ACADGILD

SESSION 11: Linear Models

Assignment 1

Table of Contents

1. Introduction

This assignment will help you understand the concepts learnt in the session.

2. Objective

This assignment will test your skills on the basics of Regression Analysis and Modeling.

3. Prerequisites

Not applicable.

4. Associated Data Files

Not applicable.

5. Problem Statement

1. Use the link given below and locate the bank marketing dataset.
https://archive.ics.uci.edu/ml/machine-learning-databases/00222/

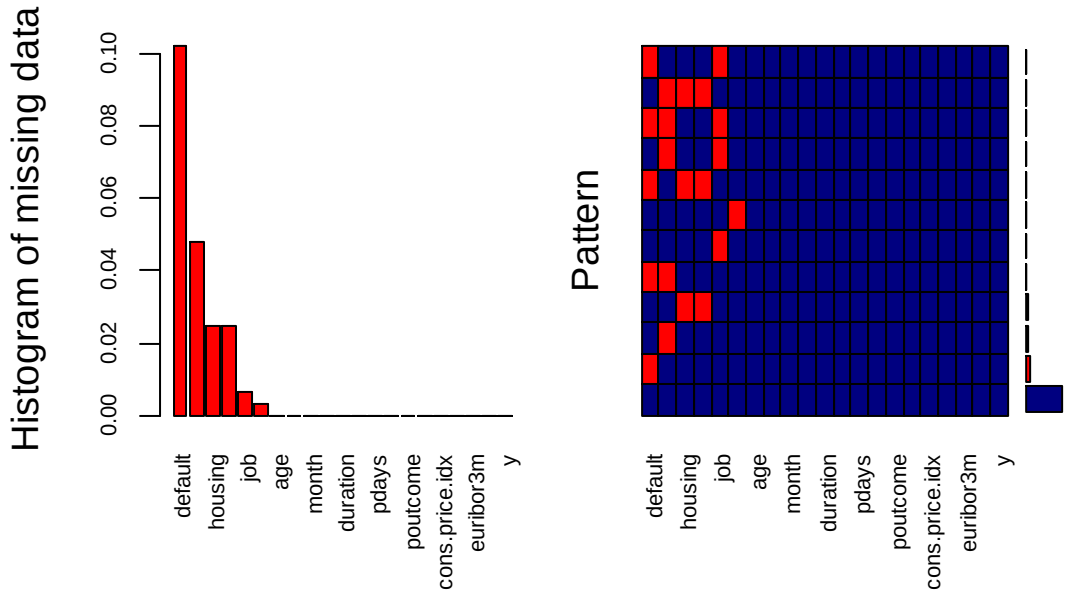
Perform the below operations:

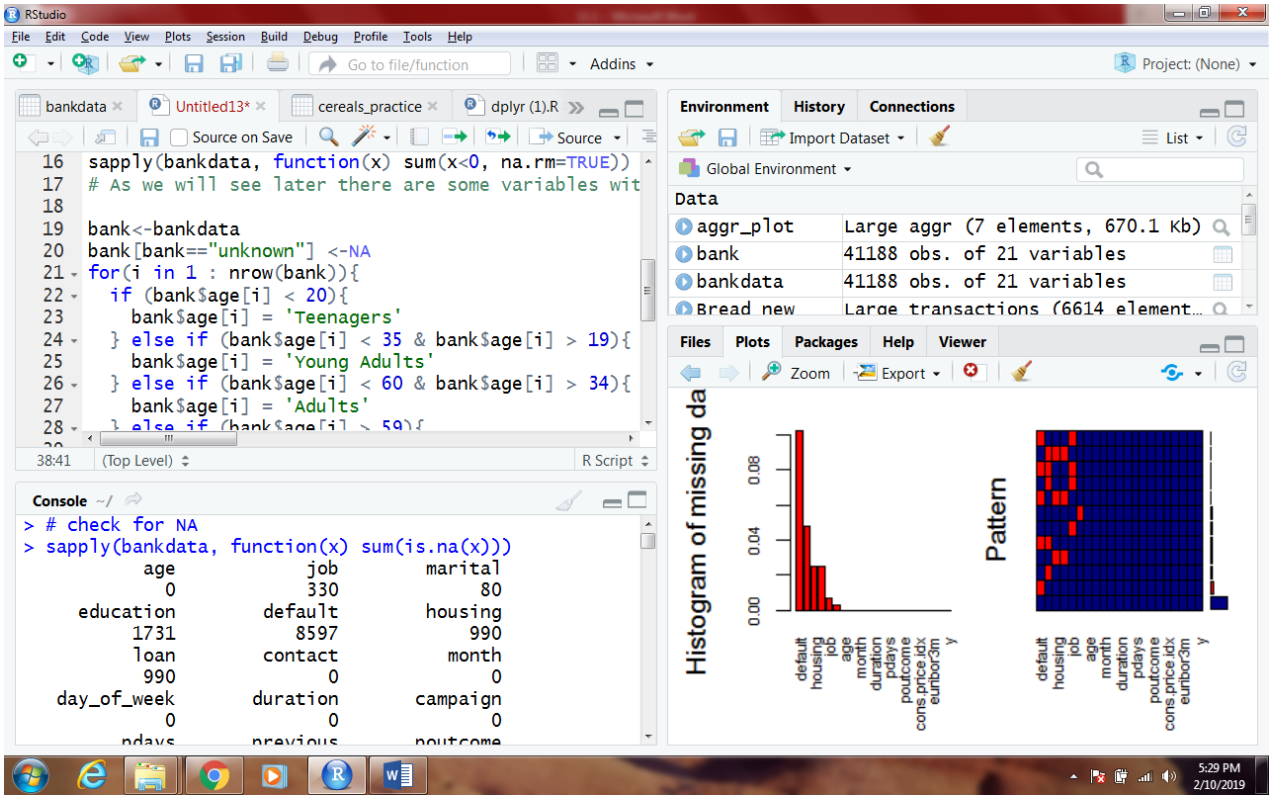
a. Create a visual for representing missing values in the dataset.
b. Show a distribution of clients based on a job.

c. Check whether is there any relation between Job and Marital Status?
d. Check whether is there any association between Job and Education?

```
A- library(readr)
B- bankdata <- read_delim("C:/Users/archana/Desktop/archana/GITS/bankdata.csv",
C-                        ";", escape_double = FALSE, trim_ws = TRUE)
D- View(bankdata)
E- dim(bankdata)
F- str(bankdata)
G- # check for NA
H- is.na(bankdata)
I- sum(is.na(bankdata))
J- sapply(bankdata, function(x) sum(is.na(x)))
K- library(readr)
L- library(dplyr)
M- library(reshape2)
N- library(ggplot2)
O- library(pander)
P- sapply(bankdata, function(x) sum(x<0, na.rm=TRUE))
Q- # As we will see later there are some variables with 'unknown' values (not equal to NAs).
   There are two variables with negative values by default
R-
S- bank<-bankdata
T- bank[bank=="unknown"] <-NA
U- for(i in 1 : nrow(bank)){
V-   if (bank$age[i] < 20){
W-     bank$age[i] = 'Teenagers'
X-   } else if (bank$age[i] < 35 & bank$age[i] > 19){
Y-     bank$age[i] = 'Young Adults'
Z-   } else if (bank$age[i] < 60 & bank$age[i] > 34){
AA-    bank$age[i] = 'Adults'
BB-  } else if (bank$age[i] > 59){
CC-    bank$age[i] = 'Senior Citizens'
```

```
DD-          }
EE-
FF- }
GG-          bank$age<-as.factor(bank$age)
HH-
II-   bank$y<-ifelse(bank$y =='yes', 1,0)
JJ-   bank$y<-as.factor(bank$y)
KK-
LL-  old_Cust_bank<-subset(bank, bank$poutcome != "nonexistent")
MM-          new_Cust_bank<-subset(bank, bank$poutcome == "nonexistent")
NN-          library(VIM)
OO-          library(ggplot2)
PP- aggr_plot <- aggr(old_Cust_bank, col=c('navyblue','red'), numbers=TRUE,
      sortVars=TRUE, labels=names(bank), cex.axis=.7, gap=3, ylab=c("Histogram of missing
      data","Pattern"))
```






```
Variables sorted by number of missings:
       Variable       Count
        default 0.102222222
      education 0.048000000
        housing 0.024711111
           loan 0.024711111
            job 0.006577778
        marital 0.003200000
            age 0.000000000
        contact 0.000000000
          month 0.000000000
    day_of_week 0.000000000
       duration 0.000000000
       campaign 0.000000000
          pdays 0.000000000
       previous 0.000000000
       poutcome 0.000000000
   emp.var.rate 0.000000000
cons.price.idx 0.000000000
 cons.conf.idx 0.000000000
```

```
      euribor3m 0.000000000
    nr.employed 0.000000000
              y 0.000000000
```

