**Module 1 – Report**

**Name: Sunil Raj Thota**

**NUID: 001099670**

**Date: 01/21/2021**

**Title: Descriptive Statistics and Regression Analysis with R**

**ALY 6015 – Intermediate Analytics**

**Prof. Roseanna Hopper**

**Introduction**

In this Assignment, I used R and R studio to perform Descriptive Statistics and Regression Analysis on various preloaded datasets provided by CRAN. We can perceive a list of all data sets in R using the function called data().

The dataset used for the Part A analysis is the Trees. It has the information of 3 variables as Height, Girth, and Volume. Part A is mostly focused on Descriptive Statistics, Exploratory Data Analysis, Linear Regression, and Plots.

The other datasets used for Part B analysis are the Rubber dataset from the MASS package which has information of 3 variables as loss, hard, and tens, and the Odd Books dataset from the DAAG package which has information of 4 variables as weight, height, breadth, and width. Part B is all about the Linear Regression and Data Visualization of 2 various datasets.

Descriptive Statistics contains Mean, Standard Deviation, Median, Quartiles, Variance, Kurtosis, Max, Skewness, and Min values. It provides us the basic information and synopsis of the given data set. Inferential Analysis provides us with insights about the performance of the data and helps us to forecast the movements for the future.

Linear Regression is mainly utilized for performing projecting analysis. The outcomes are for understanding the relation between 1 dependent and 1 or more independent variables. This relationship is analyzed by using Scatter Plots and Co-variance Plots/ Matrix. The Linear model of Y = mx + c, where x is the independent variable, Y is the dependent variable, c is the intercept, and m is the slope. I have used the regression techniques and visualized the data using different plotting techniques that are available in R.

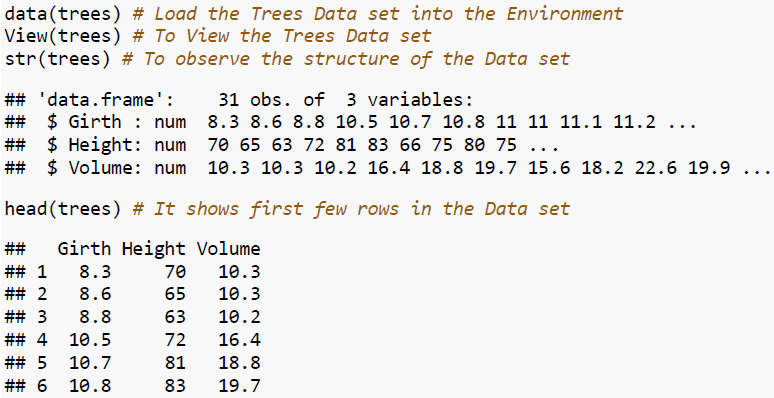
**Analysis**

**PART A**

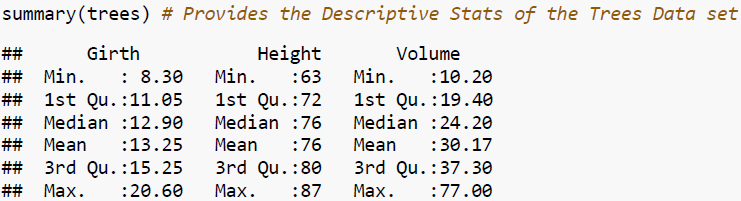
The Tree dataset is comprised of R’s dataset package and loaded for the Data Analysis. It contains 31 observations and 3 variables:Girth (in.)**,** Height (ft.), and Volume (cubic ft.)

**Steps:**

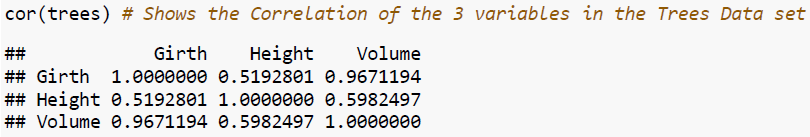
1. Trees Dataset is loaded using the command data(trees). We can view the dataset by using the command View(tress). To find the structure of the dataset we use the command str(trees) and the first six rows are printed in the console by using the command head(trees)



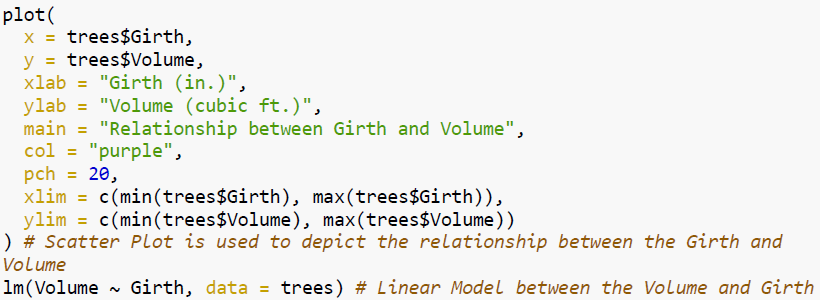
1. Summary of the Dataset has been used to breakdown the complex understanding of the dataset into simple and easy to analyze way. It can be attained using the command summary(trees).

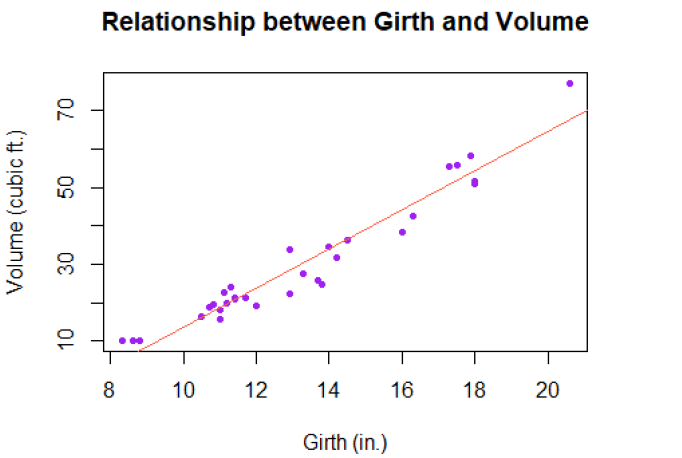


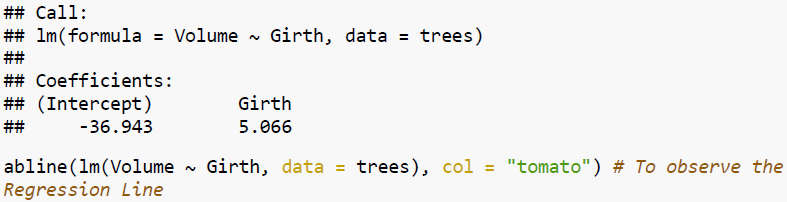
1. Correlation Matrix is used to find the relationship between the variables using the command cor(). The correlation coefficient provides understandings of the relationship between the variables. Girth and Volume has a strong relationship than Height and Volume or Girth and Height



1. By using the Scatter Plot, let's depict the relationship between the Girth and Volume variables. As the Volume increases, the Girth also increases. So, it is a Linear relationship and it can be obtained by using the command lm()

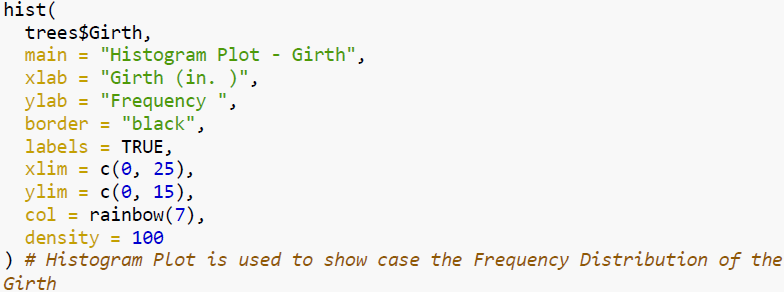




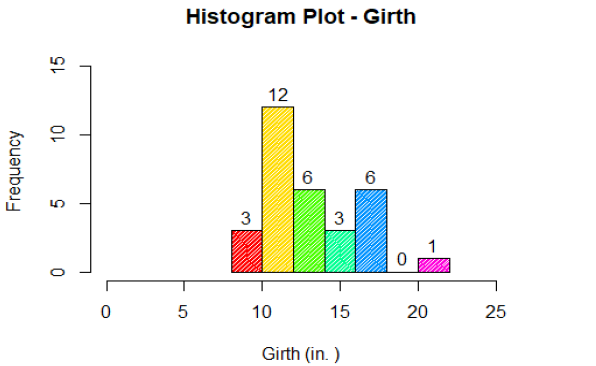


1. Now let’s plot Histogram, Density, Box Plot, and Normal Probability plots for Girth, Height, and Volume variables of the Trees dataset

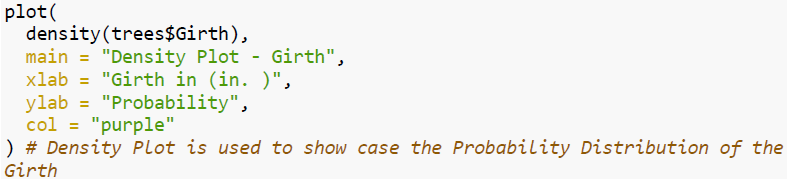
**Histogram Plot - Girth:**

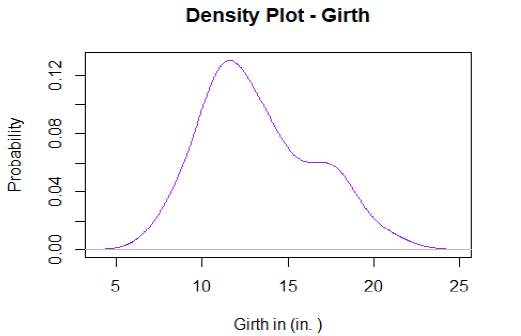


Histogram can be plotted by using the hist() function. In this, I used trees$Girth data to see on a histogram. I have chosen main, xlab, ylab, border, labels, xlim, ylim, col, and density as parameters that outline the title, x-axis label, y-axis label, border color, x-axis limits, y-axis limits, color, and density accordingly.



**Density Plot - Girth:**

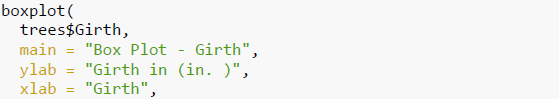


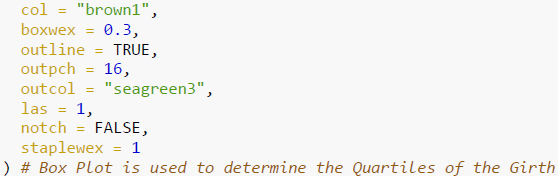


Density can be plotted by using the plot() function. In this, I used the density() method inside the plot function to invoke the trees$Girth data. I have chosen main, xlab, ylab, and col as parameters that outline the title, x label, y label, and color accordingly.

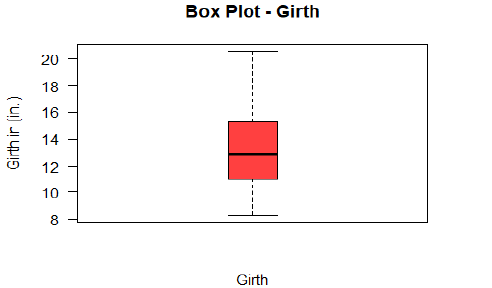
**Box Plot - Girth:**

Box Plot can be plotted by using the boxplot() function. In this, the plot function calls the trees$Girth data. I have chosen main, xlab, ylab, col, boxwex, outline, outpch, las, notch, and staplewex as parameters that outline the title, x label, y label, color, optimal power transformation, outliers, outline point character, axis, to draw notches on the sides of boxes, and staple width expansion accordingly.

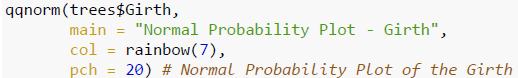




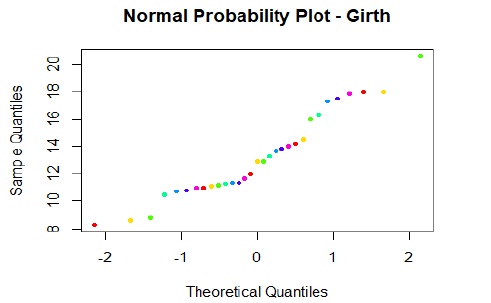
It demonstrates the Min, 1st, 3rd Quartile, Median, and Max values.



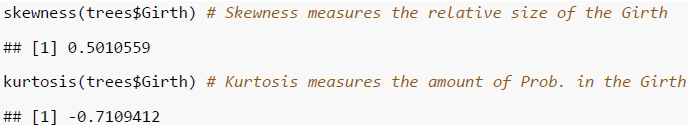
**Normal Probability Plot - Girth:**



Normal probability Plot can be plotted by using the qqnorm() function. In this, the plot function calls the trees$Girth data. I have chosen main, col, and pch as parameters that outline the title, color, and point character accordingly.



**Skewness and Kurtosis - Girth:**

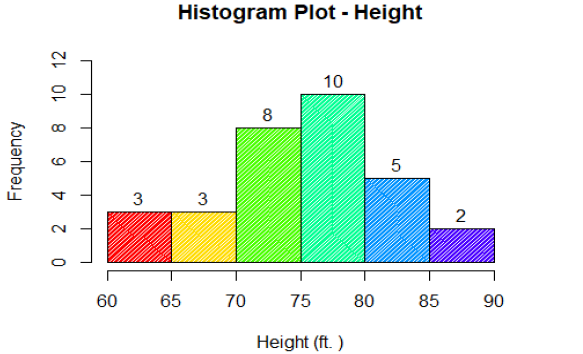


Skewness measures the relative size of the Girth and obtained the value as 0.5010559. Kurtosis measure the amount of probability in the Girth and obtained the value as -0.7109412.

Similarly, we can obtain these plots for Height and Volume variables of the Trees Dataset.

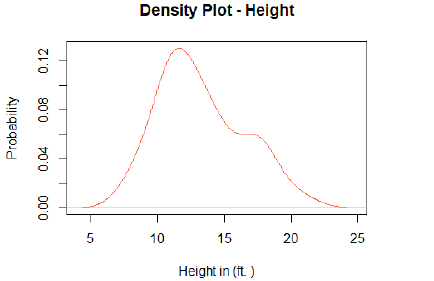
**Histogram Plot - Height:**

Let’s see the Histogram plot for Height variable of the Trees data set



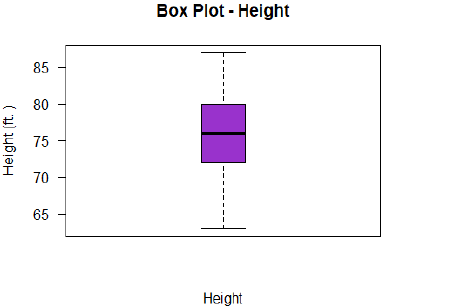
**Density Plot - Height:**

Let’s see the Density plot for Height variable of the Trees data set



**Box Plot - Height:**

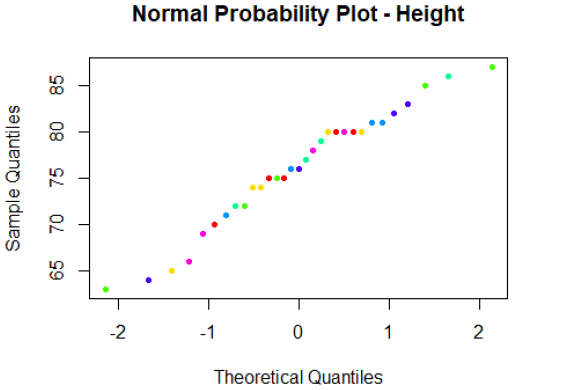
Let’s see the Box plot for Height variable of the Trees data set



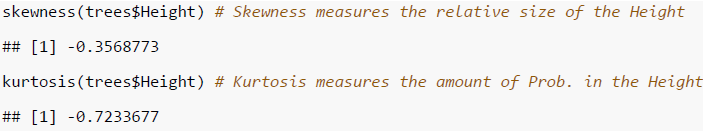
It shows the position of Min, 1st, 3rd Quartile, Max and Median values.

**Normal Probability Plot - Height:**

Let’s see the Normal Probability plot for Height variable in the Trees dataset



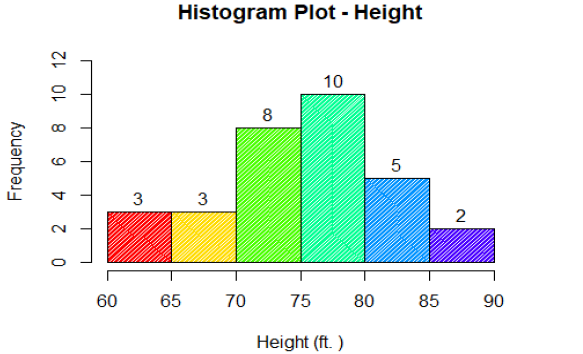
**Skewness and Kurtosis - Height:**



Skewness measures the relative size of the Height and obtained the value as 0.3568773. Kurtosis measure the amount of probability in the Height and obtained the value as -0.7233677.

