MA 2611 Lab 2

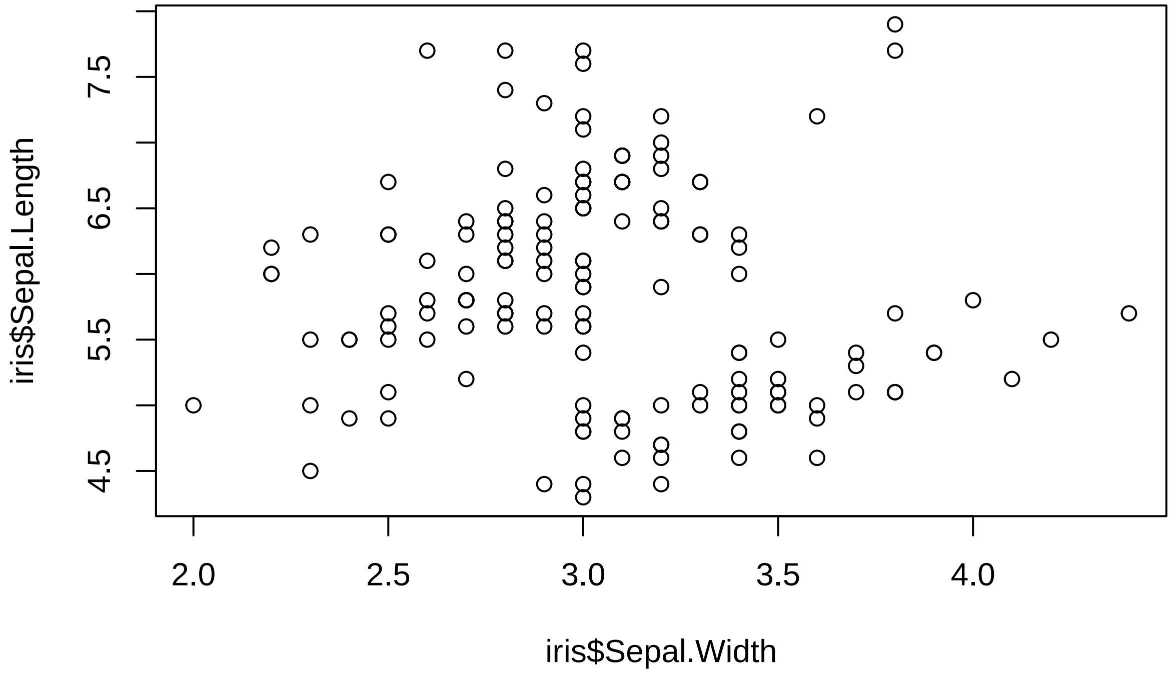
Jacob Nguyen

2022-09-08

1. Using the iris data frame, create a scatterplot that meets the following criteria:

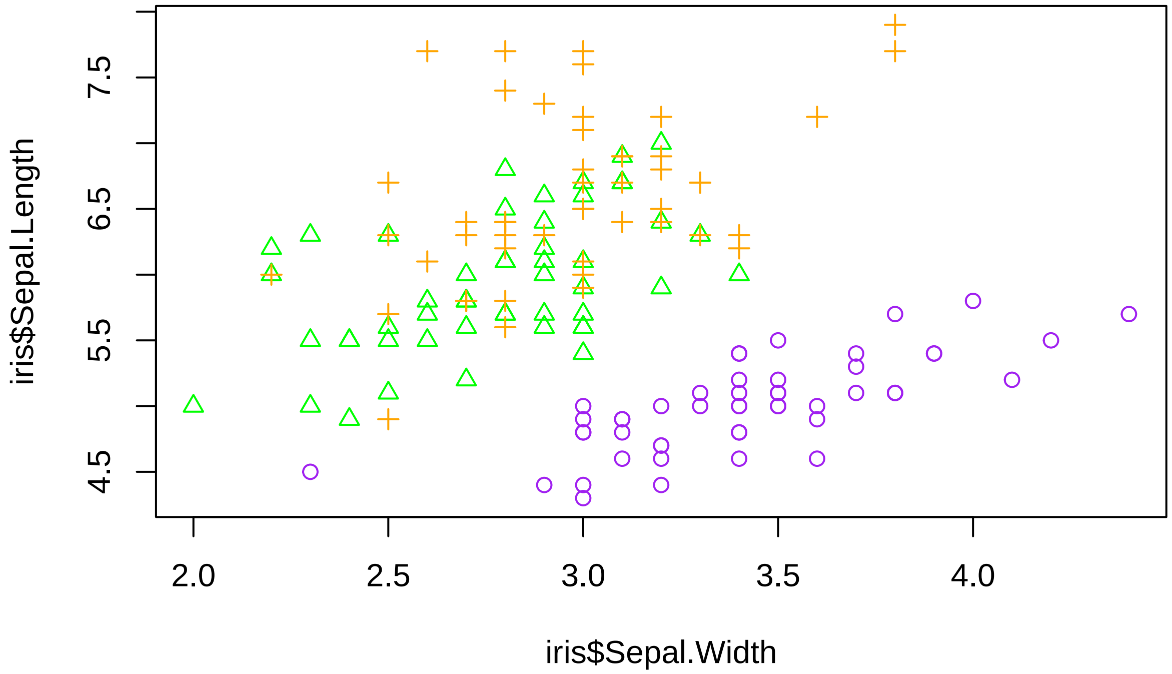
a. Plot "sepal width" on the x-axis and "sepal length" on the y-axis.

```
plot(iris$Sepal.Width,iris$Sepal.Length)
