Name: Steve G. Mwangi
Date: October 15, 2019
Assignment: TMATH 390 R Lab 3 Document.

Continuing to use R for empirical exploratory data analysis

C1. (4) Submit the R script you used to complete this computer lab assignment.
Submitted.

Read your dataset into the current R session. For demonstration this lab assumes you name your R data object data.df. Navigate to the directory in which your dataset is located using setwd, then read the dataset into R.

```
> #Read your dataset into the current R session. For demonstration this lab a
ssumes you name your R data
> #object data.df. Navigate to the directory in which your dataset is located
using setwd, then read the
> #dataset into R.
> #Determine working directory with
> getwd()
[1] "C:/Users/steve/Desktop/UWT/Fall Classes/TMATH 390/R Documents/R Assignme
nts/R_Lab_2"
> #Change working directory to: C:\Users\steve\Desktop\UWT\Fall Classes\TMATH
390\R Documents\R Assignments\R_Lab_3
> setwd("C:/Users/steve/Desktop/UWT/Fall Classes/TMATH 390/R Documents/R Assi
gnments/R_Lab_3")
> #Read csv file of my data
> data.df = read.csv("data.csv")
```

C2. (1) Describe your second quantitative variable, including its units of measurement.

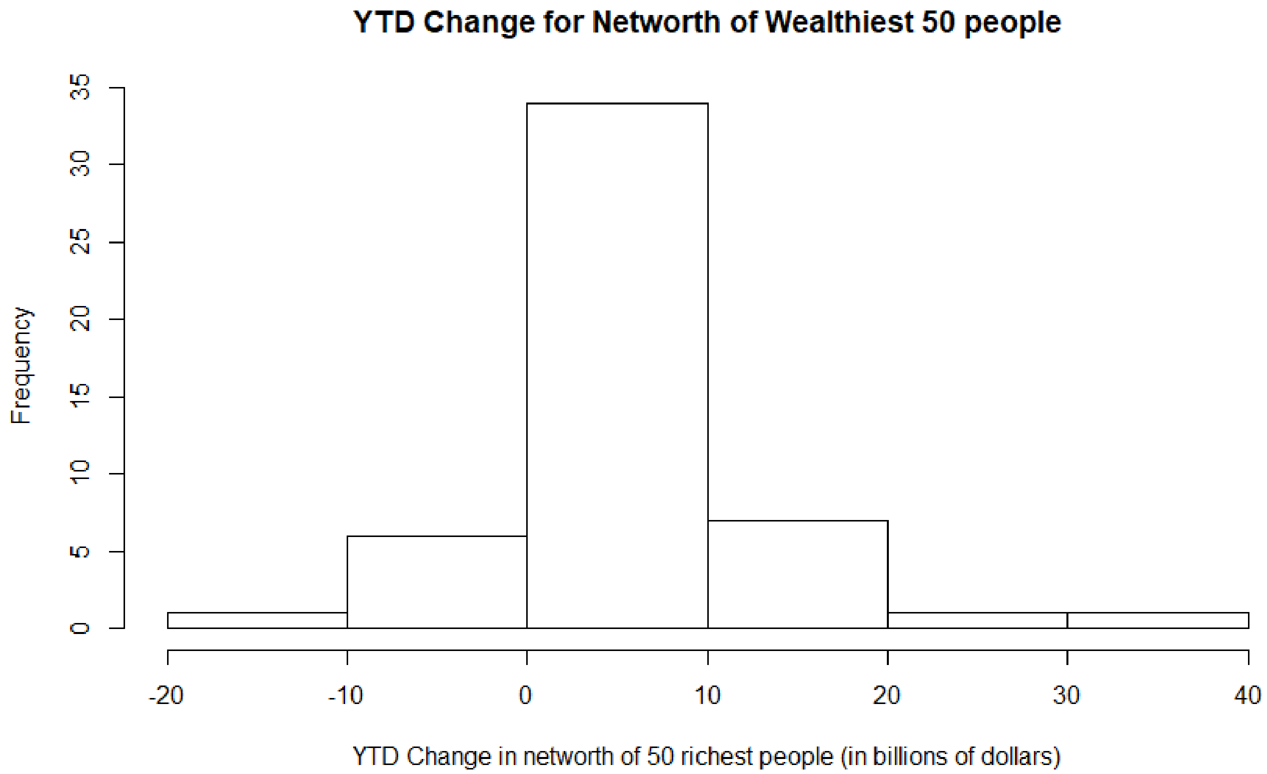
Second quantitative variable is **YTD Change** which is the change in net-worth for an individual, for the year, to the date the data was collected. Its units of measurement is **Billions of Dollars**.

C3. (1) Use the R summary command to produce a summary of your second quantitative variable. Copy and paste the results of the summary command in your R document.

```
#C3.
> #Quantitative variable chosen: YTD Change
> #Producing a summary of Quantitative YTD Change column
> summary(data.df$YTD.Change)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-16.200   1.548   4.145   5.307   8.485  35.000 
```

C4. (1) Create a histogram for your second quantitative variable. Make sure that the histogram has a properly labeled axis and title. Copy and paste the histogram into your lab document.

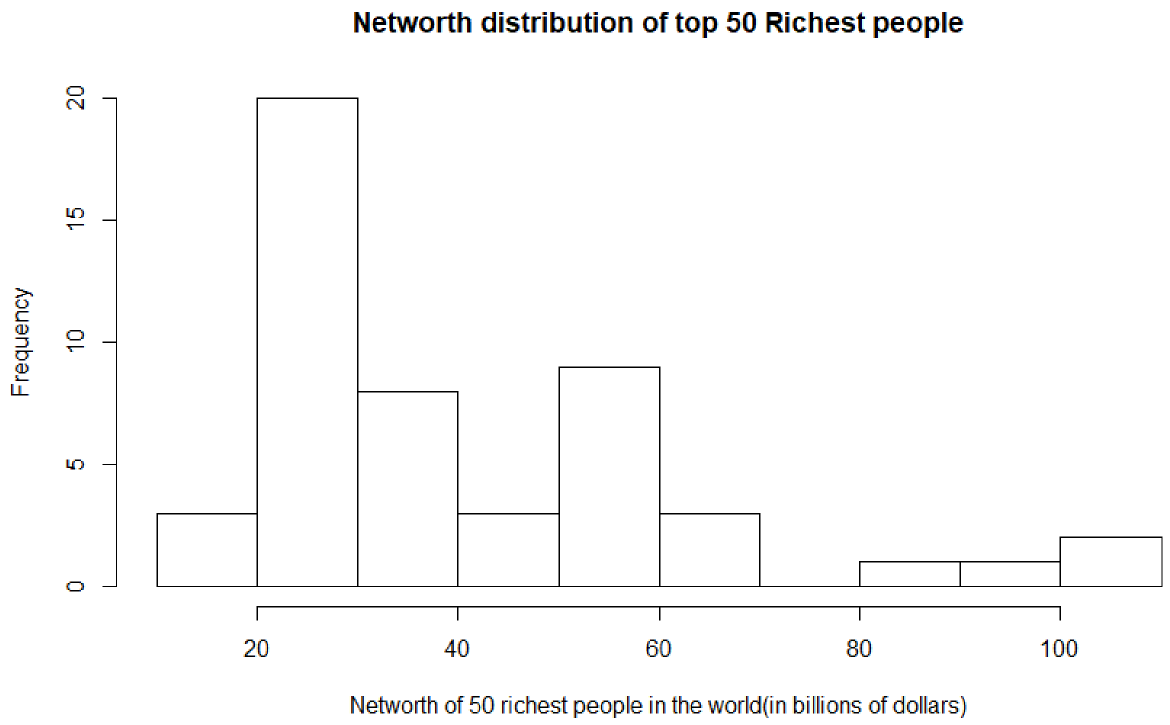
```
#C4.
> #Creating Histogram with properly labelled axes and title.
> #Command creates a histogram of chosen column, in the dataframe data.df
> #the xlab argument writes test to label the x-axis
> # main argument gives it title
> hist(data.df$YTD.Change, xlab="YTD Change in networth of 50 richest people
(in billions of dollars)", main="YTD Change for Networth of Wealthiest 50 peo
ple")
```



C5. (2) Describe the histogram you produce. What shape does the distribution take? Are there any peculiarities in the distribution?

