ALY 6015 M1 Report - Thota, Sunil Raj.R

# Intermediate Analytics  
# ALY 6015  
# Module 1 - Descriptive Statistics and Regression Analysis with R  
# 01/21/2021  
# Sunil Raj Thota  
# NUID: 001099670

# Get and set the working directories  
getwd()

## [1] "G:/NEU/Coursework/2021 Q1 Winter/ALY 6015 IA/Discussions & Assignments"

setwd('G:/NEU/Coursework/2021 Q1 Winter/ALY 6015 IA/Discussions & Assignments')  
getwd()

## [1] "G:/NEU/Coursework/2021 Q1 Winter/ALY 6015 IA/Discussions & Assignments"

# Installed the above packages into the workspace  
install.packages("datasets")  
install.packages("plyr")  
install.packages("dplyr")  
install.packages("tidyr")  
install.packages("tidyverse")  
install.packages("ggplot2")  
install.packages("ggcorrplot")  
install.packages("e1071")  
install.packages("DAAG")  
install.packages("MASS")  
install.packages("GGally")  
  
# Loaded the below libraries into the workspace  
library(plyr)

library(dplyr)

library(tidyr)

library(tidyverse)

library(ggplot2)

library(e1071)

library(MASS)

library(DAAG)

library(ggcorrplot)

library(GGally)

require(grDevices)

require(datasets)

# Part A

data(trees) # Load the Trees Data set into the Environment  
View(trees) # To View the Trees Data set  
str(trees) # To observe the structure of the Data set

## 'data.frame': 31 obs. of 3 variables:  
## $ Girth : num 8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...  
## $ Height: num 70 65 63 72 81 83 66 75 80 75 ...  
## $ Volume: num 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...

head(trees) # It shows first few rows in the Data set

## Girth Height Volume  
## 1 8.3 70 10.3  
## 2 8.6 65 10.3  
## 3 8.8 63 10.2  
## 4 10.5 72 16.4  
## 5 10.7 81 18.8  
## 6 10.8 83 19.7

summary(trees) # Provides the Descriptive Stats of the Trees Data set

## 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

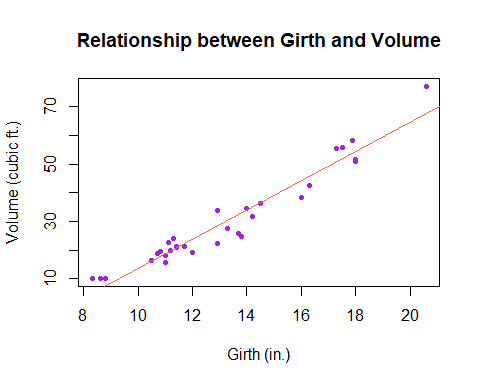
cor(trees) # Shows the Correlation of the 3 variables in the Trees Data set

## Girth Height Volume  
## Girth 1.0000000 0.5192801 0.9671194  
## Height 0.5192801 1.0000000 0.5982497  
## Volume 0.9671194 0.5982497 1.0000000

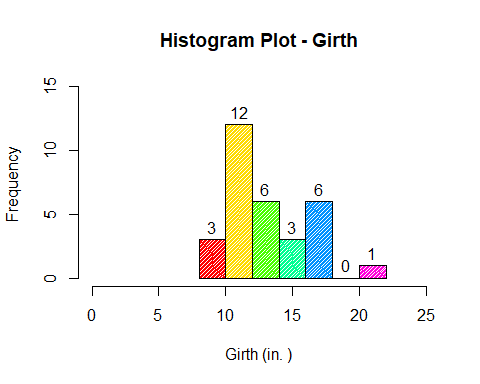
plot(  
 x = trees$Girth,  
 y = trees$Volume,  
 xlab = "Girth (in.)",  
 ylab = "Volume (cubic ft.)",  
 main = "Relationship between Girth and Volume",  
 col = "purple",  
 pch = 20,  
 xlim = c(min(trees$Girth), max(trees$Girth)),  
 ylim = c(min(trees$Volume), max(trees$Volume))  
) # Scatter Plot is used to depict the relationship between the Girth and Volume  
lm(Volume ~ Girth, data = trees) # Linear Model between the Volume and Girth

