Correlation And Visualization

Marien Bernard

12/02/2021

# \*\*Scatter Plot\*\*  
trees

## 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  
## 7 11.0 66 15.6  
## 8 11.0 75 18.2  
## 9 11.1 80 22.6  
## 10 11.2 75 19.9  
## 11 11.3 79 24.2  
## 12 11.4 76 21.0  
## 13 11.4 76 21.4  
## 14 11.7 69 21.3  
## 15 12.0 75 19.1  
## 16 12.9 74 22.2  
## 17 12.9 85 33.8  
## 18 13.3 86 27.4  
## 19 13.7 71 25.7  
## 20 13.8 64 24.9  
## 21 14.0 78 34.5  
## 22 14.2 80 31.7  
## 23 14.5 74 36.3  
## 24 16.0 72 38.3  
## 25 16.3 77 42.6  
## 26 17.3 81 55.4  
## 27 17.5 82 55.7  
## 28 17.9 80 58.3  
## 29 18.0 80 51.5  
## 30 18.0 80 51.0  
## 31 20.6 87 77.0

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.3

## -- Attaching packages --------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.1 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.2

## Warning: package 'tibble' was built under R version 4.0.2

## Warning: package 'tidyr' was built under R version 4.0.2

## Warning: package 'readr' was built under R version 4.0.2

## Warning: package 'purrr' was built under R version 4.0.2

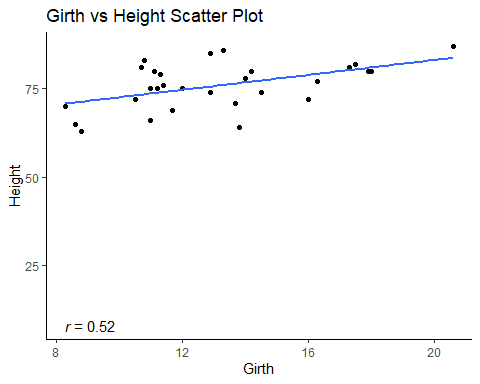
## Warning: package 'dplyr' was built under R version 4.0.2

## Warning: package 'stringr' was built under R version 4.0.2

## Warning: package 'forcats' was built under R version 4.0.2

## -- Conflicts ------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

ggplot(trees, aes(x = Girth, y = Height)) +  
   
 geom\_point() +  
 geom\_smooth(formula = y ~ x, method = "lm", se = FALSE) +  
 ggpubr::stat\_cor(aes(label = tolower(..r.label..)), label.y = 8.0) +  
 theme\_classic() +  
 theme(panel.spacing = unit(1, "lines")) +  
 labs(x = "Girth",  
 y = "Height",  
 title = "Girth vs Height Scatter Plot")



## Here The Scatter plot shows that there is a Positive Correlation  
##(i.e, Girth and Height moves along the same direction, as girth increases , height also increases)  
  
# \*\*Karl Pearson Method\*\*  
my\_data=trees  
my\_data

## 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  
## 7 11.0 66 15.6  
## 8 11.0 75 18.2  
## 9 11.1 80 22.6  
## 10 11.2 75 19.9  
## 11 11.3 79 24.2  
## 12 11.4 76 21.0  
## 13 11.4 76 21.4  
## 14 11.7 69 21.3  
## 15 12.0 75 19.1  
## 16 12.9 74 22.2  
## 17 12.9 85 33.8  
## 18 13.3 86 27.4  
## 19 13.7 71 25.7  
## 20 13.8 64 24.9  
## 21 14.0 78 34.5  
## 22 14.2 80 31.7  
## 23 14.5 74 36.3  
## 24 16.0 72 38.3  
## 25 16.3 77 42.6  
## 26 17.3 81 55.4  
## 27 17.5 82 55.7  
## 28 17.9 80 58.3  
## 29 18.0 80 51.5  
## 30 18.0 80 51.0  
## 31 20.6 87 77.0

cor\_1=cor.test(my\_data$Height, my\_data$Volume, method = "pearson")  
cor\_1

##   
## Pearson's product-moment correlation  
##   
## data: my\_data$Height and my\_data$Volume  
## t = 4.0205, df = 29, p-value = 0.0003784  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3095235 0.7859756  
## sample estimates:  
## cor   
## 0.5982497

## Karl Peasorn method is used to find the correlation between height and volume . This shows that it has a intermediate direct Correlation  
  
