**Mid project**

***Introduction to data science***

**Name: Nisma Hossain**

**Id:20-41982-1 section:D**

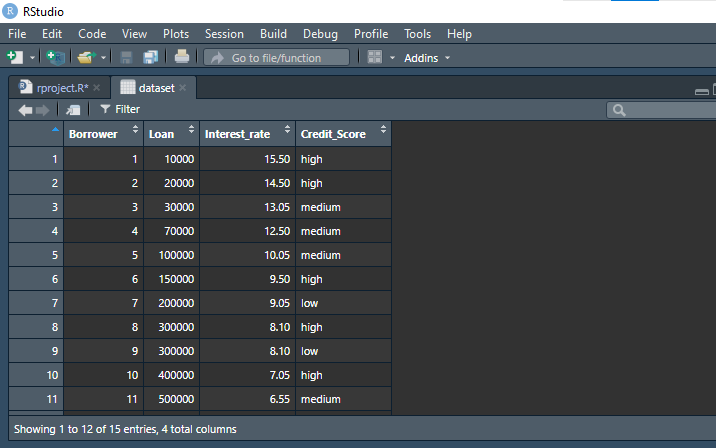
**Tasks for data\_1:**

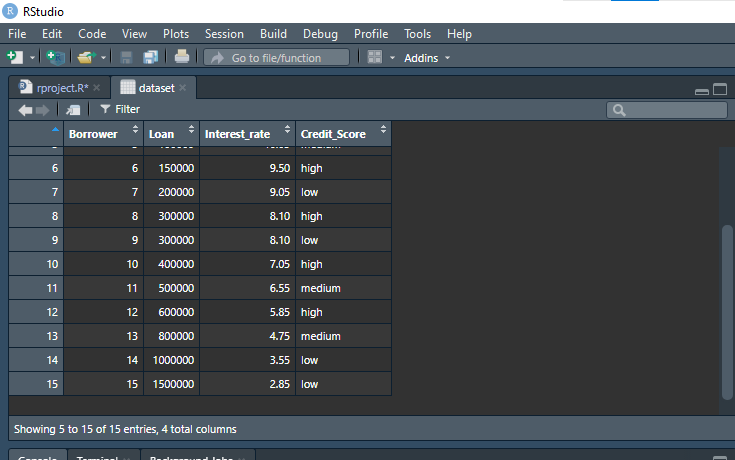
**•Import the data set(data\_1) as CSV and print the data set.**

***Code:***

dataset = read.csv('Dataset\_1.csv')

***Output:***

******

******

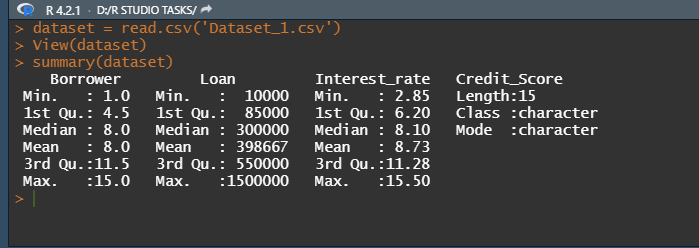
**•Find the shape of the data set.**

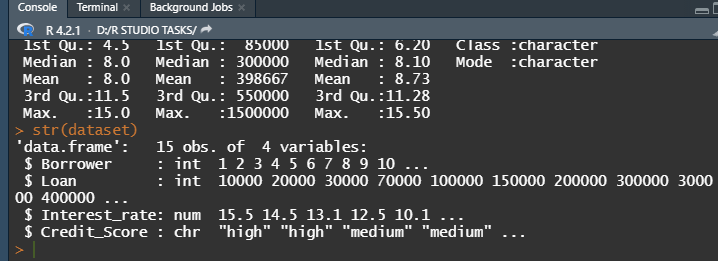
***Code:***

summary(dataset)

str(dataset)

***Output:***

******

******

**•Show the attributes name of the data set**

***Code:***

ls(dataset)