##   
## Call:  
## lm(formula = Volume ~ Girth, data = trees)  
##   
## Coefficients:  
## (Intercept) Girth   
## -36.943 5.066

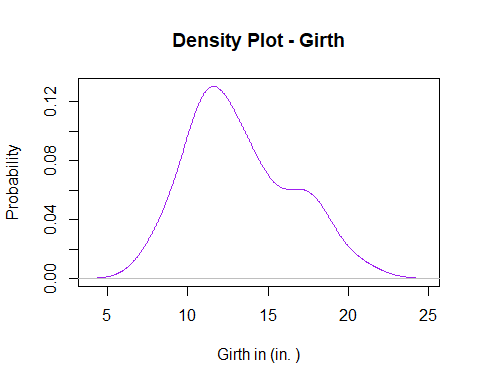
abline(lm(Volume ~ Girth, data = trees), col = "tomato") # To observe the Regression Line



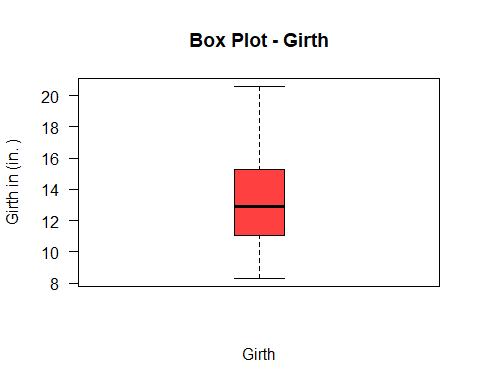
hist(  
 trees$Girth,  
 main = "Histogram Plot - Girth",  
 xlab = "Girth (in. )",  
 ylab = "Frequency ",  
 border = "black",  
 labels = TRUE,  
 xlim = c(0, 25),  
 ylim = c(0, 15),  
 col = rainbow(7),  
 density = 100  
) # Histogram Plot is used to show case the Frequency Distribution of the Girth



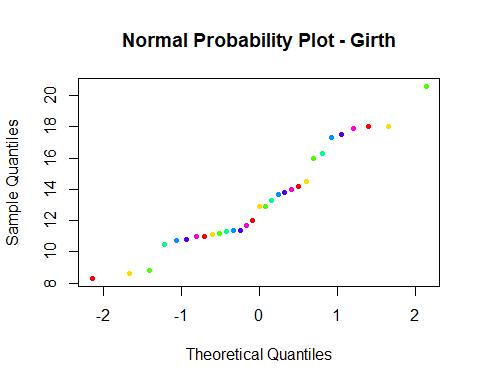
plot(  
 density(trees$Girth),  
 main = "Density Plot - Girth",  
 xlab = "Girth in (in. )",  
 ylab = "Probability",  
 col = "purple"  
) # Density Plot is used to show case the Probability Distribution of the Girth



boxplot(  
 trees$Girth,  
 main = "Box Plot - Girth",  
 ylab = "Girth in (in. )",  
 xlab = "Girth",  
 col = "brown1",  
 boxwex = 0.3,  
 outline = TRUE,  
 outpch = 16,  
 outcol = "seagreen3",  
 las = 1,  
 notch = FALSE,  
 staplewex = 1  
) # Box Plot is used to determine the Quartiles of the Girth



qqnorm(trees$Girth,  
 main = "Normal Probability Plot - Girth",  
 col = rainbow(7),  
 pch = 20) # Normal Probability Plot of the Girth



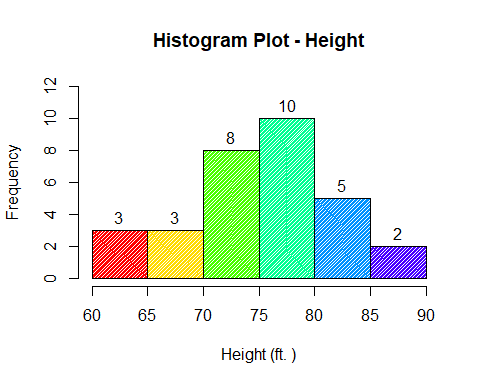
skewness(trees$Girth) # Skewness measures the relative size of the Girth

