DESCRIPTIVE STATISTICS AND REGRESSION ANALYSIS WITH R

Bhrugu Thakkar

College of Professional Studies, Northeastern University

ALY 6015: Intermediate Analytics

Dr. Steward Huang

April 11, 2020

INTRODUCTION

In the first part of assignment a data set (trees) is used to describe the data graphically and numerically. In second part of assignment data sets (rubber and oddbooks) is used to build multiple regression model using R functions to predict influential variables.

Keywords: ggplot, ggcorrplot, data analysis, lm, summary, multiple linear regression, histogram, boxplot, density plot.

ANALYSIS

Part - A

1) **Invoke R and use Tree Dataset**

To find information about tree data set

Code: ?trees

attach(trees)

The above line all the information regarding trees data set in R studio which includes its description, usage, format, source, references and examples. Trees data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Here diameter is used as Girth.

2) Find the 5 summary number in data.

Code:

summary(trees)

sapply(trees, is.factor)

```
Output:   Girth           Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
```

```
1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
```

This is a way to get the summary of Girth (diameter), Height and Volume of the trees dataset. The output of summary provides Minimum value, 1st Quartile, Median, Mean, 3rd Quartile and Maximum Value. For example in the column Girth, min : 8.30 shows that it is the minimum value in that column, 11.05 is 1st quartile of Girth column, median for Girth is 12.9, mean is 13.25 and highest value in Girth is 20.60. And it follows the same for Height and Volume too.

3) Graph a straight-line regression.

(a) Straight line regression for Girth and Volume.

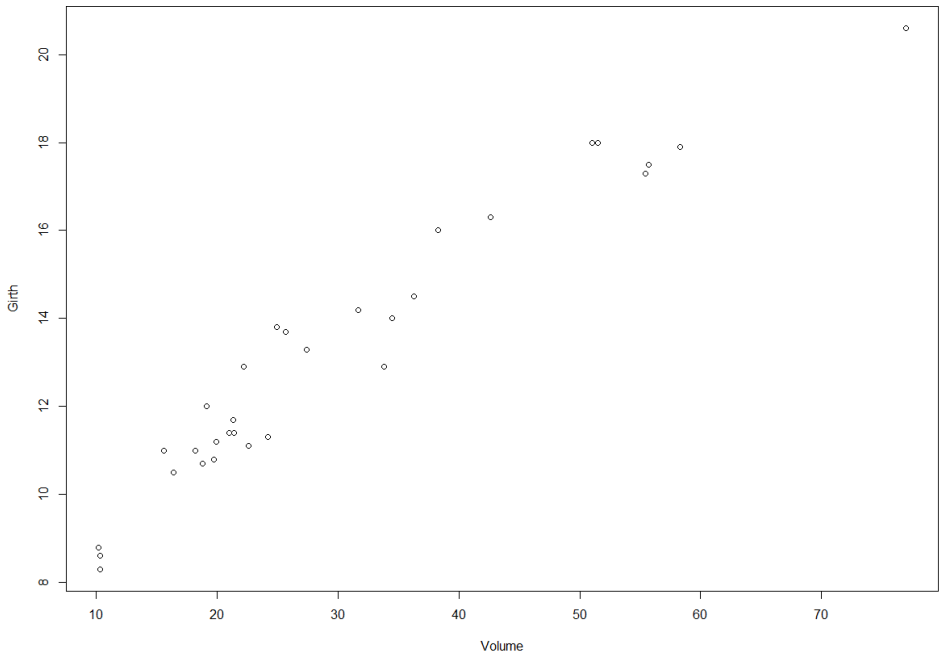
Code: regression <- lm(Girth~Volume,data = trees)

with(trees,plot(Girth~Volume))

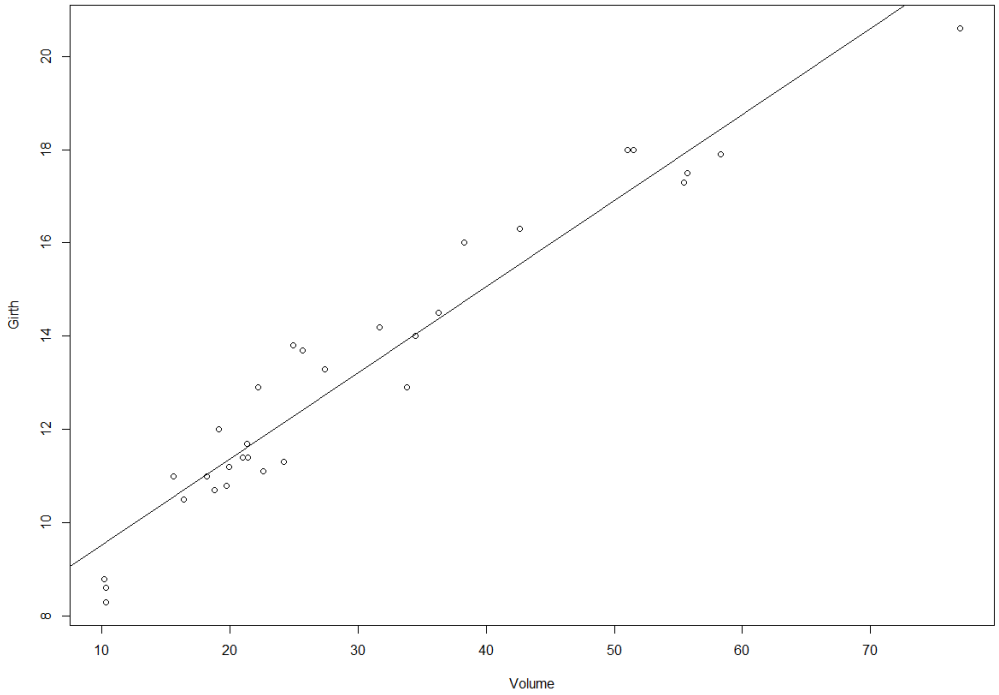
abline(regression)

The first line compares the average value of Girth and volume from data set Trees using lm method.

The second line adds a scatter plot of Girth and Volume.



Third line provides a regression line using the (abline) function.



This shows that there is positive and strong correlation between girth and volume. As the girth increases volume also increases.

