Week 1 Assignment

ALY 6015 Intermediate Analytics

Submitted to:

<ins>**Joseph Manseau**</ins>

Date :02/25/2020

Submitted by:

Ashlesha Kshirsagar(001082234)

Introduction:

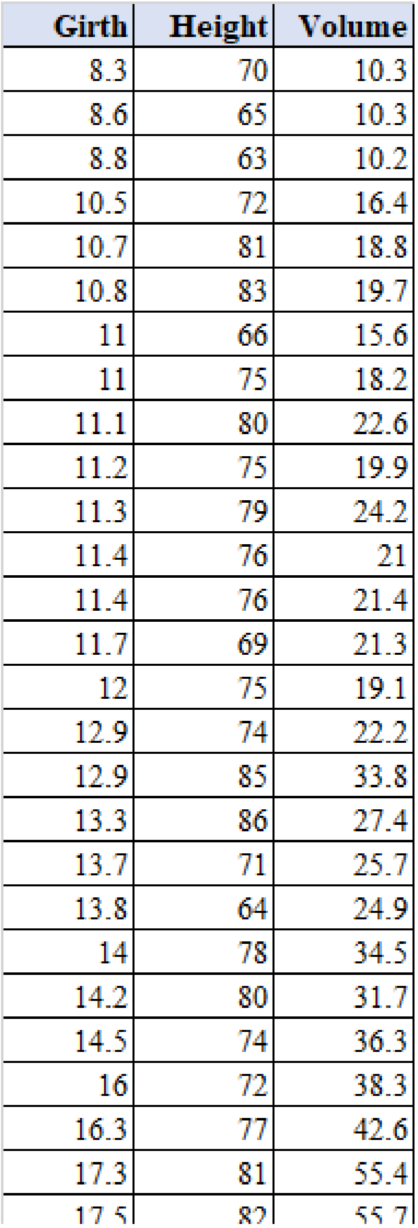
Data analysis involves search for patterns and these patterns can be found using statistics. Various statistical concepts like Descriptive Statistics, correlation, simple linear regression and multiple linear regression are widely used for data analysis. Descriptive statistics is used to understand the Measures of Central tendency and Measures of spread. Simple linear regression involves one predictor labelled as x and a dependent variable that is labelled as y. As per this regression model, the relationship between x and y can be summarized as a straight-line graph. (Godfrey, 2020).An extension to simple linear is multiple linear regression and it has more than one independent variable that can be predictors for a dependent variable. With the use of correlation, we can check how the independent variables are correlated within themselves as well as how is the correlation between the independent and dependent variable.

The packages used were psych for saving the table in Csv format, ggplot2, ggcorrplot and MASS for rubber dataset. With the help of statistical concepts and R outputs we will understand how to analyze and interpret trees, Rubber and oddbooks dataset that are available in R

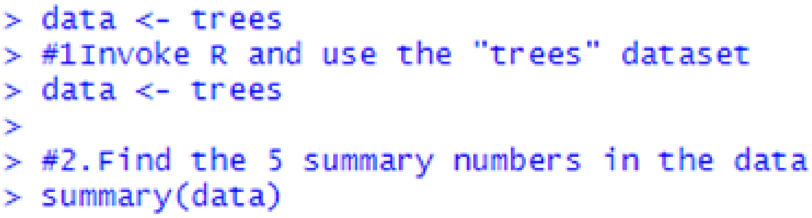
PART A

1.1 Invoke R and use the "trees" dataset

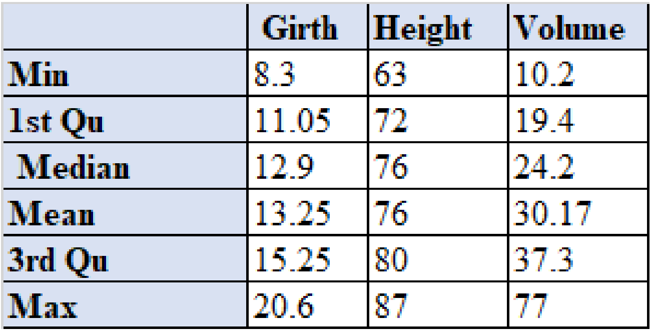
1.2 Find the 5 summary numbers in the data



Girth	Height	Volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11	66	15.6
11	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21
11.4	76	21.4
11.7	69	21.3
12	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14	78	34.5
14.2	80	31.7
14.5	74	36.3
16	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7



```
> data <- trees
> #1Invoke R and use the "trees" dataset
> data <- trees
>
> #2.Find the 5 summary numbers in the data
> summary(data)
```



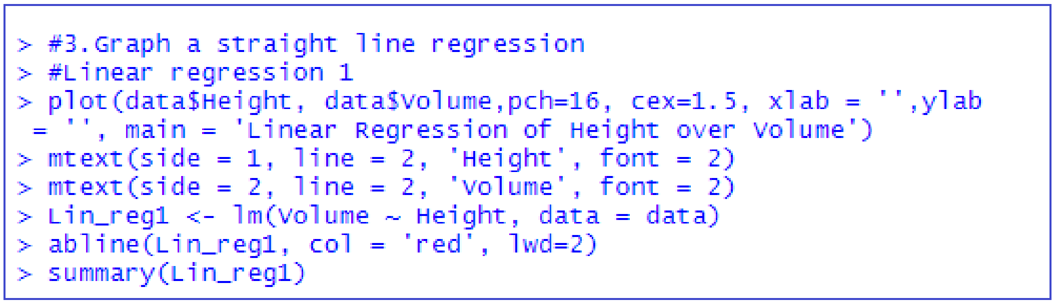
	Girth	Height	Volume
Min	8.3	63	10.2
1st Qu	11.05	72	19.4
Median	12.9	76	24.2
Mean	13.25	76	30.17
3rd Qu	15.25	80	37.3
Max	20.6	87	77

Fig 1.1 Summary of the Trees dataset

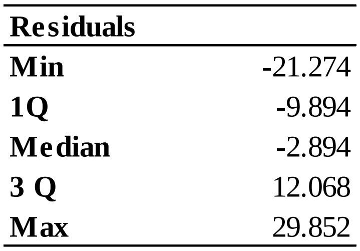
Observation:

The results of the above code show the descriptive statistics of the trees data. For each numeric variable in the data it will tell you the Minimum, Maximum, Mean, Median, 1st Quartile and 3rd Quartile values. Median and mean values will help in understanding the distribution of data i.e. whether it is normally distributed or not. Since the Median value of Girth and volume is **greater than** the mean value it clearly shows that the distribution is **Right skewed** whereas for the Height of the trees the Median and Mean values are equal which shows it is **Normally**

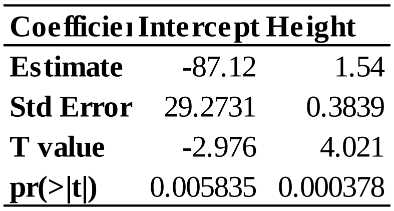
1.3 Graph a straight-line Regression



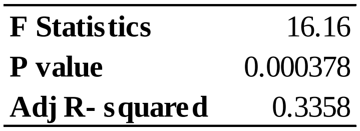
```
> #3.Graph a straight line regression
> #Linear regression 1
> plot(data$Height, data$Volume,pch=16, cex=1.5, xlab = '',ylab
 = '', main = 'Linear Regression of Height over Volume')
> mtext(side = 1, line = 2, 'Height', font = 2)
> mtext(side = 2, line = 2, 'Volume', font = 2)
> Lin_reg1 <- lm(Volume ~ Height, data = data)
> abline(Lin_reg1, col = 'red', lwd=2)
> summary(Lin_reg1)
```