## [1] 0.5010559

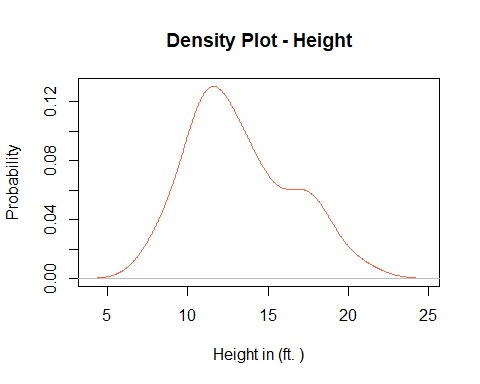
kurtosis(trees$Girth) # Kurtosis measures the amount of Prob. in the Girth

## [1] -0.7109412

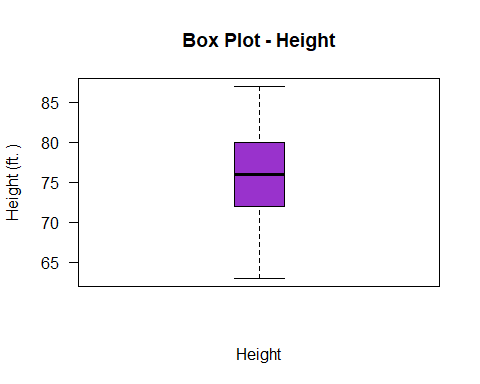
hist(  
 trees$Height,  
 main = "Histogram Plot - Height",  
 xlab = "Height (ft. )",  
 ylab = "Frequency",  
 border = "black",  
 labels = TRUE,  
 xlim = c(60, 90),  
 ylim = c(0, 12),  
 col = rainbow(7),  
 density = 100,  
) # Histogram Plot is used to show case the Frequency Distribution of the Height



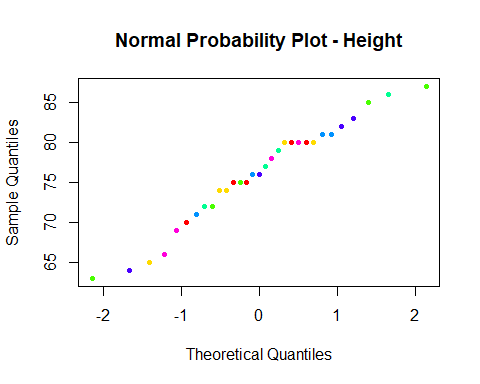
plot(  
 density(trees$Girth),  
 main = "Density Plot - Height",  
 xlab = "Height in (ft. )",  
 ylab = "Probability",  
 col = "tomato"  
) # Density Plot is used to show case the Probability Distribution of the Height



boxplot(  
 trees$Height,  
 main = "Box Plot - Height",  
 ylab = "Height (ft. )",  
 xlab = "Height",  
 col = "darkorchid",  
 boxwex = 0.3,  
 outline = TRUE,  
 outpch = 16,  
 outcol = "seagreen3",  
 las = 1,  
 notch = FALSE,  
 staplewex = 1  
) # Box Plot is used to determine the Quartiles of the Height



qqnorm(trees$Height,  
 main = "Normal Probability Plot - Height",  
 col = rainbow(7),  
 pch = 20) # Normal Probability Plot of the Height



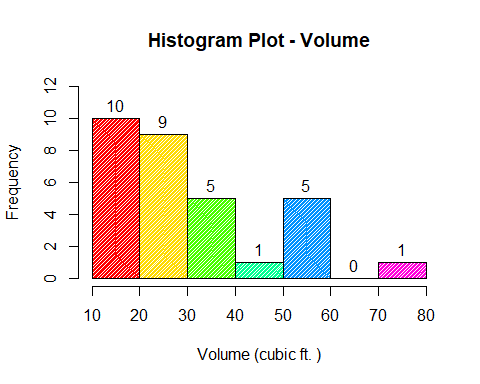
skewness(trees$Height) # Skewness measures the relative size of the Height

