**Midterm Project**

**Introduction To Data Science**

Name: Shakib Sadat Shanto

Id: 20-43074-1

Section: D

**Tasks for data\_1:**

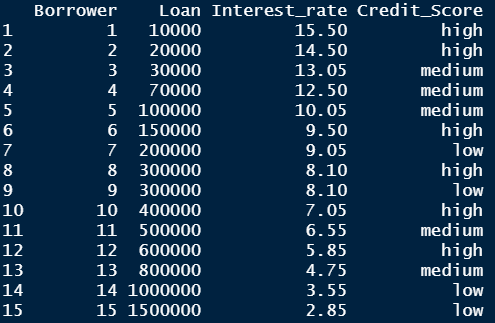
•Import the data set(data\_1) as csv and print the data set:

Code:

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

dataset

Output:



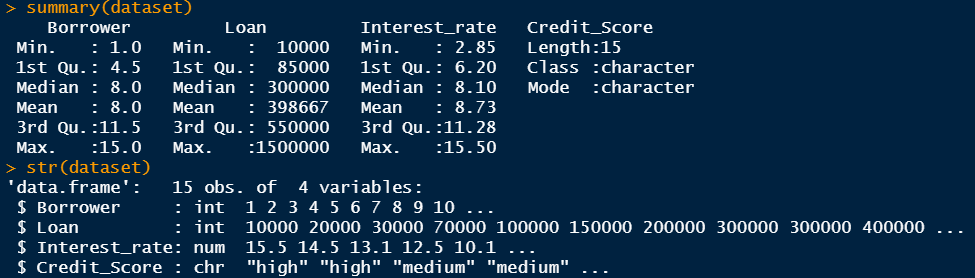
•Find the shape of the data set:

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:

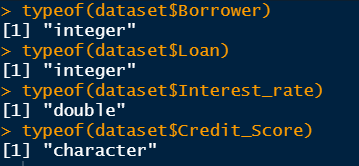
typeof(dataset$Borrower)

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:

install.packages('dplyr')

library(dplyr)

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

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

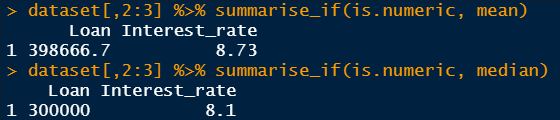
install.packages("DescTools")

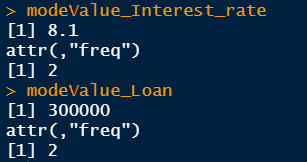
library(DescTools)

modeValue\_Loan <- Mode(dataset$Loan)

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

Output:





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

Code:

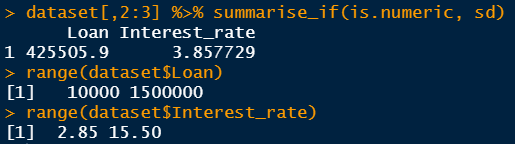
library(dplyr)

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

range(dataset$Loan)

range(dataset$Interest\_rate)

Output:



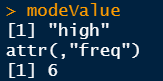
•Find the mode for Credit\_Score attribute:

Code:

library("DescTools")

modeValue <- Mode(dataset$Credit\_Score)

Output:



**Tasks for data\_2:**

•Import the data set 2(data\_2) as csv and print the data set:

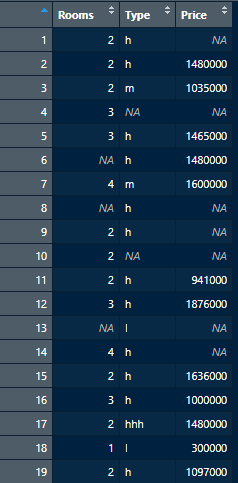
•Find the missing value for all attributes:

Code:

dataset2$Type[dataset2$Type==""] <- NA

colSums(is.na(dataset2))

Output:



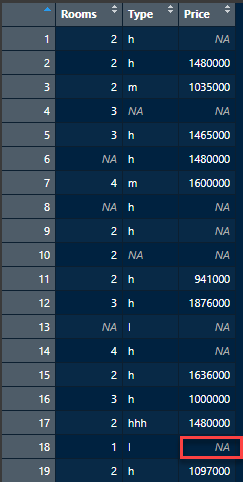


•Detect the outlier as a missing value:

Code:

dataset2$Price[dataset2$Price<=300000] <- NA

Output:



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

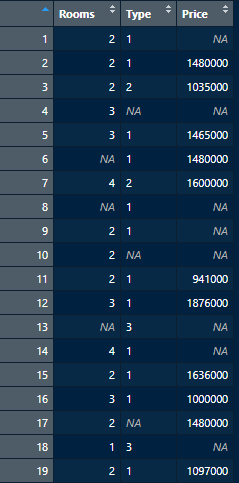
Code:

dataset2$Type = factor(dataset2$Type,

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

labels = c(1,2,3))

Output:



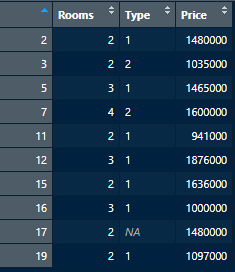
•Recover missing values by the following strategies for Rooms and Price attributes:

1. Delete the rows with missing values:

Code:

remove\_row = dataset2[complete.cases(dataset2$Rooms,dataset2$Price), ]

Output:



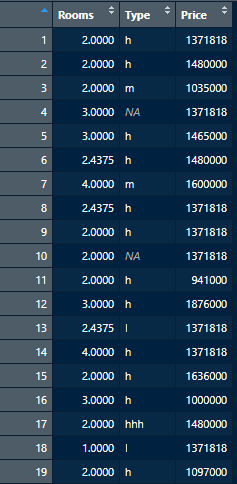
1. Recover missing values with the mean value:

Code:

dataset2$Rooms[is.na(dataset2$Rooms)] = mean(dataset2$Rooms, na.rm = TRUE)

dataset2$Price[is.na(dataset2$Price)] = mean(dataset2$Price, na.rm = TRUE)

Output:



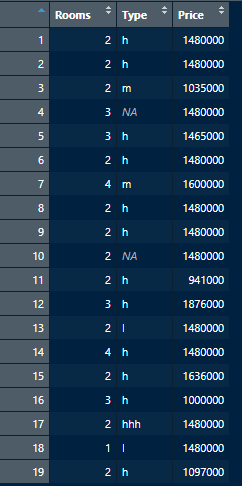
1. Recover missing values with the median value:

Code:

dataset2$Price[is.na(dataset2$Price)] = median(dataset2$Price, na.rm = TRUE)

dataset2$Rooms[is.na(dataset2$Rooms)] = median(dataset2$Rooms, na.rm = TRUE)

Output:



1. Recover missing values with the mode value:

Code:

dataset2$Rooms[is.na(dataset2$Rooms)] <- Mode(dataset2$Rooms, na.rm = TRUE)

dataset2$Price[is.na(dataset2$Price)] <- Mode(dataset2$Price, na.rm = TRUE)

Output:

