R Notebook

# Download data

path='C:/indeks.csv'  
f <- read.csv2(file = path, header = TRUE, encoding = 'UNICOD')  
#Connect library  
library (dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#Have a look at the data  
glimpse(f)

## Rows: 98  
## Columns: 6  
## $ t <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1...  
## $ Y <dbl> 0.826, 0.812, 0.844, 0.782, 1.532, 1.564, 1.657, 1.575, 1.092, 1...  
## $ X1 <dbl> 1.007, 1.012, 1.009, 1.008, 1.538, 1.597, 1.607, 1.509, 1.017, 1...  
## $ X2 <dbl> 3.856, 3.887, 3.896, 3.857, 5.508, 5.510, 5.459, 5.508, 6.673, 6...  
## $ X3 <dbl> 51.2, 51.8, 52.1, 50.5, 45.2, 45.4, 45.6, 46.0, 48.2, 48.1, 48.6...  
## $ X4 <dbl> 0.435, 0.437, 0.433, 0.438, 0.225, 0.226, 0.231, 0.224, 0.467, 0...

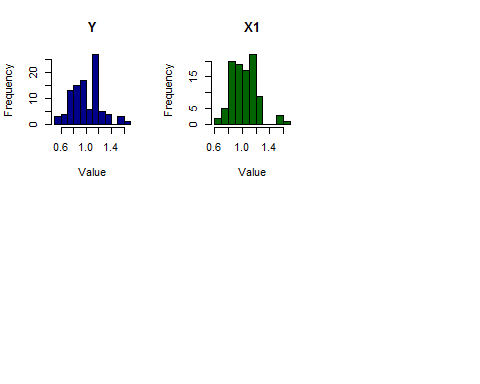
head(f)

## t Y X1 X2 X3 X4  
## 1 1 0.826 1.007 3.856 51.2 0.435  
## 2 2 0.812 1.012 3.887 51.8 0.437  
## 3 3 0.844 1.009 3.896 52.1 0.433  
## 4 4 0.782 1.008 3.857 50.5 0.438  
## 5 5 1.532 1.538 5.508 45.2 0.225  
## 6 6 1.564 1.597 5.510 45.4 0.226

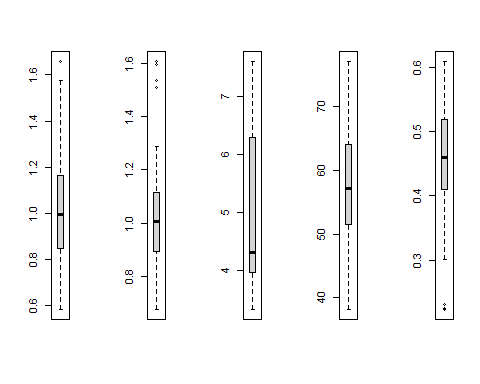
## Висновок: кількість спостережень:98 кількість змінних:6

## Histogram

library(ggplot2)  
par(mfrow = c(2, 3))  
hist(f$Y, col = 'dark blue', main = 'Y', xlab = 'Value')  
hist(f$X1, col = 'dark green', main = 'X1', xlab = 'Value')

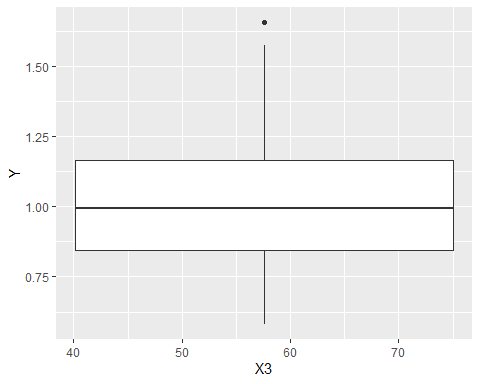
 ## Висновок:Розподіл змінних нормальний, не має довгих хвостів ## Box-plot

par(mfrow = c(1, 5))  
boxplot(f$Y)  
boxplot(f$X1)  
boxplot(f$X2)  
boxplot(f$X3)  
boxplot(f$X4)

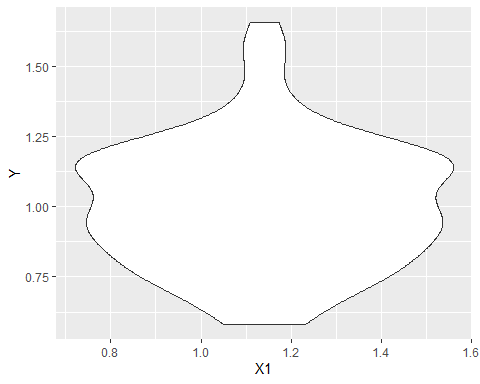


qplot(data = f,   
 x = X3,   
 y = Y,   
 geom = "boxplot")

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?

 ## Висновок:Змінні У та Х4 мають невеликі викиди, змінна Х1 значні ## Violin

qplot(data = f,   
 x = X1,   
 y = Y,   
 geom = "violin")

 ## Висновок:Розподіл нормальний, не має пропущенних данних # Statistics

library (psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

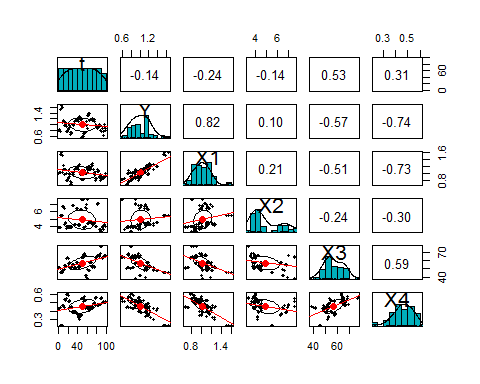
describe(f)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## t 1 98 49.50 28.43 49.50 49.50 36.32 1.00 98.00 97.00 0.00 -1.24  
## Y 2 98 1.01 0.22 1.00 1.00 0.25 0.58 1.66 1.08 0.37 0.04  
## X1 3 98 1.02 0.18 1.01 1.01 0.17 0.68 1.61 0.93 0.83 1.40  
## X2 4 98 4.92 1.27 4.32 4.80 0.63 3.34 7.60 4.26 0.81 -0.91  
## X3 5 98 57.00 8.08 57.10 56.97 9.34 38.20 77.00 38.80 0.19 -0.35  
## X4 6 98 0.46 0.08 0.46 0.46 0.09 0.22 0.61 0.38 -0.69 0.40  
## se  
## t 2.87  
## Y 0.02  
## X1 0.02  
## X2 0.13  
## X3 0.82  
## X4 0.01

## Висновок:Пропущених данних немає,всі змінні кількісні,розраховані медіана, середне відхилення, середнє,максимум та мінімум по кожній із змінних

# Correlations

pairs.panels(f, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



# Splitting the dataset into the TRAIN set and TEST set

## Висновок:Більше всього на змінну У впливає змінна Х1

set.seed(123)  
library(caTools)  
split = sample.split(f$Y, SplitRatio = 0.7)  
f\_train = subset(f, split == TRUE)  
f\_test = subset(f, split == FALSE)  
#Write prepared data to the file  
write.csv2(f\_train, file = "indeks\_train.csv")  
write.csv2(f\_test, file = "indeks\_test.csv")