So for this dataset we face only the missing value problem, which can be solved by imputation method We will impute these missing values using MICE package

```
library(mice)
old_bank<-mice(old_Cust_bank)
```

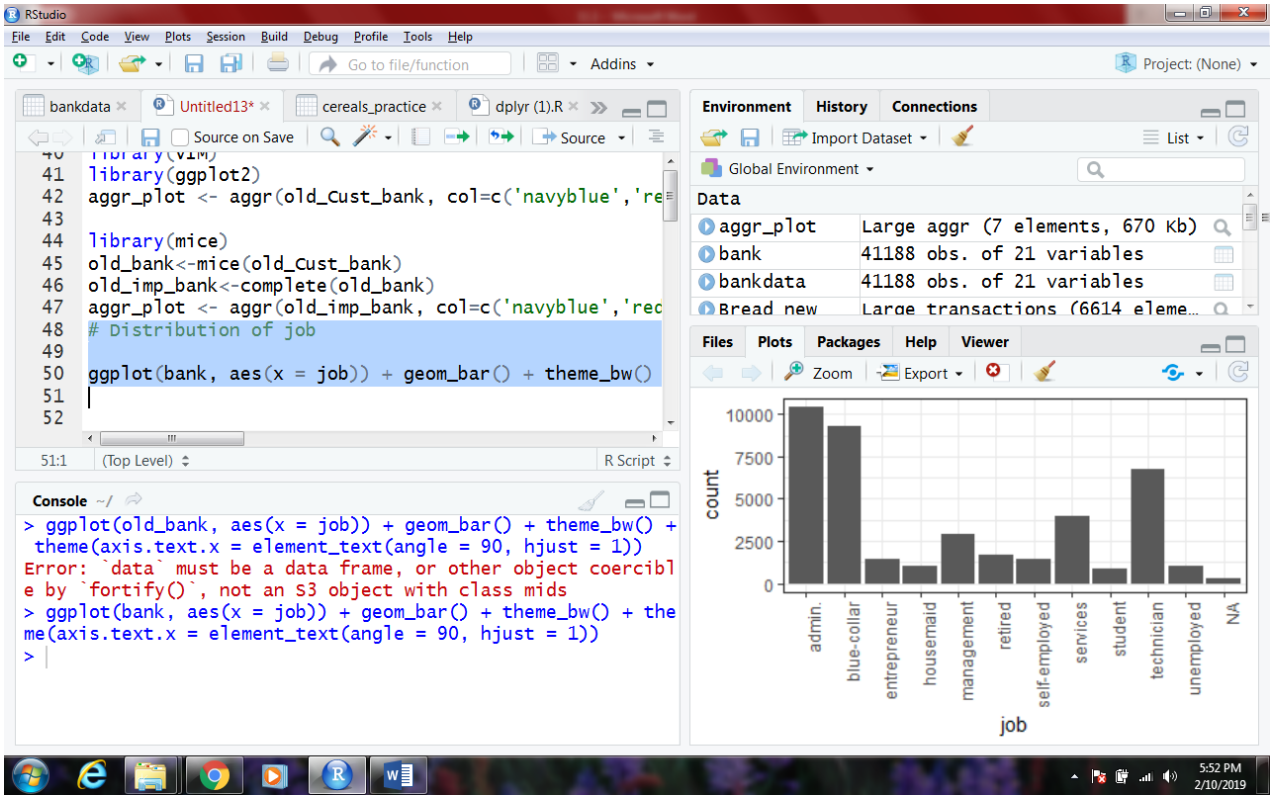
Check if missing values exists

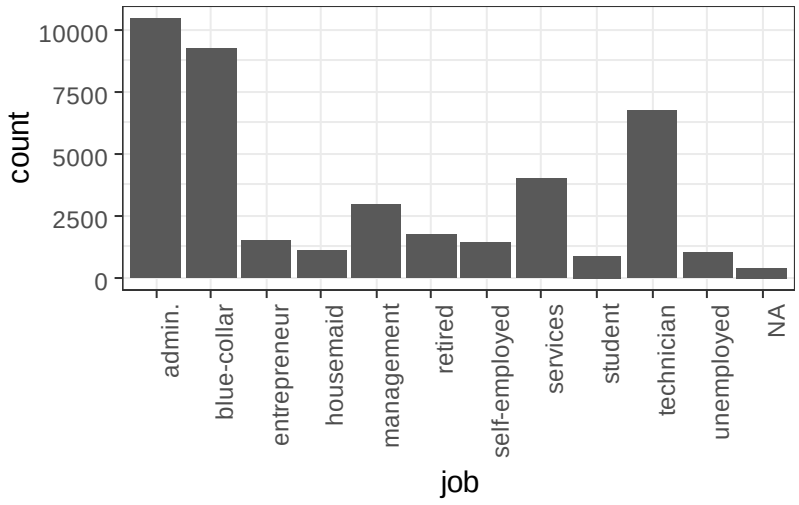
B- Show a distribution of clients based on a job

```
# Distribution of job

ggplot(bank, aes(x = job)) + geom_bar() + theme_bw() + theme(axis.text.x =
element_text(angle = 90, hjust = 1))
```

Job variable gives information on the job profile of targeted customers. There are 11 specified and 1 unspecified categories.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Project: (None)
bankdata
Untitled13*
cereals_practice
dplyr (1).R
Source on Save
Source
41 library(ggplot2)
42 aggr_plot <- aggr(old_Cust_bank, col=c('navyblue','re
44 library(mice)
45 old_bank<-mice(old_Cust_bank)
46 old_imp_bank<-complete(old_bank)
47 aggr_plot <- aggr(old_imp_bank, col=c('navyblue','rec
48 # Distribution of job
50 ggplot(bank, aes(x = job)) + geom_bar() + theme_bw()
51:1
(Top Level)
R Script
Console ~/
> ggplot(old_bank, aes(x = job)) + geom_bar() + theme_bw() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
Error: `data` must be a data frame, or other object coercibl
e by `fortify()`, not an S3 object with class mids
> ggplot(bank, aes(x = job)) + geom_bar() + theme_bw() + the
me(axis.text.x = element_text(angle = 90, hjust = 1))
Environment
History
Connections
Import Dataset
List
Global Environment
Data
aggr_plot
Large aggr (7 elements, 670 Kb)
bank
41188 obs. of 21 variables
bankdata
41188 obs. of 21 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export
10000
7500
5000
2500
0
count
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
NA
job
5:52 PM
2/10/2019


10000
7500
5000
2500
0
count