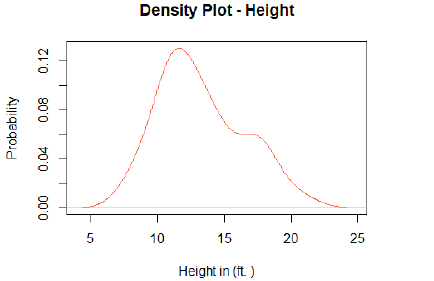
**Histogram Plot - Height:**

Let’s see the Histogram plot for Height variable of the Trees data set



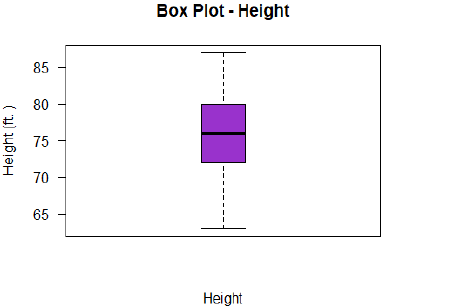
**Density Plot - Height:**

Let’s see the Density plot for Height variable of the Trees dataset



**Box Plot - Height:**

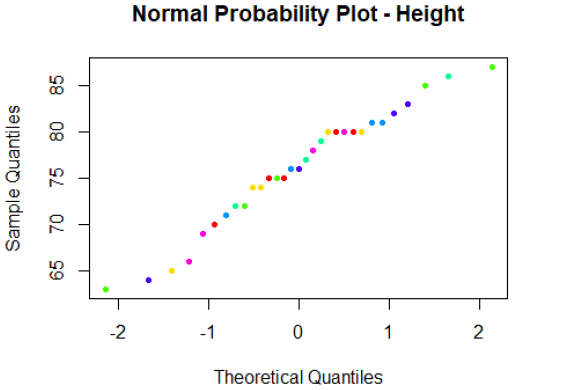
Let’s see the Box plot for Height variable of the Trees data set



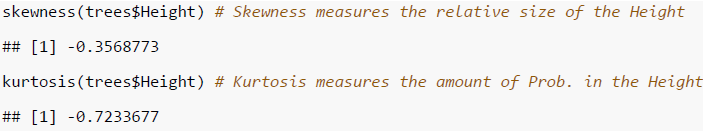
It depicts the position of Min, 1st, 3rd Quartile, Max, and Median values.

**Normal Probability Plot - Height:**

Let’s see the Normal Probability plot for Height variable of the Trees data set



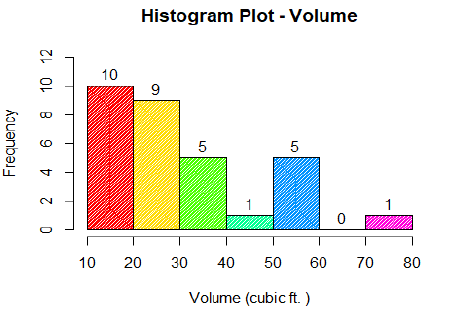
**Skewness and Kurtosis - Height:**



Skewness measures the relative size of the Height and obtained the value as 0.3568773. Kurtosis measure the amount of probability in the Height and obtained the value as -0.7233677.

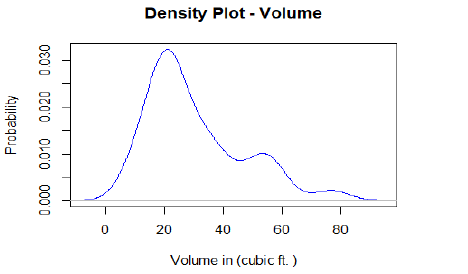
**Histogram Plot - Volume:**

Let’s see the Histogram plot for Volume variable of the Trees data set



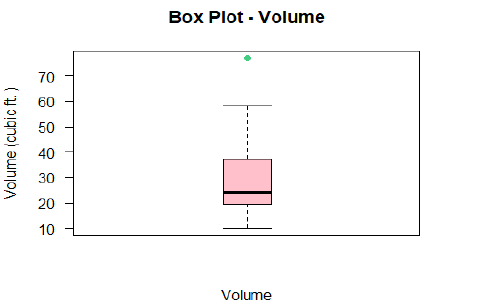
**Density Plot - Volume:**

Let’s see the Density plot for Volume variable of the Trees data set



**Box Plot - Volume:**

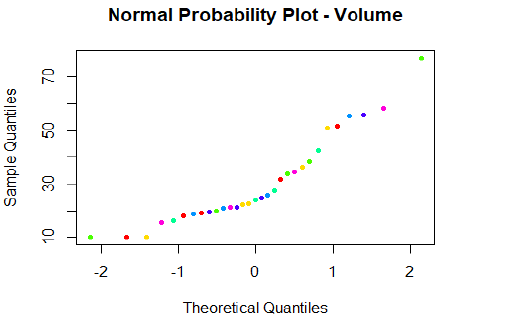
Let’s see the Box plot for Volume variable of the Trees data set



