Assignments

Contents

Assignment 1 **1**

Assignment 2 **6**

Assignment 1

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
```

Problem 2

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

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

murder, assault, urbanpop, rape, and state. The last, state, is notably excluded from the knitted pdf yet is featured in the console.

```
names(dat)
```

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

Problem 3

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

Answer: According to the DVB reading, murder would be a quantitative variable within this particular context.

What R Type of variable is it?

Answer: Murder is a UNIVARIATE NON-GRAPHICAL variable in R.

Problem 4

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

Answer: The dataset's variables compare the incidents or rates of murder, assault, and rape along state lines, as well as those of urban population sizes. Given the tenths decimal places featured in the categories of murder and rape, it is safe to assume those numbers present rates of reported incidents as neither murder nor rape can occur to a tenth degree. Assault, on the contrary, appears to be reported as raw, reported incidents given the data is presented as whole integers. The state category lists the various American states, adjacent to each states' urban population. As a note, the urban population data is presented seemingly as either proportional integers or on a scale; to illustrate, any one state's urban population data may identify what percentage of the state's population lives in an urban center.

Problem 5

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

```
hist(dat$Murder, main="Frequency of State Murder Rates",
     xlab="Murder Rate Per [X] People", ylab="Number of States",
     xlim=c(0,20), col="maroon", breaks=10, xaxp=c(0,20,10))
```

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)
```

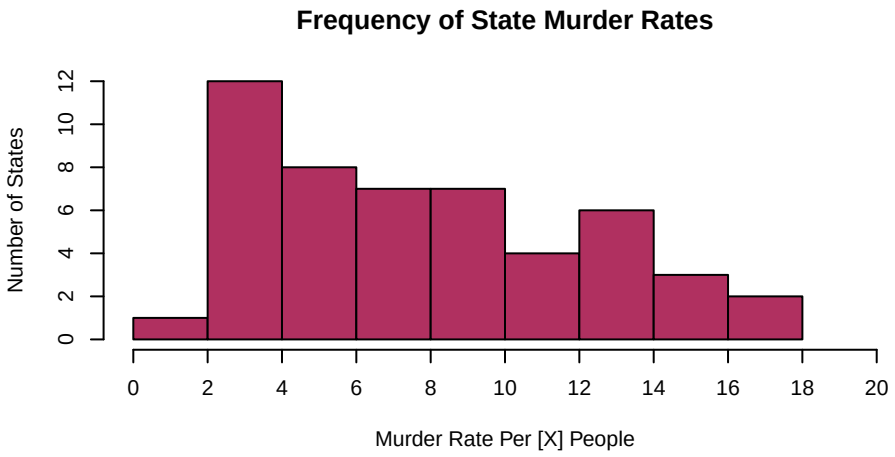
```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   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
##     state
##  Length:50
##  Class :character
##  Mode  :character
##
##
##
```

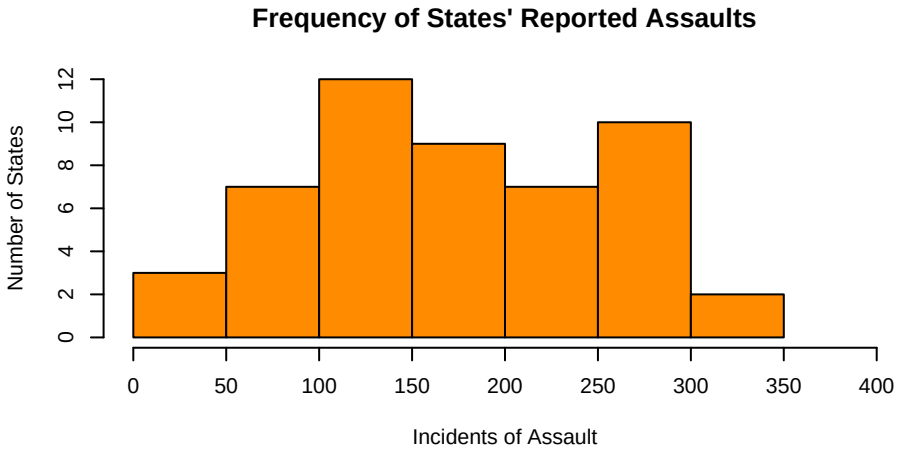
The mean for murder is 7.788, which is the average of all state murder rates; the median of 7.25 is in contrast the figure at the middle of the string of variables if sequenced. The quartiles are the percentiles of the data. Combined, both the 1st and 3rd quartile depict where half the data lies betwee. That is, 25% of states have murder rates below 4.075, the 1st Quartile, while 75% of states have murder rates are less than 11.25. Likewise, 75% of states have murder rates higher than 4.075 and 25% of states have a murder rate higher than 11.25; 50% of states would thus have a murder rate between 4.075 and 11.25. R must provide the percentiles or quartiles to depict where half of the data lies between, presenting a numerical bell curve of sorts. *The .rmd file notably produces a summary of the state variable as well, with a length of 50 and a class and mode of character. The calculation, however, is cut out in the knitted pdf document.

