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)
```

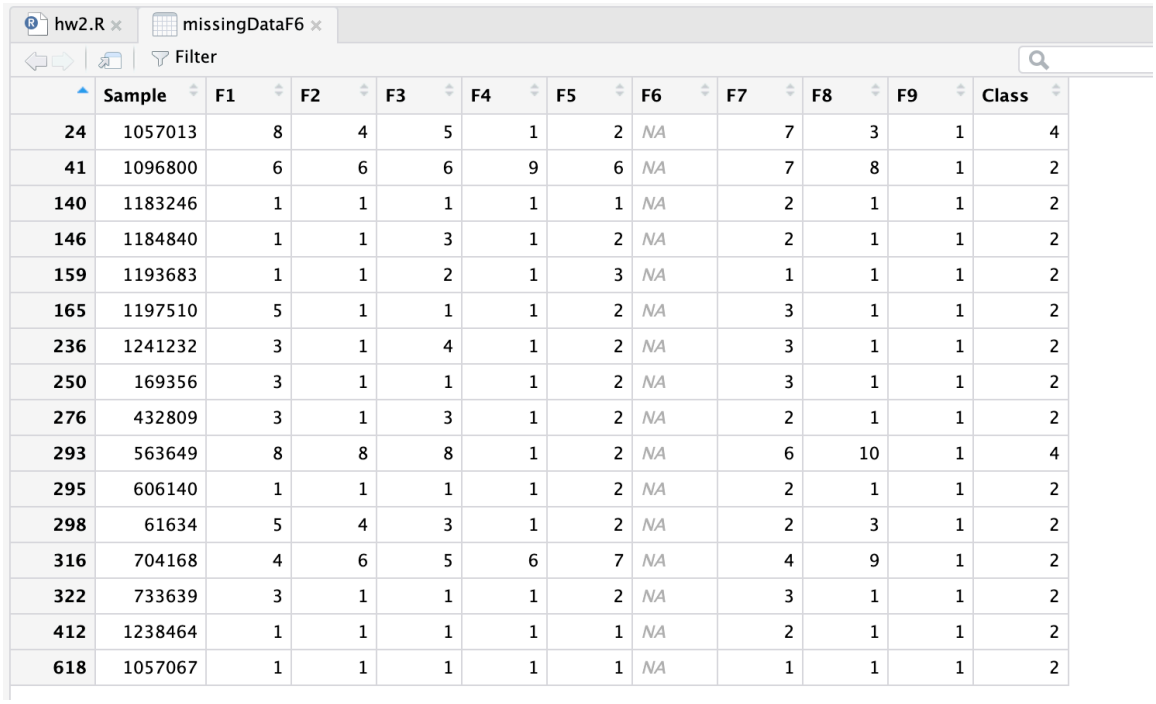
I. Summarizing each column (e.g. min, max, mean)

```
?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
```

II. Identifying missing values

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



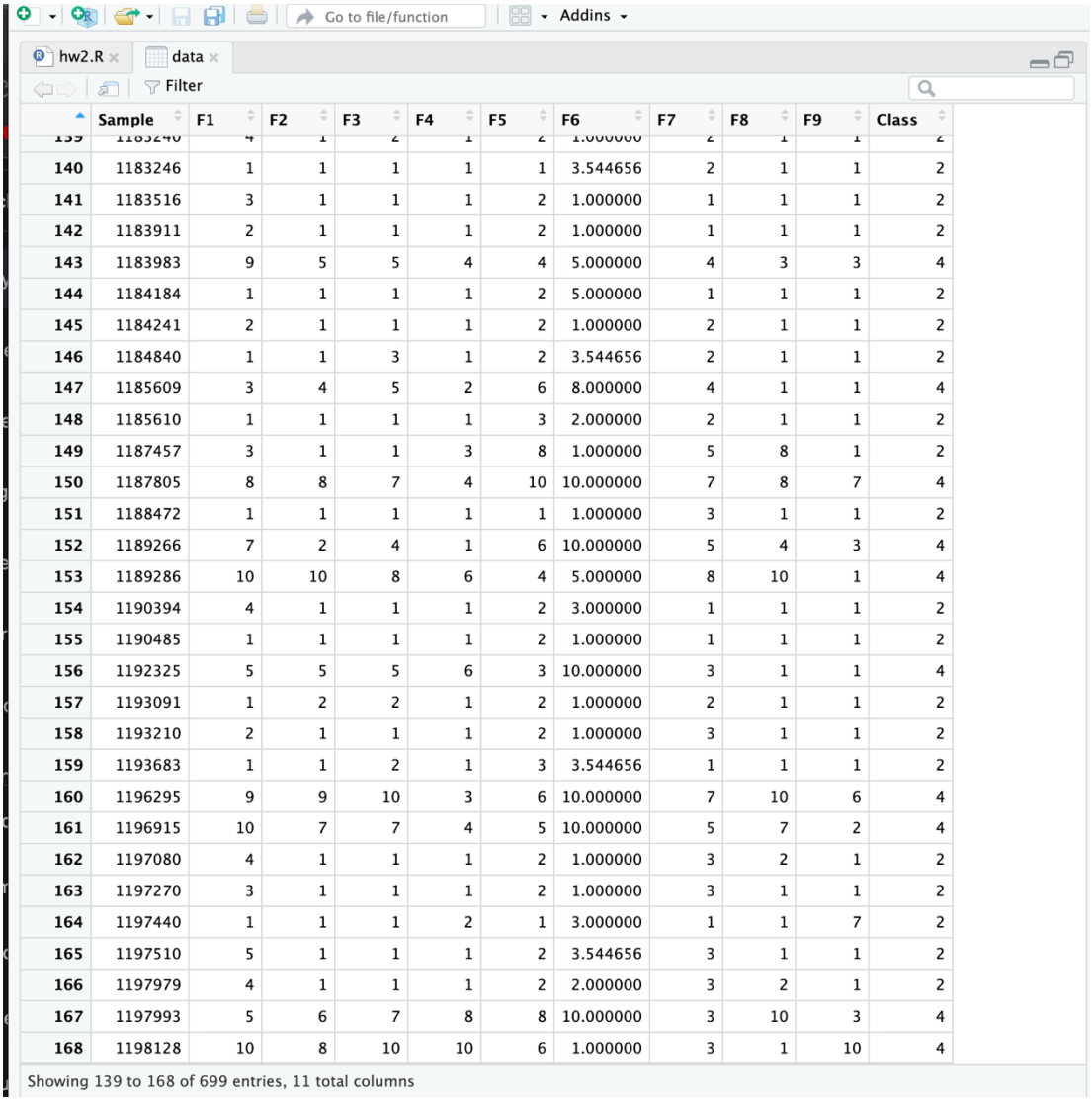
hw2.R × missingDataF6 ×

Filter

	Sample	F1	F2	F3	F4	F5	F6	F7	F8	F9	Class
24	1057013	8	4	5	1	2	NA	7	3	1	4
41	1096800	6	6	6	9	6	NA	7	8	1	2
140	1183246	1	1	1	1	1	NA	2	1	1	2
146	1184840	1	1	3	1	2	NA	2	1	1	2