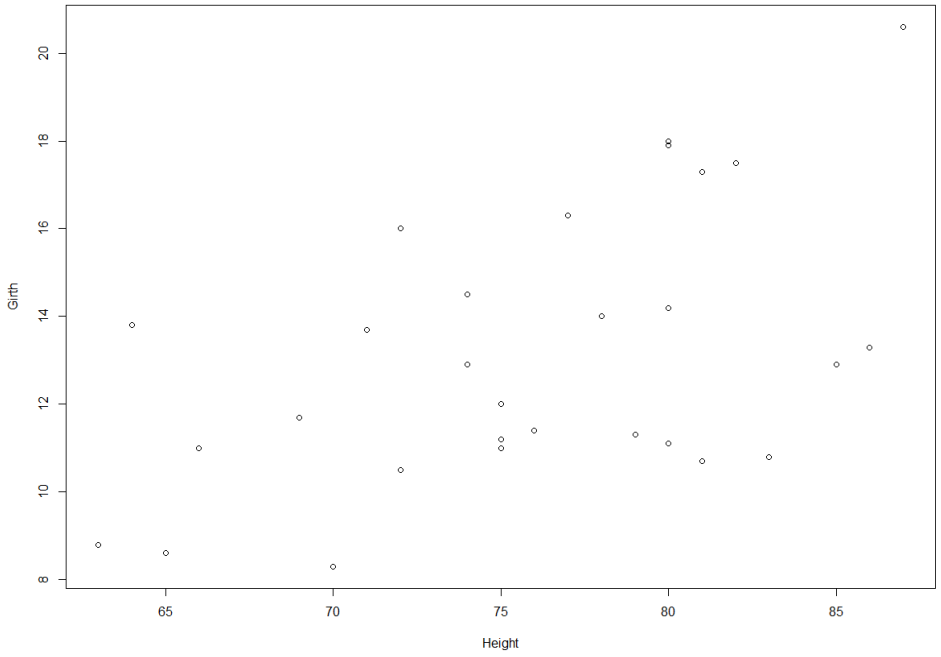
(b) Girth and height

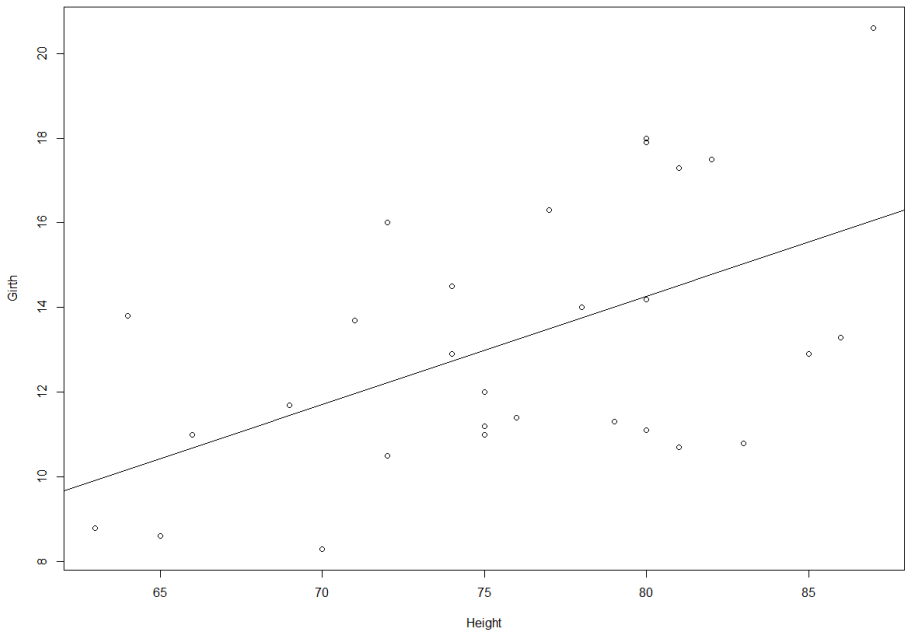
Code: regression1 <- lm(Girth~Height,data = trees)

with(trees,plot(Girth~Height))

abline(regression1)

The regression line will compare mean values of Girth and Height and 2nd and 3rd line will plot them with straight line.


Girth
Height
8
10
12
14
16
18
20
65
70
75
80
85


Girth
Height
8
10
12
14
16
18
20
65
70
75
80
85

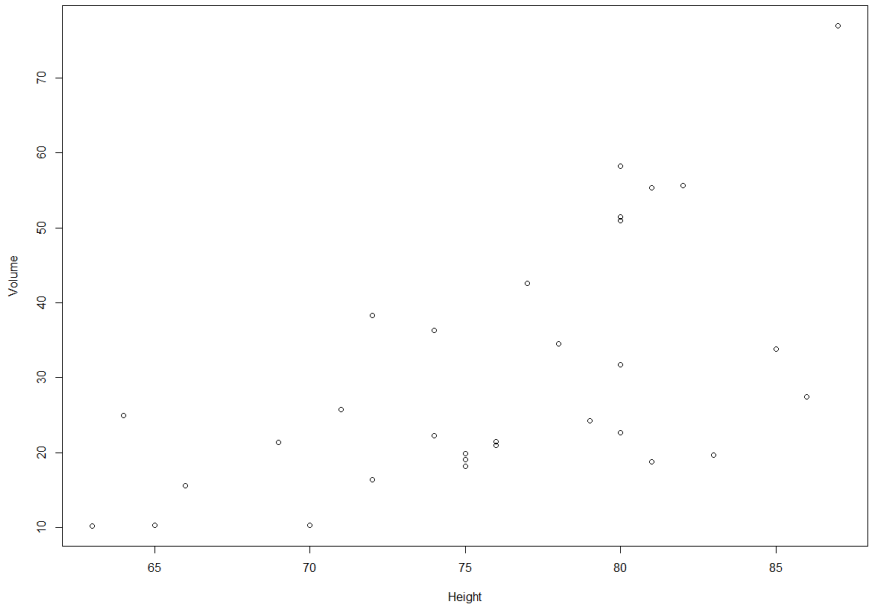
The above plot shows that there is moderately positive relation between girth and height. It can be difficult to interpret the relation between girth and height as they have moderate relationship.

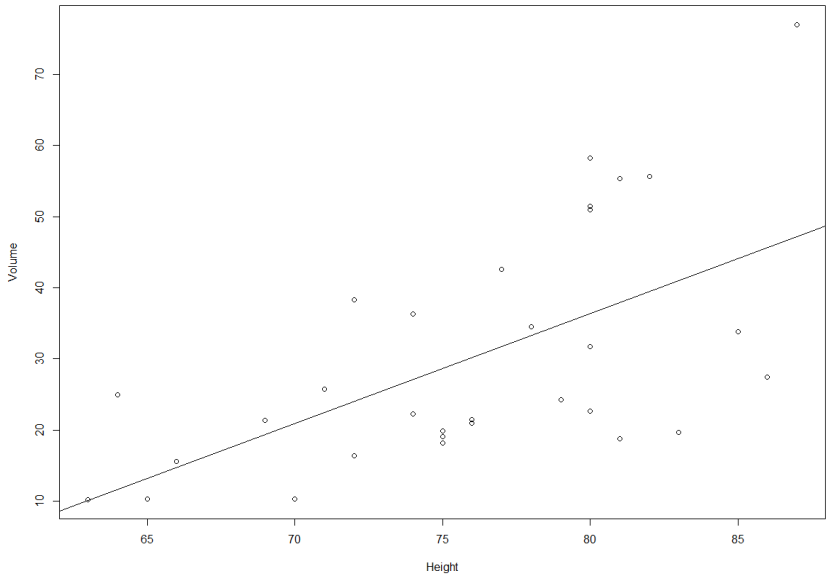
(c)Volume and Height

Code :- regression2 <- lm(Volume~Height,data = trees)

with(trees,plot(Volume~Height))

abline(regression2)





The above plot depicts moderate positive relationship between volume and height.

(4) Create Histograms and Density Plot.

(a) Histogram and density plot of Girth

```
s <- ggplot(data = trees, aes(x = Girth))

s + geom_histogram(binwidth = 2,aes(fill = Girth),colour = "blue")

plot(density(trees$Girth),type = "l")
```

Here, arguments are stored for plotting into a vector “s” and thus adding more parameters along with histogram becomes easy.

➔ ggplot package is used to plot histograms.