The histogram has a roughly symmetric unimodal distribution. There are no outliers. From the data, we garner that majority of the billionaires' net-worth increased by a value between 0-10 billion dollars Year to Date.

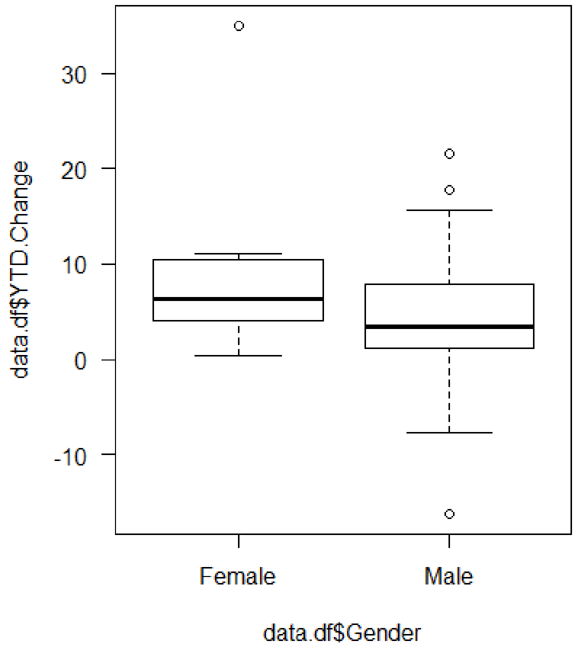
How does it compare to the shape of the distribution for your first quantitative variable that you graphed in lab 2?



The two histograms are similar in that they are unimodal.
Even though it is apparent that the histogram for the net-worths is right skewed, the one for the YTD Change is roughly symmetric.

C6. (2) Provide the boxplot that compares the distribution of your second quantitative variable across the different levels (categories) of your qualitative variable (using the same qualitative variable as in lab2).

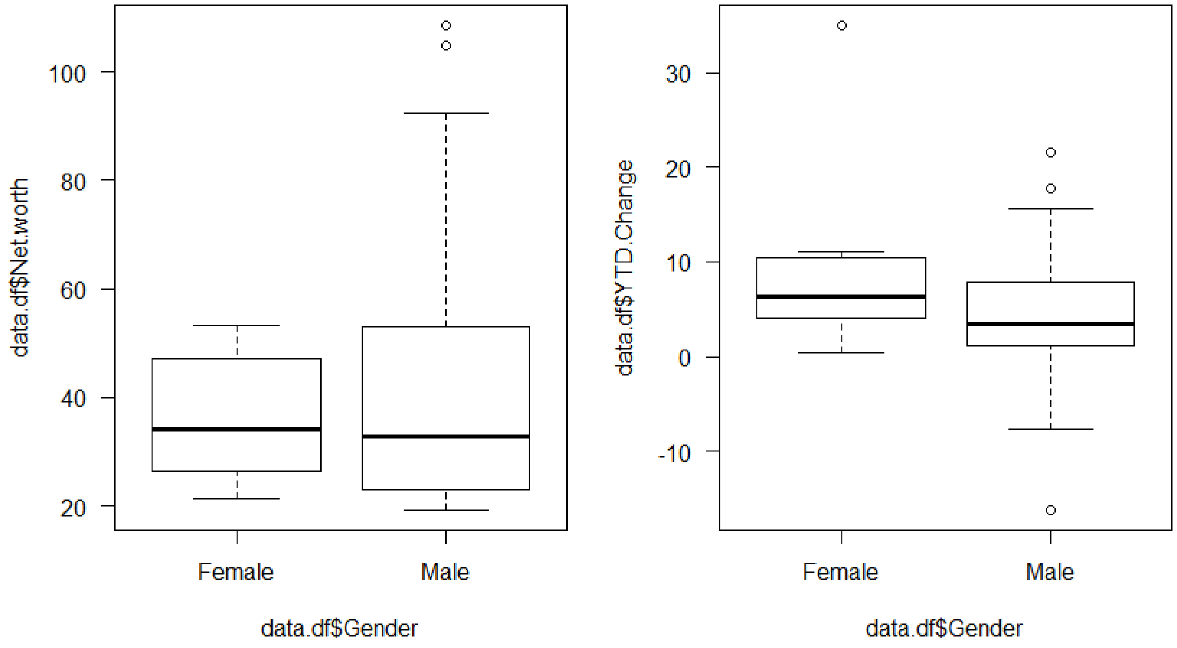
```
#Boxplot
> # Establishing a graphing window with 1 rows and 2 columns,
> # and las = 1 sets axis labels to be horizontal
> par(mfrow=c(1,2), las = 1)
> #Creating a boxplot.
> # ~ represents a relationship between two variable, with Y on left side, X
on right side.
> # Networth(quantitative) across Gender(Qualitative)
> boxplot(data.df$YTD.Change~data.df$Gender)
```



