Assignments

Contents

This page contains all of my assignments for the class.

Assignment 1

Collaborators: Tori Borlase and Halle Wasser.

Problem 1

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

```
# install.packages('datasets')
library(datasets)
```

Now, it's installed!

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
```

Answer: It is useful to rename the dataset because it allows for replication. Renaming the dataset serves essentially the same function as "Save as" on a word document, which allows you to save the current version separately and still have access to previous versions. In other words, renaming the dataset allows you to perform functions without contaminating the original data.

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"     "state"
```

Answer: The variables contained in the dataset are Murder, Assualt, UrbanPop (Urban Population), Rape, and state.

Problem 3

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

Answer: According to the DVB chapter, Murder is a quantitative variable because it measures how many murder arrests were made per 100,000 people

What R Type of variable is it?

```
class(dat$Murder)
```

```
## [1] "numeric"
```

Answer: The typeof() function just tells us what type of object is within the parenthesis. Therefore, although the R output states that "Murder" is a character, it is actually a numeric variable

Problem 4

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

?USArrests

```
head(dat)
```

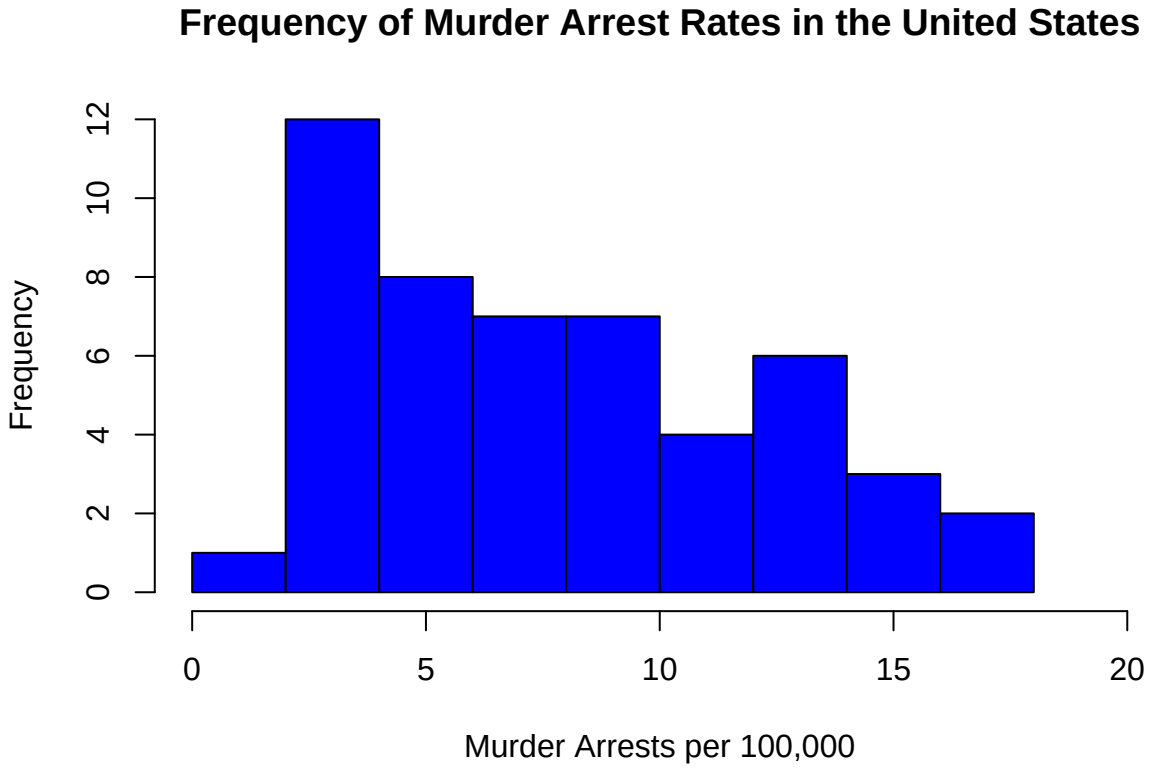
```
##            Murder Assault UrbanPop Rape       state
## Alabama      13.2     236       58 21.2     alabama
## Alaska       10.0     263       48 44.5      alaska
## Arizona       8.1     294       80 31.0     arizona
## Arkansas      8.8     190       50 19.5    arkansas
## California    9.0     276       91 40.6  california
## Colorado      7.9     204       78 38.7    colorado
```

Answer: In general, information about how many arrests were made for murder, assault, and rape per 100,000 people in each state is contained in this dataset. This dataset also contains the percent urban population for each state.