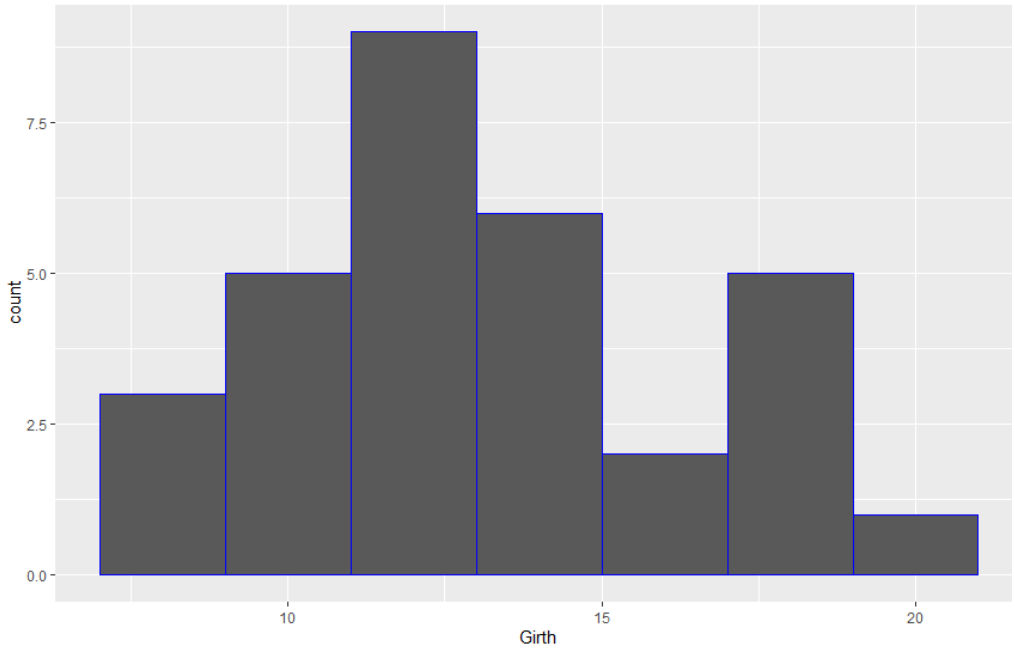
➔ (aes) stands for aesthetics, which states that Girth is plotted on X-axis.

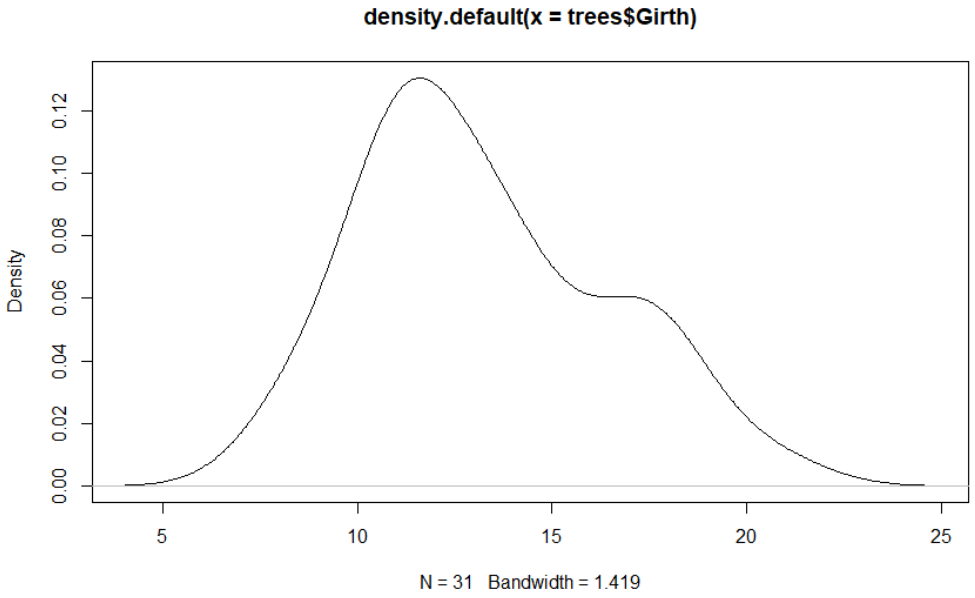
➔ Geom_histogram is the geometry for plotting a histogram

➔ binwidth we can decide how long the bins can be of a histogram.

➔ Bins of histogram will be filled with black color and blue border.

➔ Third line plots the density plot. Trees$Girth takes the girth from trees dataset and plots it. Below are the histogram and density plots of Girth.





This distribution is moderately positive as the tail is inclined towards the right of x-axis. More values are clustering towards the left side of x-axis.

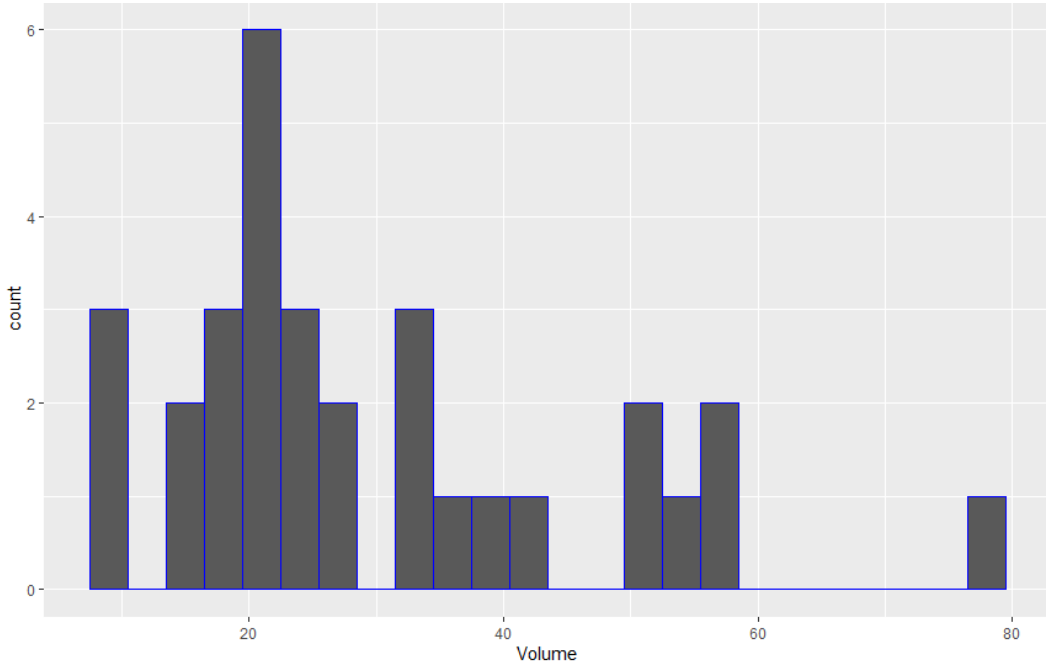
(b) Histogram and density plot of Volume

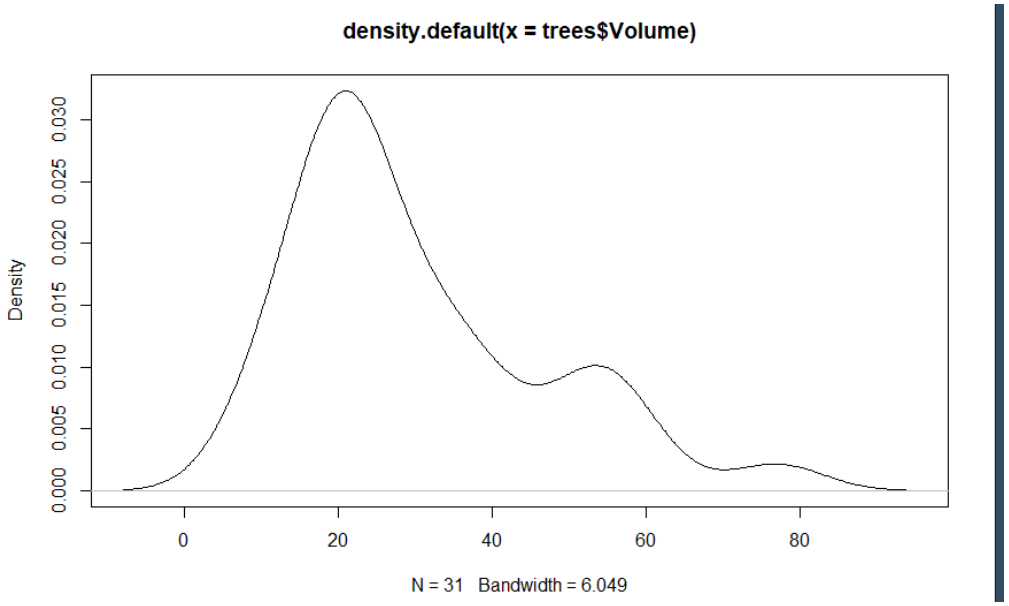
Here the code is same as above, just girth is replaced with volume and binwidth is increased to 3.