Residuals	
Min	-21.274
1Q	-9.894
Median	-2.894
3 Q	12.068
Max	29.852



Coefficiei	Intercept	Height
Estimate	-87.12	1.54
Std Error	29.2731	0.3839
T value	-2.976	4.021
pr(>|t|)	0.005835	0.000378



F Statistics	16.16
P value	0.000378
Adj R- squared	0.3358

In Figure 1.2, we can see that there exists a positive correlation between height and volume. But we cannot say that there exists a strong relationship as the data points are not clustered tightly around the line of best fit. Also the Adjusted R squared value is very less which is 0.3358

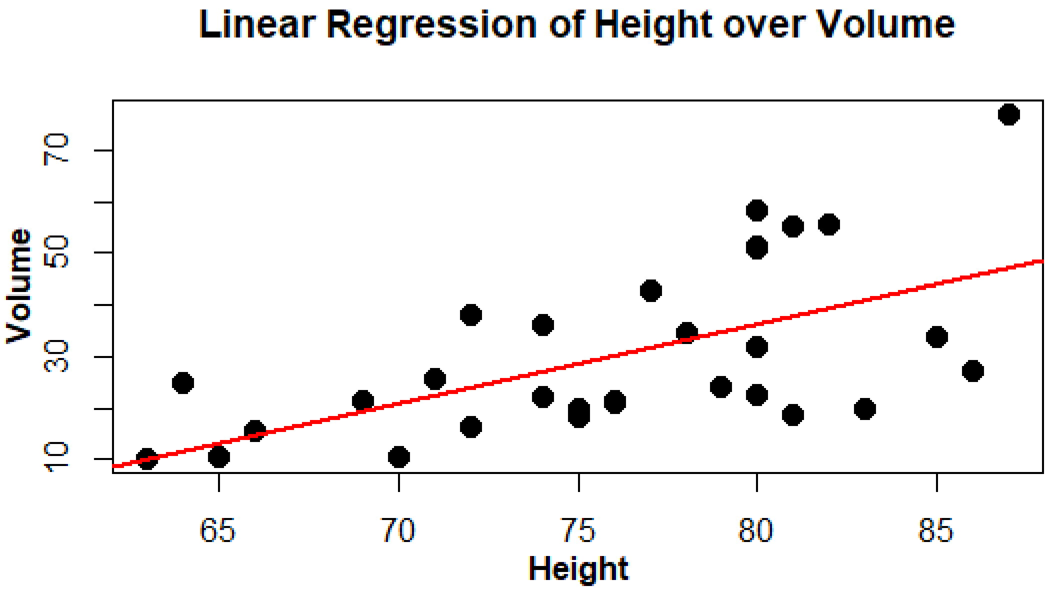
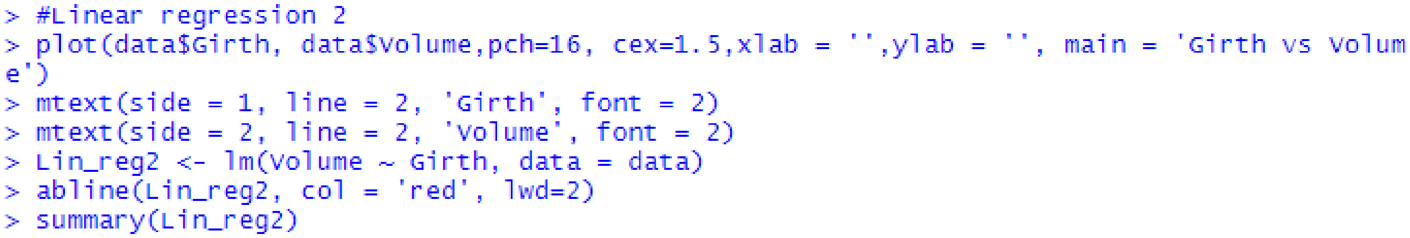
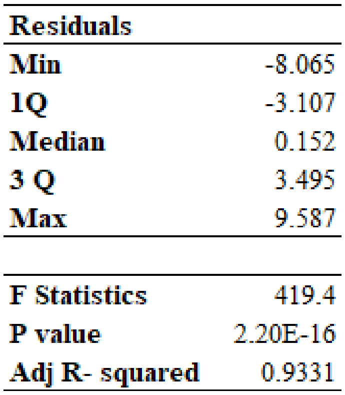


Fig 1.2: Scatter plot between Height and Volume (Dependent variable)

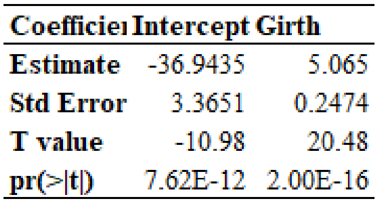


```
> #Linear regression 2
> plot(data$Girth, data$Volume,pch=16, cex=1.5,xlab = '',ylab = '', main = 'Girth vs Volum
e')
> mtext(side = 1, line = 2, 'Girth', font = 2)
> mtext(side = 2, line = 2, 'Volume', font = 2)
> Lin_reg2 <- lm(Volume ~ Girth, data = data)
> abline(Lin_reg2, col = 'red', lwd=2)
> summary(Lin_reg2)
```



Residuals	
Min	-8.065
1Q	-3.107
Median	0.152
3 Q	3.495
Max	9.587

F Statistics	419.4
P value	2.20E-16
Adj R- squared	0.9331



Coefficiei	Intercept	Girth
Estimate	-36.9435	5.065
Std Error	3.3651	0.2474
T value	-10.98	20.48
pr(>|t|)	7.62E-12	2.00E-16

In Figure 1.3, we can see that there exists a positive correlation between diameter and volume. Also, there exists a strong relationship as the data points are clustered tightly around the line of best fit. Also, we can see that the adjusted R squared value is 0.9331 which means this Tree diameter causes significant change in the volume.

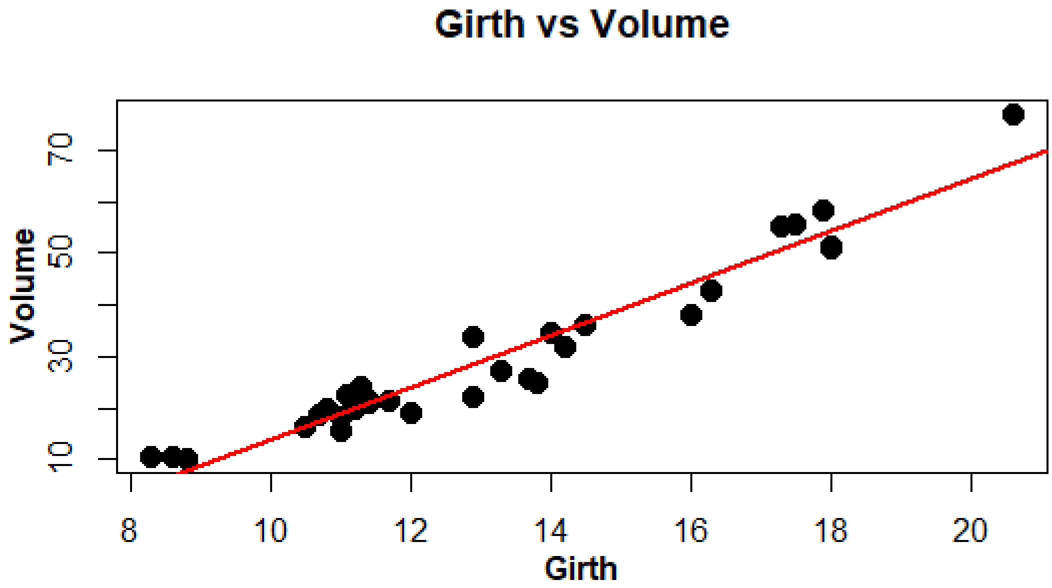
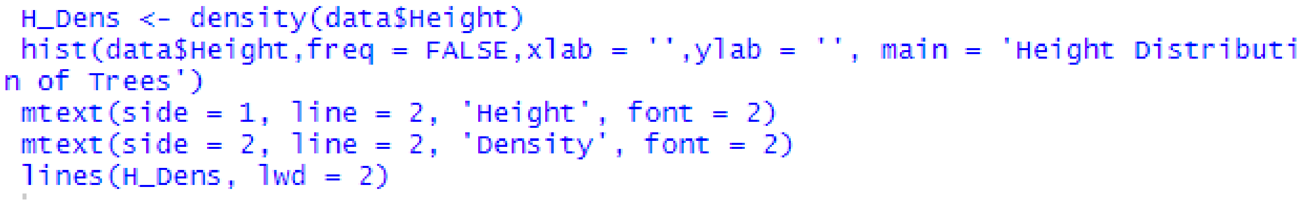