C7. (2) Describe similarities and differences in the distributions of your second quantitative variable at different levels of your qualitative variable.
Both boxplots have a roughly symmetric distribution. We have more male outliers, with 2 males gaining about 20 billion dollars, and one other one losing about 16 billion dollars, Year to Date.

For the female boxplot, the distribution is comparatively short and bound between 0 to about 12 billion dollars, and one outlier having gained over 30 billion dollars Year to date. Even though both boxplots have a somewhat equal interquartile range, the male boxplot is comparatively taller, due to its long whiskers; this could be attributed to the fact that there are 42 males to 8 females, hence the larger disagreement in male values.

Also compare this to what you observed in the boxplot of the first quantitative variable that you produced in lab 2.



The boxplots are similar, in that the male boxplots are comparatively taller in both cases compared to the female boxplots. In addition, the male boxplots both have outliers. All this could be attributed to the fact that there were more males compared to females (42 vs. 8), which could result in the female boxplots being comparatively shorter, with one or no outliers.

C8. (2) Compute the sample mean, median, and sample standard deviation for your second quantitative variable for each level of your qualitative variable. Copy and paste the R outputs that calculate the mean and standard deviation for your data at each level of your qualitative variable.

```
#C8.
> #Individual summary statistics
> #Command to get individual summary statistics
> #FEMALE
> #For mean
> mean(data.df$YTD.Change[data.df$Gender=="Female"])
[1] 9.6215
> #For median
> median(data.df$YTD.Change[data.df$Gender=="Female"])
[1] 6.28
> #Standard Dev.
> sd(data.df$YTD.Change[data.df$Gender=="Female"])
[1] 10.82883
> #MALE
> #For mean
> mean(data.df$YTD.Change[data.df$Gender=="Male"])
[1] 4.484619
> #For median
> median(data.df$YTD.Change[data.df$Gender=="Male"])
[1] 3.375
> #Standard Dev.
> sd(data.df$YTD.Change[data.df$Gender=="Male"])
[1] 6.531767
```

C9. (1) We can also use R to compute sample quantiles. Compute the 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 quantiles of you second quantitative variable. Use the code below to guide you. Report the values for the quantiles in your lab document.

```
# the quantile function returns values for the
> # quantiles indicated by probs
> # this example uses a data frame called data.df
> # and calculates the quantiles for the column called YTD Change
> # Here the seq command is used. This produces a sequence
> # that starts at the first number, ends at the second
> # number, at the increment indicated by the third number
> # you can try running the seq command by itself to see the result.
> quantile(data.df$YTD.Change, probs=seq(0.1,0.9,0.1))
    10%     20%     30%     40%     50%     60%     70%     80%     90% 
-0.9176  0.8954  1.7050  3.1640  4.1450  5.0480  7.7050  9.7280 11.3500 
```

