First Assignment:

Write a simple R script to execute the following:

**Preprocessing**  
1. Load the file “6304 Module 1 Assignment Data.xlsx” into R. This file contains   
information on 13,750 boat sales across the United States. Variables are:  
a. Price: The final sales price of the boat.  
b. State: The state in which the sale was completed.  
c. Length: The length of the boat in feet.  
d. Age: The age of the boat in years with 0 indicating a new boat.  
2. Create a data object which is a subset of the original, including only the states of   
Washington and Georgia. There are several ways this can be done, but one way is to use   
a subset command of the form shown below. You can search online for any assistance   
you need with formatting this command for your work.  
beaver=subset(wally,variable=="One"|variable=="Two")  
3. Using the numerical portion of your U number as a random number seed, take a random   
sample of 60 boats from the subset data using the method presented in class. This will be   
the data used for your assignment.

**Analysis**  
Using R calculate and report the following:  
1. The structure of the data object using the str() command.  
2. Mean, Median, Standard Deviation, Skewness, and Kurtosis of the price variable. Based   
on these descriptive measurements how closely do you think this data conforms to a   
theoretical normal distribution?  
3. A boxplot of the age variable. Based on this boxplot what can you say about the   
symmetry/skewness of this variable?  
4. Quartiles of the age variable. Show your quartiles running from the minimum to   
maximum values for the variable, incrementing by .20.   
5. A simple histogram of the length variable. Color your histogram blue and give it an   
appropriate main title. Make sure the bottom axis of your histogram covers a range from   
0 to 150. Would you say from this histogram the distribution of length follows a   
symmetric distribution, or a skewed distribution?  
6. A simple stem and leaf plot of the age variable.  
7. Two comparative boxplots for the price variable, one for each of the states included in   
your analysis. Your boxplots should be colored red and shown side by side with an   
appropriate main title and labels for the states on the bottom axis. Based on these   
boxplots what can you say about the similarity in price between the two states based on   
your sampled data?

**Preprocessing:**

**#Varun Teja Kolluru**

**#remove the list in environment window**

**rm(list=ls())**

**#Load the file from the directory**

**my\_data = import('6304 Module 1 Assignment Data.xlsx')**

**#change the column names to lower case**

**colnames(my\_data) = tolower(make.names(colnames(my\_data)))**

**#Creating a new subset for states with Washington and Georigia**

**sub\_data = subset(my\_data,state=='WA'|state=='GA')**

**#Creating a seed with my U number and create a sample from subset**

**set.seed('97')**

**my\_sample = sub\_data[sample(1:nrow(sub\_data),60),]**

**attach(my\_sample)**

**my\_sample**

The above code is a preprocessing code, which gets the data from the given file and creating a subset and using the subset data a new sample is created with seeding number.

My sample data is attached in the bottom of the file for the future reference.

**Analysis:**

1. The structure of the data object using the str() command.

**#Structure of the data object**

**str(my\_sample)**

**> str(my\_sample)**

**'data.frame':** **60 obs. of 4 variables:**

**$ price : num 149900 124900 189900 36000 44000 ...**

**$ state : chr "GA" "GA" "GA" "WA" ...**

**$ length: num 58 53 60 56 55 53 54 64 55 108 ...**

**$ age : num 10 13 9 42 20 8 15 5 28 18 …**

2. Mean, Median, Standard Deviation, Skewness, and Kurtosis of the price variable. Based   
on these descriptive measurements how closely do you think this data conforms to a   
theoretical normal distribution?

**> #Mean Median SD skewness kurtosis**

**> mean(price)**

**[1] 121699.8**

**> median(price)**

**[1] 107000**

**> sd(price)**

**[1] 79516.84**

**> skewness(price)**

**[1] 0.5520654**

**> kurtosis(price)**

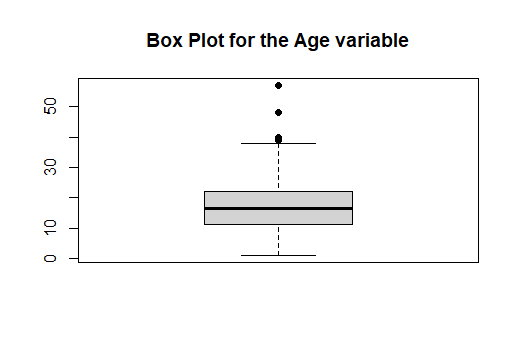
**[1] 2.140651**

We got skewness as positive value, which is a right skewed graph and the kurtosis is below 3, which will be a platokurtic graph and there is a lot of standard deviation from the standard symmetric or bell shapes curve graph.

3. A boxplot of the age variable. Based on this boxplot what can you say about the   
symmetry/skewness of this variable?

**#Boxplot for age variable**

**boxplot(age,pch=19,main="Box Plot for the Age variable")**



Based on the above boxplot for the age variable, we can say that the deviation in the standard curve to left side which is a right skewness curve. This is due to the 4 Outliers ( 4 Dots) after the 40 age limit in the Y axis in the graph. In detail, 25% quartile starts after ‘10’ and the 75% quartile ends at some where ‘24’ and the median is around ‘16’. And based on this graph we can infer that most of the boats are aged in between 10 to 25.

4. Quartiles of the age variable. Show your quartiles running from the minimum to   
maximum values for the variable, incrementing by .20.