Fig 1.3 Scatter plot between Girth and Volume (Dependent variable)

Observation: Scatter plot mainly indicates whether there exists correlation between the dependent and independent variable and also shows the cluster of data on specific intervals. One of main assumptions of linear regression is that dependent variable will have Linear relationship with the independent variables. In the Fig 1.2 and Fig 1.3 it is clearly showing the Girth and Height has linear relationship (positive correlation) with Volume.

1.4 Create Histograms and density plots



```
 H_Dens <- density(data$Height)
 hist(data$Height,freq = FALSE,xlab = '',ylab = '', main = 'Height Distributi
n of Trees')
 mtext(side = 1, line = 2, 'Height', font = 2)
 mtext(side = 2, line = 2, 'Density', font = 2)
 lines(H_Dens, lwd = 2)
```

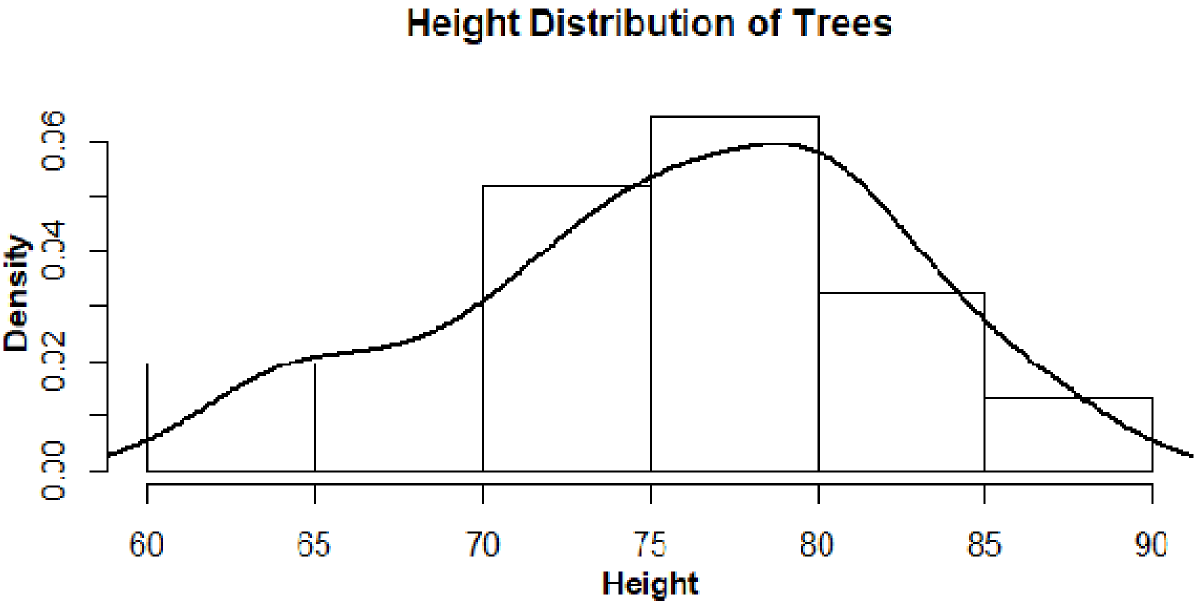
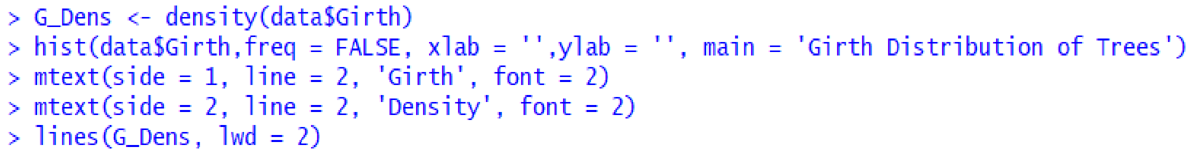


Fig 1.4: Histogram and Density plot on Height Distribution of trees



```
> G_Dens <- density(data$Girth)
> hist(data$Girth,freq = FALSE, xlab = '',ylab = '', main = 'Girth Distribution of Trees')
> mtext(side = 1, line = 2, 'Girth', font = 2)
> mtext(side = 2, line = 2, 'Density', font = 2)
> lines(G_Dens, lwd = 2)
```

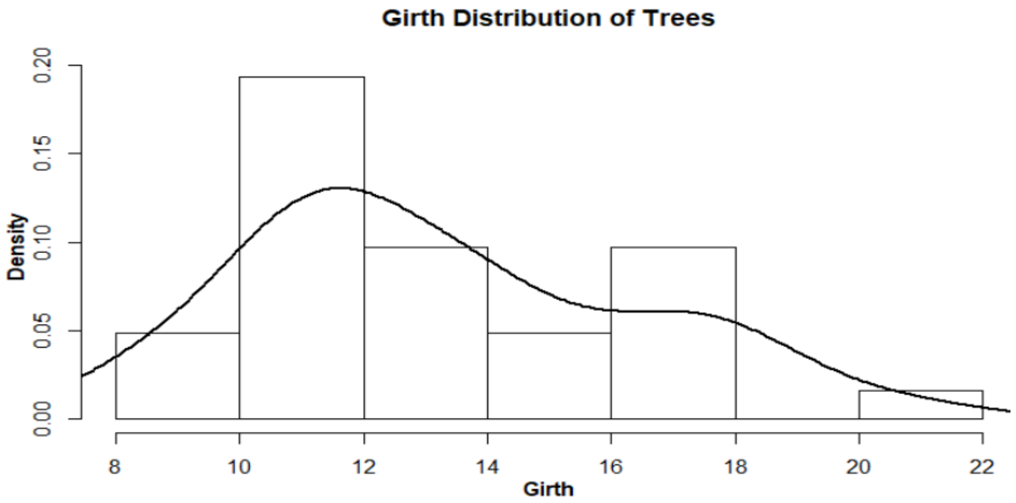
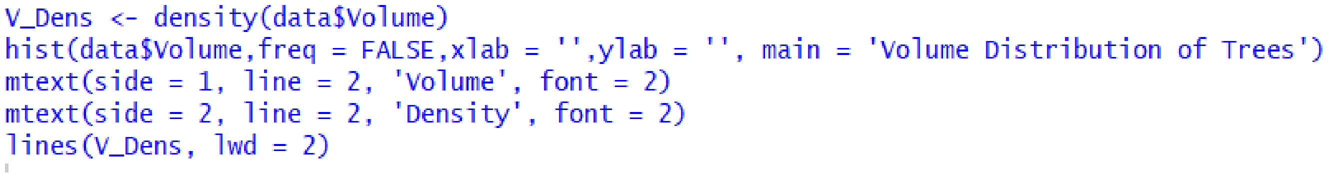