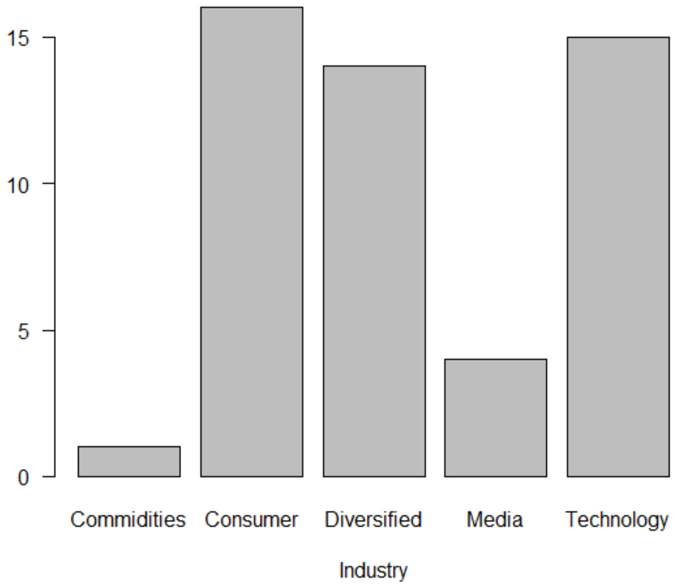
C10. (1) We can also produce summaries of our qualitative data. The table command returns a tally of the number of observations for each unique value of the variable. Use the table command to report the number of observations for each level of your qualitative variable. Use the code below.

```
> # Chosen QUalitative Variable: Gender Column
> #Producing summaries of qualitative data.
> #Data frame is called data.df, and
> #the column for qualitative variable is Gender
> table(data.df$Gender)

Female   Male 
     8     42 
```

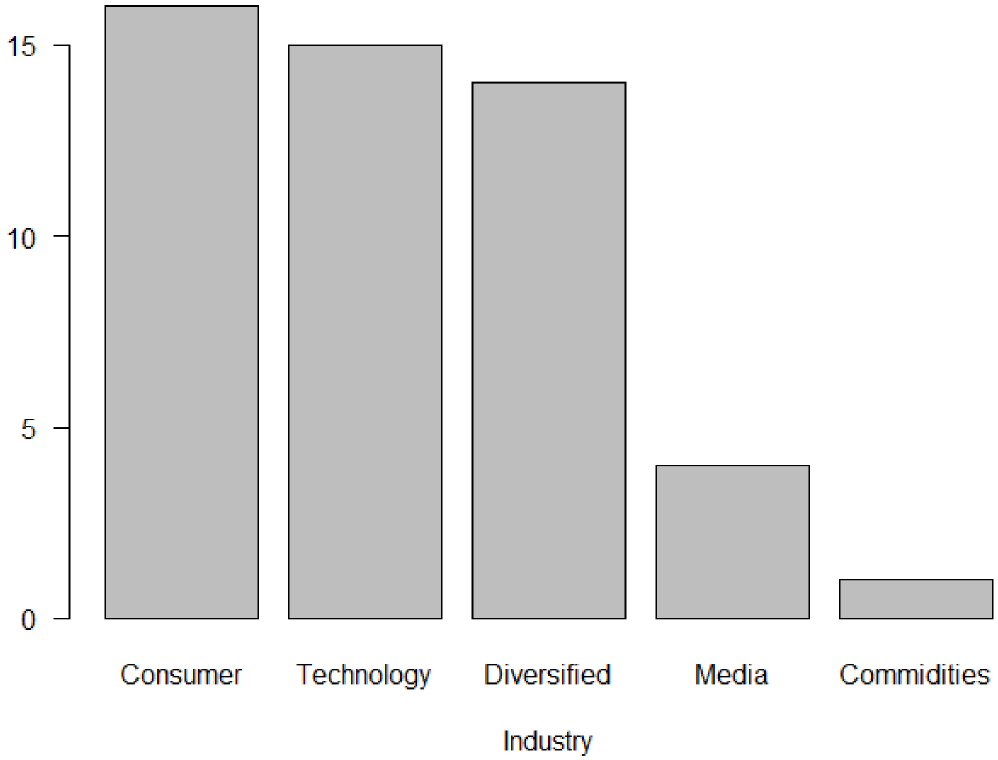
```
# Chosen QUalitative Variable: Industry Column
> #Producing summaries of qualitative data.
> #Data frame is called data.df, and
> #the column for qualitative variable is Industry
> table(data.df$Industry)

Commidities    Consumer Diversified       Media  Technology 
          1          16          14           4          15 
```