Problem 5

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

```
hist(dat$Murder, main = "Frequency of Murder Arrest Rates in the United States",
    xlab = "Murder Arrests per 100,000", border = "black", col = "blue",
    xlim = c(0, 20))
```



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.?

```
mean(dat$Murder)
```

```
## [1] 7.788
```

```
median(dat$Murder)
```

```
## [1] 7.25
```

```
quantile(dat$Murder)
```

```
##     0%    25%    50%    75%   100%
##  0.800  4.075  7.250 11.250 17.400
```

Answer: The mean and median of "Murder" are 7.788 and 7.25, respectively. The mean is the average of all the data points (i.e. the sum of all the data values divided by the number of values). The median is the middle value when the data is arranged in order and it provides information regarding robustness. A quartile is a type of quantile that divides the data set into 4 (roughly) equally-sized parts (i.e. quarters) when the data is arranged from smallest to largest value. Therefore, quartiles serve to measure the spread of values above and below the mean. R gives us the 1st and 3rd quartiles because the 2nd quartile is the same as the median. As you can see in the output above, the value associated with 50% is the same as the median (because 50% of the values in the data set are above this value).

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.

```r
# Assault
mean(dat$Assault)
```

```
## [1] 170.76
```

```r
median(dat$Assault)
```

```
## [1] 159
```

```r
quantile(dat$Assault)
```

```
##   0%  25%  50%  75% 100%
##   45  109  159  249  337
```

```r
# Rape
mean(dat$Rape)
```

```
## [1] 21.232
```