```



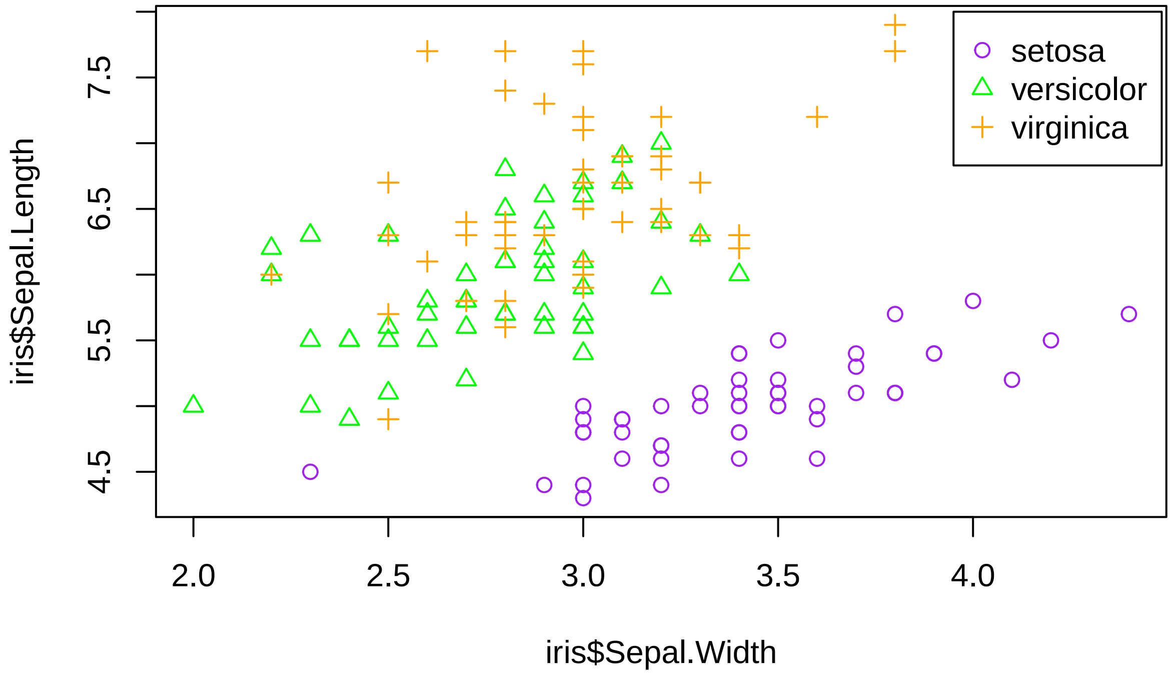
b. Add different colors and shapes, of your choosing, to distinguish between the different species of iris.

```
plot(iris$Sepal.Width,iris$Sepal.Length,pch=c(1,2,3)[unclass(iris$Species)],
col=c("purple","green","orange")[unclass(iris$Species)])
```



