

HW_02_EDA.R

Owner

2023-02-17

```
#knowledge Discovery and Data Mining (CS 513) Homework 2: Exploratory Data Analysis
#Course : CS 513-A
# First Name : Poorvi
#Last Name : Raut
# ID : 20009560
# Purpose : Homework 2: Exploratory Data Analysis (EDA)
#Problem 1 : Load the "breast-cancer-wisconsin.data.csv" from canvas into R and perform the EDA analysis by:
```

```
#clearing object environment
rm(list = ls())
#get working directory
getwd()
```

```
## [1] "C:/Users/Owner/Desktop/Spring 2023/CS 513 KDD"
```

```
#Load the "breast-cancer-wisconsin.data.csv" from canvas into R and perform the EDA analysis
dataSet<-read.csv("/Users/Owner/Desktop/Spring 2023/CS 513 KDD/breast-cancer-wisconsin.csv",na.strings = "?")
#View Breast Cancer Dataset
View(dataSet)

# I. Summarizing each column (e.g. min, max, mean )
summary(dataSet)
```

```

##      Sample          F1          F2          F3
## Min.   : 61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
## 1st Qu.: 870688  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
## Median : 1171710 Median : 4.000   Median : 1.000   Median : 1.000
## Mean    : 1071704 Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
## 3rd Qu.: 1238298 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
## Max.    :13454352 Max.   :10.000  Max.   :10.000  Max.   :10.000
##
##      F4          F5          F6          F7
## Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
## 1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000
## Median : 1.000   Median : 2.000   Median : 1.000   Median : 3.000
## Mean    : 2.807   Mean   : 3.216   Mean   : 3.545   Mean   : 3.438
## 3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000
## Max.    :10.000  Max.   :10.000  Max.   :10.000  Max.   :10.000
##
##                  NA's :16
##      F8          F9          Class
## Min.   : 1.000   Min.   : 1.000   Min.   :2.00
## 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
## Median : 1.000   Median : 1.000   Median :2.00
## Mean    : 2.867   Mean   : 1.589   Mean   :2.69
## 3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
## Max.    :10.000  Max.   :10.000  Max.   :4.00
##

```

#II. Identifying missing values

```
is.na(dataSet)
```



```
## [676,] FALSE  
## [677,] FALSE  
## [678,] FALSE  
## [679,] FALSE  
## [680,] FALSE  
## [681,] FALSE  
## [682,] FALSE  
## [683,] FALSE  
## [684,] FALSE  
## [685,] FALSE  
## [686,] FALSE  
## [687,] FALSE  
## [688,] FALSE  
## [689,] FALSE  
## [690,] FALSE  
## [691,] FALSE  
## [692,] FALSE  
## [693,] FALSE  
## [694,] FALSE  
## [695,] FALSE  
## [696,] FALSE  
## [697,] FALSE  
## [698,] FALSE  
## [699,] FALSE FALSE
```

```
print("Number of Missing Values")
```

```
## [1] "Number of Missing Values"
```

```
print(sum(is.na(dataSet)))
```

```
## [1] 16
```

```
View(dataSet)
```

#III. Replacing the missing values with the “mean” of the column.

```
for(i in 1:ncol(dataSet)){  
  dataSet[is.na(dataSet[,i]),i]<-mean(dataSet[,i],na.rm=TRUE)  
}  
summary(dataSet)
```

```

##      Sample          F1          F2          F3
## Min.   : 61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
## 1st Qu.: 870688  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
## Median : 1171710 Median : 4.000   Median : 1.000   Median : 1.000
## Mean    : 1071704 Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
## 3rd Qu.: 1238298 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
## Max.   :13454352  Max.   :10.000  Max.   :10.000  Max.   :10.000
##          F4          F5          F6          F7
## Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
## 1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000
## Median : 1.000   Median : 2.000   Median : 1.000   Median : 3.000
## Mean    : 2.807   Mean   : 3.216   Mean   : 3.545   Mean   : 3.438
## 3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 5.000   3rd Qu.: 5.000
## Max.   :10.000  Max.   :10.000  Max.   :10.000  Max.   :10.000
##          F8          F9          Class
## Min.   : 1.000   Min.   : 1.000   Min.   :2.00
## 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
## Median : 1.000   Median : 1.000   Median :2.00
## Mean    : 2.867   Mean   : 1.589   Mean   :2.69
## 3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
## Max.   :10.000  Max.   :10.000  Max.   :4.00