Fig 1.5: Histogram and Density plot on Girth Distribution of trees



```
V_Dens <- density(data$Volume)
hist(data$Volume,freq = FALSE,xlab = '',ylab = '', main = 'Volume Distribution of Trees')
mtext(side = 1, line = 2, 'Volume', font = 2)
mtext(side = 2, line = 2, 'Density', font = 2)
lines(V_Dens, lwd = 2)
```

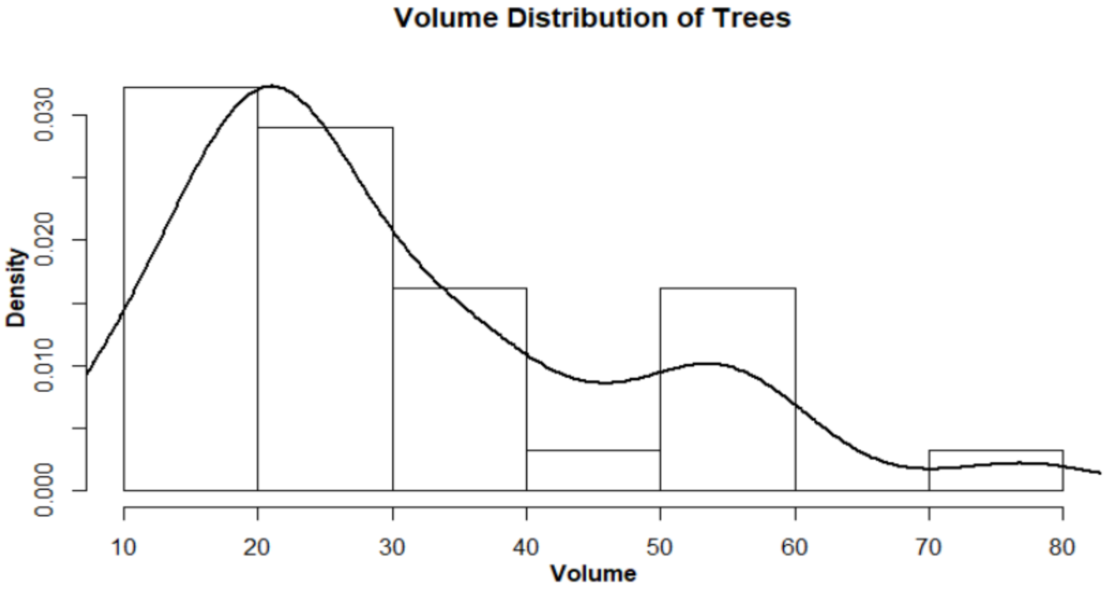
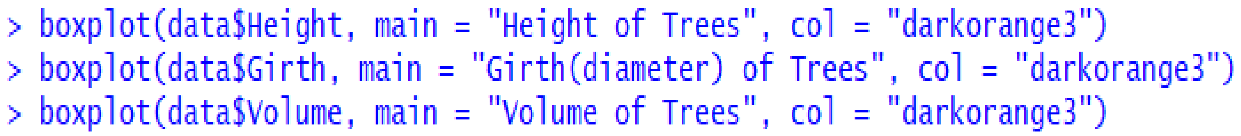


Fig 1.6 Histogram and Density plot on Volume Distribution of trees

Observation:

The above figures 1.4,1.5 and 1.6 shows the distribution of data in the trees data. Histograms are very helpful in tell the distribution, but the distribution is mostly changing if we change the bins. But Density plot is the variation of histogram which uses **kernel smoothing** to plot values and very helpful in understanding the real distribution of the data by eliminating the noise. Peak of density curve correlates with the larger bins in histogram which indicate the more frequent value(mode). From Fig 1.4 we can tell that the height values are normally distributed because the median and mean of the data is **same**. Whereas the volume and girth values as on fig 1.5and Fig 1.6 follows Right skewed distribution. Volume values more skewed right because there is **high difference** in the median and mean value.

1.5 Create Boxplots



```
> boxplot(data$Height, main = "Height of Trees", col = "darkorange3")
> boxplot(data$Girth, main = "Girth(diameter) of Trees", col = "darkorange3")
> boxplot(data$Volume, main = "Volume of Trees", col = "darkorange3")
```

