Assignments

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

Assignment 1

Collaborators: Theodora Athanitis 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)
```

Answer: 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: Renaming data sets ensures that you do not edit any of your original data on accident, and it is also easier to type.

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 are Murder, Assault, Urban Population, and Rape, in addition to the new variable that we just created, state.

Problem 3

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

Answer: Murder is a quantitative variable according to the DVB chapter because we are measuring the amount of murder arrests per 100,000.

What R Type of variable is it?

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

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

Answer: In R, the data points in the Murder column are numeric variables according to the class() function.

Problem 4

What information is contained in this data set, 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
```

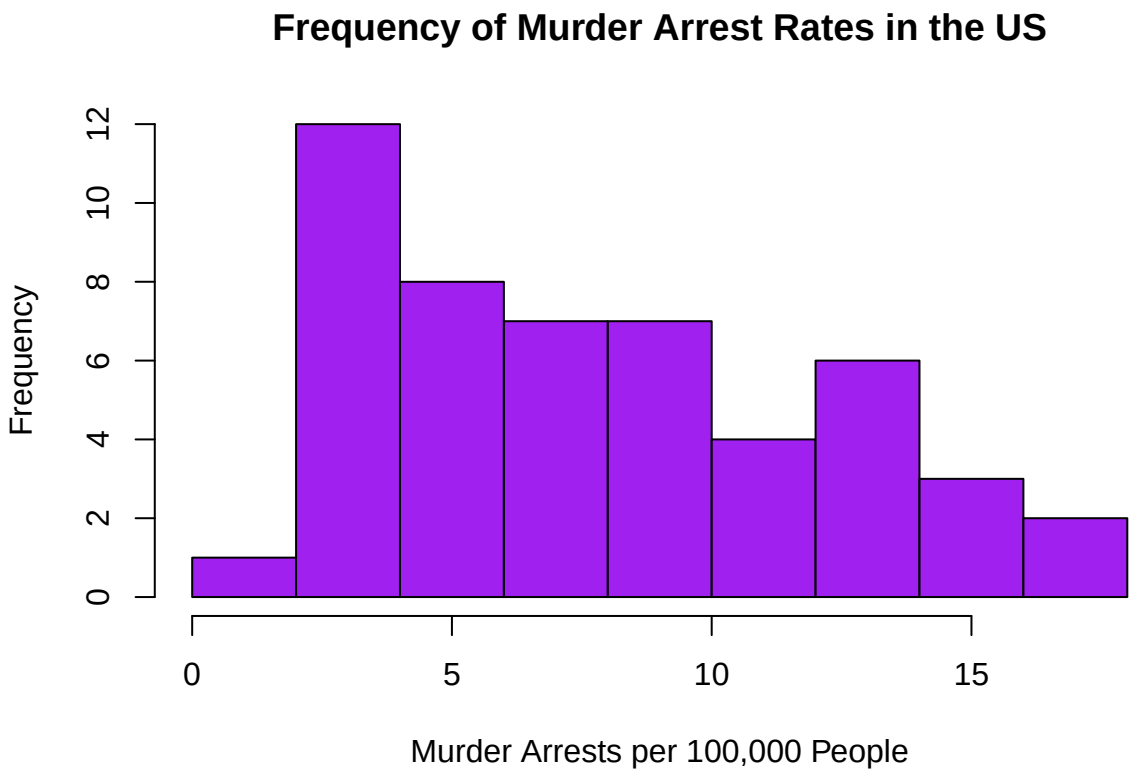
```
?USArrests
```

Answer: Generally, this data set contains rates for specific crime arrests per 100,000 people collected by D.R. McNeil, as well as percentages of urban population per state in 1973. The numbers in the Murder, Assault, and Rape columns are equivalent to the calculated arrest rates for Murder, Assault, and Rape. The Urban Population numbers are the percentages of the population in each state that live in an urban area.