**> #qunatile for age variable**

**> quantile(age,probs = seq(0,1,0.20))**

**0% 20% 40% 60% 80% 100%**

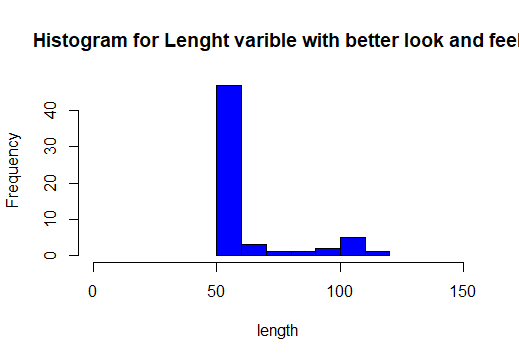
**1 10 15 18 29 57**

5. A simple histogram of the length variable. Color your histogram blue and give it an   
appropriate main title. Make sure the bottom axis of your histogram covers a range from   
0 to 150. Would you say from this histogram the distribution of length follows a   
symmetric distribution, or a skewed distribution?

**#Histogram for the length variable**

**hist(length,col="blue",main="Histogram for Lenght varible with better look and feel",**

**xlim = c(0,150))**

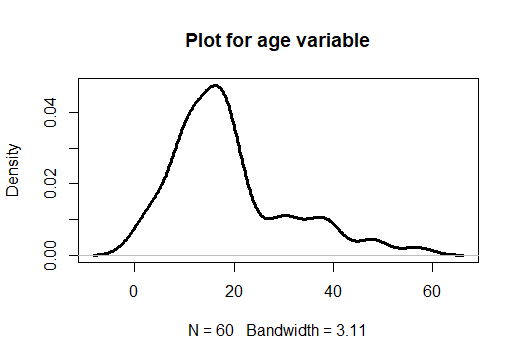


By viewing this histogram graph for length variable, we can say that this do not follow a symmetric shaped curve. We can say that this is a right skewed curve. And if we closely observe the graph, after the ‘100’, there is a raise in the graph and this might be a double curved graph.

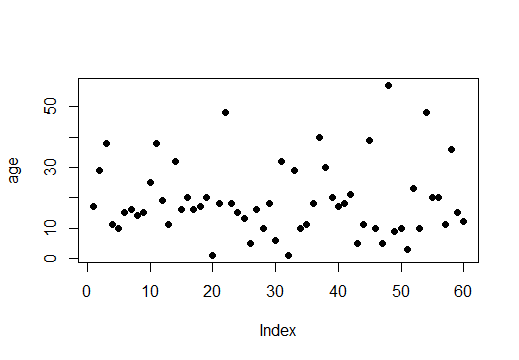
6. A simple stem and leaf plot of the age variable.

**#stem and leaf plot for age variable**

**plot(density(age),lwd=3,main="Plot for age variable")**



**plot(age,pch=19)**

  
7. Two comparative boxplots for the price variable, one for each of the states included in   
your analysis. Your boxplots should be colored red and shown side by side with an   
appropriate main title and labels for the states on the bottom axis. Based on these   
boxplots what can you say about the similarity in price between the two states based on   
your sampled data?

**#Boxplot comparing prices for GA and WA state**

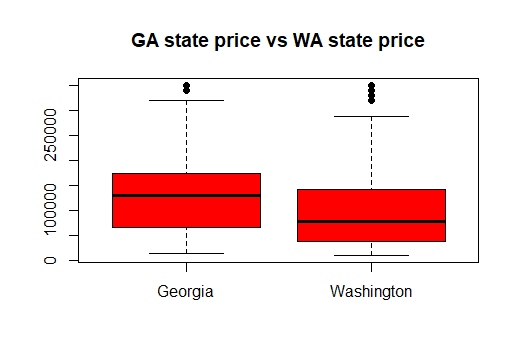
**my.ga = subset(my\_data,state=='GA')**

**my.wa = subset(my\_data,state=='WA')**

**boxplot(my.ga$price,my.wa$price,pch=19,**

**main="GA state price vs WA state price",**

**col=c("red","red"),names=c("Georgia","Washington"))**



**P.T.O**

My Sample Set:

> my\_sample

price state length age

4135 124990 GA 60 17

10322 29995 WA 53 29

7811 39000 WA 55 38

18 299000 GA 117 11

12140 99000 WA 51 10

5167 229900 WA 58 15

77 159900 GA 103 16

6423 84950 GA 57 14

110 227000 GA 101 15

236 109900 GA 94 25

7602 22500 WA 55 38

11597 34999 GA 52 19

12773 214000 WA 50 11

7340 49900 GA 56 32

13657 54995 GA 50 16

13231 29995 GA 50 20

65 12346 WA 105 16

7129 79900 GA 56 17

4307 69900 GA 60 20

13135 211633 GA 50 1

4137 129900 GA 60 18

1288 269000 WA 71 48

4974 99500 GA 59 18

5647 229900 WA 58 15

9683 124900 GA 53 13

7878 159900 GA 55 5

11510 119997 WA 52 16

4256 134900 GA 60 10

102 164900 GA 101 18

9918 154900 WA 53 6

2970 29900 GA 62 32

9532 279995 WA 53 1

4902 54900 GA 59 29

13715 242000 WA 50 10

13272 214000 WA 50 11

322 109000 GA 89 18

9264 32800 WA 54 40

13090 29995 WA 50 30

10039 44000 GA 53 20

6667 79900 GA 56 17

36 149000 GA 108 18

12252 29000 GA 51 21

7309 249000 GA 56 5

10425 129900 GA 53 11

4303 69000 WA 60 39

5047 149900 GA 58 10

6715 249000 GA 56 5

3889 29500 WA 61 57

218 249000 GA 96 9

11741 99000 WA 51 10

11744 239000 WA 51 3

7694 139500 WA 55 23

5798 149900 GA 58 10

1325 69500 WA 70 48

4787 69900 GA 59 20

9160 69900 GA 54 20

6990 105000 WA 56 11

7434 29500 WA 56 36

6956 59900 GA 56 15

5968 79500 GA 57 12