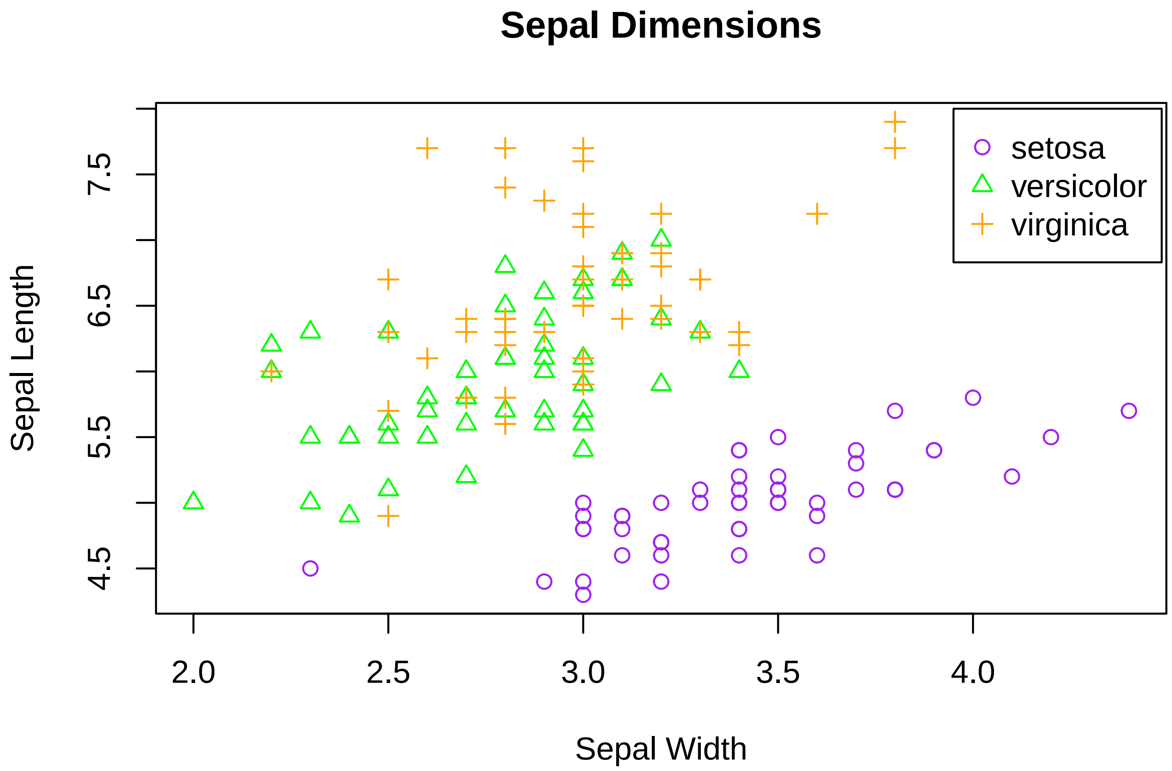
c. Add a legend with a title to the top right corner of the plot and make sure the correct colors and shapes correspond to the species on the plot.

```
plot(iris$Sepal.Width,iris$Sepal.Length,pch=c(1,2,3)[unclass(iris$Species)],
col=c("purple","green","orange")[unclass(iris$Species)])
legend(3.95,8,legend=as.character(unique(iris$Species)),
col=c("purple","green","orange"),pch=1:3)
```



d. Add labels to the x and y axes and a plot title.

```
plot(iris$Sepal.Width,iris$Sepal.Length,pch=c(1,2,3)[unclass(iris$Species)],
xlab="Sepal Width",ylab="Sepal Length",main="Sepal Dimensions",
col=c("purple","green","orange")[unclass(iris$Species)])
legend(3.95,8,legend=as.character(unique(iris$Species)),
col=c("purple","green","orange"),pch=1:3)
```