Code :- s1 <- ggplot(data = trees, aes(x = Volume))

s1 + geom_histogram(binwidth = 3,aes(fill = Volume),colour = "blue")

plot(density(trees$Volume),type = "l")



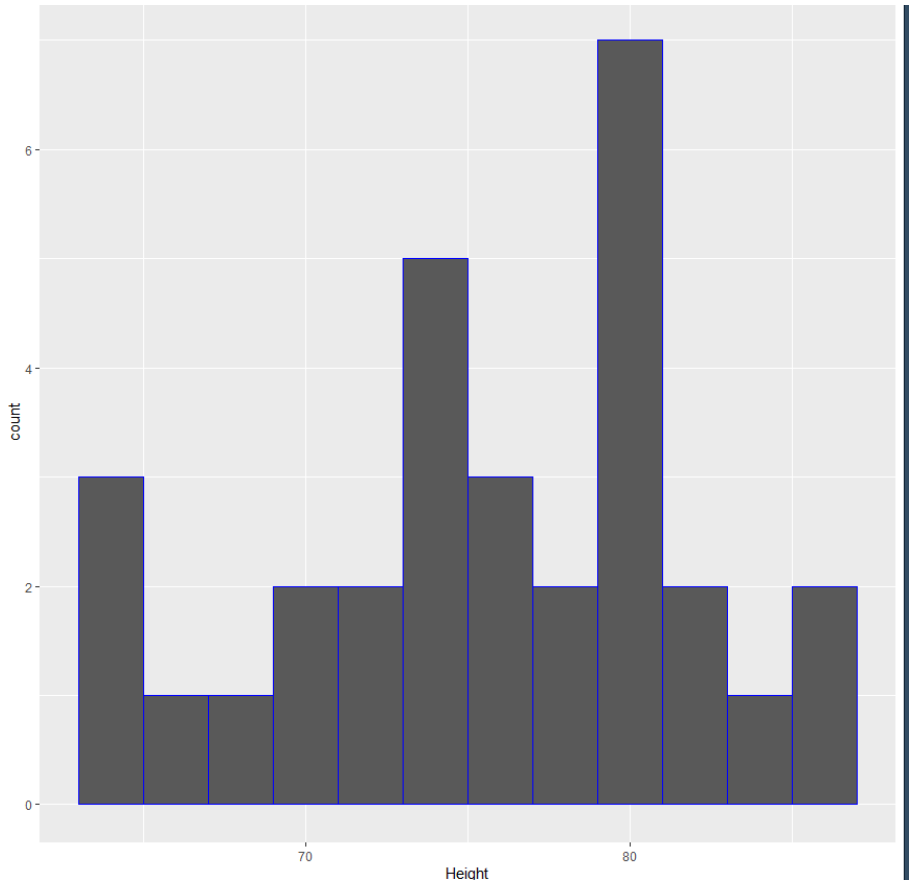


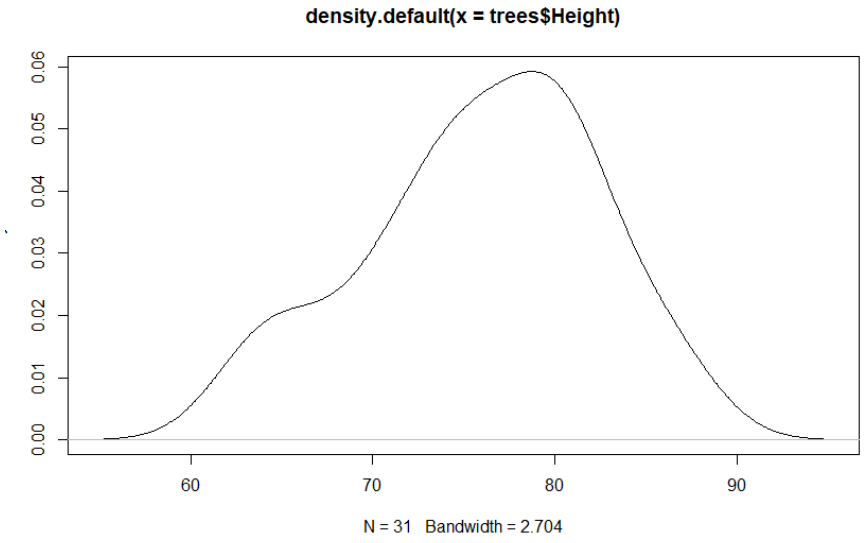
The distribution in this plot is positively skewed as the tail on right is longer than left. Most of the values cluster toward left of the x-axis and less values on the right.

c) Histogram and Density plots of Height

Code :-

```
s2 <- ggplot(data = trees, aes(x = Height))
s2 + geom_histogram(binwidth = 2,aes(fill = Height),colour = "blue")
plot(density(trees$Height),type = "l")
```





The distribution in this plot is moderately negative skewed as the tail is longer on the left side of x-axis. Larger values tend to cluster towards the right side of x-axis.

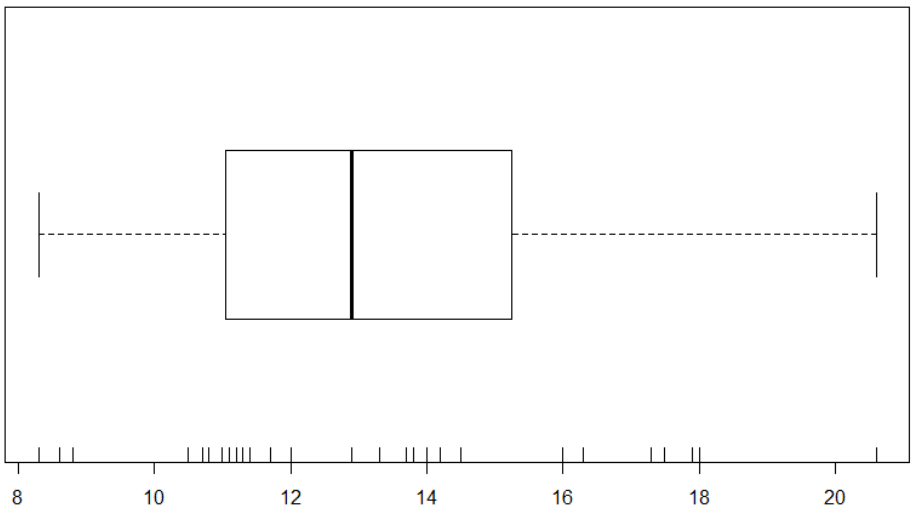
(5) Create Boxplots

(a) Boxplot for Girth

Code:

```
boxplot(trees$Girth,horizontal = T)
rug(trees$Girth,side = 1)
```

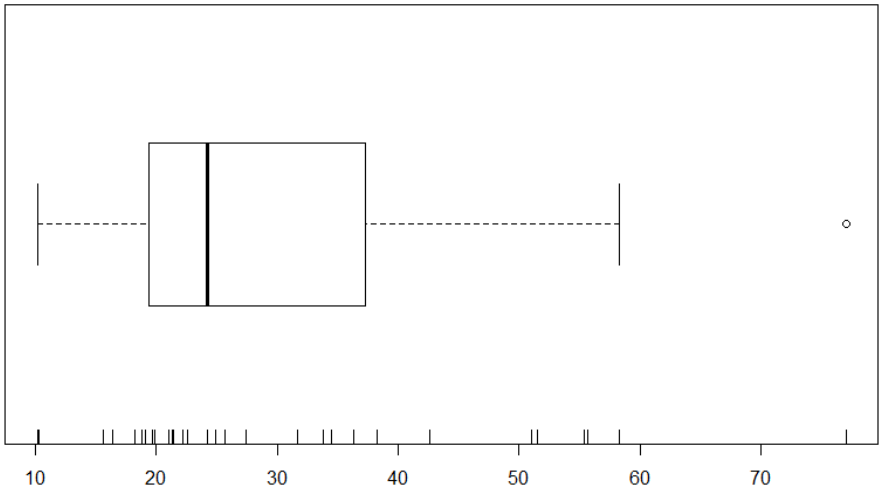
Trees$Girth selects Girth column from trees dataset and then it is boxplotted. Rug displays the marks on axis. Basically, boxplot displays how the data is spread out.