Problem 5

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

```
hist(dat$Murder, xlab = "Murder Arrests per 100,000 People", main = "Frequency of Murder Arrest Rates i
```



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)
```

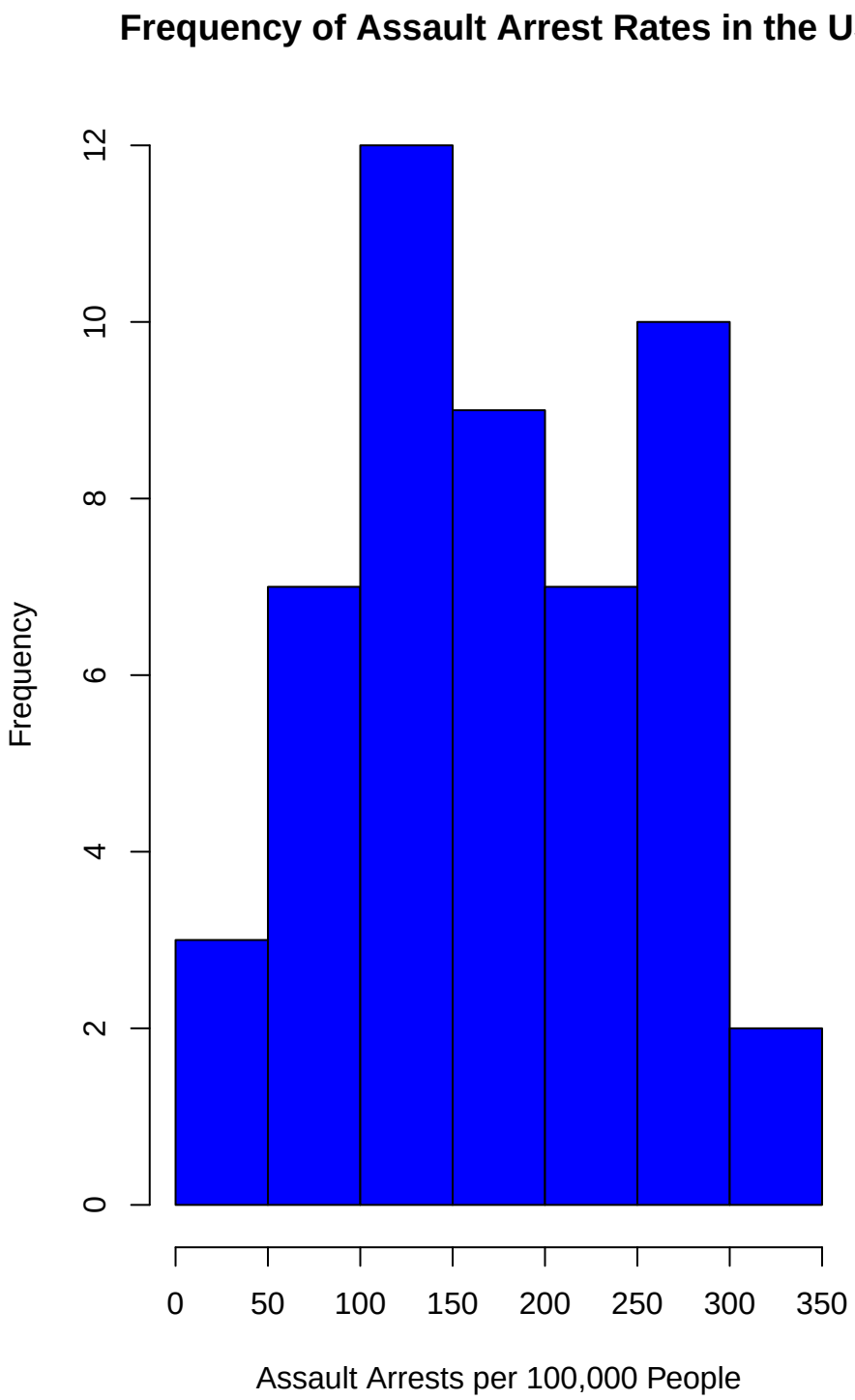
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

Answer: The mean is 7.788. The median is 7.250. Q1 is 4.075 and Q3 is 11.250. The minimum value is .800 and the maximum value is 17.400. Mean and median are both measures of center; however, the mean is the average value of a data set, which is calculated by adding all of the values and dividing by the number of total data points, and the mean is calculated by finding the middle value when the data set is arranged in order. A quartile is essentially the same as a median, but with 25% and 75% of the data, rather than 50%. This means that 25% of the data in a data set is below the 1st quartile, 50% is below the second quartile (the median), and 75% is below the 3rd quartile. R gives you quartiles in order to help visualize the data and see if it is distributed symmetrically, and only gives you quartiles 1 and 3 because the median is the 2nd quartile.

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, xlab = "Assault Arrests per 100,000 People", main = "Frequency of Assault Arrest Rate
```