***Output:***

******

**.•Find the types of data for all attributes.**

***Code:***

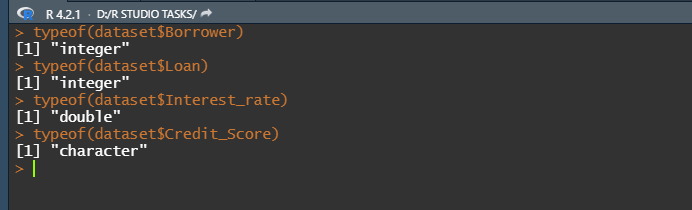
**typeof(dataset$Borrower)**

**typeof(dataset$Loan)**

**typeof(dataset$Interest\_rate)**

**typeof(dataset$Credit\_Score)**

***Output:***

******

**•Measure of center(mean, median and mode) for Loan and Interest\_rate attributes.**

***Code:***

mean(dataset$Loan)

mean(dataset$Interest\_rate)

install.packages("DescTools")

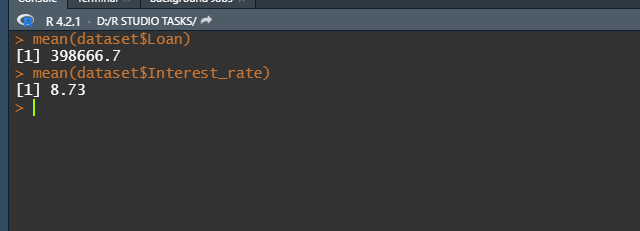
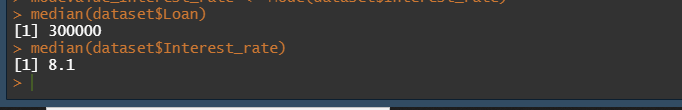
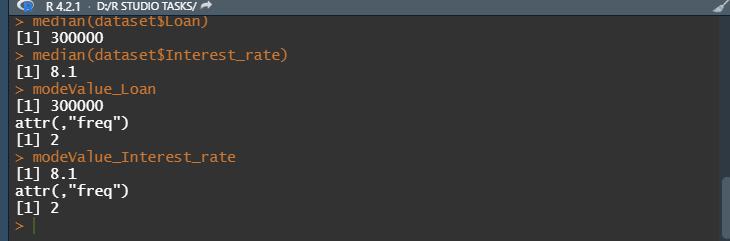
library(DescTools)

modeValue\_Loan <- Mode(dataset$Loan)

modeValue\_Interest\_rate <- Mode(dataset$Interest\_rate)

median(dataset$Loan)

median(dataset$Interest\_rate)

***Output:***

**•Measure of Spread (range and standard Deviation) for Loan and Interest\_rate attributes.**

***Code:***

install.packages('dplyr')

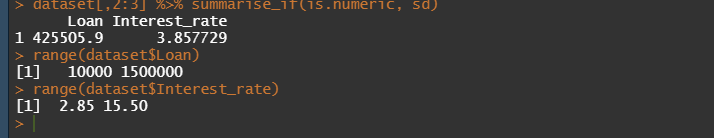
library(dplyr)

dataset[,2:3] %>% summarise\_if(is.numeric, sd)

range(dataset$Loan)

range(dataset$Interest\_rate)

***Output:***

******

**•Find the mode for Credit\_Scoreattribute.**

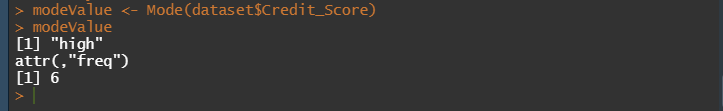
***Code:***

library("DescTools")

modeValue <- Mode(dataset$Credit\_Score)

modeValue

***Output:***

******

**Tasks for dataset2:**

Importing the dataset2:

dataset1 = read.csv('Dataset\_2.csv')

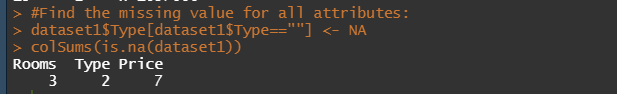
**•Find the missing value for all attributes.**

***Code:***

dataset1$Type[dataset1$Type==""] <- NA

colSums(is.na(dataset1))

