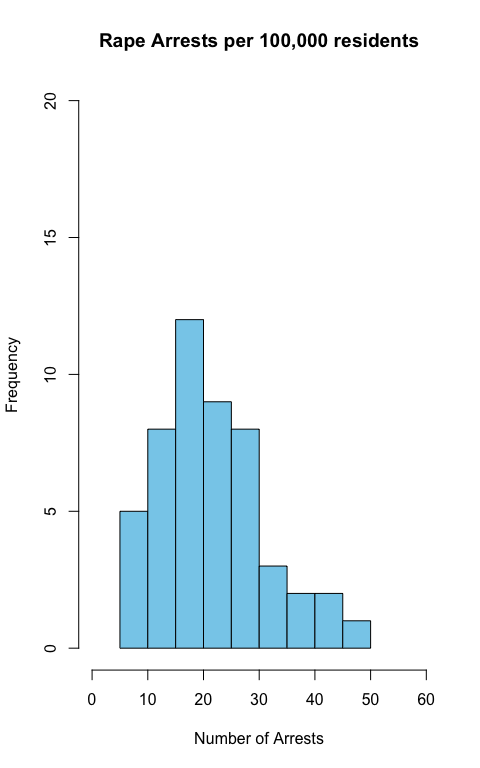
Repeat the same steps you followed for Murder, for the variables Assault and Rape. Now plot all three histograms together. You can do this by using the command par(mfrow=c(3,1)) and then plotting each of the three.

hist(dat$Assault, main='Assault Arrests per 100,000 residents', xlab="Number of Arrests", ylab="Frequency", col="yellow", xlim=c(0,400), ylim=c(0,20),)



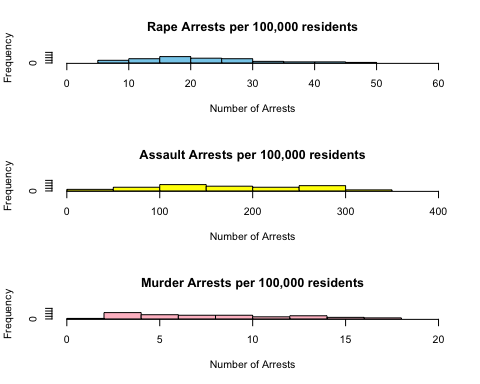
#The average number of assault arrests per 100,000 residents is 170.8. The median for this data set is 159.0.

#Histogram of Rape Arrests per 100,000 Residents   
hist(dat$Rape, main='Rape Arrests per 100,000 residents', xlab="Number of Arrests", ylab="Frequency", col="sky blue", xlim=c(0,60), ylim=c(0,20))



#The average number of rape arrests per 100,000 residents is 21.23 and the median for this data set is 20.10.

par(mfrow=c(3,1))  
hist(dat$Rape, main='Rape Arrests per 100,000 residents', xlab="Number of Arrests", ylab="Frequency", col="sky blue", xlim=c(0,60), ylim=c(0,20))  
hist(dat$Assault, main='Assault Arrests per 100,000 residents', xlab="Number of Arrests", ylab="Frequency", col="yellow", xlim=c(0,400), ylim=c(0,20),)  
hist(dat$Murder, main='Murder Arrests per 100,000 residents', xlab="Number of Arrests", ylab="Frequency", xlim=c(0,20), ylim=c(0,20),col="pink",breaks=10)



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

Answer:The command par is used to to set graphical parameters. In this case, it can be used to put multiple graphs on a single plot.

What can you learn from plotting the histograms together?

Answer: From plotting the histograms together, I can compare frequency of the arrest rates for these three crimes. I can see that the number of assault arrests per 100,000 residents has a much wider spread than rape arrests and murder arrests. The number of murder and rape arrests per 100,000 residents has a much smaller spread. This shows that on average, there are more assault arrests per 100,000 residents in the United States than murder or rape arrests. Looking at histograms graphed together, we can visually compare how different, or similar, the statistical characteristics, such as mean, median, spread, of the different data sets are.

### 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.

#install.packages("ggplot2")  
#install.packages("maps")  
#'maps' and 'ggplot2' were already downloaded, so I was getting an error message when I had the code in for it.

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 library code loads the desired packages. In this case, it is loading the packages called ‘map’ and ‘ggplot2’. The combination of the ‘map’ package and the ‘ggplot2’ package allows for the display of maps. The next line is giving the code for the map that we want to generate of the United States. The quoting function aes is for aesthetic mappings and the code that follows is describing what is making up the aesthetic mapping. We are pulling data from the USArrests data set, which was shortened to dat. The fill=murder command is telling R that we want the states to be shaded in according to the“Murder” data set. The expand\_limits code describes the x and y axis by longitude and latitude.

$$\\[2in]$$