**American International University-Bangladesh**



**Course:** INTRODUCTION TO DATA SCIENCE

**Assignment Title:** Mid assignment

**Submitted by:**

Name: Md. Shamim Siddiky  
Id: 20-42649-1  
Section: D

**Submitted to:**

**TOHEDUL ISLAM**

Assistant professor, American International University-Bangladesh.

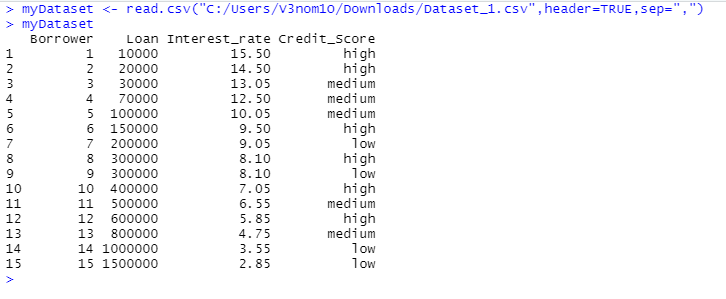
**Submission Date:** 8th November, 2022

**For Dataset\_1:**

1. **To import the dataset and print**

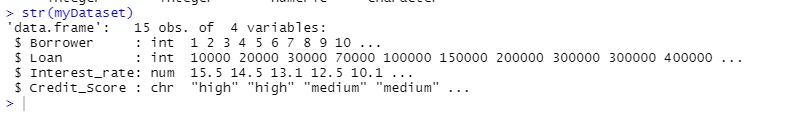
myDataset <- read.csv ("C:/Users/V3nom1O/Downloads/Dataset\_1.csv”, header=TRUE, sep=",")

myDataset



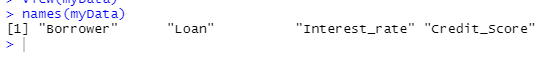
1. **To find the shape of the dataset**

str(myDataset)



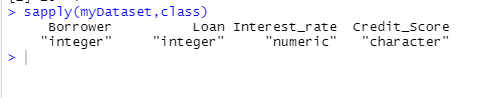
1. **To show the attributes names of the dataset**

names(myDataset)



1. **To find the types of data for all attributes**

Sapply (myDataset, class)



1. **To measure of center (mean, median, mode) for Loan and Interest\_rate**

#loan

loan.mean <- mean(myDataset$Loan)

loan.median <- median(myDataset$Loan)

#mode; creating mode function

val <- as.numeric(myDataset$Loan)

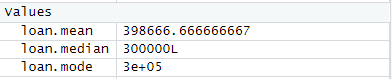
getLoanMode <- function(val){

uniqv <- unique(val)

uniqv[which.max(tabulate(match(val, uniqv)))]

}

loan.mode <- getLoanMode(val)



#interest rate

rate.mean <- mean(myDataset$Interest\_rate)

rate.median <- median(myDataset$Interest\_rate)

#mode; creating mode function

val2 <- as.numeric(myDataset$Interest\_rate)

getRateMode <- function(val2){

uniqv <- unique(val2)

uniqv[which.max(tabulate(match(val2, uniqv)))]

}

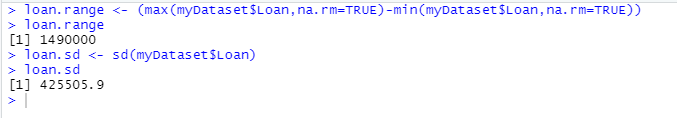
rate.mode <- getRateMode(val2)



1. **To measure of Spread (range and standard deviation) for loan and interest\_rate**

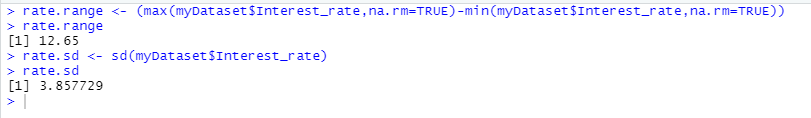
loan.range <- (max(myDataset$Loan,na.rm=TRUE)-min(myDataset$Loan,na.rm=TRUE))

loan.sd <- sd(myDataset$Loan)



rate.range <- (max(myDataset$Interest\_rate,na.rm=TRUE)- min(myDataset$Interest\_rate,na.rm=TRUE))

rate.sd <- sd(myDataset$Interest\_rate)



1. **To find the mode for Credit\_score attribute**

val3 <- as.character(myDataset$Credit\_Score)

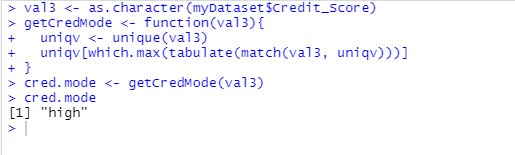
getCredMode <- function(val3){

uniqv <- unique(val3)

uniqv[which.max(tabulate(match(val3, uniqv)))]

}

cred.mode <- getCredMode(val3)



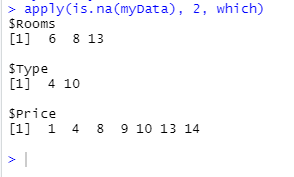
**For Dataset\_2:**

myData <- read.csv("C:/Users/V3nom1O/Downloads/Dataset\_2.csv",header=TRUE,sep=",")