## [1] -0.3568773

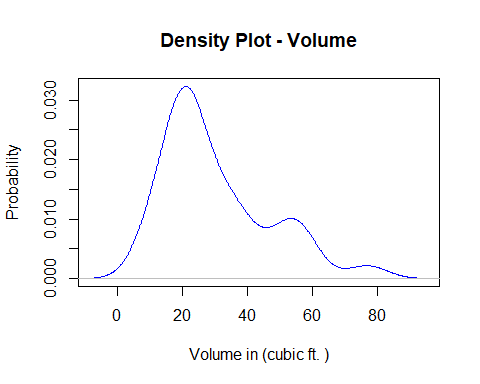
kurtosis(trees$Height) # Kurtosis measures the amount of Prob. in the Height

## [1] -0.7233677

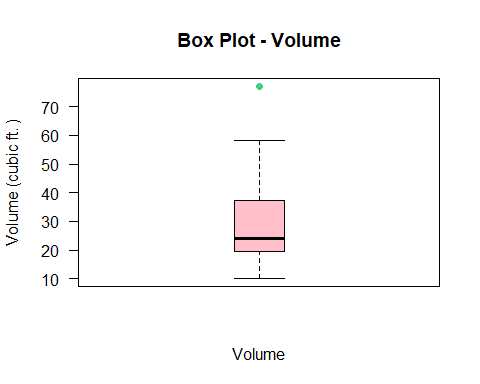
hist(  
 trees$Volume,  
 main = "Histogram Plot - Volume",  
 xlab = "Volume (cubic ft. )",  
 ylab = "Frequency",  
 border = "black",  
 labels = TRUE,  
 xlim = c(10, 80),  
 ylim = c(0, 12),  
 col = rainbow(7),  
 density = 100,  
) # Histogram Plot is used to show case the Frequency Distribution of the Volume



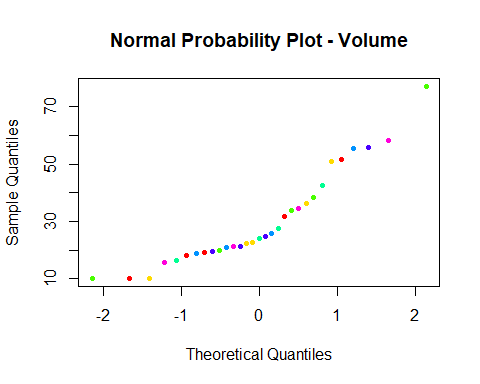
plot(  
 density(trees$Volume),  
 main = "Density Plot - Volume",  
 xlab = "Volume in (cubic ft. )",  
 ylab = "Probability",  
 col = "blue"  
) # Density Plot is used to show case the Probability Distribution of the Volume



boxplot(  
 trees$Volume,  
 main = "Box Plot - Volume",  
 ylab = "Volume (cubic ft. )",  
 xlab = "Volume",  
 col = "pink",  
 boxwex = 0.3,  
 outline = TRUE,  
 outpch = 16,  
 outcol = "seagreen3",  
 las = 1,  
 notch = FALSE,  
 staplewex = 1  
) # Box Plot is used to determine the Quartiles of the Volume



qqnorm(trees$Volume,  
 main = "Normal Probability Plot - Volume",  
 col = rainbow(7),  
 pch = 20) # Normal Probability Plot of the Volume



skewness(trees$Volume) # Skewness measures the relative size of the Volume

## [1] 1.013274

kurtosis(trees$Volume) # Kurtosis measures the amount of Prob. in the Volume

## [1] 0.2460393

# Part B

data(Rubber) # Load the Rubber Data set into the Environment  
View(Rubber) # To View the Rubber Data set  
str(Rubber) # To observe the structure of the Data set