2. Using the quakes data frame, complete the following exercises:

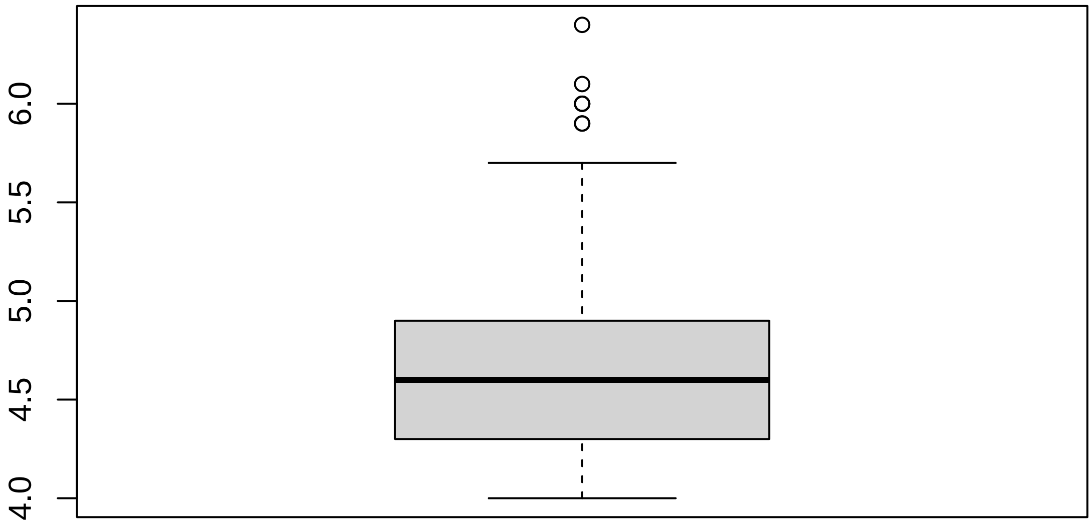
a. Are there outliers for the "magnitude" measured? Support your answer with both a five number summary and boxplot of the data set. If there are outliers, would they be considered outliers in the practical sense? Why or why not?

There are outliers for the "magnitude" measured.

```
fivenum(quakes$mag)
```

```
## [1] 4.0 4.3 4.6 4.9 6.4
```

```
boxplot(quakes$mag)
```



These are not considered outliers in the practical sense because in the real world these magnitudes of earthquakes are possible and can happen.