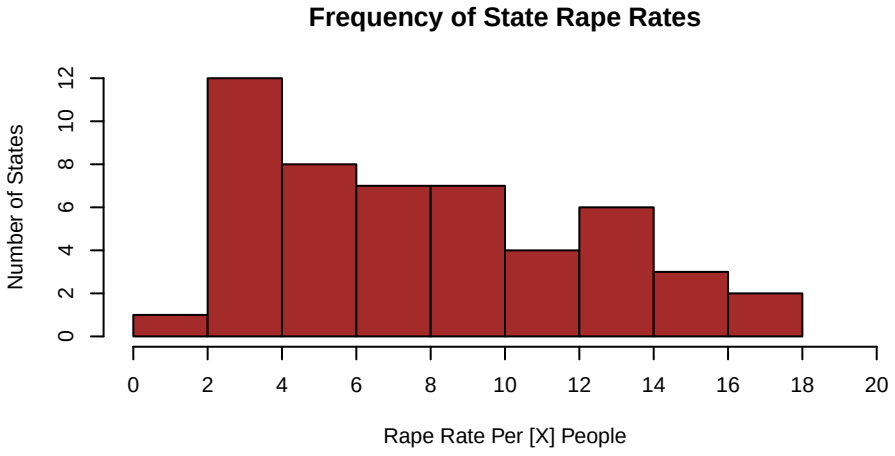
Problem 7

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.

```
par(mfrow=c(3,1))
hist(dat$Murder, main="Frequency of State Murder Rates",
     xlab="Murder Rate Per [X] People", ylab="Number of States",
     xlim=c(0,20), col="maroon", breaks=10, xaxp=c(0,20,10))
hist(dat$Assault, main="Frequency of States' Reported Assaults",
     xlab="Incidents of Assault", ylab="Number of States",
     xlim=c(0,400), col="darkorange", breaks=10, xaxp=c(0,400,8))
hist(dat$Murder, main="Frequency of State Rape Rates",
     xlab="Rape Rate Per [X] People", ylab="Number of States",
     xlim=c(0,20), col="brown", breaks=10, xaxp=c(0,20,10))
```







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

Answer: Command par allows you define the parameters of a graph or computation. In this specific case, we set the parameters to the entirety of three graphs, which are all included in one plot.

What can you learn from plotting the histograms together?

Answer: By plotting all three plots adjacent of one another, the combined graphic allows for a quick assessment of the number of states with varying assault, murder, and rape rates. The graph thus allows you for a

quick comparison of central tendency, spread, skewness, and kurtosis for the various crimes.

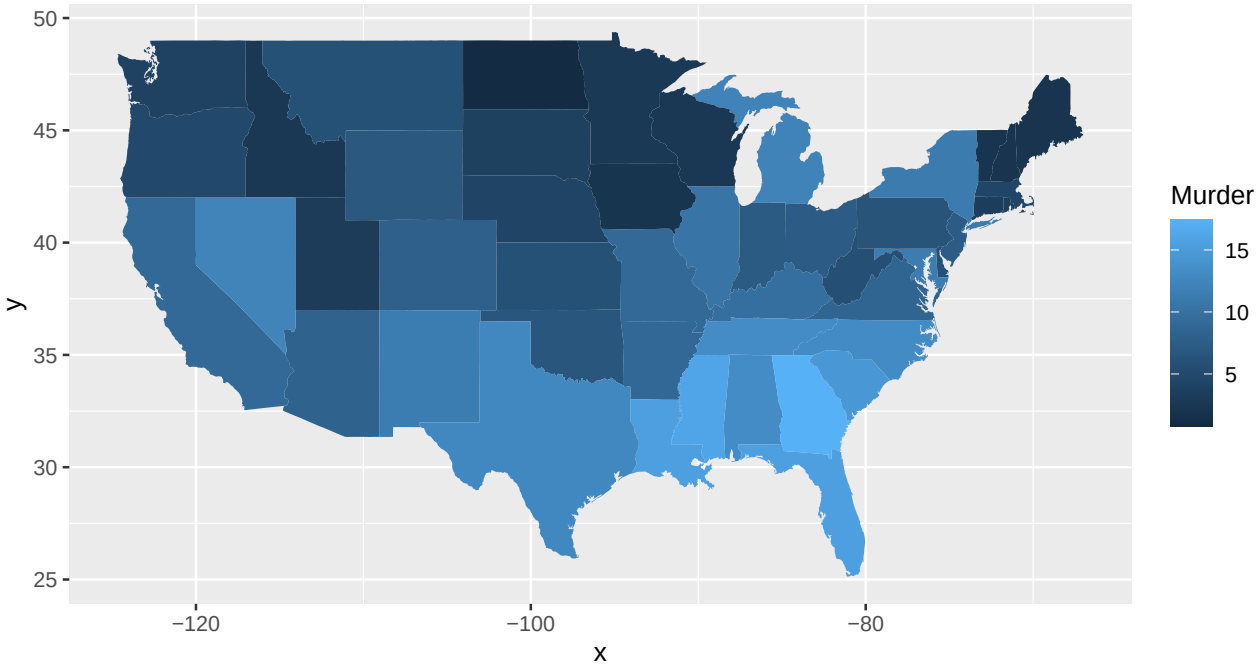
Problem 8

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:

```
# install.packages("maps")
# install.packages("ggplot2")
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 code generates a map of the United States where each state is shaded according to their murder rate. States with darker shadings have lower murder rates as those with ligher shades have higher murder rates. The install commands download the graphed data, that is they download a map and hold the manipulated data. The ggplot command then creates the graph by loading the "dat" variable, which was predefined as the US Arrest data. The map is then divided along state lines and shaded according to the murder rate, per the map_id and fill command respectively. The geom_map command allows for additional manipulation to the states' borders, defining each state alongside existing state boundaries. Finally, the expand_limits command expands the graph by manipulating the x and y axes.

Assignment 2

(Coming soon)