  
# \*\*Correlation Matrix\*\*  
my\_data=trees[,c(1,2,3)]  
my\_data

## 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  
## 7 11.0 66 15.6  
## 8 11.0 75 18.2  
## 9 11.1 80 22.6  
## 10 11.2 75 19.9  
## 11 11.3 79 24.2  
## 12 11.4 76 21.0  
## 13 11.4 76 21.4  
## 14 11.7 69 21.3  
## 15 12.0 75 19.1  
## 16 12.9 74 22.2  
## 17 12.9 85 33.8  
## 18 13.3 86 27.4  
## 19 13.7 71 25.7  
## 20 13.8 64 24.9  
## 21 14.0 78 34.5  
## 22 14.2 80 31.7  
## 23 14.5 74 36.3  
## 24 16.0 72 38.3  
## 25 16.3 77 42.6  
## 26 17.3 81 55.4  
## 27 17.5 82 55.7  
## 28 17.9 80 58.3  
## 29 18.0 80 51.5  
## 30 18.0 80 51.0  
## 31 20.6 87 77.0

head(my\_data,5)

## 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

cor\_2=round(cor(my\_data,method="spearman"),4)  
cor\_2

## Girth Height Volume  
## Girth 1.0000 0.4408 0.9547  
## Height 0.4408 1.0000 0.5787  
## Volume 0.9547 0.5787 1.0000

## The correlation matrix is used to investigate the dependence between multiple variables at the same time. The result is a table containing the correlation coefficients between each variable and the others.  
  
# \*\*Visualizing Correlation Matrix\*\*  
  
# \*\*USING CORRELOGRAM\*\*  
#install.packages(corrplot)  
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.2

## corrplot 0.84 loaded

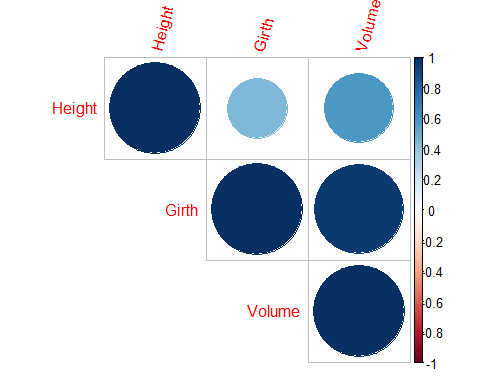
my\_data=trees[,c(1,2,3)]  
my\_data

## 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  
## 7 11.0 66 15.6  
## 8 11.0 75 18.2  
## 9 11.1 80 22.6  
## 10 11.2 75 19.9  
## 11 11.3 79 24.2  
## 12 11.4 76 21.0  
## 13 11.4 76 21.4  
## 14 11.7 69 21.3  
## 15 12.0 75 19.1  
## 16 12.9 74 22.2  
## 17 12.9 85 33.8  
## 18 13.3 86 27.4  
## 19 13.7 71 25.7  
## 20 13.8 64 24.9  
## 21 14.0 78 34.5  
## 22 14.2 80 31.7  
## 23 14.5 74 36.3  
## 24 16.0 72 38.3  
## 25 16.3 77 42.6  
## 26 17.3 81 55.4  
## 27 17.5 82 55.7  
## 28 17.9 80 58.3  
## 29 18.0 80 51.5  
## 30 18.0 80 51.0  
## 31 20.6 87 77.0

head(my\_data,5)

## 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

cor\_3=corrplot(cor\_2, type = "upper",order = "hclust",tl.col = "red", tl.srt = 75)



cor\_3

## Height Girth Volume  
## Height 1.0000 0.4408 0.5787  
## Girth 0.4408 1.0000 0.9547  
## Volume 0.5787 0.9547 1.0000

##Positive correlation are indicated by blue color, Negative correlation is indicated bu red color.  
  
##Color intensity and size of circles are also propotional to correlation.  
  
##Here we can see that there are no negative correlations, also the correlation between "height and girth" and "Height and volume " are positive and intermediate whereas all other correlations are positive and strong.  
  