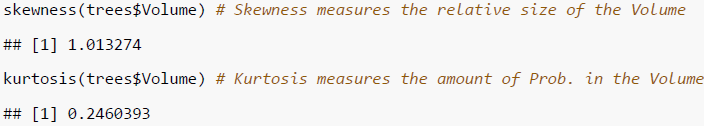
It shows the position of Min, 1st, 3rd Quartile, Median and Max values.

**Normal Probability Plot - Volume:**

Let’s see the Normal Probability plot for Volume variable of the Trees data set



**Skewness and Kurtosis - Volume:**

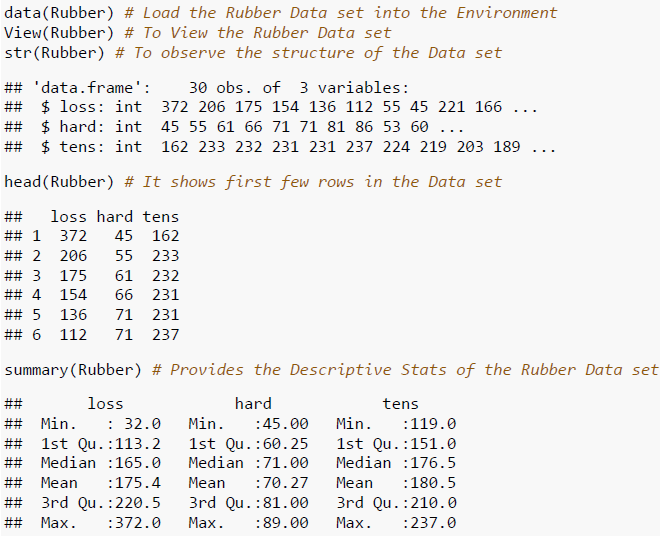


Skewness measures the relative size of the Volume and obtained the value as 1.013274. Kurtosis measure the amount of probability in the Volume and obtained the value as 0.2460393.

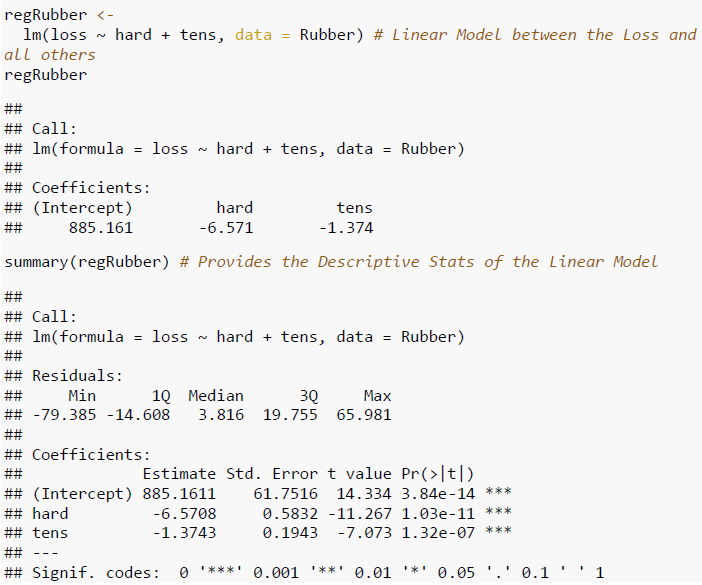
**PART B**

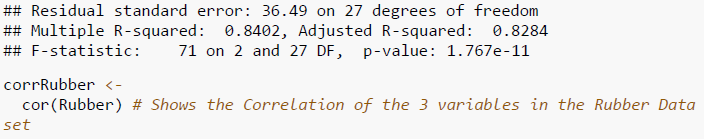
**Rubber:**

In this, Rubber and oddbooks datasets were used which are loaded by using MASS and DAAG Packages using CRAN. The Rubber dataset consists of 30 observations and 3 variables as loss, hard, and tens.