1. **To find the missing values of all attributes**

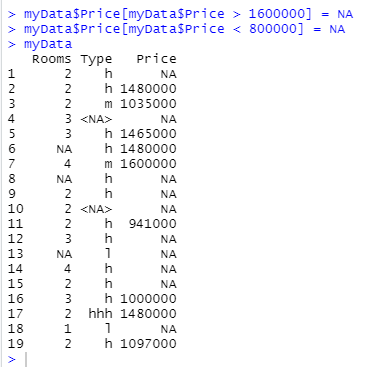
apply(is.na(myData), 2, which)



1. **Detect outliers as a missing value**

myData$Price[myData$Price > 1600000] = NA

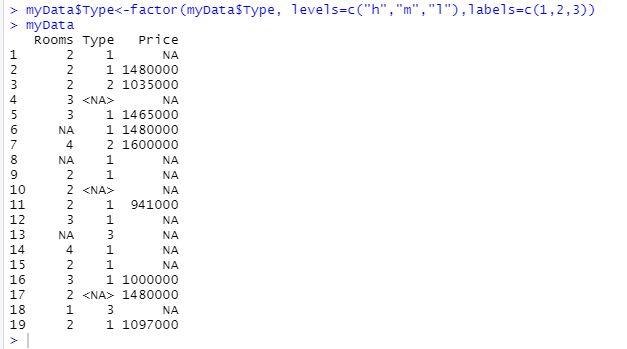
myData$Price[myData$Price < 800000] = NA



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

myData$Type<-factor(myData$Type, levels=c("h","m","l"),labels=c(1,2,3))

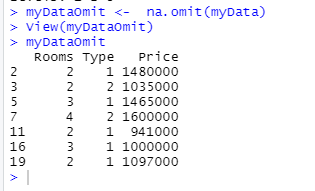
myData



1. **Recover missing values by the following strategies for Rooms and Price Attributes**
   1. **Delete the rows with missing value**

myDataOmit <- na.omit(myData)

myDataOmit



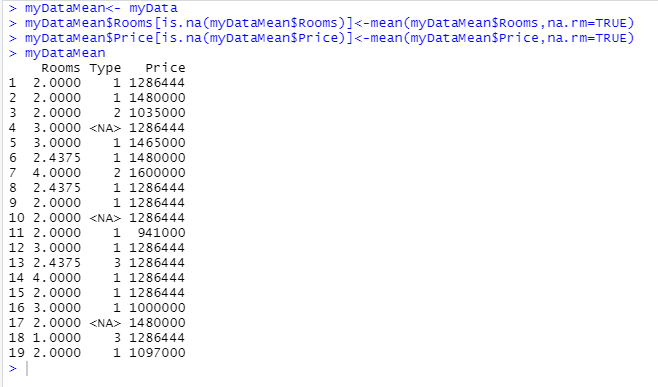
* 1. **Recover missing values with mean value**

myDataMean<- myData

myDataMean$Rooms[is.na(myDataMean$Rooms)]<-mean(myDataMean$Rooms,na.rm=TRUE)

myDataMean$Price[is.na(myDataMean$Price)]<-mean(myDataMean$Price,na.rm=TRUE)

myDataMean



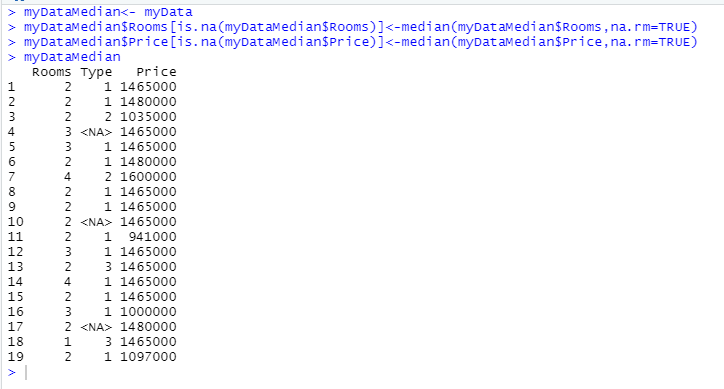
* 1. **Recover missing values with median value**

myDataMedian<- myData

myDataMedian$Rooms[is.na(myDataMedian$Rooms)]<-median(myDataMedian$Rooms,na.rm=TRUE)

myDataMedian$Price[is.na(myDataMedian$Price)]<-median(myDataMedian$Price,na.rm=TRUE)

myDataMedian



* 1. **Recover missing values with mode value**

rooms <- myData$Rooms

getRoomMode <- function(rooms){

uniqv <- unique(rooms)

uniqv[which.max(tabulate(match(rooms, uniqv)))]

}

room.mode <- getRoomMode(rooms)

myData$Rooms[is.na(myData$Rooms)]<-room.mode

myDataPrice <- na.omit(myData)

price <- myDataPrice$Price

getPriceMode <- function(price){

uniqv <- unique(price)

uniqv[which.max(tabulate(match(price, uniqv)))]

}

price.mode <- getPriceMode(price)

myData$Price[is.na(myData$Price)]<-price.mode

myData