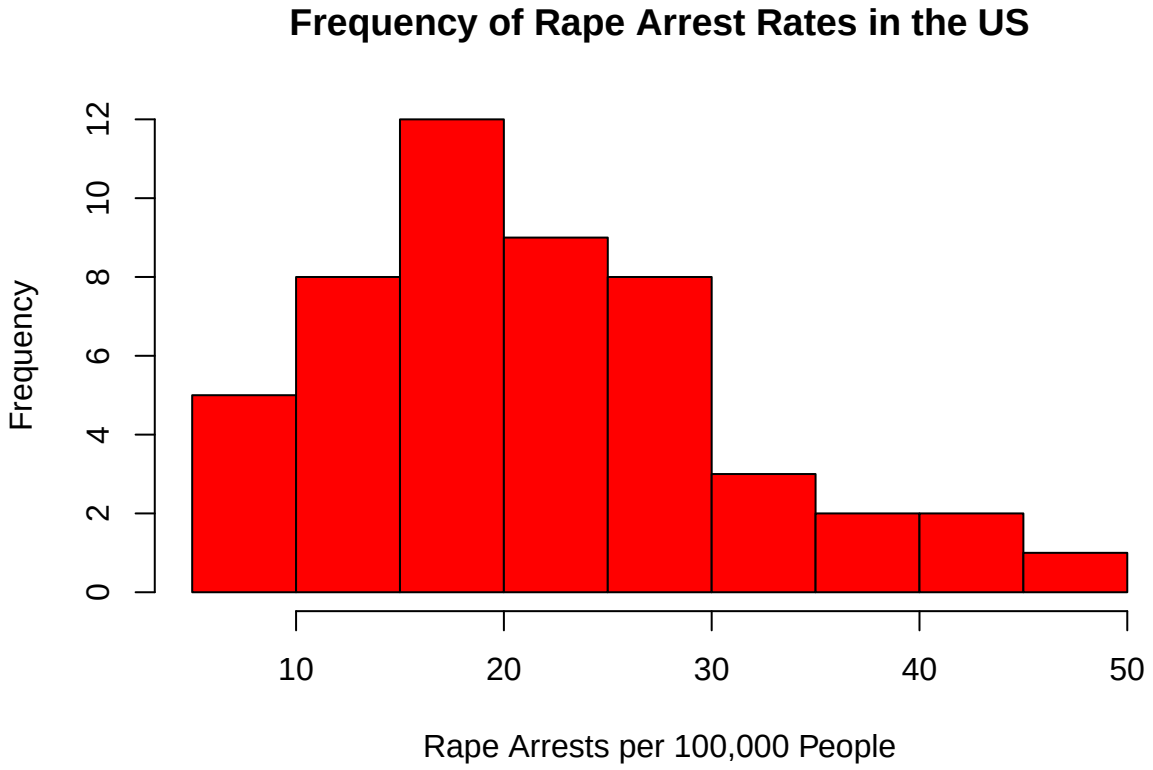



```
summary(dat$Assault)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

Answer: The mean is 170.8. The median is 159.0. Q1 is 109.0 and Q3 is 249.0. The minimum value is 45.0 and the maximum value is 337.0.