C11. (1) We can also create a bar graph of a qualitative variable using the R function barplot. Use the code below and provide a bar graph for your qualitative variable in your lab document.

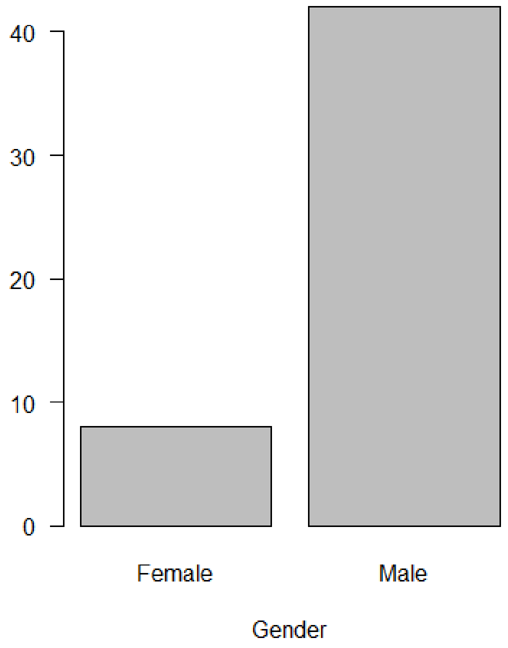
```
#Creating a bar graph of a qualitative variable using the R function barplot.
> # the barchart command takes the results of the table
> # command and creates a barplot from them
> barplot(table(data.df$Industry),
+         xlab="Industry")
```

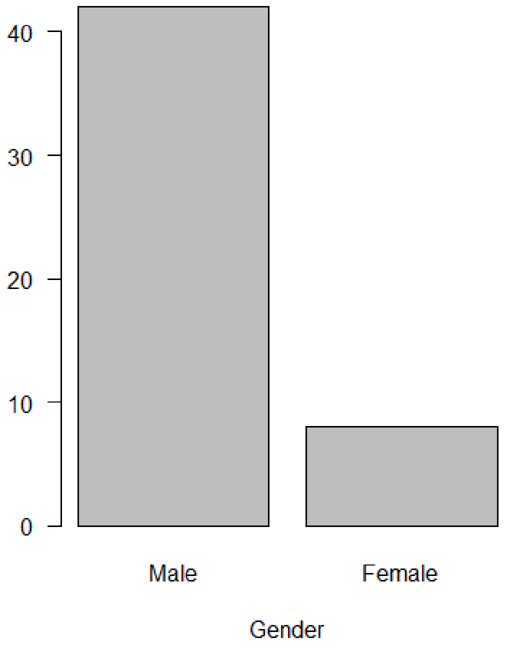


```
> # if you want to make a Pareto graph, you can sort your
> # table then create a bar graph using that.
> # here we create an object for the table
> table1.tab=table(data.df$Industry)
> # the sort function will rearrange this, and the decreasing
> # argument tells the function to sort in descending order
> table1a.tab=sort(table1.tab,decreasing=TRUE)
> # Now create a Pareto graph using the sorted table object
> barplot(table1a.tab,
+         xlab="Industry")

>
```

