1-Load the “breast-cancer-wisconsin.data.csv” from canvas into R and perform the EDA analysis by:

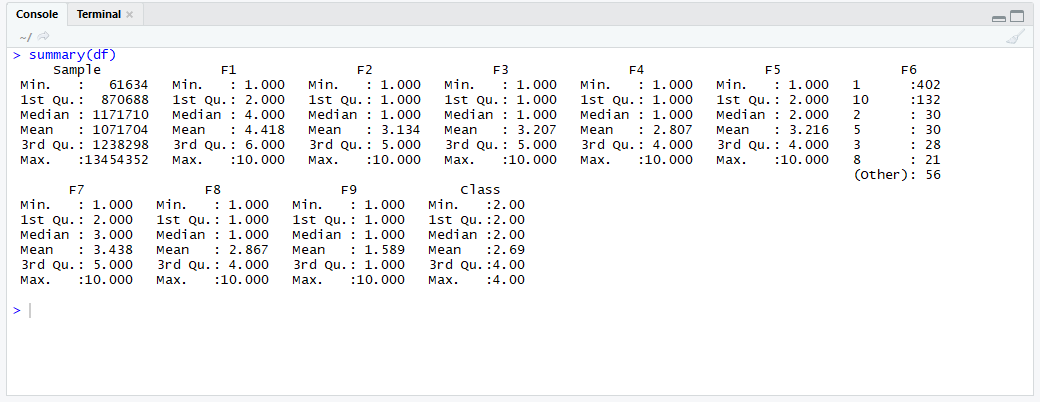
df<-read.csv("C:/Users/rajpu/Desktop/KDDM/breast-cancer-wisconsin.data.csv",header = TRUE,sep = ",")

df.

I. Summarizing each column (e.g. min, max, mean)

summary(df)

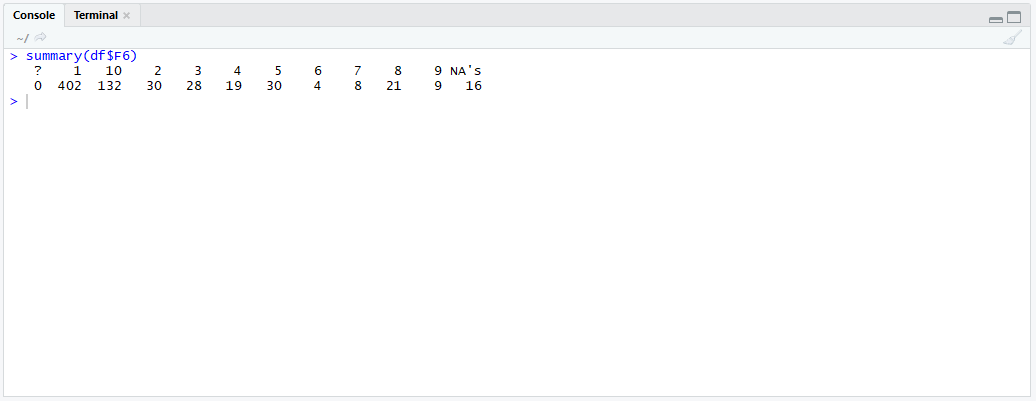
Output:



II.Identifying missing values

df[(df == "?")] <- NA

summary(df$F6)



III.Replacing the missing values with the “mode” (most frequent value) of the column.

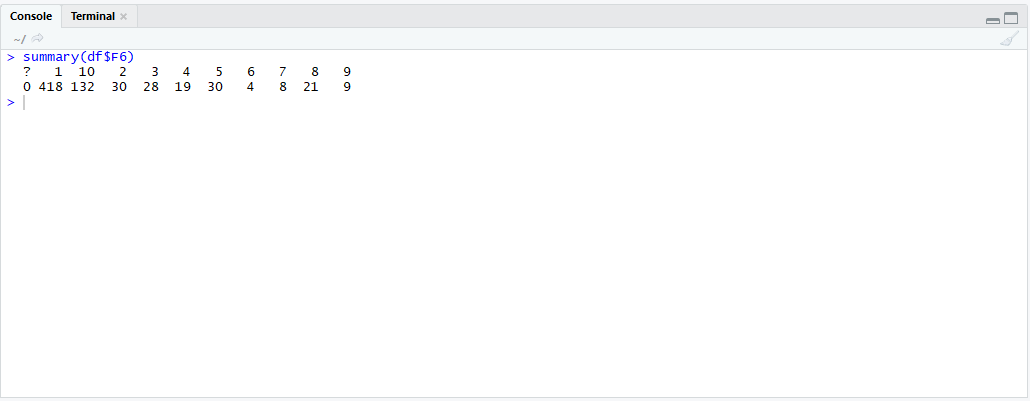
install.packages("modeest")

library(modeest)

x <- mfv(df$F6)

df[is.na(df)] <- x

summary(df$F6)