159	1193683	1	1	2	1	3	NA	1	1	1	2
165	1197510	5	1	1	1	2	NA	3	1	1	2
236	1241232	3	1	4	1	2	NA	3	1	1	2
250	169356	3	1	1	1	2	NA	3	1	1	2
276	432809	3	1	3	1	2	NA	2	1	1	2
293	563649	8	8	8	1	2	NA	6	10	1	4
295	606140	1	1	1	1	2	NA	2	1	1	2
298	61634	5	4	3	1	2	NA	2	3	1	2
316	704168	4	6	5	6	7	NA	4	9	1	2
322	733639	3	1	1	1	2	NA	3	1	1	2
412	1238464	1	1	1	1	1	NA	2	1	1	2
618	1057067	1	1	1	1	1	NA	1	1	1	2

III. 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)
```



hw2.R | data

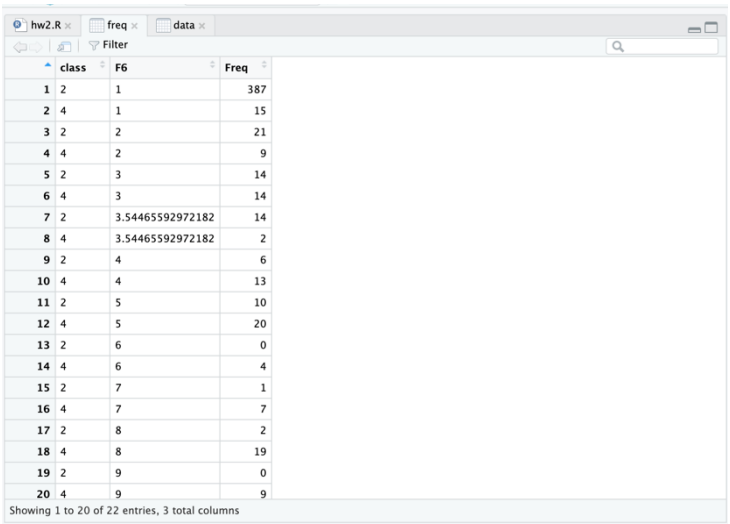
Filter

	Sample	F1	F2	F3	F4	F5	F6	F7	F8	F9	Class
