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))
```

Now, the state names have become a new 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"     "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: By using the class() function, we find that Murder is a numeric variable in R.

Problem 4

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

```
head(dat)
```

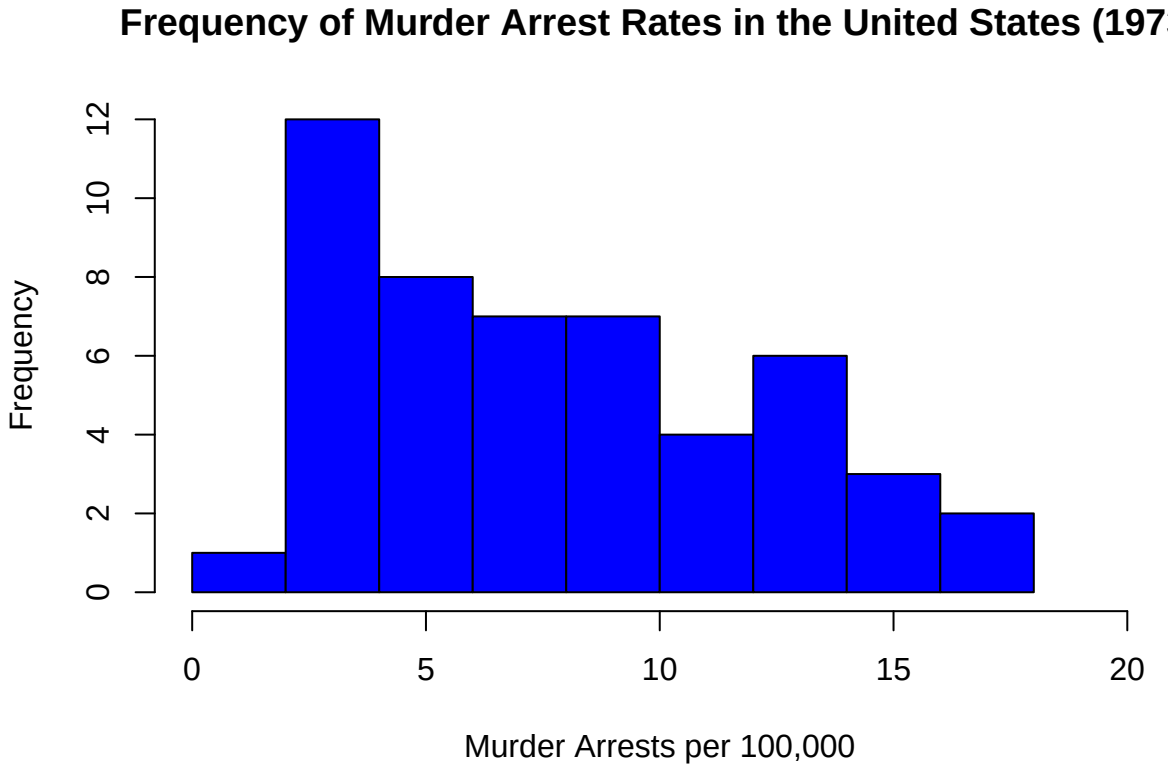
```
##            Murder Assault UrbanPop Rape      state
## Alabama      13.2     236       58 21.2    alabama
## Alaska       10.0     263       48 44.5     alaska
## Arizona       8.1     294       80 31.0    arizona
## Arkansas      8.8     190       50 19.5   arkansas
## California    9.0     276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

Answer: Firstly, it's important to note that this data was collected/released in 1973. 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 percentages of the population that reside in urban areas within each state. Although this data set is just a data set that comes with R and is generally used for practice, it teaches beginners to be sensitive about the analyses that they conduct using R. In this case specifically, the data points are not just numbers; rather, they represent people in the real world that were arrested for specific crimes. Thus, this data teaches us to be cognizant of the conclusions and inferences that are drawn from our calculations/analyses.

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 (1973)",
    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.?

```
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 and median of "Murder" are 7.788 and 7.250, respectively. The minimum value is .800, Q1 is 4.075, Q3 is 11.250, and the maximum is 17.400. 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 in a given data set (these quartiles include 25%, 50%, 75% and 100%). R only 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 [and below] 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.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    45.0   109.0   159.0   170.8   249.0   337.0
```