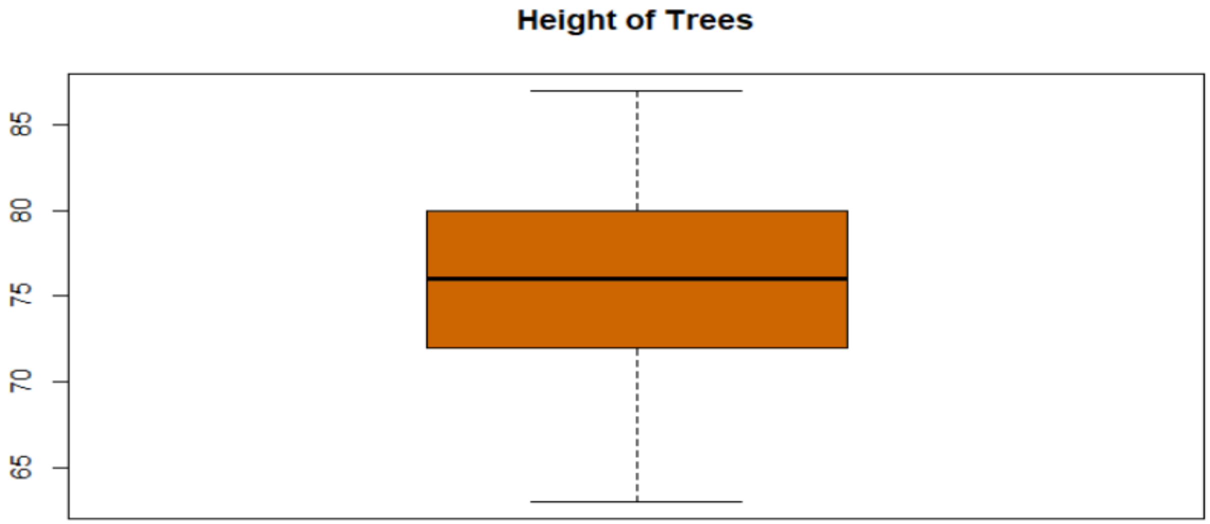


Fig 1.7 Box Plot showing the distribution of volume data of trees

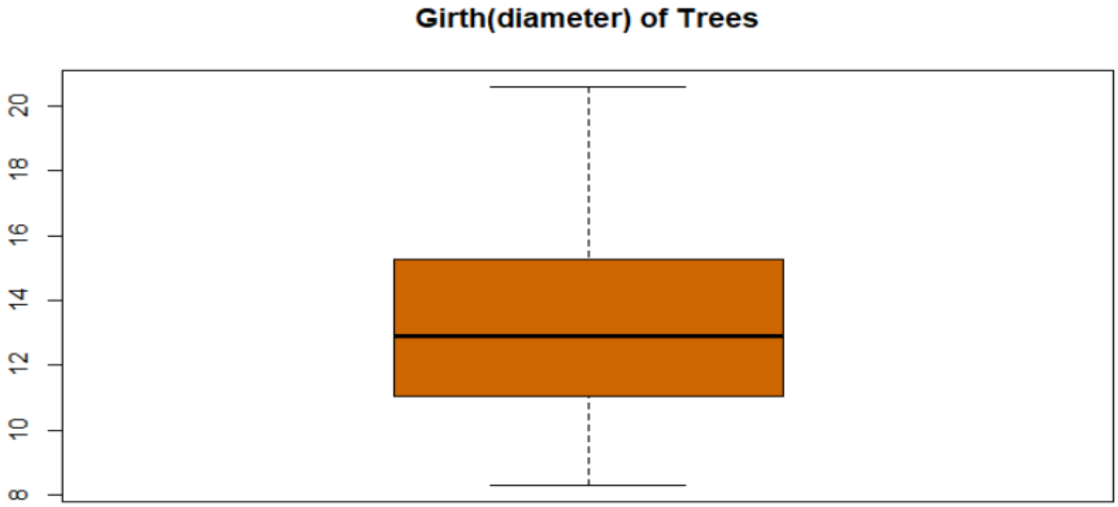


Fig 1.8 Box Plot showing the distribution of Diameter data of trees

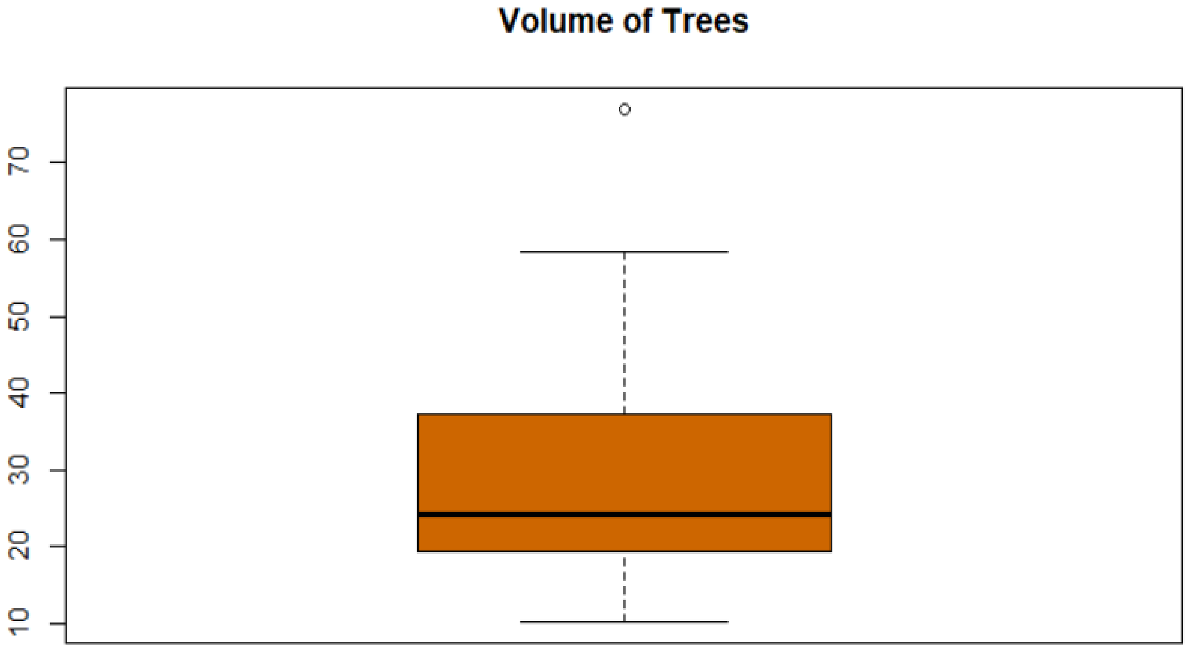
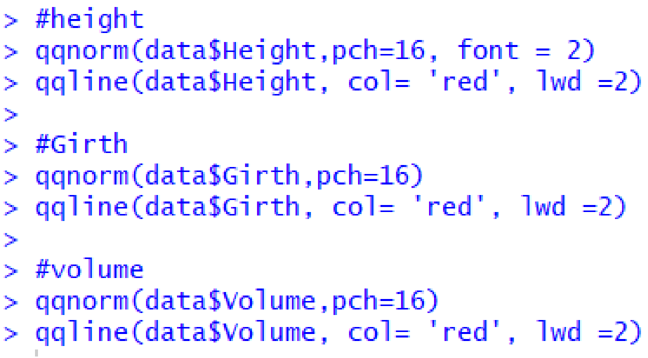


Fig1.9.Box Plot showing the distribution of Volume data of trees

Observation:

Box plots are useful as they provide a visual summary of the data which enables to quickly identify minimum, maximum, median, percentile values, the dispersion of the data set, and signs of skewness. The box in the graph shows the inter-quartile range where most of the data is present. Fig 1.7.clearly shows the five-point summary of the height of trees and the median line(50th percentile) equally splits the box which indicates that data is normally distributed. In Fig 1.8 and Fig 1.9 it clearly seen that the median line is closer the 25th percentile which indicates that the diameter and volume values are **Right skewed.** Closer the line it gets to the 25th percentile more it skewed to right and closer it gets to the 75th percentile more it is skewed to left.

1.6.Normal probability plots



```
> #height
> qqnorm(data$Height,pch=16, font = 2)
> qqline(data$Height, col= 'red', lwd =2)
>
> #Girth
> qqnorm(data$Girth,pch=16)
> qqline(data$Girth, col= 'red', lwd =2)
>
> #volume
> qqnorm(data$Volume,pch=16)
> qqline(data$Volume, col= 'red', lwd =2)
```