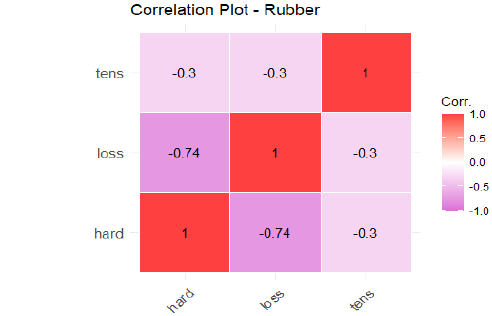


The dataset is loaded using the command data(Rubber). We can view the dataset by using the command View(Rubber). To find the structure of the dataset we use the command str(Rubber) and the first six rows are printed in the console by using the command head(Rubber). Summary(Rubber) provides the Descriptive stats of the Rubber dataset.

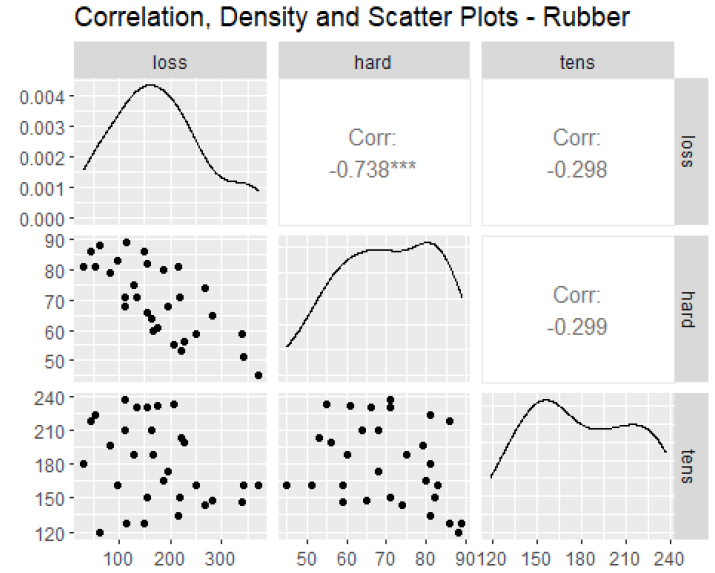




Below plot shows the Correlation between the 3 variables of the Rubber dataset

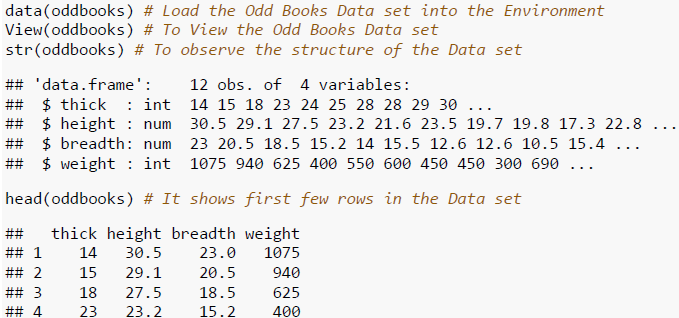


From the below plot, it states that loss and hard variables have a bad correlation.