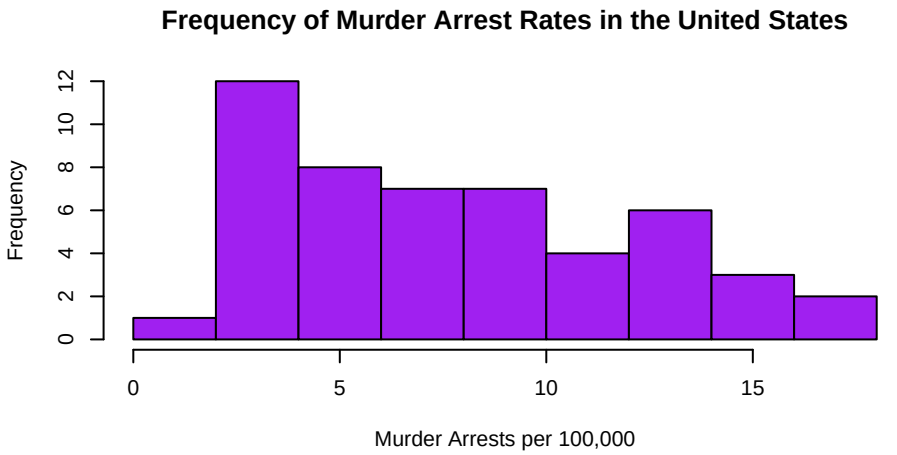
```r
median(dat$Rape)
```

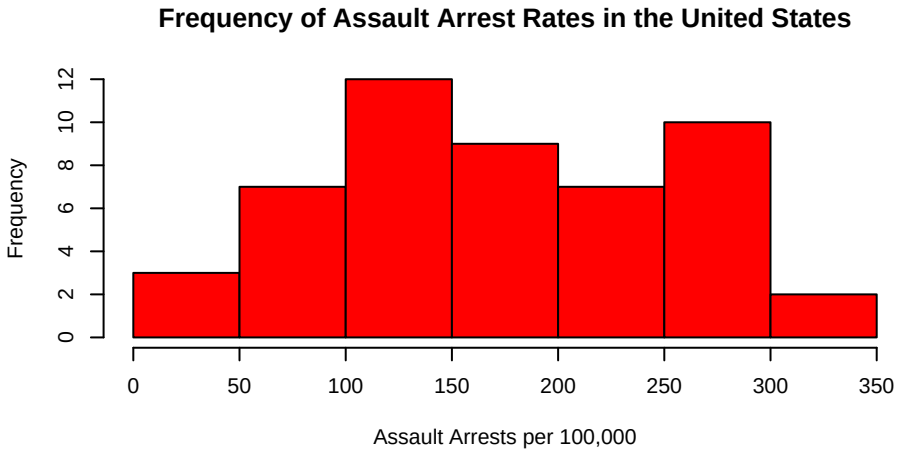
```
## [1] 20.1
```

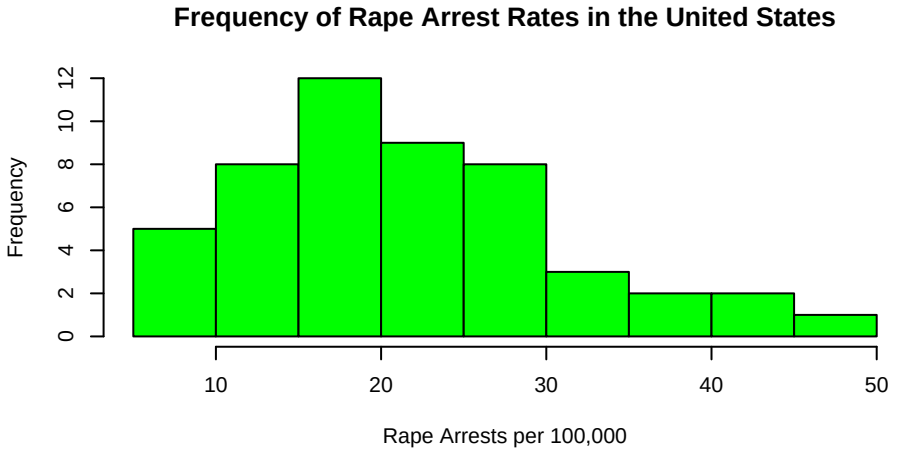
```r
quantile(dat$Rape)
```

```
##     0%    25%    50%    75%   100%
##  7.300 15.075 20.100 26.175 46.000
```

```r
par(mfrow = c(3, 1))
hist(dat$Murder, main = "Frequency of Murder Arrest Rates in the United States",
    xlab = "Murder Arrests per 100,000", border = "black", col = "Purple")
hist(dat$Assault, main = "Frequency of Assault Arrest Rates in the United States",
    xlab = "Assault Arrests per 100,000", border = "black", col = "Red")
hist(dat$Rape, main = "Frequency of Rape Arrest Rates in the United States",
    xlab = "Rape Arrests per 100,000", border = "black", col = "Green")
```







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

Answer: The R command 'par' can be used to combine plots into one cohesive graph.

What can you learn from plotting the histograms together?

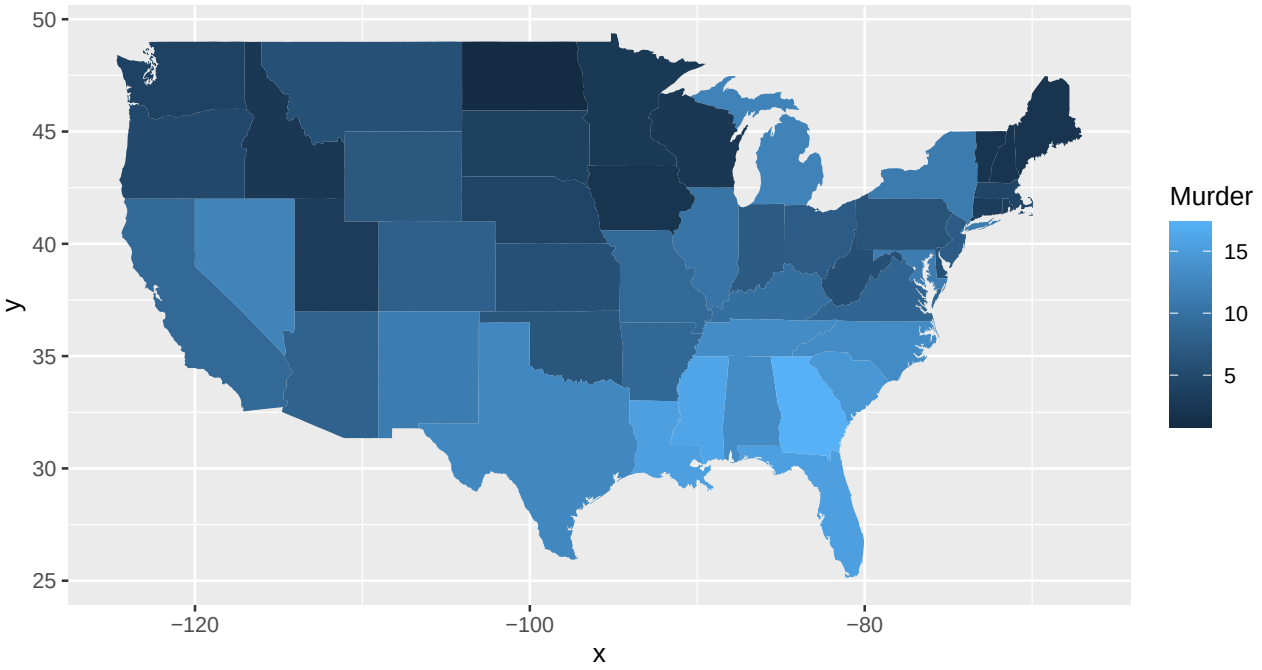
Answer: By plotting the histograms together, we can learn that histograms for Rape and Murder are (relatively) right-skewed whereas the histogram for Assault is closer to a bimodal distribution.

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') this installs the maps package,
# which allows the program to draw geographical maps.
# install.packages('ggplot2') this installs the ggplot2
# program, which allows the program to create elegant data
# visualizations using the Grammar of Graphics. In other
# words, it allows for very complex plots/graphics to be
# created within a data frame.
library(maps)
# this line loads the newly-installed 'maps' package
library(ggplot2)
# this line loads the newly-installed 'ggplot2' package
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)
```




```
# this long line of code reflects the 3 fundamental parts
# of the ggplot: data, aesthetics, and geometry. It tells
# the program to use the Murder data from dat and the new
# 'state' variable that we create in Problem 2. It also
# indicates the x and y variables such that they are
# instead reflecting latitude and longitude.
```

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

Answer is commented out above

Assignment 2

(Coming soon)