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

setwd("/Users/jiayinhuang/SIT-homework/CS/cs513-homework")

data <- read.csv("breast-cancer-wisconsin.csv", na.strings = "?")

View(data)

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

?summary

summary(data)

Sample F1 F2 F3 F4

Min. : 61634 Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.000

1st Qu.: 870688 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 1.000

Median : 1171710 Median : 4.000 Median : 1.000 Median : 1.000 Median : 1.000

Mean : 1071704 Mean : 4.418 Mean : 3.134 Mean : 3.207 Mean : 2.807

3rd Qu.: 1238298 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000 3rd Qu.: 4.000

Max. :13454352 Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

F5 F6 F7 F8 F9

Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.000

1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000

Median : 2.000 Median : 1.000 Median : 3.000 Median : 1.000 Median : 1.000

Mean : 3.216 Mean : 3.545 Mean : 3.438 Mean : 2.867 Mean : 1.589

3rd Qu.: 4.000 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 4.000 3rd Qu.: 1.000

Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

NA's :16

Class

Min. :2.00

1st Qu.:2.00

Median :2.00

Mean :2.69

3rd Qu.:4.00

Max. :4.00

1. Identifying missing values

sum(is.na(data))

missingDataF6 = data[which(is.na(data$F6)),]

View(missingDataF6)

Graphical user interface, application, table, Excel

Description automatically generated

1. Replacing the missing values with the “mean” of the column.

# Calculate the mean of the "F6" column

mean\_F6 <- mean(data$F6, na.rm = TRUE)

# Replace missing values with the mean of the "F6" column

data$F6 <- replace(data$F6, is.na(data$F6), mean\_F6)

View(data)

Graphical user interface, table

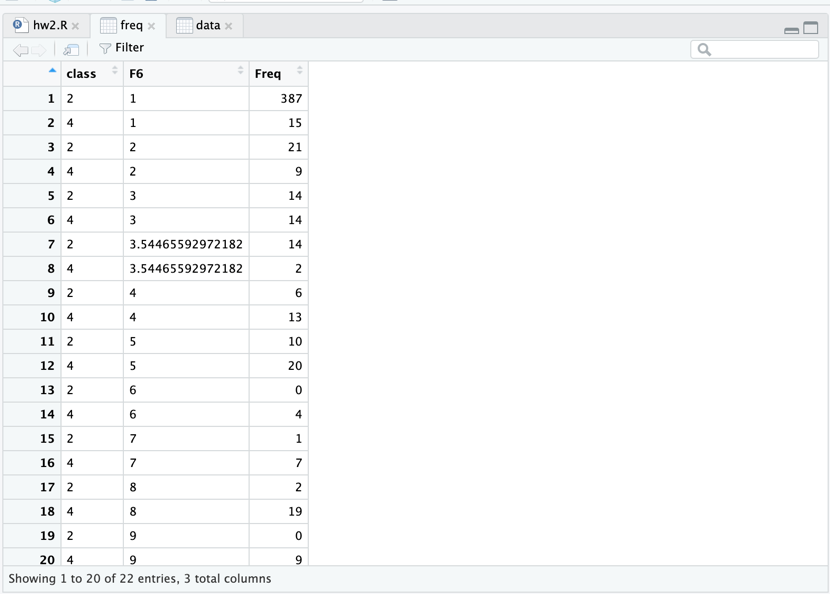
Description automatically generated

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

# Create a frequency table for "Class" and "F6"

freq <- table(class=data$Class, F6=data$F6)

View(freq)



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

?pairs

pairs(data[c(2,3)], main = "F1 & F2")

pairs(data[c(2,4)], main = "F1 & F3")

pairs(data[c(2,5)], main = "F1 & F4")

pairs(data[c(2,6)], main = "F1 & F5")

pairs(data[c(2,7)], main = "F1 & F6")

pairs(data[c(3,4)], main = "F2 & F3")

pairs(data[c(3,5)], main = "F2 & F4")

pairs(data[c(3,6)], main = "F2 & F5")

pairs(data[c(3,7)], main = "F2 & F6")

pairs(data[c(4,5)], main = "F3 & F4")

pairs(data[c(4,6)], main = "F3 & F5")

pairs(data[c(4,7)], main = "F3 & F6")

pairs(data[c(5,6)], main = "F4 & F5")

pairs(data[c(5,7)], main = "F4 & F6")

Calendar

Description automatically generatedpairs(data[c(6,7)], main = "F5 & F6")

Calendar

Description automatically generatedCalendar

Description automatically generated

Calendar

Description automatically generated

Calendar

Description automatically generatedCalendar

Description automatically generated

Calendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generated

Calendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generated

# Create scatter plots of each pair of columns

pairs(data[, 2:7], main = "Scatter Plots of F1 to F6")

Diagram, qr code

Description automatically generated

1. Show histogram box plot for columns F7 to F9

# Create a histogram for each column

hist(data$F7, main = "Histogram of F7", xlab = "F7 Values")

hist(data$F8, main = "Histogram of F8", xlab = "F8 Values")

hist(data$F9, main = "Histogram of F9", xlab = "F9 Values")

# Create a box plot of columns F7 to F9

boxplot(data[, 8:10], main = "Box Plot of F7 to F9", xlab = "Column", ylab = "Value")

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

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.

# Delete all objects from R environment

rm(list = ls())

# Load the dataset

data <- read.csv("breast-cancer-wisconsin.csv", na.strings = "?")

View(data)

# Replace missing values "?" with NA

data[data == "?"] <- NA

# Remove any rows with missing values

data <- na.omit(data)

View(data)

Graphical user interface, application, table

Description automatically generated