140	1183246	1	1	1	1	1	3.544656	2	1	1	2
141	1183516	3	1	1	1	2	1.000000	1	1	1	2
142	1183911	2	1	1	1	2	1.000000	1	1	1	2
143	1183983	9	5	5	4	4	5.000000	4	3	3	4
144	1184184	1	1	1	1	2	5.000000	1	1	1	2
145	1184241	2	1	1	1	2	1.000000	2	1	1	2
146	1184840	1	1	3	1	2	3.544656	2	1	1	2
147	1185609	3	4	5	2	6	8.000000	4	1	1	4
148	1185610	1	1	1	1	3	2.000000	2	1	1	2
149	1187457	3	1	1	3	8	1.000000	5	8	1	2
150	1187805	8	8	7	4	10	10.000000	7	8	7	4
151	1188472	1	1	1	1	1	1.000000	3	1	1	2
152	1189266	7	2	4	1	6	10.000000	5	4	3	4
153	1189286	10	10	8	6	4	5.000000	8	10	1	4
154	1190394	4	1	1	1	2	3.000000	1	1	1	2
155	1190485	1	1	1	1	2	1.000000	1	1	1	2
156	1192325	5	5	5	6	3	10.000000	3	1	1	4
157	1193091	1	2	2	1	2	1.000000	2	1	1	2
158	1193210	2	1	1	1	2	1.000000	3	1	1	2
159	1193683	1	1	2	1	3	3.544656	1	1	1	2
160	1196295	9	9	10	3	6	10.000000	7	10	6	4
161	1196915	10	7	7	4	5	10.000000	5	7	2	4
162	1197080	4	1	1	1	2	1.000000	3	2	1	2
163	1197270	3	1	1	1	2	1.000000	3	1	1	2
164	1197440	1	1	1	2	1	3.000000	1	1	7	2
165	1197510	5	1	1	1	2	3.544656	3	1	1	2
166	1197979	4	1	1	1	2	2.000000	3	2	1	2
167	1197993	5	6	7	8	8	10.000000	3	10	3	4
168	1198128	10	8	10	10	6	1.000000	3	1	10	4