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
NA
job

c. Check whether is there any relation between Job and Marital Status?

```
library(MASS)
sum(is.na(bank$job))
sum(is.na(bank$marital))
bank1 <- bank %>% filter(marital != "NA")
bank.data<-data.frame(bank1$job,bank1$education)
bank.data= table(bank1$job,bank1$marital)
print(bank.data)
print(chisq.test(bank.data))
```

```
divorced married single
  admin.            1085    4372   3459
  blue-collar        507    4092   1332
  entrepreneur       152     785    186
  housemaid          100     530     91
  management         270    1678    455
  retired            279     942     67
  self-employed      114     659    332
  services           409    1621    936
  student              9      34    705
  technician         654    2965   2017
  unemployed          95     433    225
> print(chisq.test(bank.data))

        Pearson's Chi-squared test

data:  bank.data
X-squared = 2939.9, df = 20, p-value < 2.2e-16
```

The p value is less than 0.05 so there is high correlation between job and marital status

d.Check whether is there any association between Job and Education?

```
library(MASS)
sum(is.na(bank$job))
sum(is.na(bank$education))
```

```
sum(is.na(bank$marital))
bank1 <- bank %>% filter(education != "NA")
bank.data<-data.frame(bank1$job,bank1$education)
bank.data= table(bank1$job,bank1$education)
print(bank.data)
print(chisq.test(bank.data))
```

```
                university.degree
  admin.                     5021
  blue-collar                  69
  entrepreneur                478
  housemaid                   116
  management                 1787
  retired                     234
  self-employed               677
  services                    149
  student                     132
  technician                 1550
  unemployed                  225
>
```