```

#Rounding Values to first 3 decimal places

```
dataSet[,-1]<-round(dataSet[,-1],3)
```

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

```
data<-table(dataSet$Class,dataSet$F6)
ftable(data)
```

```

##      1   2   3 3.545   4   5   6   7   8   9   10
## 
## 2  387  21  14    14   6  10   0   1   2   0   3
## 4   15   9  14     2  13  20   4   7  19   9 129

```

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

```
plot(dataSet[2:7],main="Scatter Plot of F1 to F6",ph=10,col=2)
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a  
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a  
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a  
## graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a  
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a  
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a
## graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a
## graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a
## graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter

## Warning in plot.window(...): "ph" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter

## Warning in title(...): "ph" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a  
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical  
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

```
## Warning in title(...): "ph" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ph" is not a
## graphical parameter
```

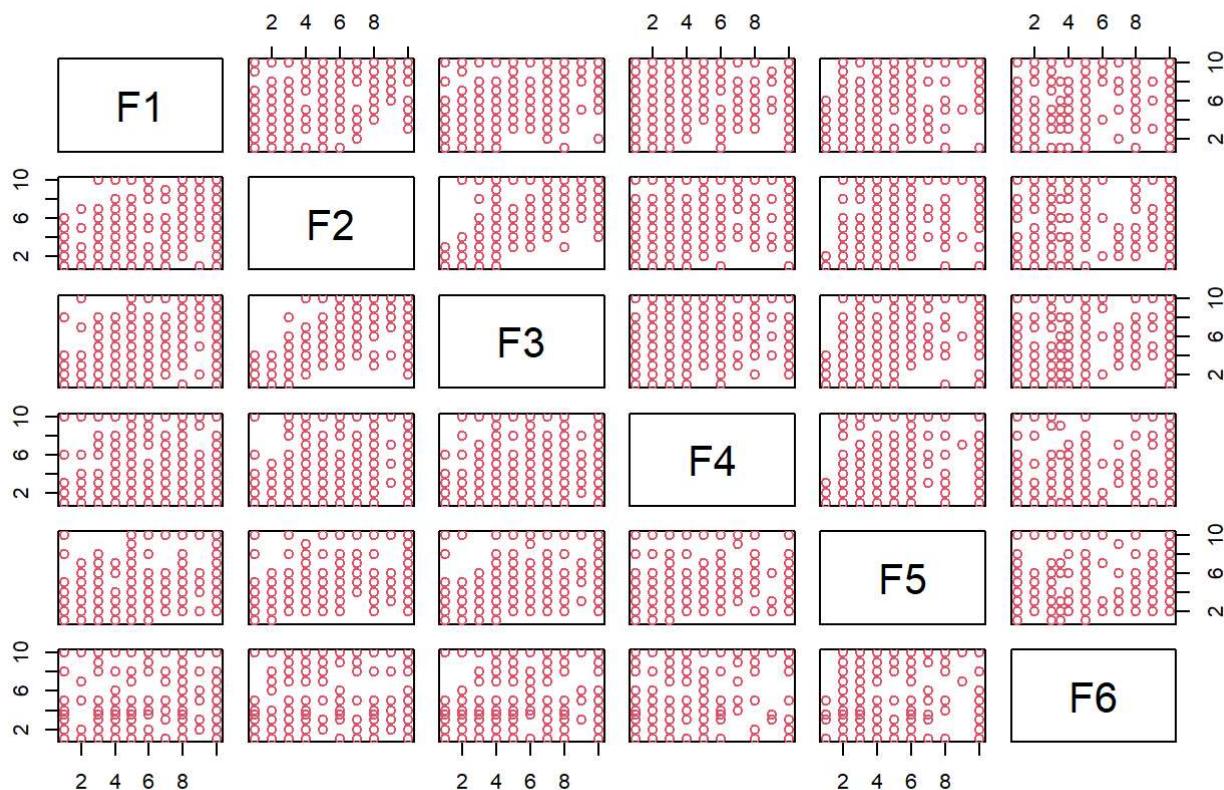
```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "ph" is not a graphical
## parameter
```