(b) Boxplot of Volume

Code:

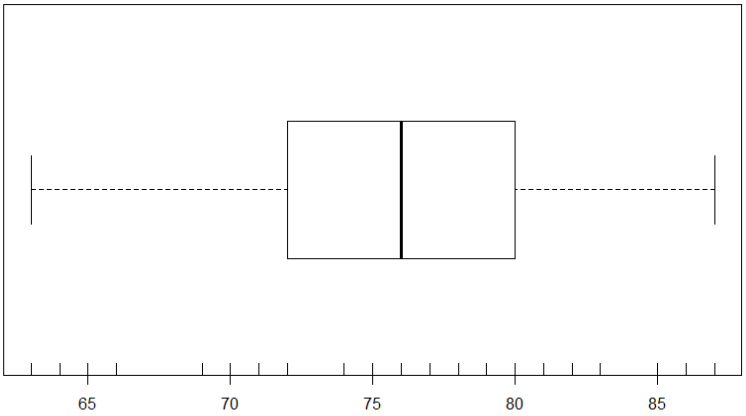
```
boxplot(trees$Volume,horizontal = T)

rug(trees$Volume,side = 1)
```



(d) Boxplot of Height

Code:

```
boxplot(trees$Height,horizontal = T)

rug(trees$Height,side = 1)
```



Values provided by boxplot are:

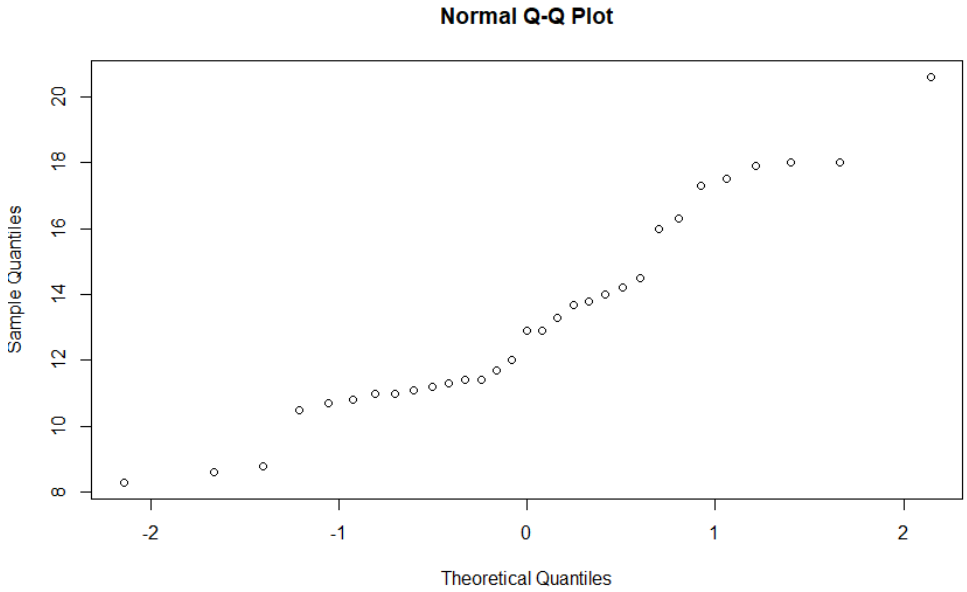
```
     Girth           Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
 Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
```

(6) Normal Probability Plots for Girth,Height and Volume.

Code:

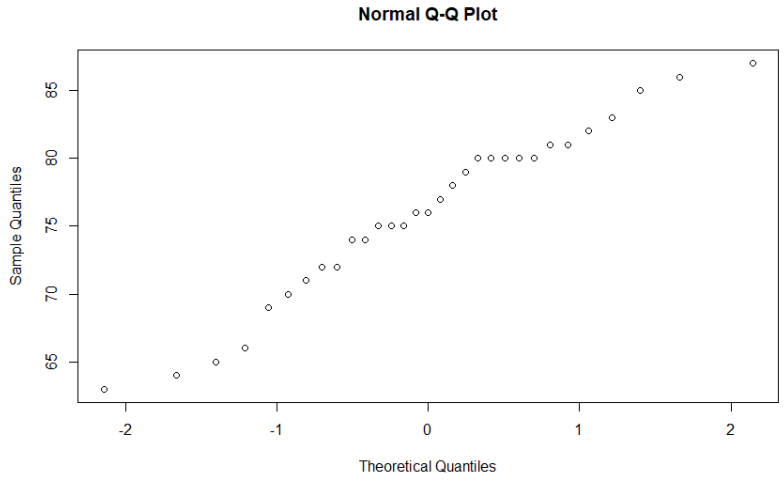
(a) Girth

qqnorm(trees$Girth)



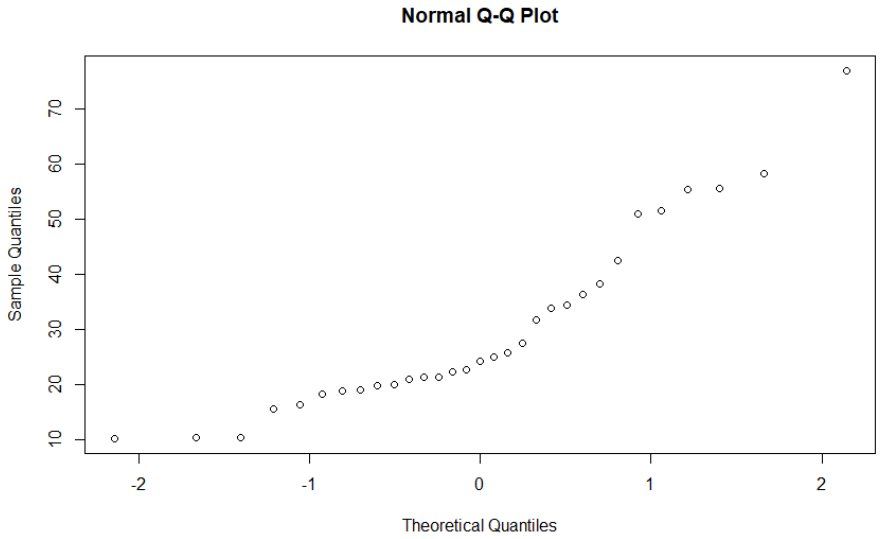
(b) Height

qqnorm(trees$Height)



(c) Volume

qqnorm(trees$Volume)



Part – B

Installing the MASS and DAAG packages for Rubber and oddbooks dataset.

RUBBER DATASET

Code:

install.packages(MASS)

library(MASS)

?Rubber

install.packages(DAAG)

library(DAAG)

?oddbooks

Following are the code to explore the dataset and obtain some information about it. Also the dataset is stored in a vector called mydata.

Code:

mydata <- Rubber

mydata

str(mydata)

ncol(mydata)

nrow(mydata)

Output:

```
> str(mydata)
'data.frame':  30 obs. of  3 variables:
 $ loss: int  372 206 175 154 136 112 55 45 221
166 ...
 $ hard: int  45 55 61 66 71 71 81 86 53 60 ...
 $ tens: int  162 233 232 231 231 237 224 219 20
3 189 ...
> ncol(mydata)
[1] 3
> nrow(mydata)
[1] 30
```

➔ Following code shows the test done to check the correlation between loss, hardness and tens using pearson method.