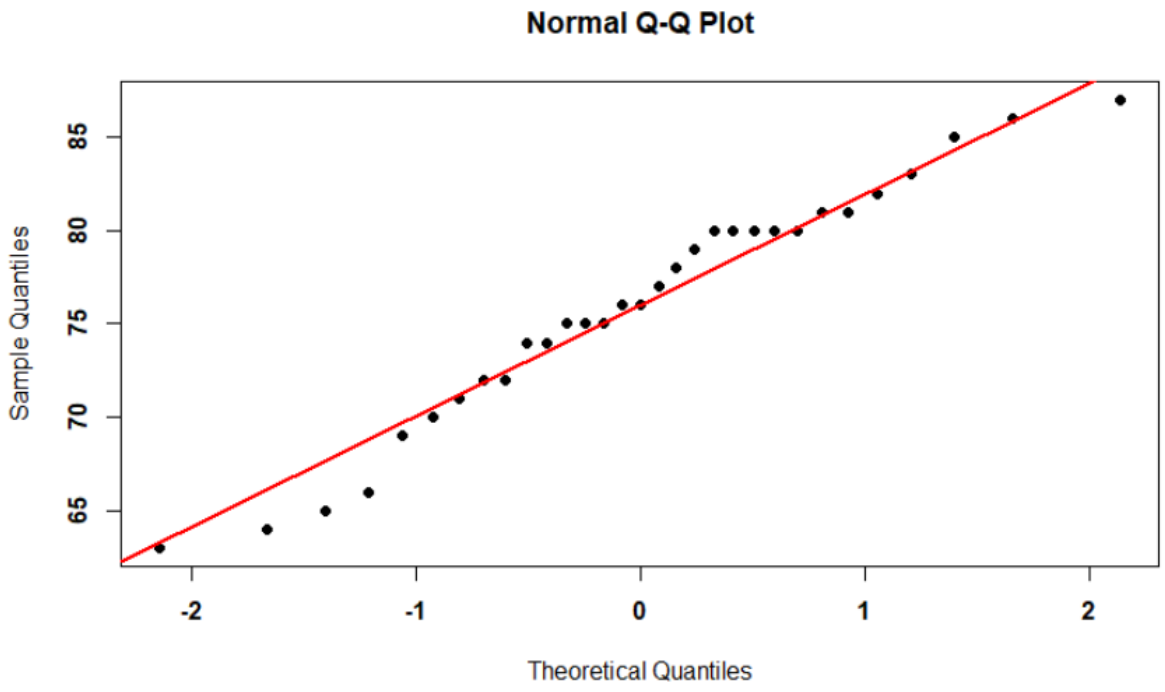


Fig 1.10: Normal Q-Q Plot for Height of Trees

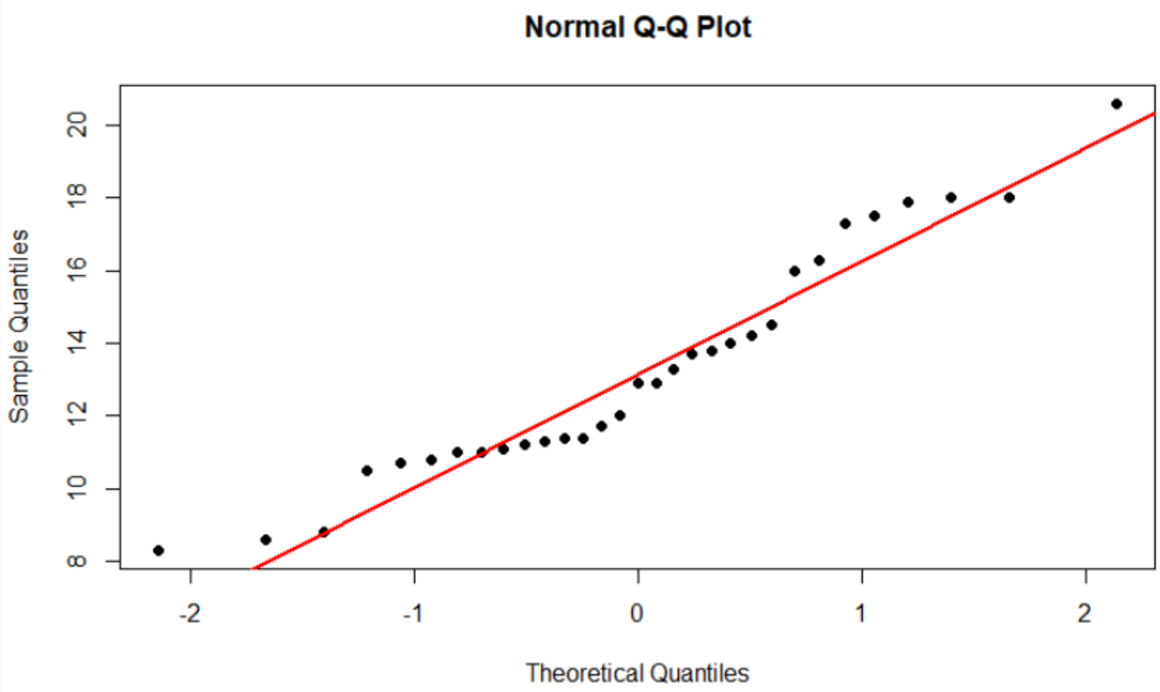


Fig 1.11 Normal Q-Q Plot for Girth(Diameter) of Trees

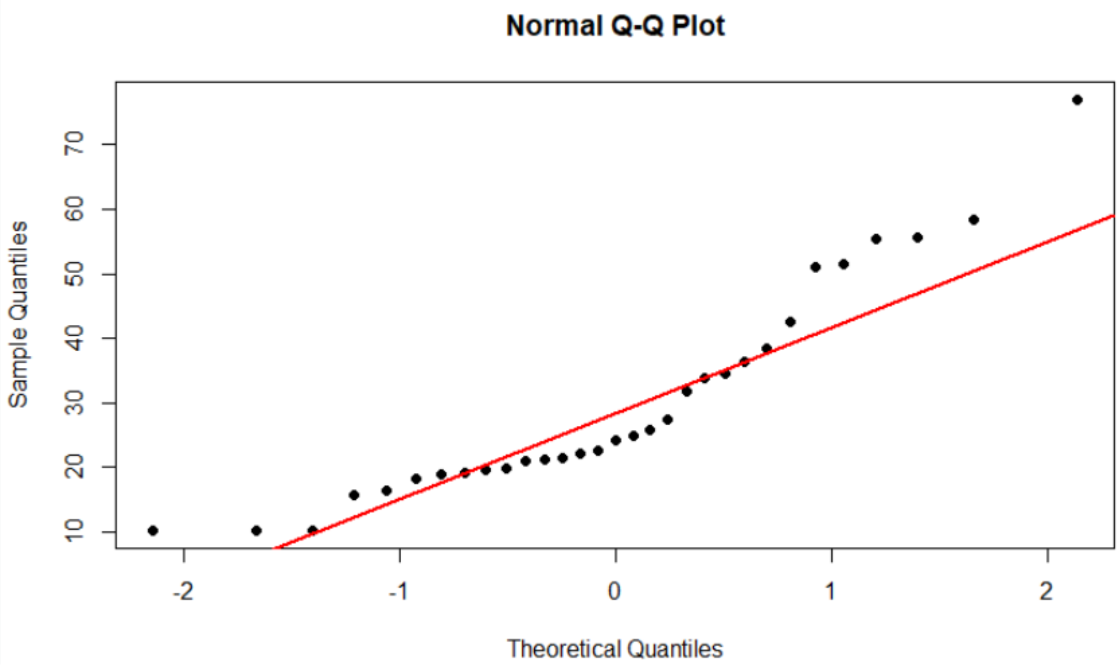
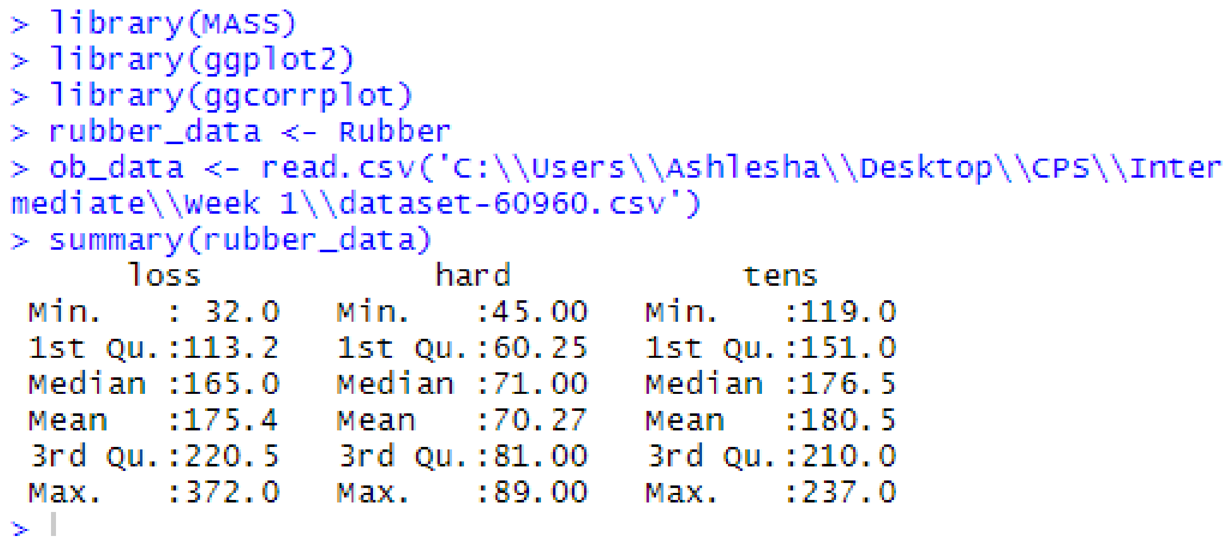


Fig 1.12 Normal Q-Q Plot for Volume of Trees

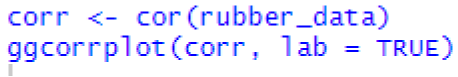
Observation

Normal Probability plot is one of the informative ways of visualizing the data to check whether the data is normally distributed or not. This plot graphs z-scores of our data and the straight diagonal line in the normal probability plot indicates the normal distribution and data points that deviate from the straight line indicates it is not a normally distributed plot. From the Fig 1.10,1.11 and 1.12 we can clearly see that height of trees is normal distributed and the volume and diameter of trees is a skewed distribution since many data points deviates from straight line.