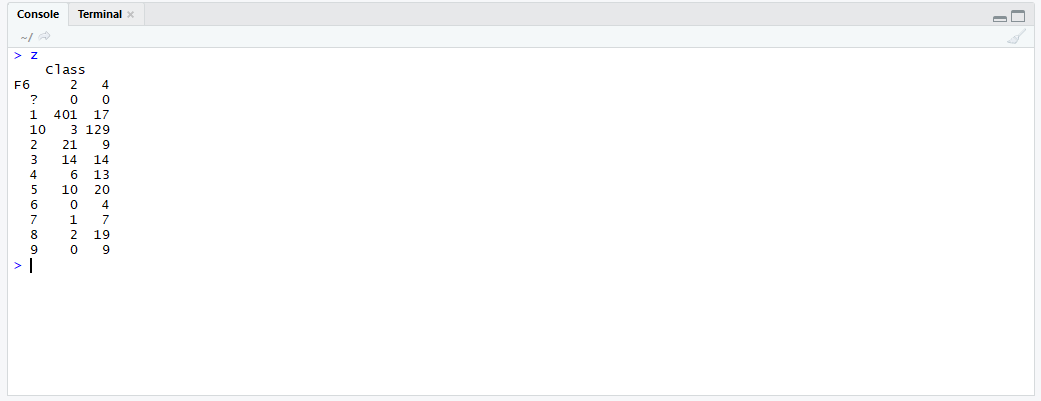


IV. Displaying the frequency table of “Class” vs. F6

attach(df)

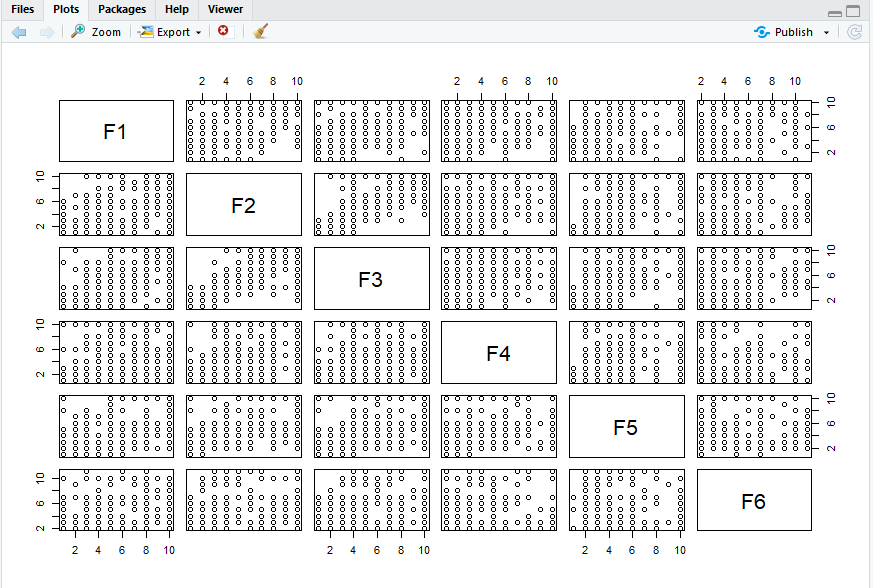
z<- table(F6,Class)

z



V. Displaying the scatter plot of F1 to F6, one pair at a time

plot(df[,2:7])



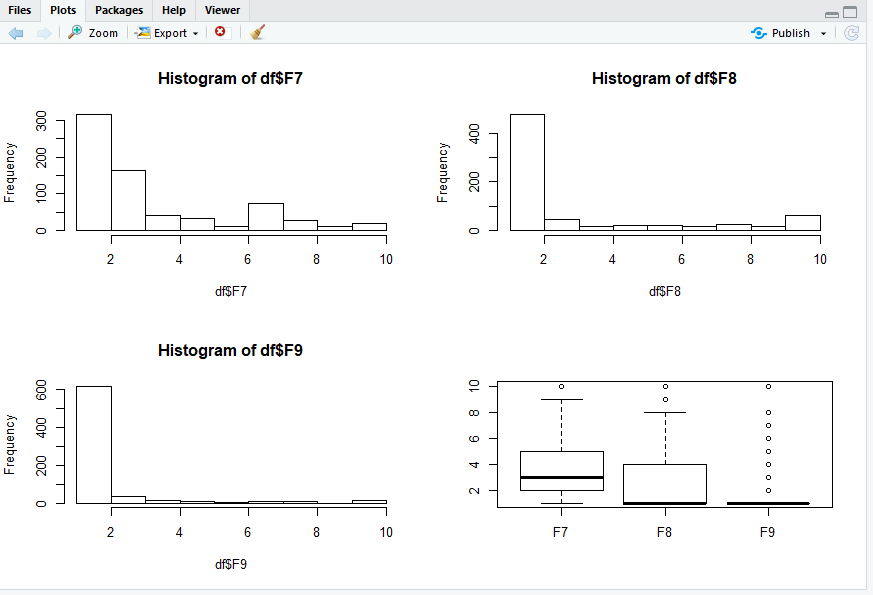
VI. Show histogram box plot for columns F7 to F9

par(mfrow=c(2,2))

hist(df$F7)

hist(df$F8)

hist(df$F9)



2- Delete all the objects from your R- environment. Reload the “breast-cancer-wisconsin.data.csv” from canvas into R. Remove any row with a missing value in any of the columns.

rm(list = ls())

df<-read.csv("C:/Users/rajpu/Desktop/KDDM/breast-cancer-wisconsin.data.csv",header = TRUE,na.strings=c("?"))

df

cs <- na.omit(df)

cs

summary(cs)