Code:

cor.test(x = mydata$loss,y = mydata$hard,method = "pearson")

Output:

```
        Pearson's product-moment correlation

data:  mydata$loss and mydata$hard
t = -5.7821, df = 28, p-value = 3.294e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8674373 -0.5140675
sample estimates:
       cor
-0.7377107
```

cor.test(x = mydata$loss, y = mydata$tens, method = " pearson ")

Output:

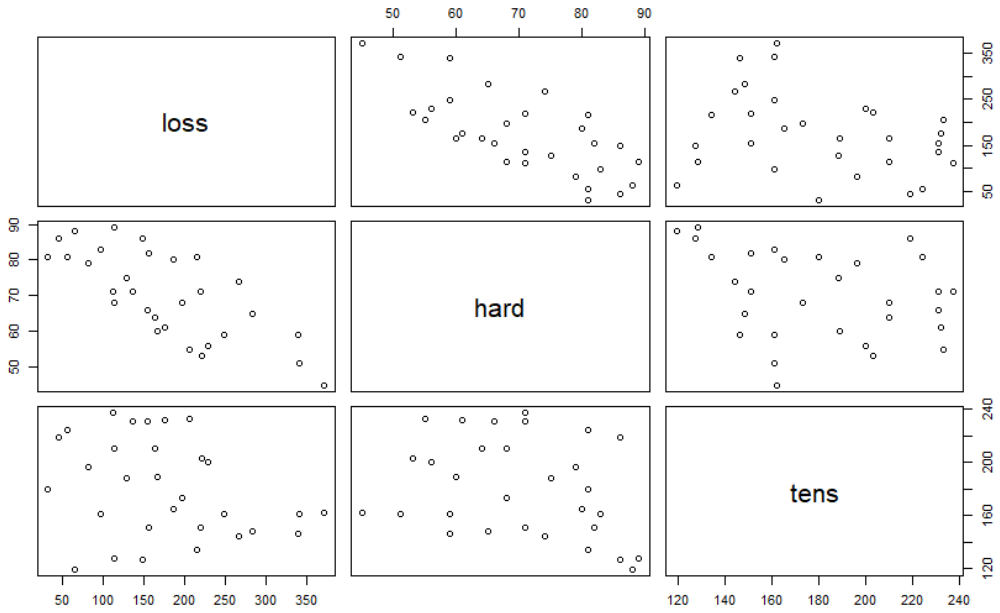
```
        Pearson's product-moment correlation

data:  mydata$loss and mydata$tens
t = -1.6543, df = 28, p-value = 0.1092
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.59472855  0.06932824
sample estimates:
       cor
-0.2983939
```

➔ So this shows that there is negative correlation between loss and hard i.e. (-0.7377107) and also negative correlation between loss and tens i.e. (-0.2983939).

➔ To visualize this data following code is used.

Code: pairs(mydata)

Output:



Code:

regressor <- lm(loss ~ hard + tens, data = mydata)

summary(regressor)

```
Output : Residuals:

    Min      1Q  Median      3Q     Max
-79.385 -14.608   3.816  19.755  65.981

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 885.1611    61.7516  14.334 3.84e-14 ***
hard         -6.5708     0.5832 -11.267 1.03e-11 ***
tens         -1.3743     0.1943  -7.073 1.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.49 on 27 degrees of freedom
Multiple R-squared:  0.8402,      Adjusted R-squared:  0.8284
F-statistic:    71 on 2 and 27 DF,  p-value: 1.767e-11
```

➔ Intercept in the output is 885.16. So we can conclude that when hardness and tensile strength = 0, if the hardness increases by 1 then loss will decrease by 6.57 and if the tensile strength increases by 1 then loss will decrease by 1.37. As the significance level shown in last column the p-value is 0.001 and the confidence level is 99.999

➔ **So there is negative correlation between loss and hardness, tensile strength.**

Code:

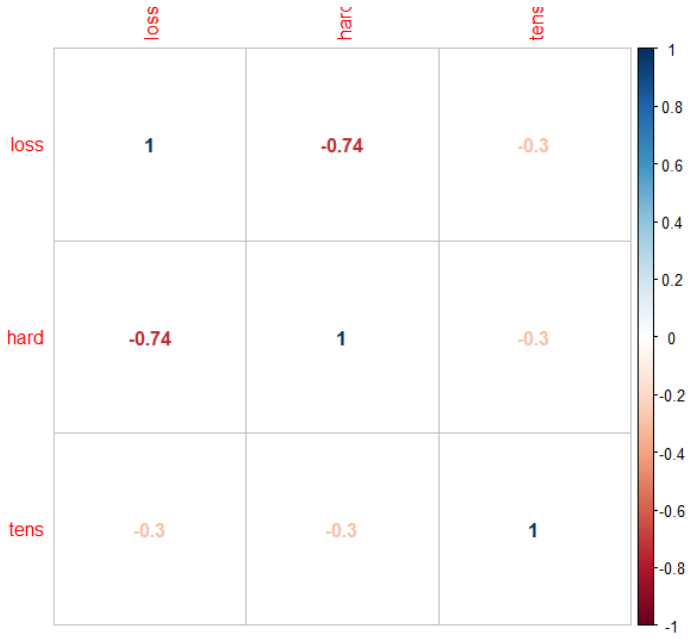
```
r <- cor(mydata, use = "complete.obs")
    r
```

Output:

```
            loss       hard       tens
loss  1.0000000 -0.7377107 -0.2983939
hard -0.7377107  1.0000000 -0.2992345
tens -0.2983939 -0.2992345  1.0000000
```

➔ To visualize it 2 graphs functions can be used which are corrplot and ggcorrplot. The code and output is as follows.

➔ Code: *corrplot(r,method = "number")*

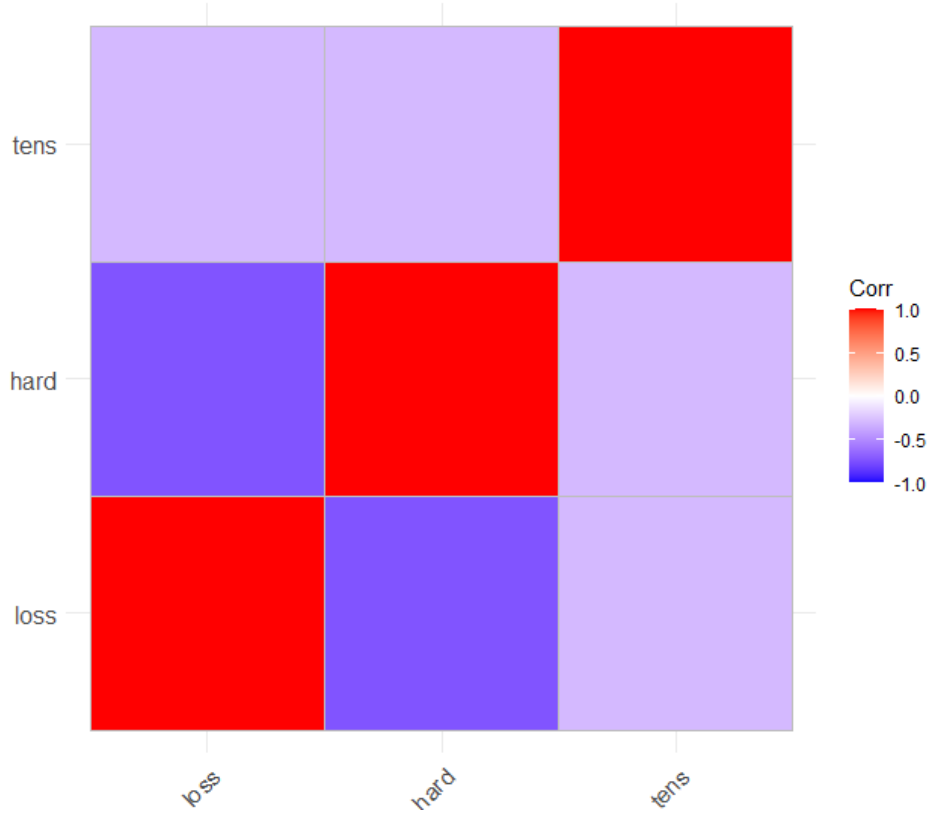


➔ Output:

➔ *The code using ggcorrplot is*

Ggcorrplot(r)

Output:



ODDBOOKS DATASET

➔ Oddbooks data represents the thickness, height, width and weight of 12 books. Following code is used to explore the dataset and understand it before making an analysis.

Code:

?oddbooks

Oddbooks

str(oddbooks)

Output:

```
'data.frame':      12 obs. of  4 variables:
 $ thick  : int  14 15 18 23 24 25 28 28 29 30 ...
 $ height : num  30.5 29.1 27.5 23.2 21.6 23.5 19.7 19.8 17.3 22
.8 ...
 $ breadth: num  23 20.5 18.5 15.2 14 15.5 12.6 12.6 10.5 15.4 .
..
 $ weight : int  1075 940 625 400 550 600 450 450 300 690 ...
```

➔ Here we can observe that as the thickness increases, height, breadth and weight decreases.

MULTIPLE REGRESSION MODELS

Code: mydata1 <- log(oddbooks)

Here we have used the log function to get the natural logarithm of the value.

Output:

```
     thick   height  breadth   weight
1  2.639057 3.417727 3.135494 6.980076
2  2.708050 3.370738 3.020425 6.845880
3  2.890372 3.314186 2.917771 6.437752
4  3.135494 3.144152 2.721295 5.991465
5  3.178054 3.072693 2.639057 6.309918
6  3.218876 3.157000 2.740840 6.396930
7  3.332205 2.980619 2.533697 6.109248
8  3.332205 2.985682 2.533697 6.109248
9  3.367296 2.850707 2.351375 5.703782
10 3.401197 3.126761 2.734368 6.536692
11 3.583519 2.879198 2.397895 5.991465
12 3.784190 2.602690 2.219203 5.521461
```

WEIGHT ~ THICK

Code:

model1 <- lm(weight~thick,data=mydata1)

model1

summary(model1)

summary(model1$coefficients)

Output:

```
Call:
lm(formula = weight ~ thick, data = mydata1)

Coefficients:
(Intercept)        thick
      9.692       -1.073



Call:
lm(formula = weight ~ thick, data = mydata1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37651 -0.12228  0.00898  0.12479  0.49276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6920     0.7076  13.697 8.35e-08 ***
thick        -1.0726     0.2190  -4.897 0.000626 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2446 on 10 degrees of freedom
Multiple R-squared:  0.7057,      Adjusted R-squared:  0.6762
F-statistic: 23.98 on 1 and 10 DF,  p-value: 0.0006263



  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.073   1.619   4.310   4.310   7.001   9.692
```

➔ When thickness is 0 the weight is 9.69 and as the weight increases by 1, thickness will decrease by 1.07.

➔ **<u>Thus, weight and thickness has *negative correlation*.</u>**

WEIGHT ~ THICK + HEIGHT

Code:

model2 <- lm(weight ~ thick + height,data = mydata1)

model2

summary(model2)

summary(model2$coefficients)

Output:

```
Call:
lm(formula = weight ~ thick + height, data = mydata1)

Coefficients:
(Intercept)       thick      height
    -1.2632       0.3129       2.1143

>

Call:
lm(formula = weight ~ thick + height, data = mydata1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37423 -0.04502 0.03677 0.10435  0.19131

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2632     3.5520  -0.356   0.7303
thick         0.3129     0.4724   0.662   0.5243
height        2.1143     0.6782   3.117   0.0124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1788 on 9 degrees of freedom
Multiple R-squared:  0.8585,      Adjusted R-squared:  0.827
F-statistic:  27.3 on 2 and 9 DF,  p-value: 0.0001509

>

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.2632 -0.4751  0.3129  0.3880  1.2136  2.1143
```

➔ When thickness and height is 1.26, if weight increases by 1 thickness will also increase by 0.3 and height will increase by 2.11.

➔ **So the correlation between weight and thickness + height is *negative.***

WEIGHT~THICK+HEIGHT+BREADTH

Code:

model3 <- lm(weight ~ thick + height + breadth, data = mydata1)

model3

summary(model3)

summary(model3$coefficients)

Output:

```
Call:
lm(formula = weight ~ thick + height + breadth, data = mydata1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33818 -0.02858  0.06164  0.07445  0.12585

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7191     3.2162  -0.224    0.829
thick         0.4648     0.4344   1.070    0.316
height        0.1537     1.2734   0.121    0.907
breadth       1.8772     1.0696   1.755    0.117

Residual standard error: 0.1611 on 8 degrees of freedom
Multiple R-squared:  0.8978,      Adjusted R-squared:  0.8595
F-statistic: 23.43 on 3 and 8 DF,  p-value: 0.000257

>
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.71912 -0.06453  0.30921  0.44412  0.81786  1.87719
```

➔ When thickness and height is 0, weight is -0.71. When weight is increases by 1 the thickness increases by 0.46, height by 0.15 and breadth by 1.87.

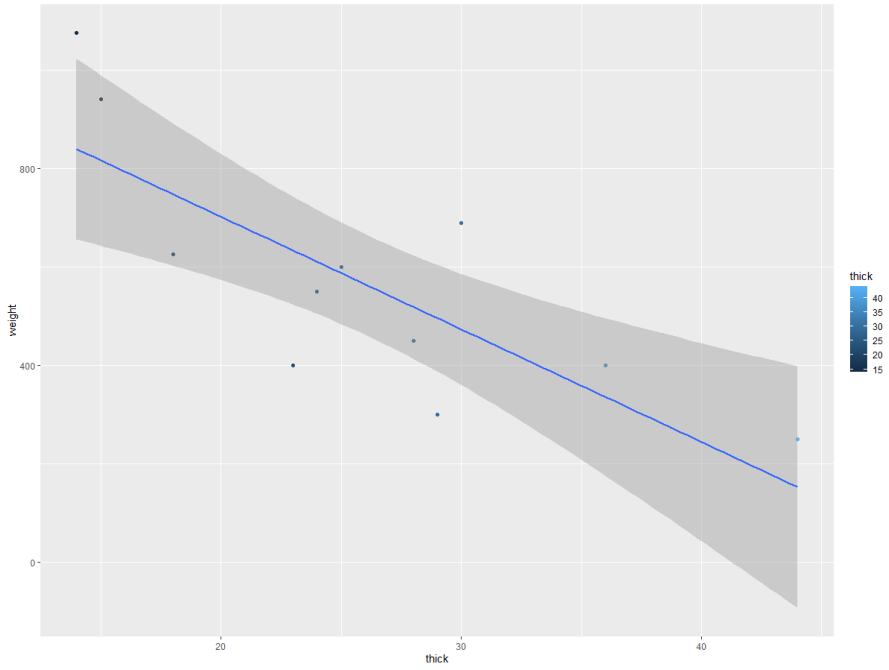
➔ **So there is negative correlation between weight and thickness + height + breadth.**

VISUALIZATION:

WEIGHT~THICK

Code:

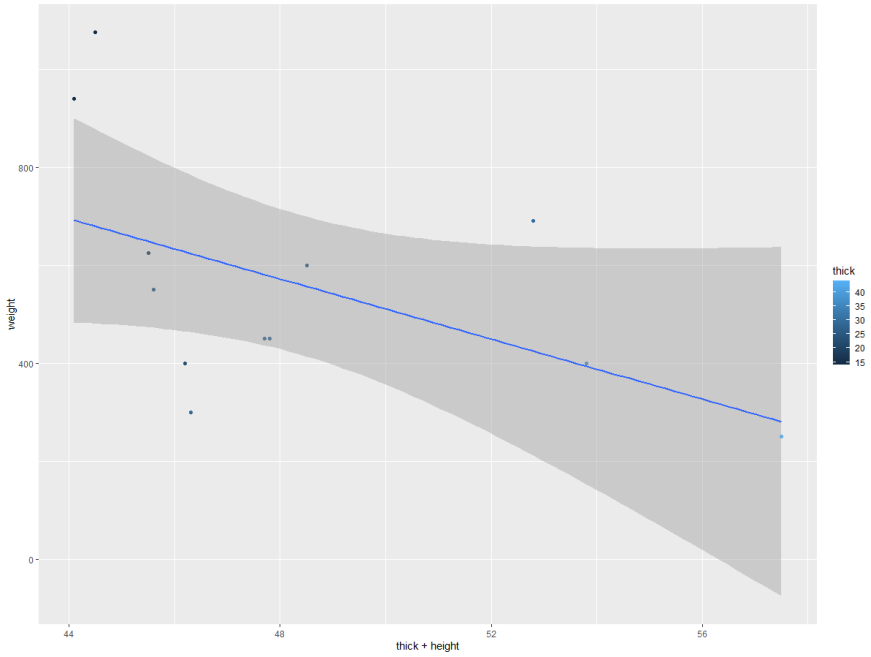
ggplot(oddbooks,aes(thick,weight)) + geom_point(aes(color = thick)) +geom_smooth(method = "lm")



WEIGHT~THICK+HEIGHT

Code:

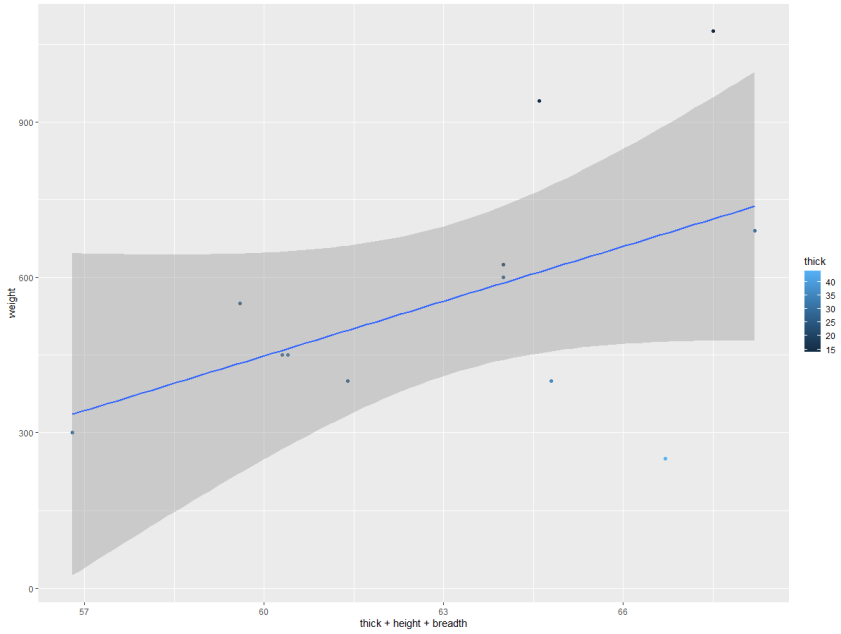
ggplot(oddbooks,aes(thick+height,weight)) + geom_point(aes(color = thick)) + geom_smooth(method = "lm")



WEIGHT~THICK+HEIGHT+BREADTH

Code:

ggplot(oddbooks,aes(thick + height + breadth, weight)) + geom_point(aes(color = thick))
+geom_smooth(method = "lm")



To plot these models 2 libraries are loaded

Code: library(corrplot)

library(visreg)

Conclusion:

➔ In total 3 data sets were used which are Trees, Rubber, and Odd books.

➔ Descriptive statistics and simple linear and multiple regression analysis were done using R on all three datasets to gain insights and understand it.

➔ Following operations were made on these data sets:

➔ (a) Exploring and Viewing data in R.2.

➔ (b) Finding the summary of the dataset which displays information such as Min value, Max value, 1stQuartile, Median, Mean and 3rdQuartile.

➔ (c) Visualizing and plotting Histograms, Density plots, Box plots and Normal probability plots to properly understand the dataset.

➔ (d) Building and visualizing multiple regression model by calculating the correlation matrix and plotting scatter plots to explore and understand the dataset.

References

Kabacoff, R. (n.d.). Multiple (Linear) Regression. Retrieved April 12, 2020, from https://www.statmethods.net/stats/regression.html

Interpreting the Intercept in a Regression Model. (2020, January 16). Retrieved April 12, 2020, from https://www.theanalysisfactor.com/interpreting-the-intercept-in-a-regression-model/.

Footnotes

[1][Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

Tables

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

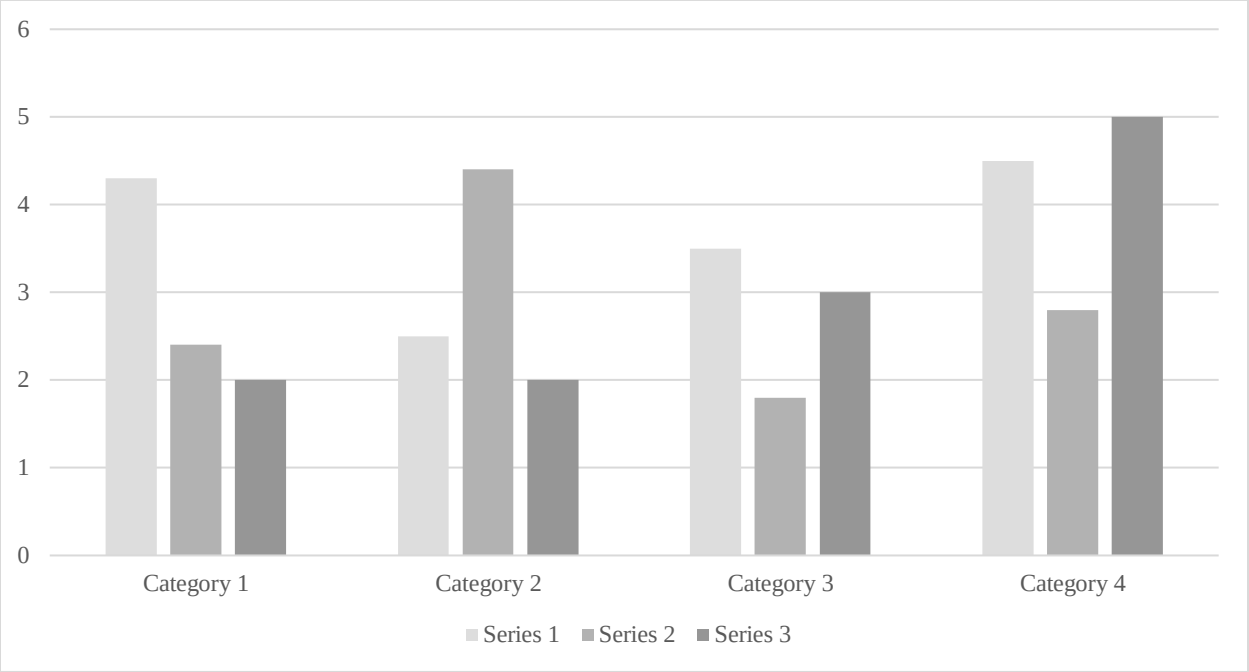


Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the *APA Style Manual, 6th Edition*.