```
hist(dat$Rape, xlab = "Rape Arrests per 100,000 People", main = "Frequency of Rape Arrest Rates in the U
```

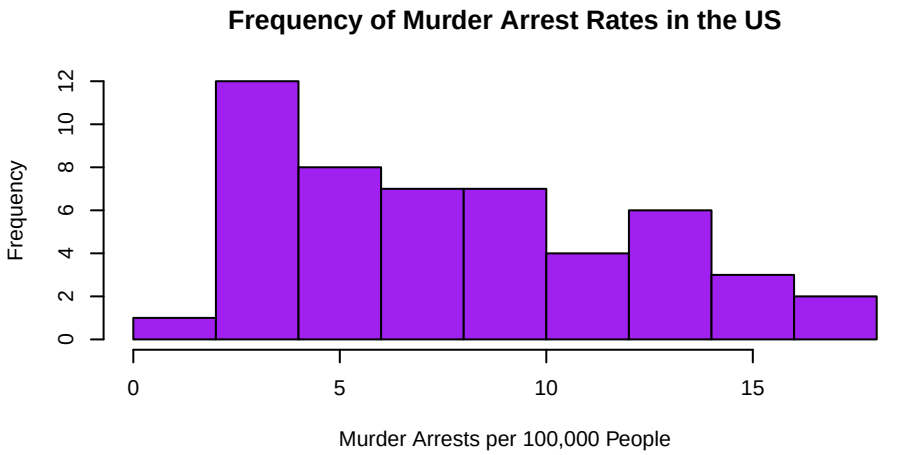


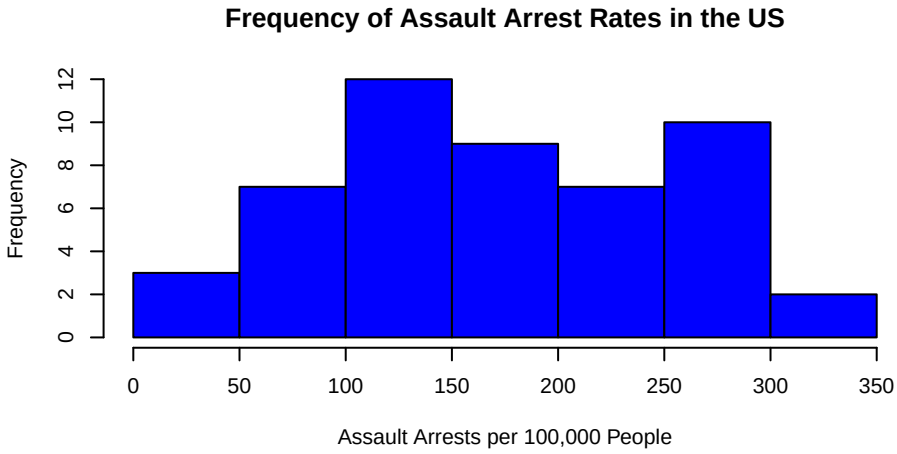

```
summary(dat$Rape)
```

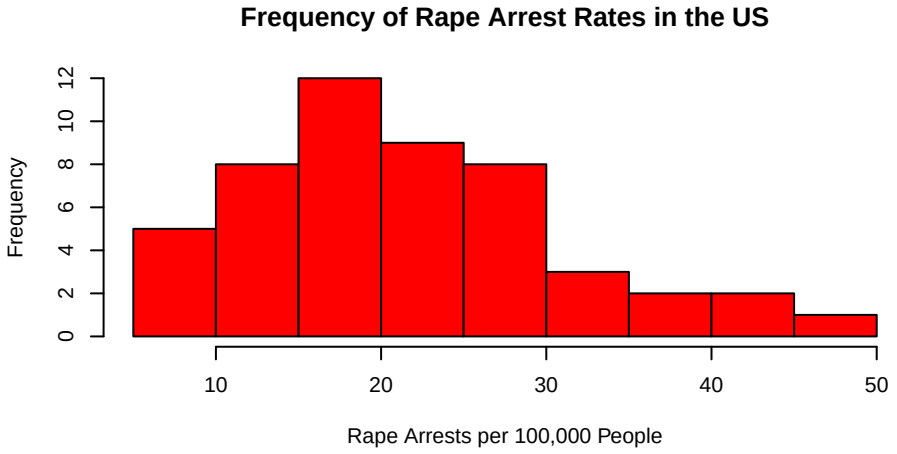
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

Answer: The mean is 21.23. The median is 20.10. Q1 is 15.07 and Q3 is 26.18. The minimum value is 7.30 and the maximum value is 46.00.

```
par(mfrow=c(3,1))
hist(dat$Murder, xlab = "Murder Arrests per 100,000 People", main = "Frequency of Murder Arrest Rates i
hist(dat$Assault, xlab = "Assault Arrests per 100,000 People", main = "Frequency of Assault Arrest Rate
hist(dat$Rape, xlab = "Rape Arrests per 100,000 People", main = "Frequency of Rape Arrest Rates in the
```







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

Answer: The 'par' command sets parameters for graphs, changes how they are displayed, and allows us to tell R to compare additional variables within the data set. 'mfrow' puts things in an array, and the parameters of (1,3) tell it to display them stacked on top of each other, rather than side by side or otherwise.

What can you learn from plotting the histograms together?

Answer: We can learn that assault arrest rates per 100,000 people are far higher than the arrest rates for

murder and rape by comparing the x axis of all three histograms. Additionally, while the frequency of arrest rates for rape and murder skew relatively right and are unimodal, the frequency of assault arrest rates is bimodal. Comparing and plotting them together can generally help us visually compare distributions overall.

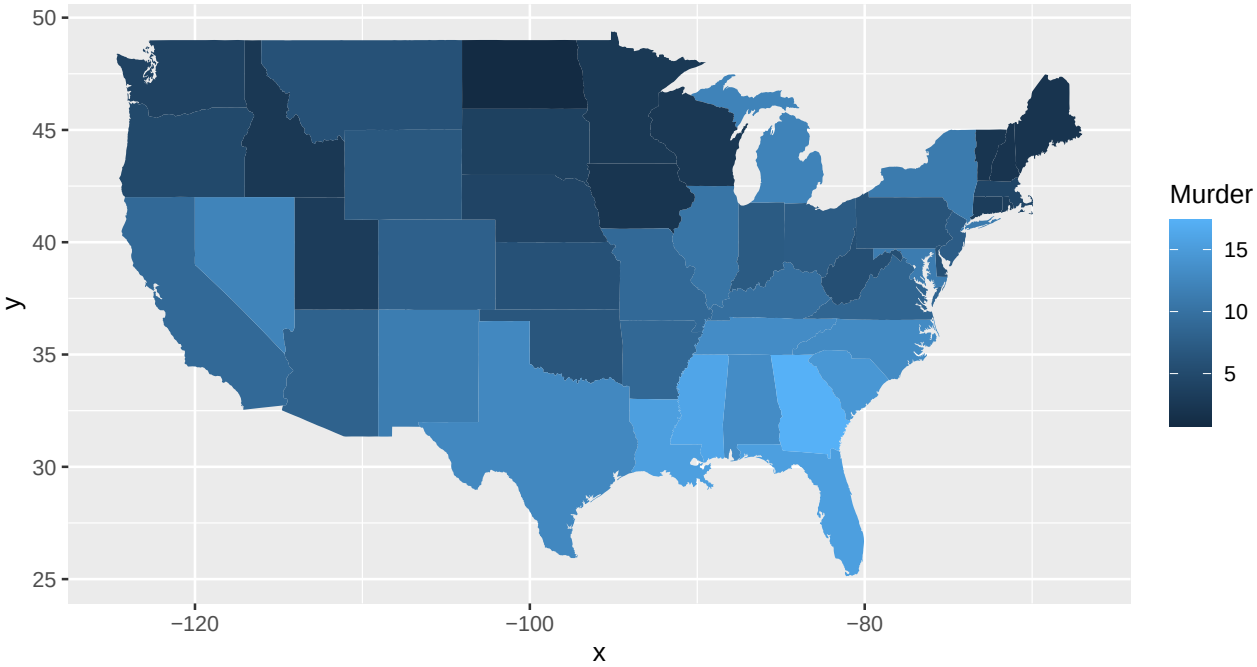
Problem 8

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:

```
#install.packages("maps")
#- This line installs a package that allows for map-drawing based on state names provided within the da
#install.packages("ggplot2")
#- This line installs a package that makes improves the data visualization so that the map is easier to
#library('maps')
#- This line tells R where to find the package to help draw the map
#library('ggplot2')
#- This line tells R where to find the package to help draw the map
#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 line tells R to use ggplot and to take information from dat variable 'state', and then to draw
```

Assignment 2