***Output:***

******

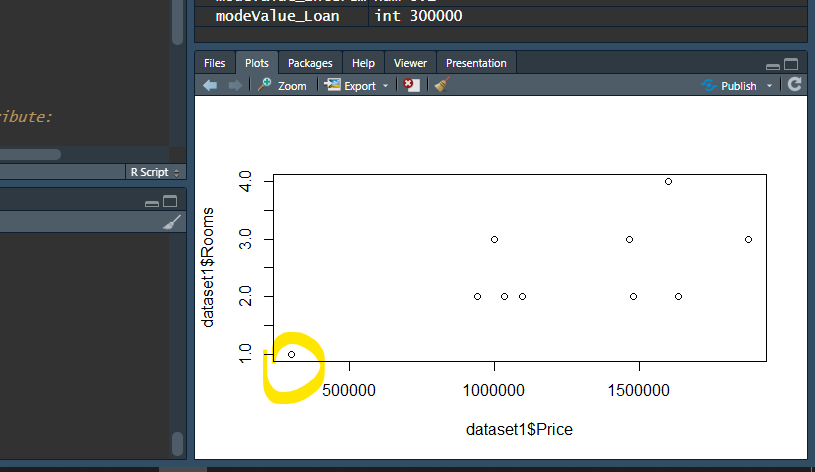
**•Detect the outlier as a missing value.**

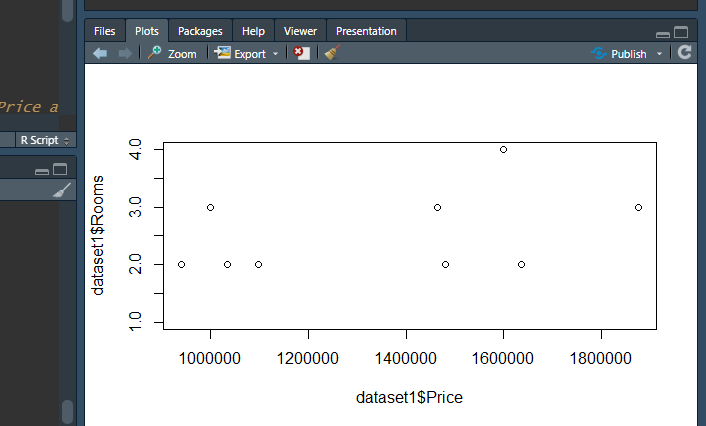
***Code:***

dataset1$Price[dataset1$Price<=300000] <- NA

plot(dataset1$Price,dataset1$Rooms)

***Output:***

******

******

**•Annotate h as 1, m as 2, and l as 3 from the “Type” attribute.**

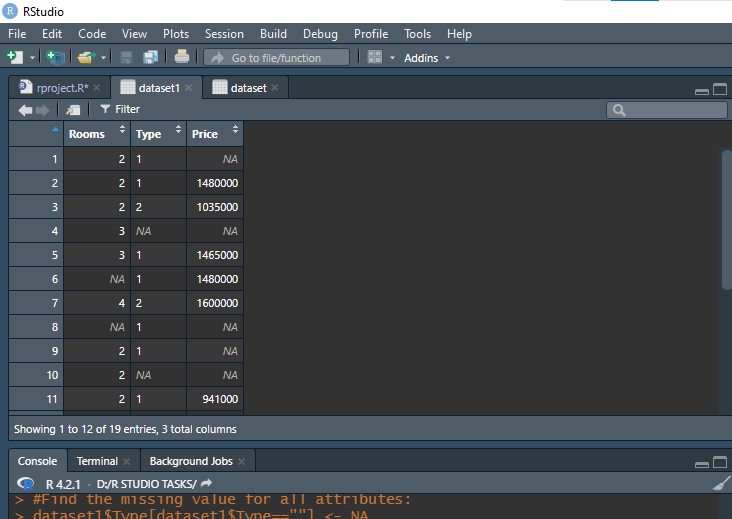
***Code:***

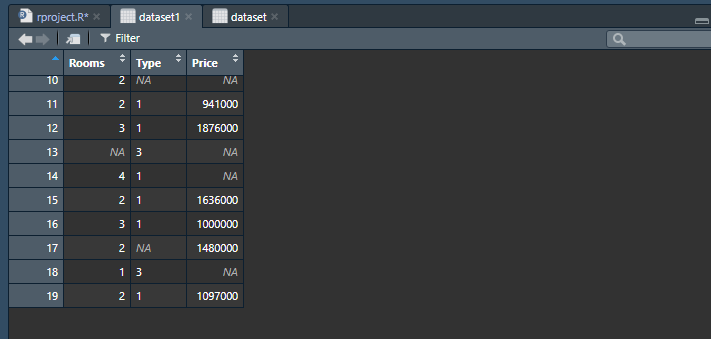
dataset1$Type = factor(dataset1$Type,

levels = c("h","m","l"),

labels = c(1,2,3))

***Output:***

******

******

**•Recover missing values by the following strategies for Rooms and Price attributes.**

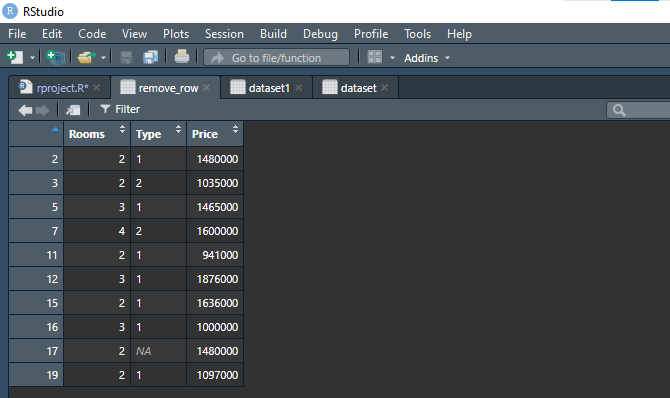
**I.Delete the rows with missing values.**

***Code:***

remove\_row =

dataset1[complete.cases(dataset1$Rooms,dataset1$Price), ]

***Output:***

******

**II. Recover missing values with the mean value.**

***Code:***

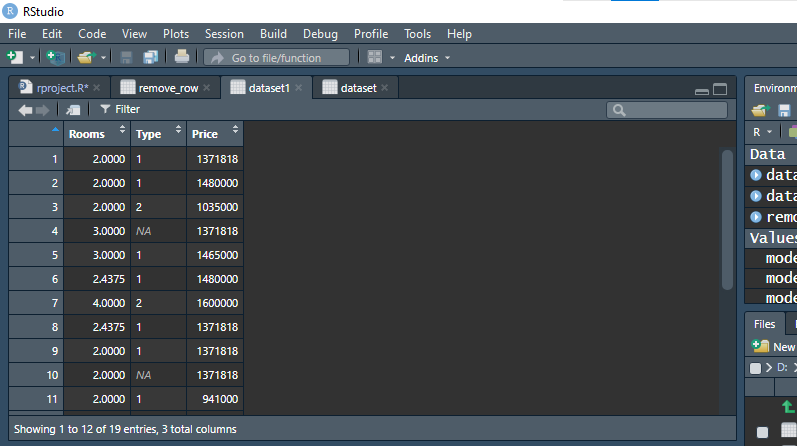
dataset1$Rooms[is.na(dataset1$Rooms)] = mean(dataset1$Rooms,

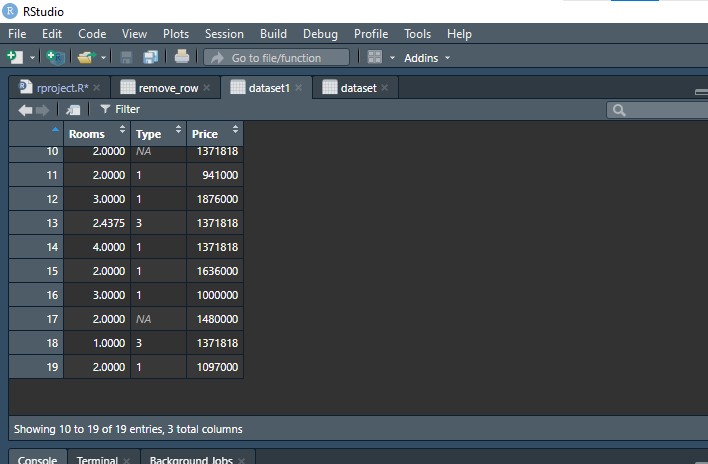
na.rm = TRUE)

dataset1$Price[is.na(dataset1$Price)] = mean(dataset1$Price, na.rm =

TRUE)

***Output:***

******

******

**III.Recover missing values with the median value.**

***Code:***

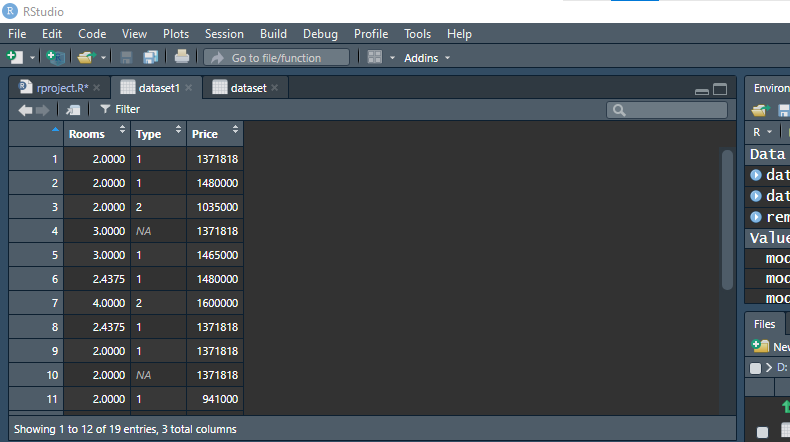
dataset1$Price[is.na(dataset1$Price)] = median(dataset1$Price, na.rm

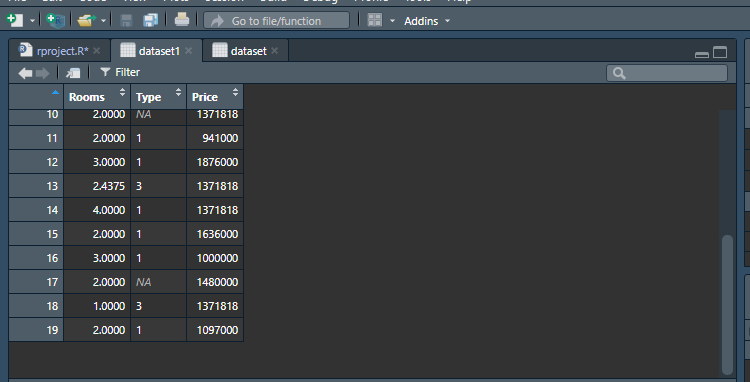
= TRUE)

dataset1$Rooms[is.na(dataset1$Rooms)] = median(dataset1$Rooms,

na.rm = TRUE)

***Output:***

******

******

**IV.Recover missing values with the mode value.**

***Code:***

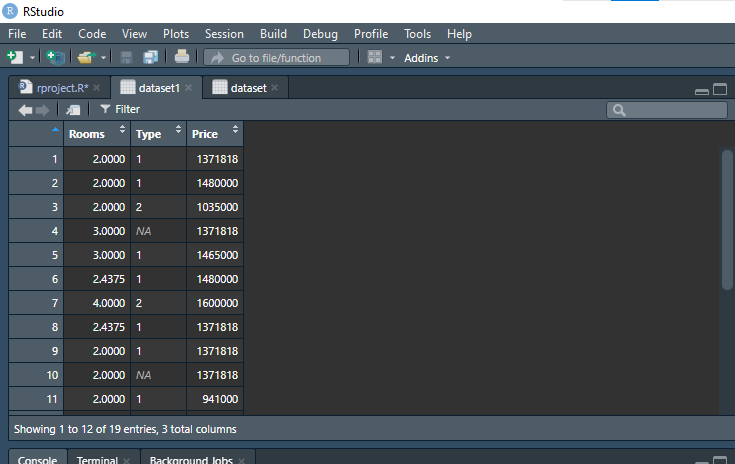
dataset1$Rooms[is.na(dataset1$Rooms)] <- Mode(dataset1$Rooms,

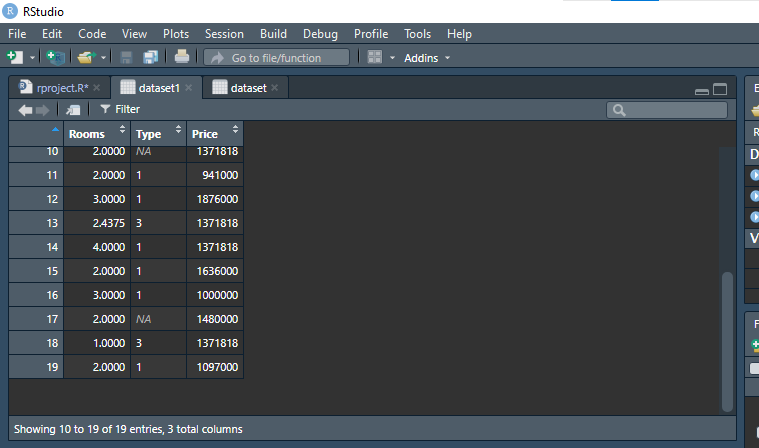
na.rm = TRUE)

dataset1$Price[is.na(dataset1$Price)] <- Mode(dataset1$Price, na.rm

= TRUE)

***Output:***

******

******