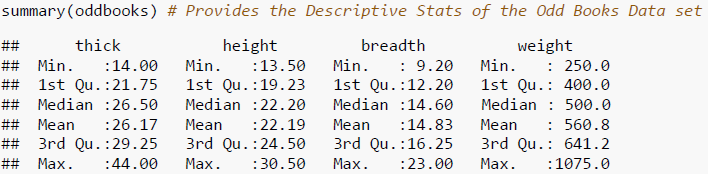


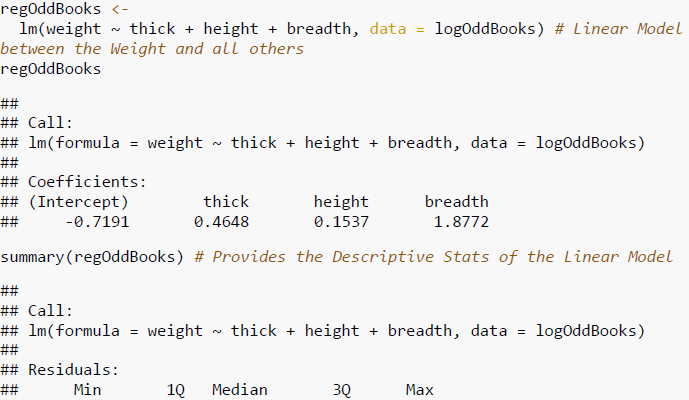
**oddBooks:**

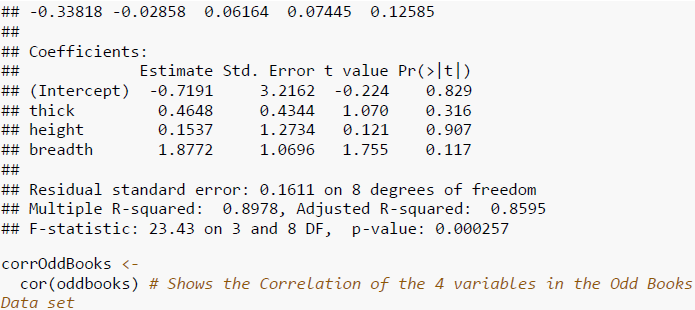
The oddBooks data set consists of 12 observations and 4 variables as thick, height, breadth, and weight.



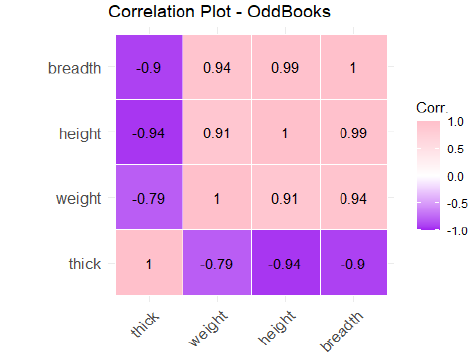
This is loaded by the command data(oddBooks). We can view the dataset by using the command View(oddBooks). To find the structure of the dataset we use the command str(oddBooks) and the first six rows are printed in the console by using the command head(oddBooks). Summary(oddBooks) provides the Descriptive stats of the oddBooks dataset.



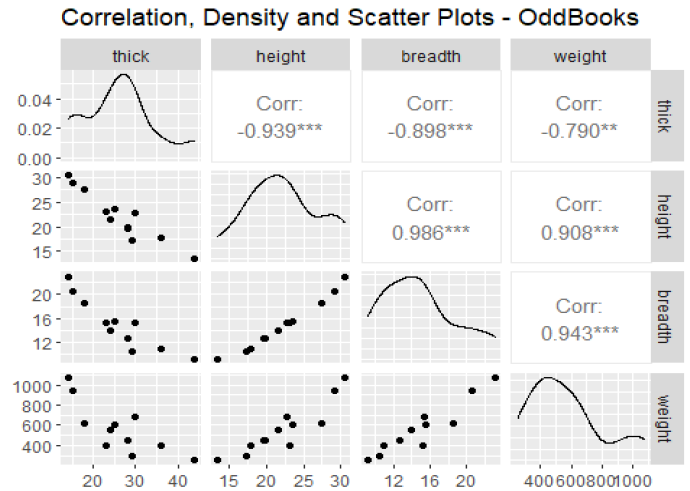




Below plot shows the Correlation between the 4 variables of the oddBooks dataset



From the below plot, it can be found that the correlations between height, thick, and breadth are good so as the coeff’s. are not definite. The data in these variables are same from the plots.



**Conclusion**

To perform multiple analytical operations, we use R on widespread range of data like Trees, Rubber, and OddBooks datasets. R is a very powerful tool to perform analysis which is mainly built by researchers, statisticians, and developers. In the beginning we did Descriptive Statistics Analysis, Exploratory Data Analysis, and continued with the Linear Regression and Inferential Analysis to get understandings from the provided data. Plotted Histograms, Density, Normal Probability, and Box which helps us to understand the data in a clear way. Used ggpairs function from the GGally module for plotting.

**References**

[1] Ggpairs - A Ggplot2 Generalized Pairs Plot is retrieved from <https://www.rdocumentation.org/packages/GGally/versions/1.5.0/topics/ggpairs>

[2] J H Maindonald, Using R for Data Analysis and Graphics was retrieved from <https://cran.r-project.org/doc/contrib/usingR.pdf>

[3] Deborah R. Abrams, Tabachnick & Fidell (1989), Using multivariate statistics. (2nd edition). New York: HarperCollins was retrieved from <https://dss.princeton.edu/online_help/analysis/regression_intro.htm>