```
## Warning in plot.window(...): "ph" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ph" is not a graphical parameter
```

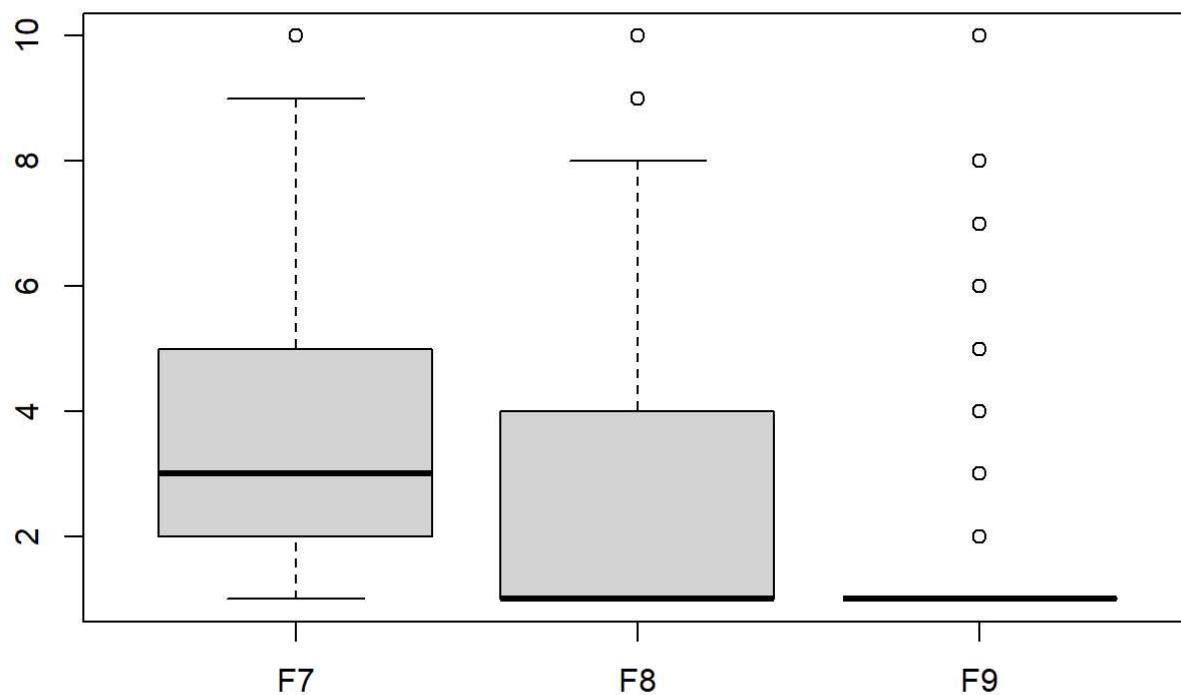
```
## Warning in title(...): "ph" is not a graphical parameter
```

Scatter Plot of F1 to F6



```
#VI. Show histogram box plot for columns F7 to F9
boxplot(dataSet[8:10],main="Box Plot for columns F7 to F9")
```

Box Plot for columns F7 to F9



#Problem 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 the objects from your R- environment.
rm(list=ls())
#Reload the “breast-cancer-wisconsin.data.csv” from canvas into R.
dataSet<-read.csv("/Users/Owner/Desktop/Spring 2023/CS 513 KDD/breast-cancer-wisconsin.csv",na.string = "?")
# Remove any row with a missing value in any of the columns.
Data_missing<-na.omit(dataSet)
View(Data_missing)
nrow(Data_missing)
```

```
## [1] 683
```

```
### END OF ASSIGNMENT ###
```