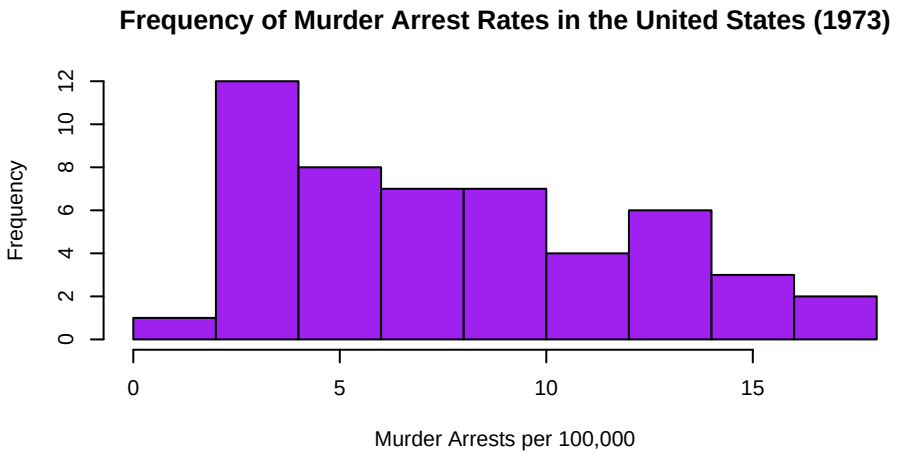
For assault arrests, the minimum value is 45.0, the maximum value is 337.0, the median is 159.0, and the mean is 170.8. Furthermore, Q1 is 109.0 and Q3 is 249.0.

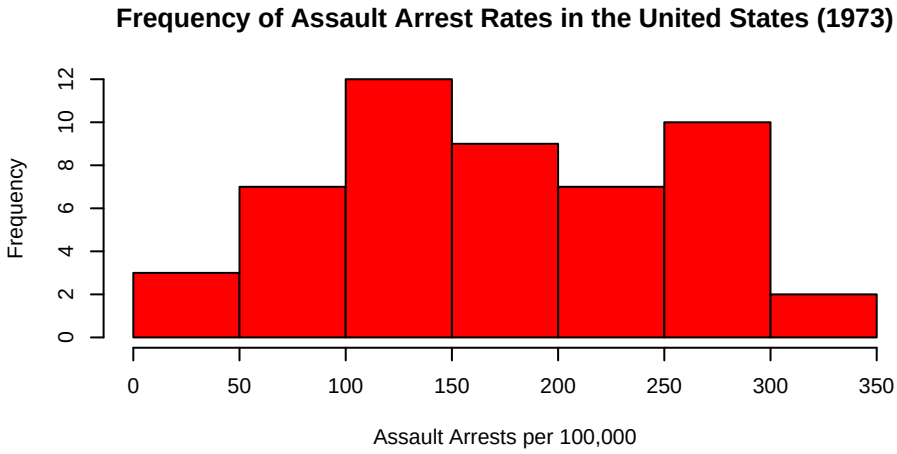
```
# Rape
summary(dat$Rape)
```

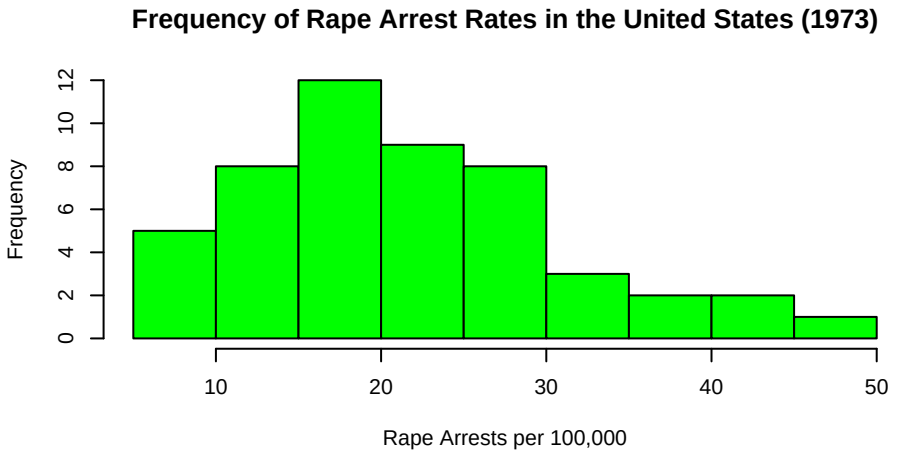
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