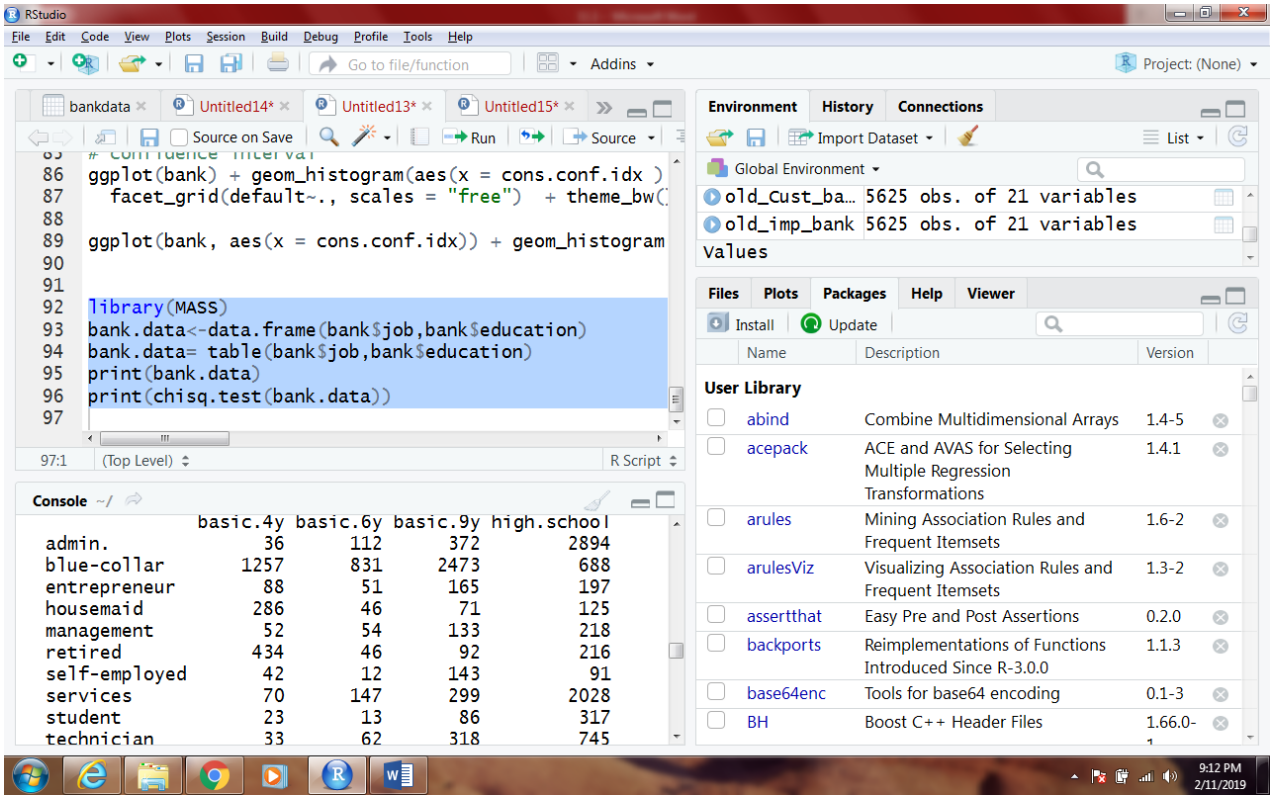
```
print(chisq.test(bank.data))

        Pearson's Chi-squared test

data:  bank.data

X-squared = 27184, df = 60, p-value < 2.2e-16
```

The p value is less than 0.05 so there is high correlation between job and education



6. Expected Format

1. R file should be submitted where applicable.
2. R file should be in PDF or in .r format
3. Proper screenshots of the outputs should be submitted as well
4. The r codes, if submitted in any other format, will be subjected to deduction in marks

Note: Your solution will not be entertained if it is any other format, e.g., .zip, .doc, .rtf etc.

7. Approximate Time to Complete Task

20 mins.