R Markdown Assignment

Varun Teja Kolluru

2022-09-10

### Remove the list in environment window

rm(list=ls())

### Add the libraries which are used in the code

library(rio)  
library(moments)

### Load the file from the directory

my\_data = import('test.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)

The above code is a pre processing 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.

### Below is the sample data

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

# Start of analysis

### Structure of the data object

str(my\_sample)

## 'data.frame': 60 obs. of 4 variables:  
## $ price : num 124990 29995 39000 299000 99000 ...  
## $ state : chr "GA" "WA" "WA" "GA" ...  
## $ length: num 60 53 55 117 51 58 103 57 101 94 ...  
## $ age : num 17 29 38 11 10 15 16 14 15 25 ...

### Mean Median SD skewness kurtosis

mean(price)

## [1] 121699.8

median(price)

## [1] 107000

sd(price)

## [1] 79516.84

skewness(price)

## [1] 0.5520654

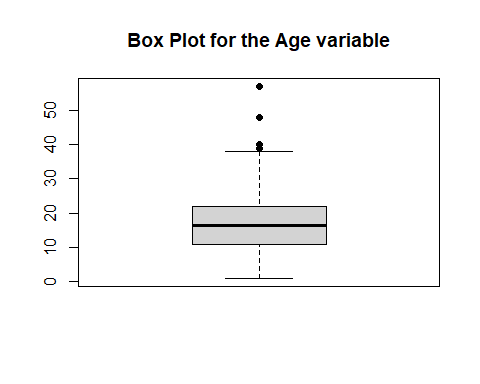
kurtosis(price)

## [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.

### 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 Right 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 1.07^{5}. And based on this graph we can infer that most of the boats are aged in between 10 to 25.

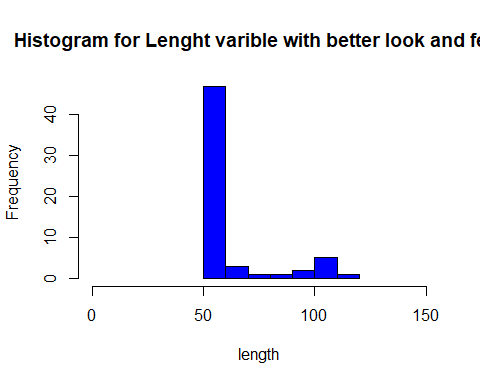
### Qunatile for age variable

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

## 0% 20% 40% 60% 80% 100%   
## 1 10 15 18 29 57

### Histogram for the length variable

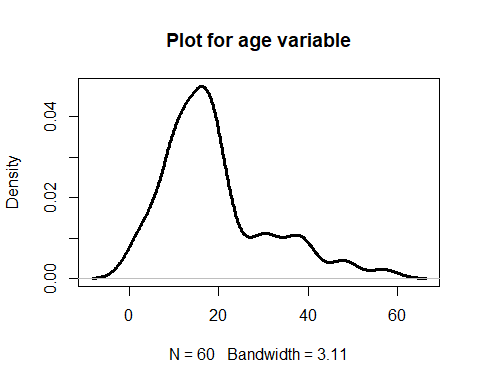
hist(length,col="blue",main="Histogram for Lenght varible with better look and feel",  
 xlim = c(0,150))



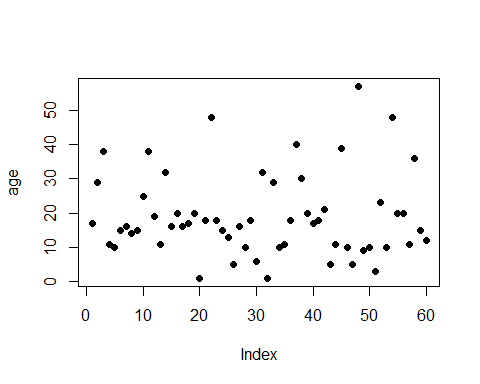
By viewing this histogram graph for length variable, we can say that this do not follow a symmetric shaped curve. We can say that this is a Right skewed curve. And if we closely observe the graph, after the ‘100’, there is a raise in the graph and this might be a double curved graph.

### stem and leaf plot for age variable

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



plot(age,pch=19)



### 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"))