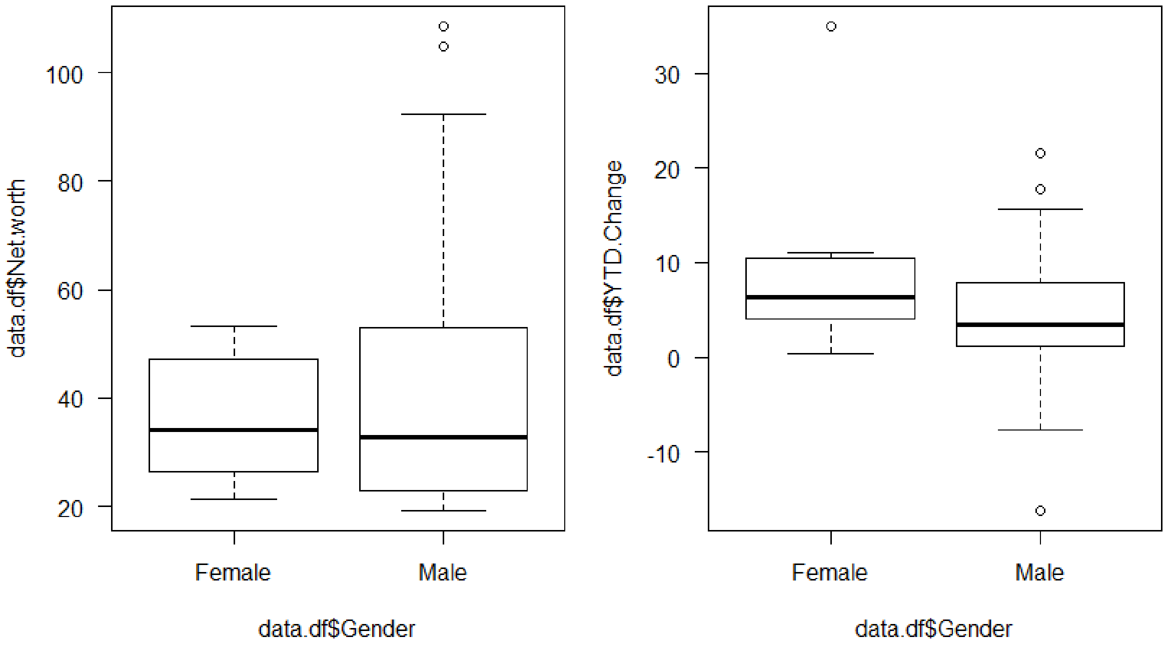


```
#Creating a bar graph of a qualitative variable using the R function barplot.
> # the barchart command takes the results of the table
> # command and creates a barplot from them
> barplot(table(data.df$Gender),
+         xlab="Gender")
> # if you want to make a Pareto graph, you can sort your
> # table then create a bar graph using that.
> # here we create an object for the table
> table1.tab=table(data.df$Gender)
> # the sort function will rearrange this, and the decreasing
> # argument tells the function to sort in descending order
> table1a.tab=sort(table1.tab,decreasing=TRUE)
> # Now create a Pareto graph using the sorted table object
> barplot(table1a.tab,
+         xlab="Gender")
```