For rape arrests, the minimum value is 7.30, the maximum value is 46.00, the median is 20.10, and the mean is 21.23. Furthermore, Q1 is 15.07 and Q3 is 26.18.

```
par(mfrow = c(3, 1))
hist(dat$Murder, main = "Frequency of Murder Arrest Rates in the United States (1973)",
    xlab = "Murder Arrests per 100,000", border = "black", col = "Purple")
hist(dat$Assault, main = "Frequency of Assault Arrest Rates in the United States (1973)",
    xlab = "Assault Arrests per 100,000", border = "black", col = "Red")
hist(dat$Rape, main = "Frequency of Rape Arrest Rates in the United States (1973)",
    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' is used to modify how graphs are displayed, allowing plots to be combined into one cohesive graph. For specifically, 'par' gives the program parameters to plot it a certain way and then 'mfrow' puts it into an array. In this case, the array has 3 rows and 1 column, as denoted in the line of code.

What can you learn from plotting the histograms together?

Answer: By plotting the histograms together, we can learn quite a few things because the distributional differences are displayed very clearly. Firstly, we notice that the number of assault arrests per 100,000 people is significantly higher than both that of murder arrests and rape arrests. Furthermore, we notice that histograms for Rape Arrests and Murder Arrests are unimodal and right-skewed (albeit to different degrees) whereas the histogram for Assault Arrests 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.

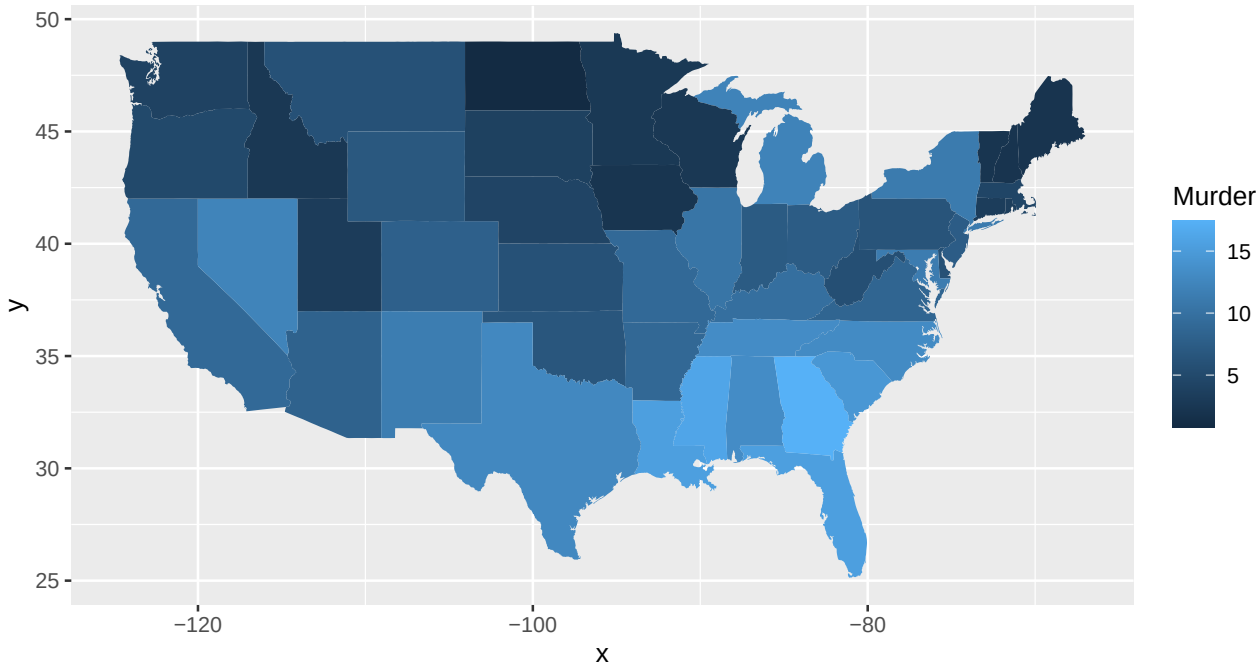
Run this code:

```
# install.packages('maps')

# 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 is shown below. (Unfortunately, it would not allow me to fix the display of the comments next to each line, so if there is difficulty understanding which portion is code and which portion is my explanation, please let me know).

```
# 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. The first
# parameter tells the program to use the 'dat' dataset. It
# then tells the program to use the Murder data from 'dat'
# and the new 'state' variable that we create in Problem 2
# to create the aesthetic element of the plot. Together, it
# uses essentially signals that the mapping aesthetic
# layout should be based on the state variable while the
# color of each state (i.e. the respective shade of blue)
# is determined by the value associated with the number of
# murder arrests. The last part of this line of code
# indicates that the x and y limits of the plot should
# reflect the latitude and longitude values of the states.
# Collectively, each portion of this line of code plays a
# role in displaying a colorful map of the United States,
# whether different shades of blue represent different
# numbers of Murder arrests across the United States in
# 1973.
```

Assignment 2

(Coming soon)