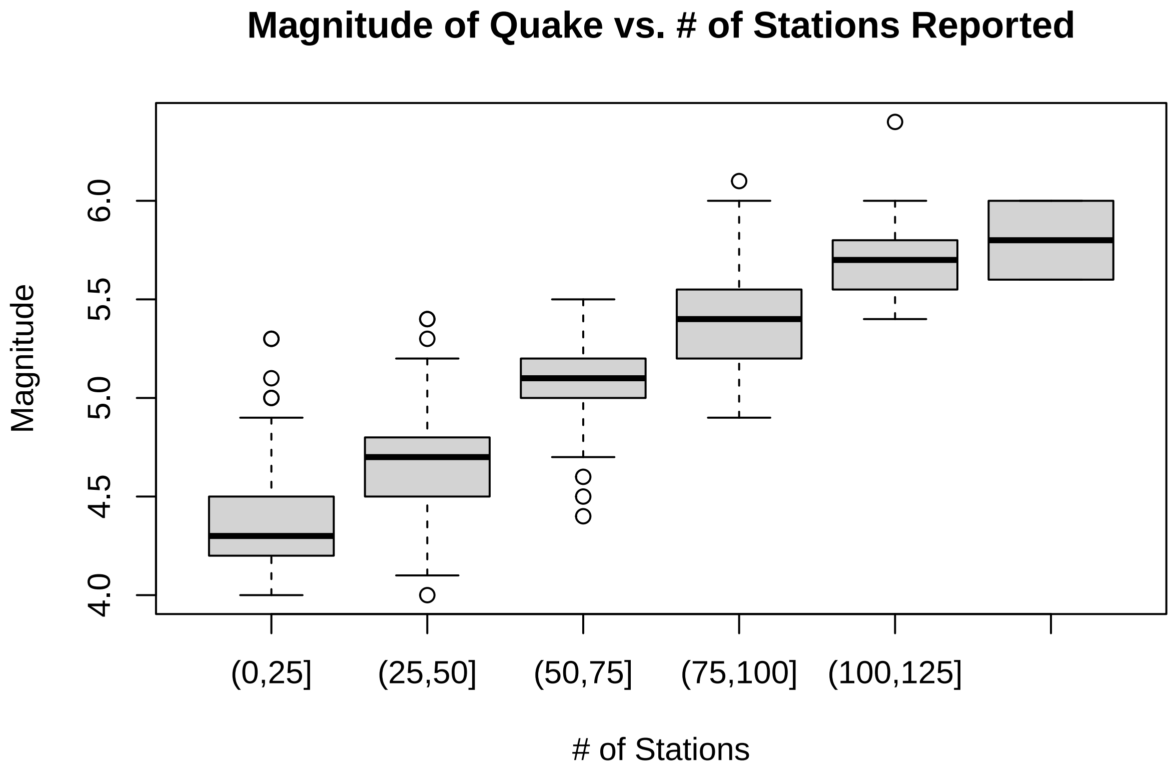
b. Work through the following steps:

c. Create and save a vector that breaks the “stations” data into categorical groups in increments of 25.

```
stations.groups=cut(quakes$stations,breaks=c(0,25,50,75,100,125,150))
```

ii. Create a single plot with multiple boxplots of the “magnitude” data grouped by the “stations” categorical groups (from part (i)) to compare the differences in magnitude by groups of station counts. Be sure to add axis labels and a title to the plot.

```
boxplot(quakes$mag~stations.groups, xlab="# of Stations",ylab="Magnitude",
main="Magnitude of Quake vs. # of Stations Reported")
```



iii. What does the plot in part (ii) tell you about the relationship between "magnitude" and "stations"? Are there differences in magnitude depending on the station counts? Why or why not? How does this compare to your takeaways from the scatterplots in the activity?

The more stations that report an earthquake, the greater the magnitude of the earthquake. There are differences in the median magnitude depending on the number of stations that are reporting the quake. This makes sense because the larger the magnitude of an earthquake, the greater the amount of people that can feel the quake. I learned that the context matters when looking at boxplots, as certain statistical outliers can still exist in a practical sense. Also with boxplots, you can gain interpret different things compared to scatterplots, such as how the magnitude seems to increase as more stations report the quake.

3. Make a vector using the data set of the highest points (in feet) in each US state:

2413, 20310, 12637, 2753, 14505, 14440, 2379, 447, 345, 4784, 13803, 12668, 1235, 1257, 1671, 4041, 4145, 535, 5270, 3360, 3489, 1979, 2302, 807, 1772, 12807, 5427, 13147, 6288, 1803, 13167, 5343, 6684, 3508, 1549, 4975, 11249, 3213, 811, 3560, 7244, 6643, 8571, 13534, 4395, 5729, 14417, 4863, 1951, 13809

```
x = c(2413, 20310, 12637, 2753, 14505, 14440, 2379, 447, 345, 4784, 13803, 12668,
1235, 1257, 1671, 4041, 4145, 535, 5270, 3360, 3489, 1979, 2302, 807, 1772,
12807, 5427, 13147, 6288, 1803, 13167, 5343, 6684, 3508, 1549, 4975, 11249,
3213, 811, 3560, 7244, 6643, 8571, 13534, 4395, 5729, 14417, 4863, 1951, 13809)
x
```