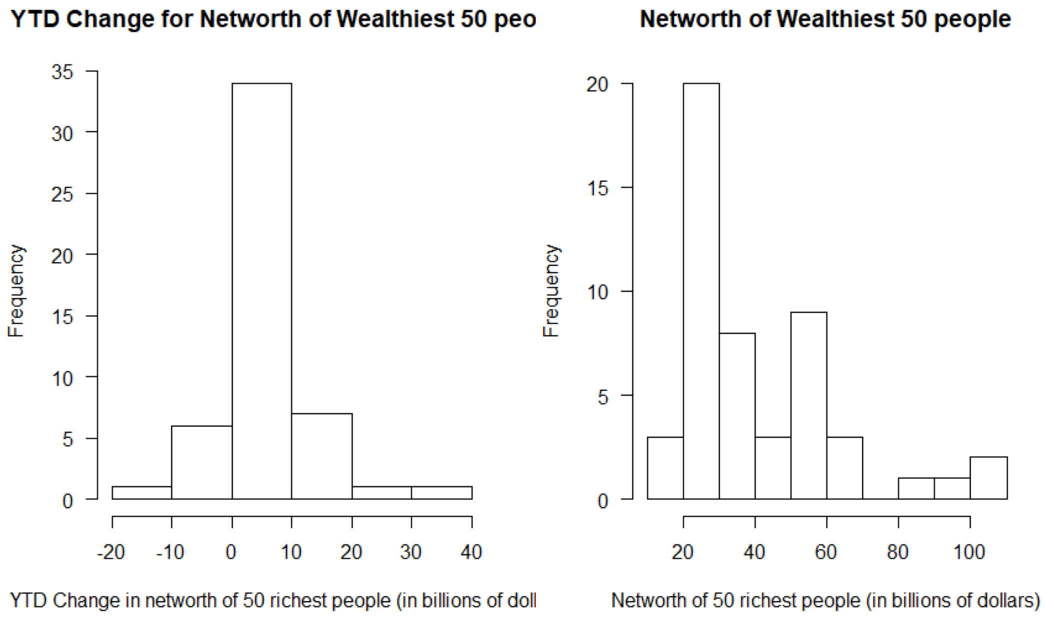




C12. (2) Use the numerical summaries and the plots to summarize your analysis. Use full sentences to write at least a paragraph and make sure to refer directly to your graphs and numerical summaries. In your summary comment on the following: How are the distributions of your two quantitative variables similar? How do they differ? Consider the shape, variability, and measure(s) of location. How do your two quantitative variables change at different levels of the qualitative variable? Do they change similarly? Were your observations evenly dispersed across the levels of your qualitative variable? What do you think explains the distribution of your qualitative variable?



Even though male and female boxplots in both cases for net-worth and YTD change show that their interquartile range are almost the same, especially in the YTD change, one outstanding feature is that the male boxplots are comparatively longer. In addition, the male boxplots have more outliers too. This higher variability in the male distribution could be due to the fact that the data had more males than females (42 males vs 8 females). Hence, data from the 8 Females seeming to be in more agreement than that from the 42 males.



For the quantitative distributions, the two histograms seem to both be unimodal and positively skewed, with the YTD Change distribution being more symmetrical than the Net Worth distribution. From both histograms, we can gather that there are more people in the middle of the distributions, with similar Net-Worth's (20-60 billion dollars) and similar gains (0-10 billion dollars) year to date. This correlation could account for the similarity in shape of the two distributions, in that the high frequency of people who gained 0-10 billion dollars YTD, are the same who have 20-60-billion-dollar net worth.

```
#C12. c
> # the quantile function returns values for the
> # quantiles indicated by probs
> # this example uses a data frame called data.df
> # and calculates the quantiles for the column called YTD Change
> # Here the seq command is used. This produces a sequence
> # that starts at the first number, ends at the second
> # number, at the increment indicated by the third number
> # you can try running the seq command by itself to see the result.
> quantile(data.df$Net.worth, probs=seq(0.1,0.9,0.1))
  10%   20%   30%   40%   50%   60%   70%   80%   90% 
20.77 22.18 25.86 28.88 34.10 37.60 51.19 53.36 67.22 

>
```

Additionally, both distributions have higher frequency at about the middle, something that is like salaries or incomes, at the standard level (sub millions of dollars). This shows that even though there is the ultra-wealthy billionaires, their wealth is also relative to other ultra-wealthy individuals, with some billionaires being outliers on the higher end or lower end, and majority in between. E.g. 90th percentile and above have net-worths of more than 67 billion, then 10-90 percentiles with between 21-60 billion dollars, and those within the 10th percentile, with less than 21 billion dollars.