#Chart Plot  
#install.packages("PerformanceAnalytics")  
library(PerformanceAnalytics)

## Warning: package 'PerformanceAnalytics' was built under R version 4.0.3

## Loading required package: xts

## Warning: package 'xts' was built under R version 4.0.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.2

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':  
##   
## first, last

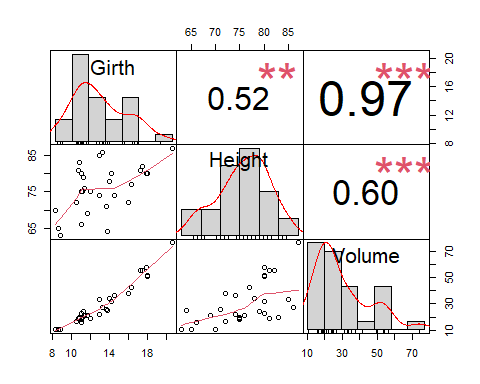
##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':  
##   
## legend

my\_data = trees[, c(1,2,3)]  
head(my\_data,5)

## 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

cor\_4=chart.Correlation(my\_data,histogram=TRUE,method="pearson")



##Diagonol contains the Distributions  
##Lower Triangle contains Scatter Diagram with fitted Line  
###Upper triangle contains correlation value with significance level stars  
##here (\*\*\*)-0.001, (\*\*)-0.01,(\*)-0.1 and so on..  
  
  
#Symbolic Number  
cor\_5=cor(trees[,c(1,2,3)])  
cor\_5

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

symnum(cor\_5, abbr.colnames = F)

## Girth Height Volume  
## Girth 1   
## Height . 1   
## Volume B . 1   
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

## 0 to 0.3 = '', 0.3 to 0.6 = '.' , 0.6 to 0.8 = ',' , 0.8 to 0.9 = '+' , 0.9 to 0.95 = '\*' , 0.95 to 1 = 'B'  
  
## volume and Girth has correlation between 0.9 to 0.95 , whereas height and girth has correlation between 0.3 to 0.6   
  
#ggcorrplot  
my\_data = trees[, c(1,2,3)]  
head(my\_data,5)

## 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

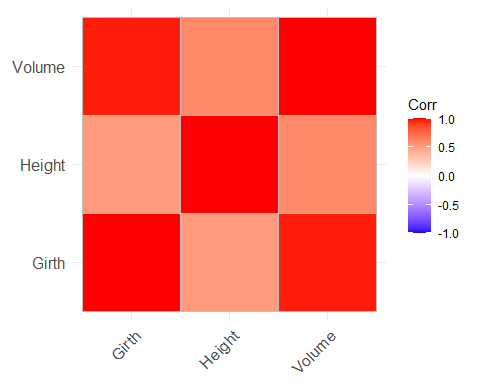
cor\_6=round(cor(my\_data),4)  
cor\_6

## Girth Height Volume  
## Girth 1.0000 0.5193 0.9671  
## Height 0.5193 1.0000 0.5982  
## Volume 0.9671 0.5982 1.0000

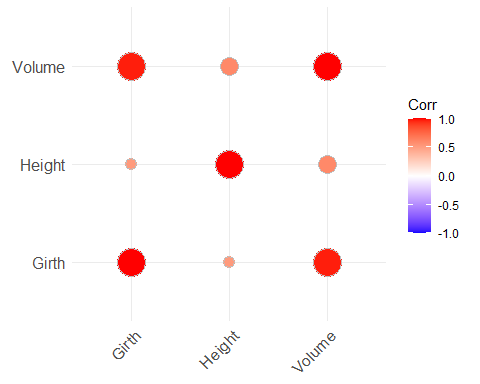
#install.packages("ggcorrplot")  
library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.0.3

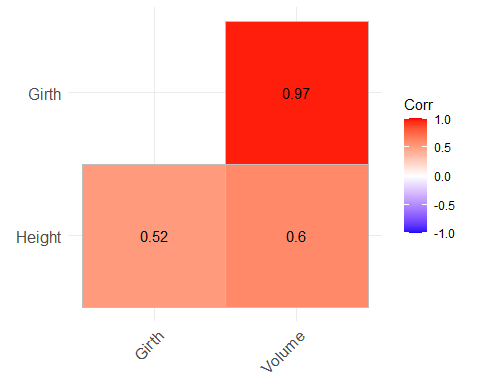
#SQUARES  
ggcorrplot(cor\_6)



#CIRCLES  
ggcorrplot(cor\_6,method="circle")



#Lower Triangle  
ggcorrplot(cor\_6,hc.order = T,type = "lower",lab = T)



##the red color indicates all correlations are positive.