Showing 139 to 168 of 699 entries, 11 total columns

IV. 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)
```



hw2.R | freq | data

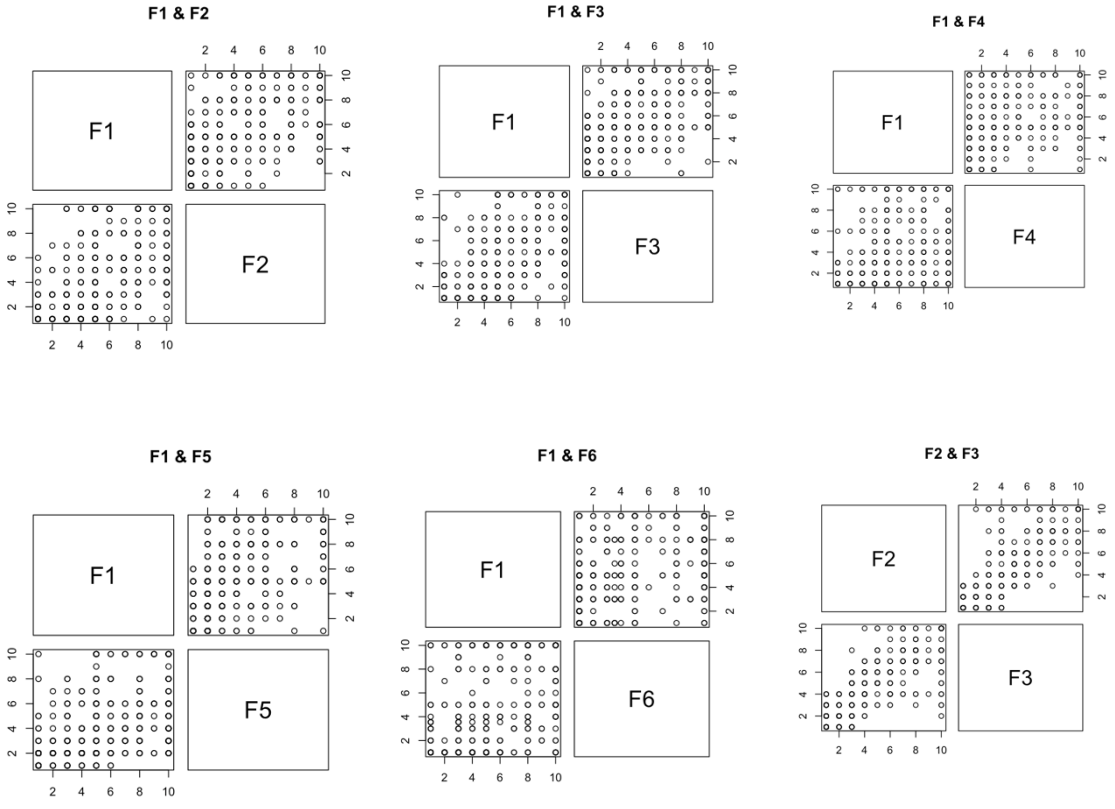
Filter

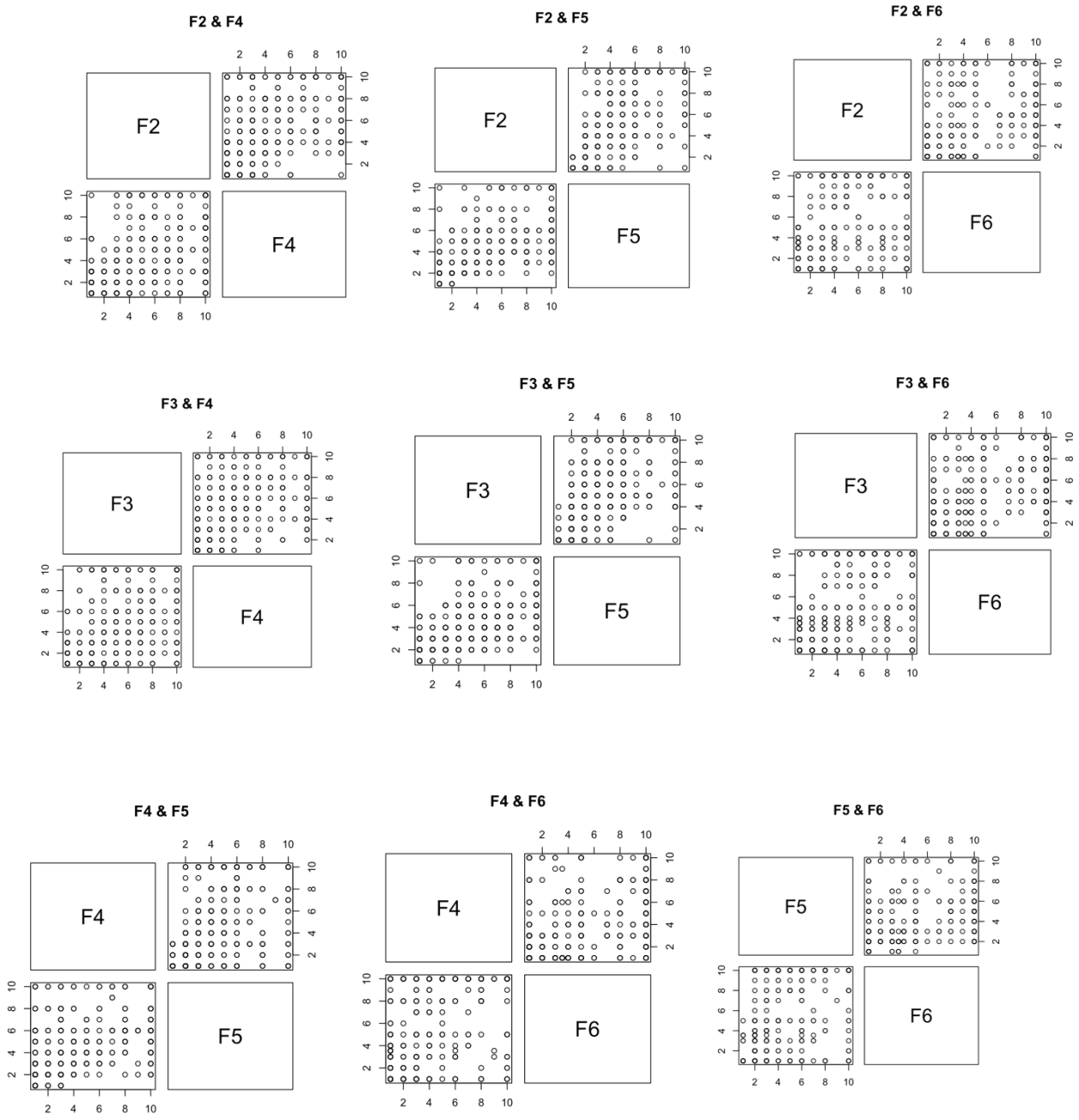
	class	F6	Freq
1	2	1	387
2	4	1	15
3	2	2	21
4	4	2	9
5	2	3	14
6	4	3	14
7	2	3.54465592972182	14
8	4	3.54465592972182	2
9	2	4	6
10	4	4	13
11	2	5	10
12	4	5	20
13	2	6	0
14	4	6	4
15	2	7	1
16	4	7	7
17	2	8	2
18	4	8	19
19	2	9	0
20	4	9	9

Showing 1 to 20 of 22 entries, 3 total columns

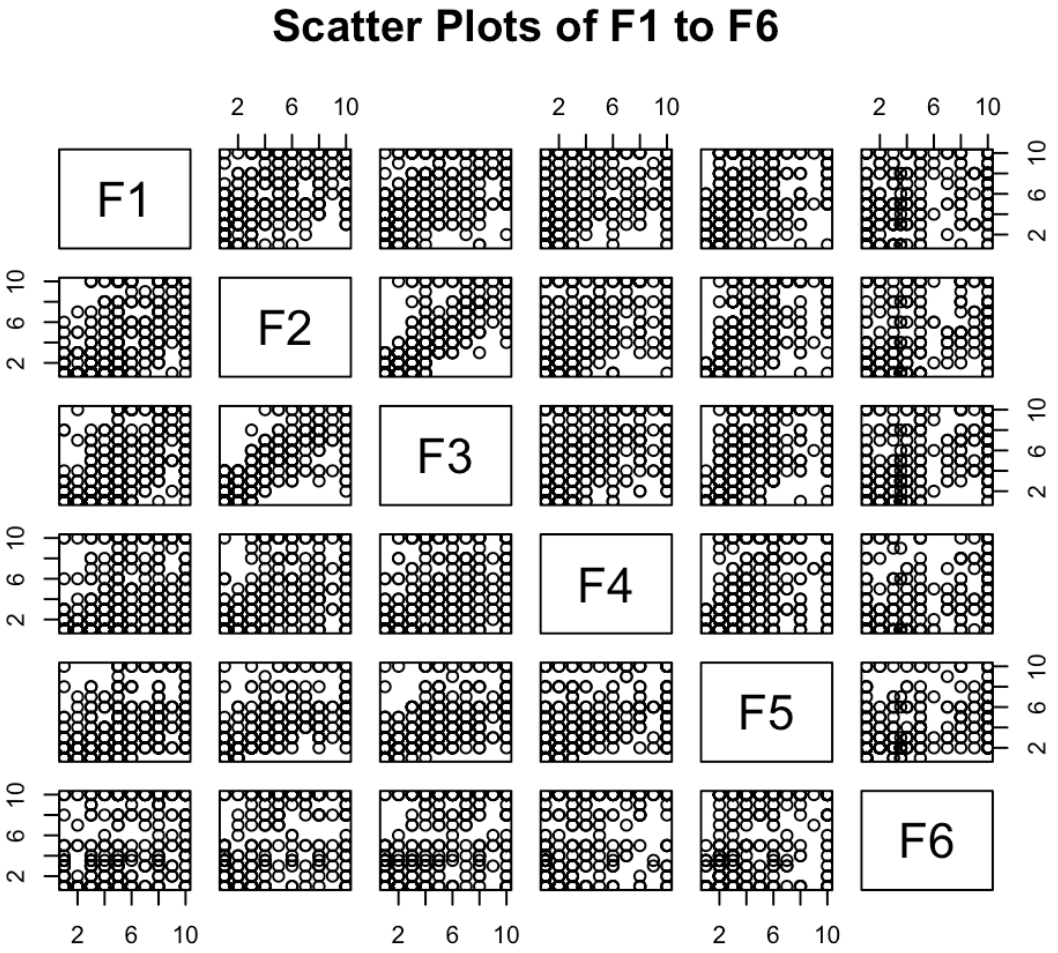