## '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 162 233 232 231 231 237 224 219 203 189 ...

head(Rubber) # It shows first few rows in the Data set

## loss hard tens  
## 1 372 45 162  
## 2 206 55 233  
## 3 175 61 232  
## 4 154 66 231  
## 5 136 71 231  
## 6 112 71 237

summary(Rubber) # Provides the Descriptive Stats of the Rubber Data set

## 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

log(Rubber) # Log computes logarithms of the Rubber Data set

## loss hard tens  
## 1 5.918894 3.806662 5.087596  
## 2 5.327876 4.007333 5.451038  
## 3 5.164786 4.110874 5.446737  
## 4 5.036953 4.189655 5.442418  
## 5 4.912655 4.262680 5.442418  
## 6 4.718499 4.262680 5.468060  
## 7 4.007333 4.394449 5.411646  
## 8 3.806662 4.454347 5.389072  
## 9 5.398163 3.970292 5.313206  
## 10 5.111988 4.094345 5.241747  
## 11 5.099866 4.158883 5.347108  
## 12 4.727388 4.219508 5.347108  
## 13 4.406719 4.369448 5.278115  
## 14 3.465736 4.394449 5.192957  
## 15 5.429346 4.025352 5.298317  
## 16 5.278115 4.219508 5.153292  
## 17 4.852030 4.317488 5.236442  
## 18 4.574711 4.418841 5.081404  
## 19 4.158883 4.477337 4.779123  
## 20 5.517453 4.077537 5.081404  
## 21 5.389072 4.262680 5.017280  
## 22 5.225747 4.382027 5.105945  
## 23 5.043425 4.406719 5.017280  
## 24 4.736198 4.488636 4.852030  
## 25 5.831882 3.931826 5.081404  
## 26 5.828946 4.077537 4.983607  
## 27 5.645447 4.174387 4.997212  
## 28 5.587249 4.304065 4.969813  
## 29 5.370638 4.394449 4.897840  
## 30 4.997212 4.454347 4.844187

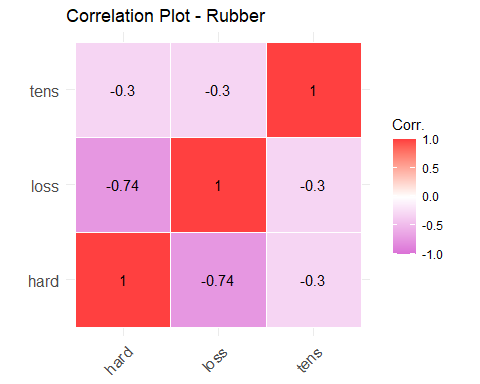
regRubber <-  
 lm(loss ~ hard + tens, data = Rubber) # Linear Model between the Loss and all others  
regRubber

##   
## Call:  
## lm(formula = loss ~ hard + tens, data = Rubber)  
##   
## Coefficients:  
## (Intercept) hard tens   
## 885.161 -6.571 -1.374

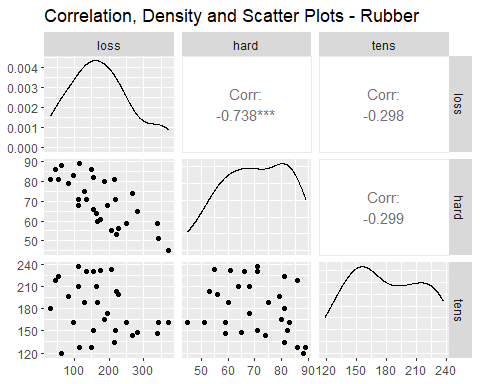
summary(regRubber) # Provides the Descriptive Stats of the Linear Model

##   
## Call:  
## lm(formula = loss ~ hard + tens, data = Rubber)  
##   
## 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

corrRubber <-  
 cor(Rubber) # Shows the Correlation of the 3 variables in the Rubber Data set  
  
ggcorrplot(  
 corrRubber,  
 ggtheme = ggplot2::theme\_minimal,  
 title = "Correlation Plot - Rubber",  
 hc.order = TRUE,  
 colors = c("orchid", "white", "brown1"),  
 outline.col = "white",  
 lab = TRUE,  
 method = "square",  
 show.legend = TRUE,  
 legend.title = "Corr.",  
 lab\_col = "black",  
 lab\_size = 4  
) # Shows the Correlation Plot of the 3 variables in the Rubber Data set



ggpairs(  
 Rubber,  
 mapping = NULL,  
 columns = 1:ncol(Rubber),  
 title = "Correlation, Density and Scatter Plots - Rubber",  
 upper = list(continuous = "cor"),  
 lower = list(continuous = "points"),  
 diag = list(continuous = "densityDiag"),  
 axisLabels = c("show", "internal", "none"),  
) # Shows the Correlation, Density, and Scatter Plots of the 3 variables in the Rubber Data set



data(oddbooks) # Load the Odd Books Data set into the Environment  
View(oddbooks) # To View the Odd Books Data set  
str(oddbooks) # To observe the structure of the Data set

## '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 ...

head(oddbooks) # It shows first few rows in the Data set

## thick height breadth weight  
## 1 14 30.5 23.0 1075  
## 2 15 29.1 20.5 940  
## 3 18 27.5 18.5 625  
## 4 23 23.2 15.2 400  
## 5 24 21.6 14.0 550  
## 6 25 23.5 15.5 600

summary(oddbooks) # Provides the Descriptive Stats of the Odd Books Data set

## thick height breadth weight   
## Min. :14.00 Min. :13.50 Min. : 9.20 Min. : 250.0   
## 1st Qu.:21.75 1st Qu.:19.23 1st Qu.:12.20 1st Qu.: 400.0   
## Median :26.50 Median :22.20 Median :14.60 Median : 500.0   
## Mean :26.17 Mean :22.19 Mean :14.83 Mean : 560.8   
## 3rd Qu.:29.25 3rd Qu.:24.50 3rd Qu.:16.25 3rd Qu.: 641.2   
## Max. :44.00 Max. :30.50 Max. :23.00 Max. :1075.0

logOddBooks <-  
 log(oddbooks) # Log computes logarithms of the Odd Books Data set  
logOddBooks

## 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

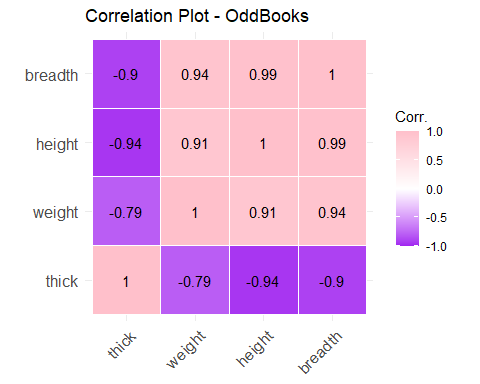
regOddBooks <-  
 lm(weight ~ thick + height + breadth, data = logOddBooks) # Linear Model between the Weight and all others  
regOddBooks

##   
## Call:  
## lm(formula = weight ~ thick + height + breadth, data = logOddBooks)  
##   
## Coefficients:  
## (Intercept) thick height breadth   
## -0.7191 0.4648 0.1537 1.8772

summary(regOddBooks) # Provides the Descriptive Stats of the Linear Model

##   
## Call:  
## lm(formula = weight ~ thick + height + breadth, data = logOddBooks)  
##   
## 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

corrOddBooks <-  
 cor(oddbooks) # Shows the Correlation of the 4 variables in the Odd Books Data set  
  
ggcorrplot(  
 corrOddBooks,  
 ggtheme = ggplot2::theme\_minimal,  
 title = "Correlation Plot - OddBooks",  
 hc.order = TRUE,  
 colors = c("purple", "white", "pink"),  
 outline.col = "white",  
 lab = TRUE,  
 method = "square",  
 show.legend = TRUE,  
 legend.title = "Corr.",  
 lab\_col = "black",  
 lab\_size = 4  
) # Shows the Correlation Plot of the 4 variables in the Odd Books Data set



ggpairs(  
 oddbooks,  
 mapping = NULL,  
 columns = 1:ncol(oddbooks),  
 title = "Correlation, Density and Scatter Plots - OddBooks",  
 upper = list(continuous = "cor"),  
 lower = list(continuous = "points"),  
 diag = list(continuous = "densityDiag"),  
 axisLabels = c("show", "internal", "none"),  
) # Shows the Correlation, Density, and Scatter Plots of the 4 variables in the Odd Books Data set