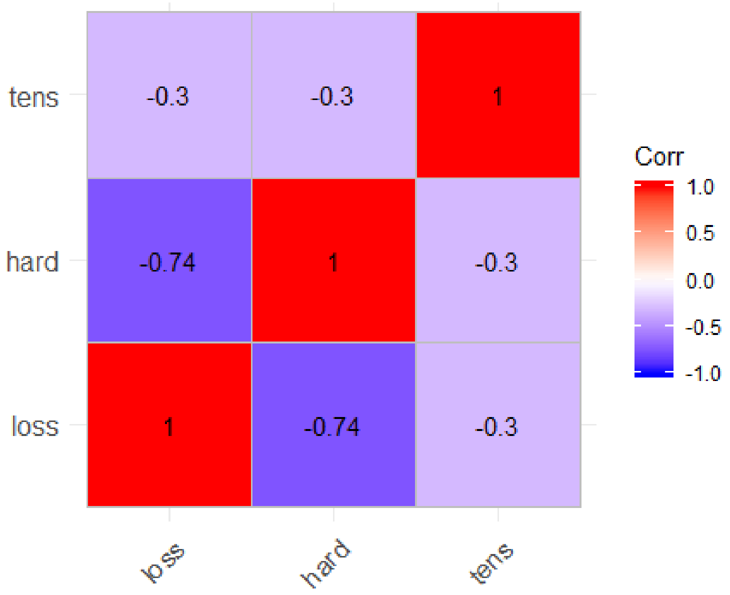
PART B



```
> library(MASS)
> library(ggplot2)
> library(ggcorrplot)
> rubber_data <- Rubber
> ob_data <- read.csv('C:\\Users\\Ashlesha\\Desktop\\CPS\\Inter
mediate\\week 1\\dataset-60960.csv')
> summary(rubber_data)
      loss            hard            tens      
 Min.   : 32.0   Min.   :45.00   Min.   :119.0  
 1st Qu.:113.2   1st Qu.:60.25   1st Qu.:151.0  
 Median :165.0   Median :71.00   Median :176.5  
 Mean   :175.4   Mean   :70.27   Mean   :180.5  
 3rd Qu.:220.5   3rd Qu.:81.00   3rd Qu.:210.0  
 Max.   :372.0   Max.   :89.00   Max.   :237.0  
> 
```



```
corr <- cor(rubber_data)
ggcorrplot(corr, lab = TRUE)
```

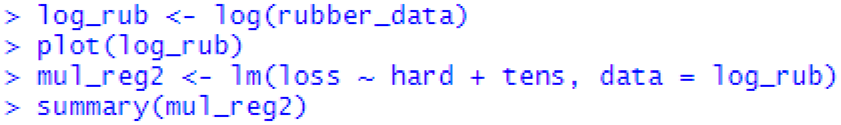


From the correlation output we can see that correlation of loss to loss is 1 which is the obvious case as correlation between the same variables will be the strongest.

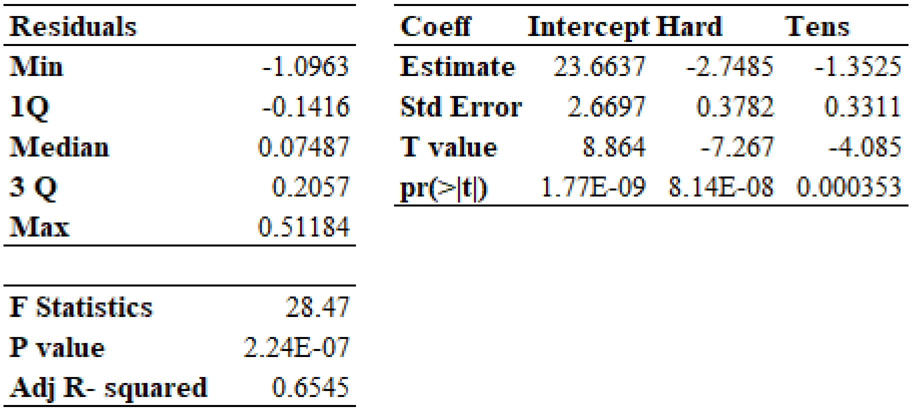
Fig 2.1 Correlation Matrix for Rubber

Observation: From Fig 2.1As we can see That red represents strong positive correlation and blue represents strong negative We can easily identify that hard and loss are strongly negatively correlated whereas tens and loss are negatively correlated but not strong.

Multiple Regression using log normalization on Rubber Data



```
> log_rub <- log(rubber_data)
> plot(log_rub)
> mul_reg2 <- lm(loss ~ hard + tens, data = log_rub)
> summary(mul_reg2)
```



Residuals	
Min	-1.0963
1Q	-0.1416
Median	0.07487
3 Q	0.2057
Max	0.51184

Coeff	Intercept	Hard	Tens
Estimate	23.6637	-2.7485	-1.3525
Std Error	2.6697	0.3782	0.3311
T value	8.864	-7.267	-4.085
pr(>|t|)	1.77E-09	8.14E-08	0.000353

F Statistics	28.47
P value	2.24E-07
Adj R- squared	0.6545

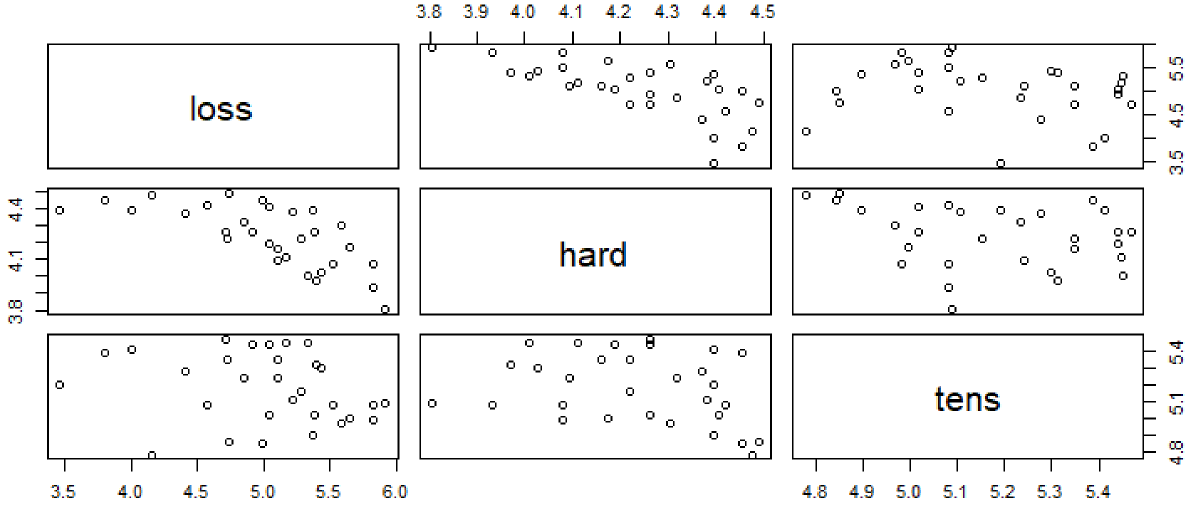
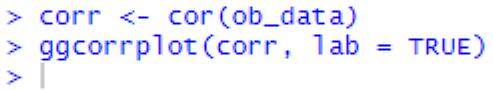
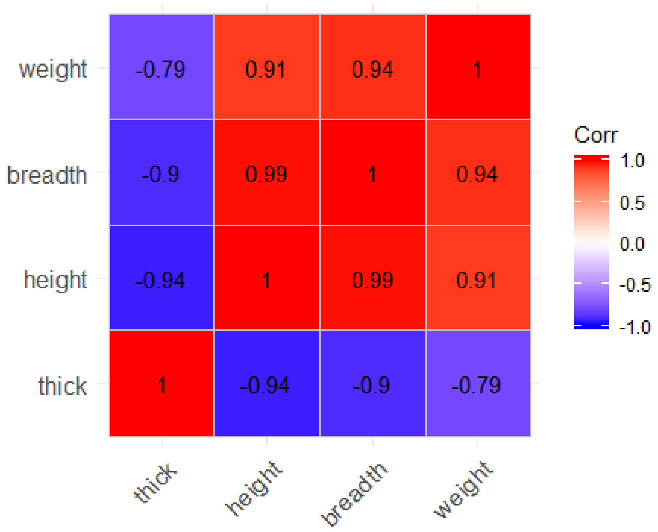