V. Displaying the scatter plot of F1 to F6, one pair at a time

```
?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")
pairs(data[c(6,7)], main = "F5 & F6")
```



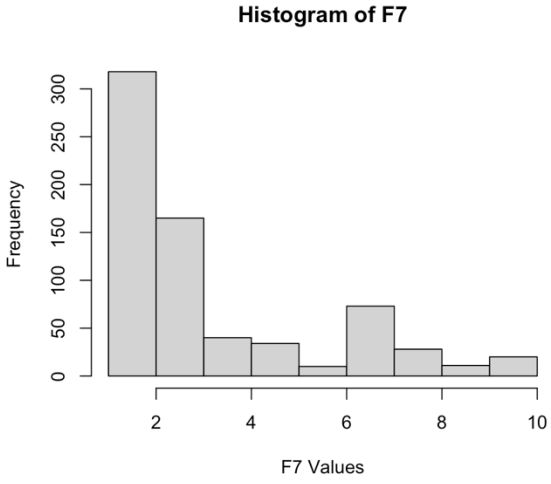

F2 & F4
F2
F4
F2 & F5
F2
F5
F2 & F6
F2
F6
F3 & F4
F3
F4
F3 & F5
F3
F5
F3 & F6
F3
F6
F4 & F5
F4
F5
F4 & F6
F4
F6
F5 & F6
F5
F6

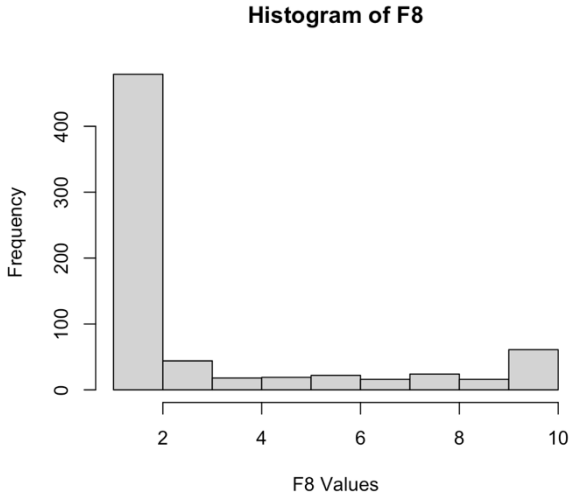
```
# Create scatter plots of each pair of columns
pairs(data[, 2:7], main = "Scatter Plots of F1 to F6")
```

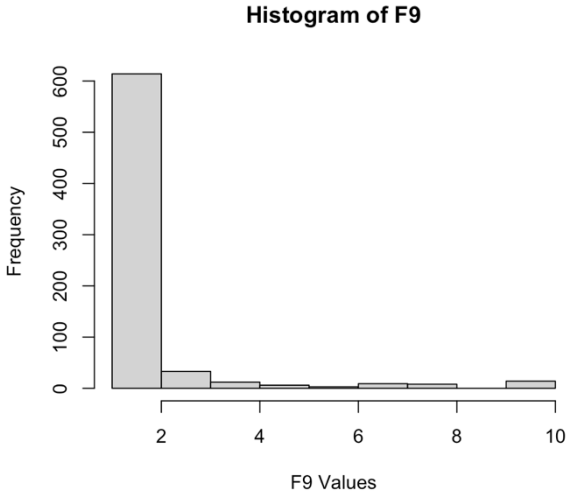


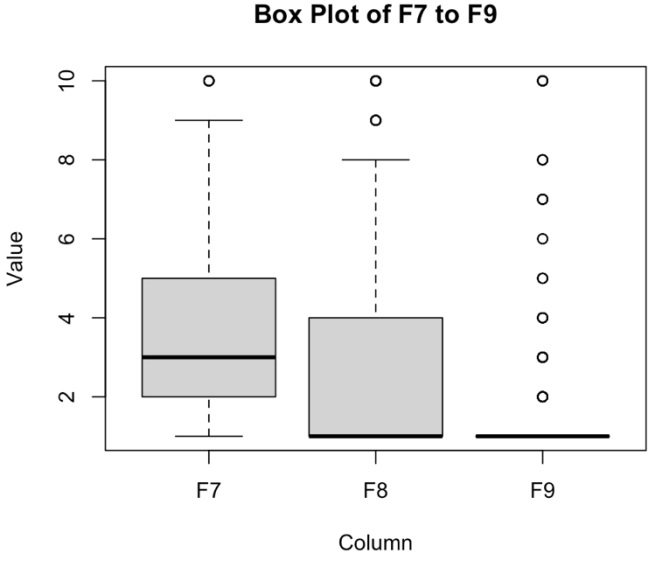
VI. Show histogram box plot for columns F7 to F9

```
# 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")
```



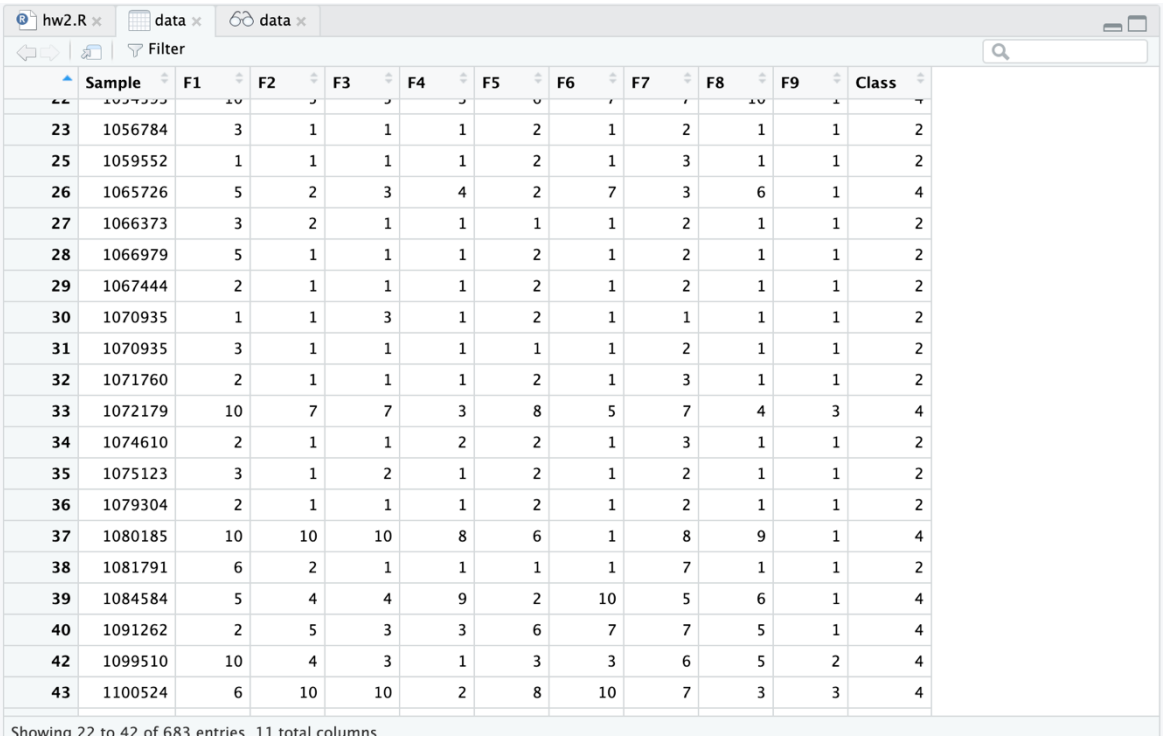






2- Delete all the objects from your R- environment. Reload the “breast-cancer-wisconsin.data.csv” from canvas into R. Remove any row with a missing value in any of the columns.

```
# 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)
```



hw2.R | data | data

Filter

	Sample	F1	F2	F3	F4	F5	F6	F7	F8	F9	Class
23	1056784	3	1	1	1	2	1	2	1	1	2
25	1059552	1	1	1	1	2	1	3	1	1	2
26	1065726	5	2	3	4	2	7	3	6	1	4
27	1066373	3	2	1	1	1	1	2	1	1	2
28	1066979	5	1	1	1	2	1	2	1	1	2
29	1067444	2	1	1	1	2	1	2	1	1	2
30	1070935	1	1	3	1	2	1	1	1	1	2
31	1070935	3	1	1	1	1	1	2	1	1	2
32	1071760	2	1	1	1	2	1	3	1	1	2
33	1072179	10	7	7	3	8	5	7	4	3	4
34	1074610	2	1	1	2	2	1	3	1	1	2
35	1075123	3	1	2	1	2	1	2	1	1	2
36	1079304	2	1	1	1	2	1	2	1	1	2
37	1080185	10	10	10	8	6	1	8	9	1	4
38	1081791	6	2	1	1	1	1	7	1	1	2
39	1084584	5	4	4	9	2	10	5	6	1	4
40	1091262	2	5	3	3	6	7	7	5	1	4
42	1099510	10	4	3	1	3	3	6	5	2	4
43	1100524	6	10	10	2	8	10	7	3	3	4

Showing 22 to 42 of 683 entries, 11 total columns