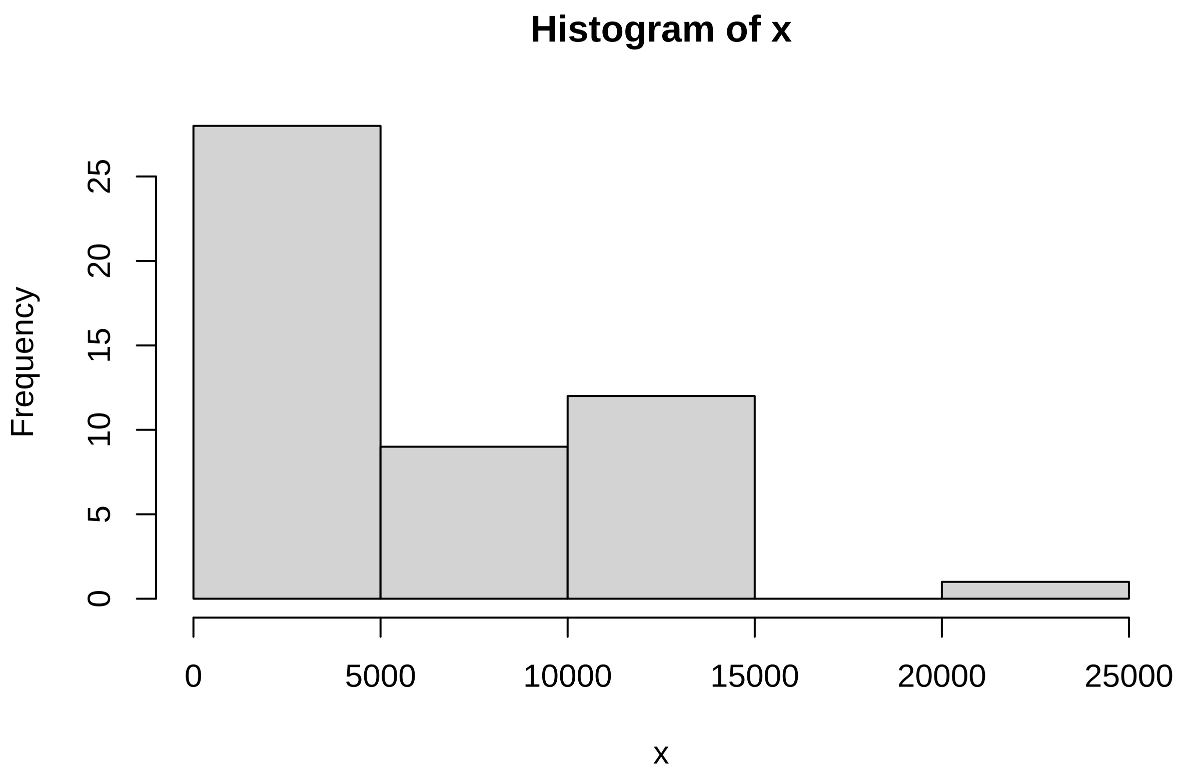
```
##  [1]  2413 20310 12637  2753 14505 14440  2379   447   345  4784 13803 12668
## [13]  1235  1257  1671  4041  4145   535  5270  3360  3489  1979  2302   807
## [25]  1772 12807  5427 13147  6288  1803 13167  5343  6684  3508  1549  4975
## [37] 11249  3213   811  3560  7244  6643  8571 13534  4395  5729 14417  4863
## [49]  1951 13809
```

a. Without graphing the data, does this data set have skewness? Why or why not? If it does, is the skewness positive or negative? Explain your reasoning.