Fig 2.2 Multivariate analysis on Rubber data

Observation: From the output we can see that the coefficient of determination is approaching 0.65. This means 65 percent of the variation of the data is covered by the regression line which would mean this model can be used to predict loss by abrasion. However, a 99 percent coverage of variation is considered to be the best regression model.



```
> corr <- cor(ob_data)
> ggcorrplot(corr, lab = TRUE)
> 
```

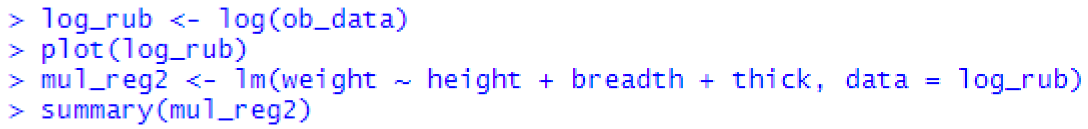


	thick	height	breadth	weight
weight	-0.79	0.91	0.94	1
breadth	-0.9	0.99	1	0.94
height	-0.94	1	0.99	0.91
thick	1	-0.94	-0.9	-0.79

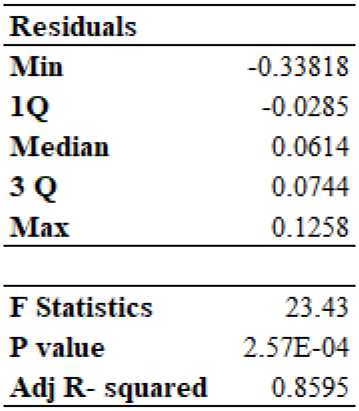
From the correlation output and the matrix above we can see that there is no strong correlation in any of the variables. Both the positive and negative correlations are weak and thus this further proves that none of the variable can be predicted using multiple linear regression.

Fig 2.3 Correlation Matrix for Odd books

Multiple Regression using log normalization on Odd Books Data

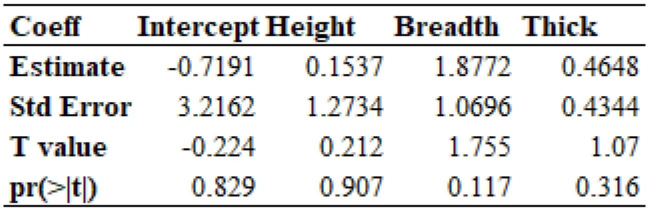


```
> log_rub <- log(ob_data)
> plot(log_rub)
> mul_reg2 <- lm(weight ~ height + breadth + thick, data = log_rub)
> summary(mul_reg2)
```



Residuals	
Min	-0.33818
1Q	-0.0285
Median	0.0614
3 Q	0.0744
Max	0.1258

F Statistics	23.43
P value	2.57E-04
Adj R- squared	0.8595



Coeff	Intercept	Height	Breadth	Thick
Estimate	-0.7191	0.1537	1.8772	0.4648
Std Error	3.2162	1.2734	1.0696	0.4344
T value	-0.224	0.212	1.755	1.07
pr(>|t|)	0.829	0.907	0.117	0.316

<u>Observation</u>: From the output we can see that the coefficient of determination is approaching 0.85. This means 85 percent of the variation of the weight will depend on the height, breadth and thickness. We can also say that the accuracy of this model is the 85%

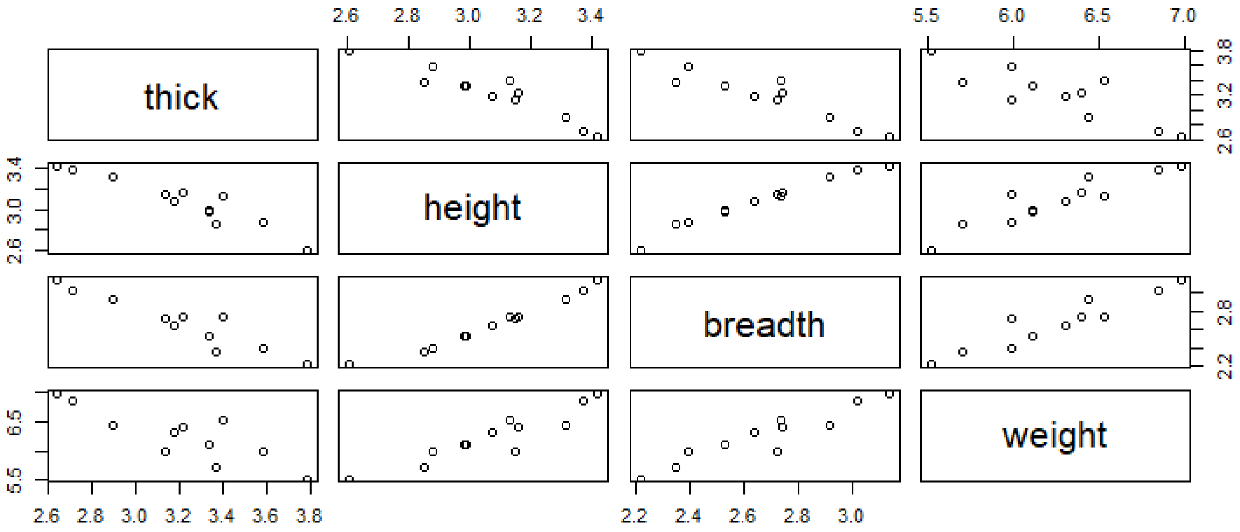


Fig 2.4 Fig 2.2 Multivariate analysis on Odd Books data

Conclusion

We have performed descriptive statistics which helps in understand the distribution of the data. We learned the functions which helps to plot the histogram, boxplots and density plot which helps to visually check how the data is distributed. Correlation matrix helps in understanding the correlation between the dependent and independent variables. Also, logarithm of one or more variables improves the fit of the model by transforming the distribution of the features

The line of best fit has maximum number of points clustered around it if the R squared value is higher.

Reference

1. Godfrey, K. (2020). Simple Linear Regression in Medical Research. Retrieved 12 January 2020
2. J H MainDonald(2008), Using R for Data Analysis and Graphics Introduction, Code and Commentary
3. Nathans, L., Oswald, F., & Nimon, K. (2020). Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. Retrieved 12 January 2020, from