This data set has positive skewness because when calculated the mean was greater than the mean, meaning that when graphed, more bars would extend to the right of the graph after the modal bar.

b. Create a histogram of the data set to check your answer in part (a). Did the skewness depicted align with your answer?

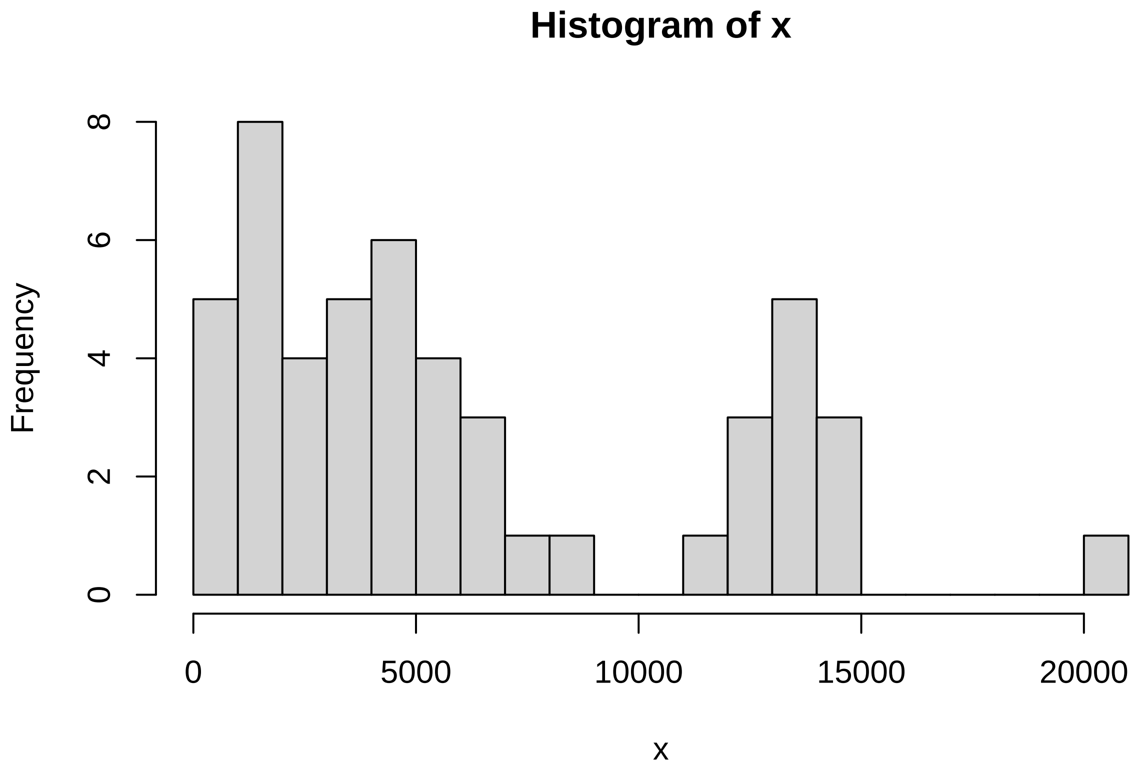
```
hist(x)
```



The skewness depicted did align with my answer given that the bars of positive height extended further to the right of the modal bar.

c. If you change the number of bins from the histogram in part (b) to 15 bins, what additional information do you gain? (hint use ?hist to get more info)

```
hist(x,breaks=15)
```



The additional information that we gain is how positively skewed the data is, and where modal bars are located. We also find out that this data has multiple modal bars and how the majority of the highest points in the US are on the smaller side, reinforcing that the data is positively skewed.

d. When comparing the histograms from parts (b) and (c), what does this tell you about the importance of using the "right" number of bins? Explain.

The importance about using the right number of bins is so you can interpret the data correctly and make the correct conclusions when analyzing. If you don't use enough the data is not portrayed well, and if you use too many, you are unable to fully understand the distribution of the data.