ACADGILD

SESSION 11: Linear Models

Assignment 1

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?

SOLUTION

a. Create a visual for representing missing values in the dataset.

The R-script for the given problem is as follows:

```
# Import Bank Marketing Data library(readr)
bank <- read.csv("E:/munmun_acadgild/acadgild data analytics/supporting files/bank-additional/bank-additional/bank-additional.csv", sep=";")
View(bank)
dim(bank)
str(bank)

  # a. Create a visual for representing missing values in the dataset.
library(psych)
psych::describe(bank)
library(VIM)
missing <- bank
missing[missing == "unknown"] <- NA

aggr(missing, col=c('blue', 'red'),
    numbers=TRUE, sortvars= TRUE,
    labels=names(missing), cex.axis=0.5,
    gap=3, ylab=c("missing data","pattern"))

sapply(missing, function(x) sum(is.na(x)))
```

The output of the R-Script (from Console window) is given as follows:

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Source

Console ~/

```
> bank <- read.csv("F:/ACADGILD - Online Course/1. DATA SETS/bankdata.csv", sep=";")
> View(bank)
> dim(bank)
[1] 41188    21
> str(bank)
'data.frame':   41188 obs. of  21 variables:
 $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
 $ job           : Factor w/ 12 levels "","admin.","blue-collar",..: 5 9 9 2 9 9 2 3 11 9 ...
 $ marital       : Factor w/ 4 levels "","divorced",..: 3 3 3 3 3 3 3 3 4 4 ...
 $ education     : Factor w/ 8 levels "","basic.4y",..: 2 5 5 3 5 4 7 1 7 5 ...
 $ default       : Factor w/ 3 levels "","no","yes": 2 1 2 2 2 1 2 1 2 2 ...
 $ housing       : Factor w/ 3 levels "","no","yes": 2 2 3 2 2 2 2 2 3 3 ...
 $ loan          : Factor w/ 3 levels "","no","yes": 2 2 2 2 3 2 2 2 2 2 ...
 $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
 $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
 $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
 $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ cons.price.idx: num  94 94 94 94 94 ...
 $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
 $ nr.employed   : num  5191 5191 5191 5191 5191 ...
 $ y             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
>
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

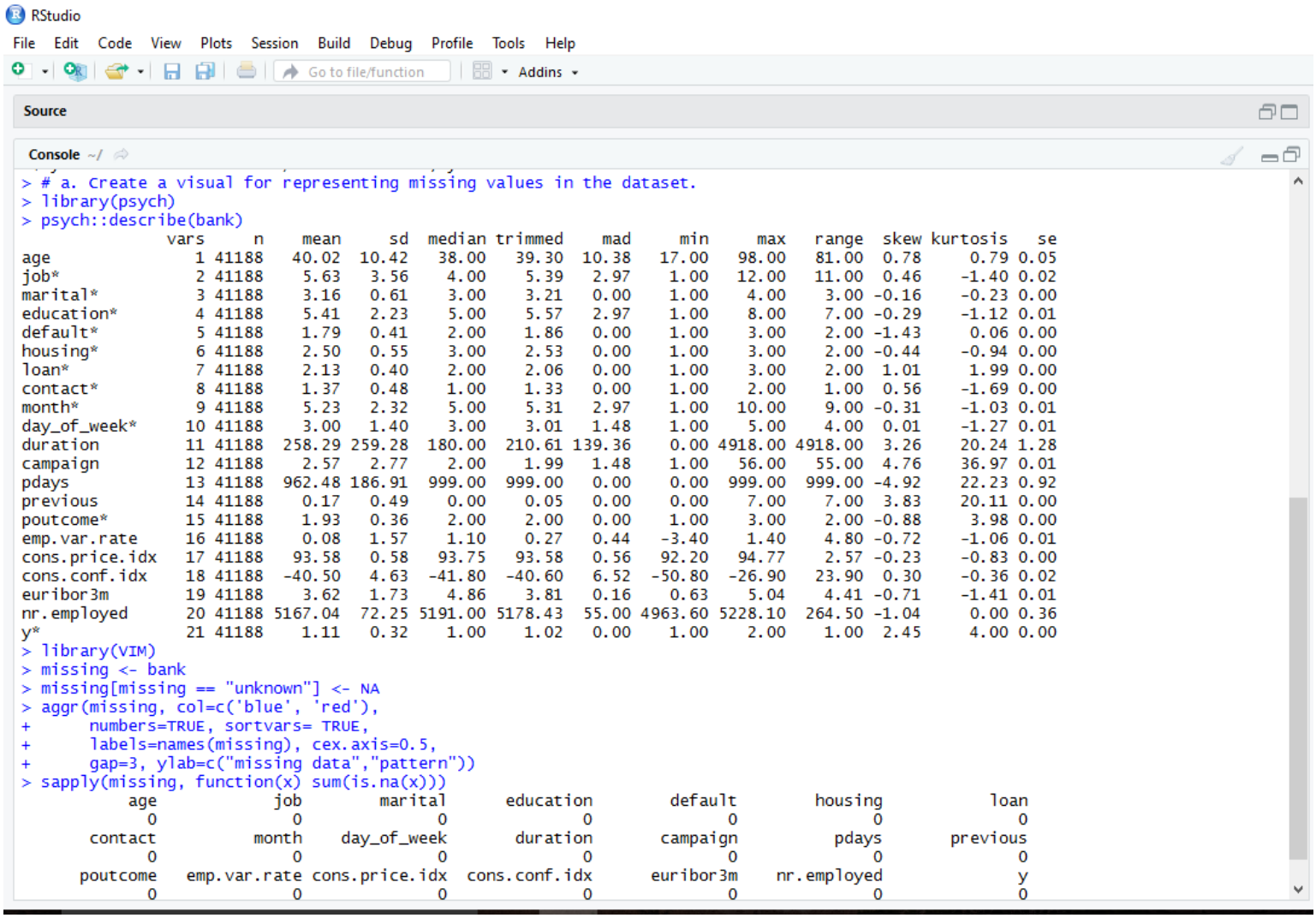
Assignment 11.1.R* | bank

Filter

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx
1	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.
2	57	services	married	high.school		no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.
3	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.
4	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.
5	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.
6	45	services	married	basic.9y		no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.
7	59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	1.1	93.
8	41	blue-collar	married			no	no	telephone	may	mon	217	1	999	0	nonexistent	1.1	93.
9	24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	1.1	93.
10	25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	1.1	93.
11	41	blue-collar	married			no	no	telephone	may	mon	55	1	999	0	nonexistent	1.1	93.
12	25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent	1.1	93.
13	29	blue-collar	single	high.school	no	no	yes	telephone	may	mon	137	1	999	0	nonexistent	1.1	93.
14	57	housemaid	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999	0	nonexistent	1.1	93.
15	35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999	0	nonexistent	1.1	93.
16	54	retired	married	basic.9y		yes	yes	telephone	may	mon	174	1	999	0	nonexistent	1.1	93.
17	35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999	0	nonexistent	1.1	93.
18	46	blue-collar	married	basic.6y		yes	yes	telephone	may	mon	440	1	999	0	nonexistent	1.1	93.
19	50	blue-collar	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999	0	nonexistent	1.1	93.
20	39	management	single	basic.9y		no	no	telephone	may	mon	195	1	999	0	nonexistent	1.1	93.
21	30	unemployed	married	high.school	no	no	no	telephone	may	mon	38	1	999	0	nonexistent	1.1	93.
22	55	blue-collar	married	basic.4y		yes	no	telephone	may	mon	262	1	999	0	nonexistent	1.1	93.

Showing 1 to 23 of 41,188 entries

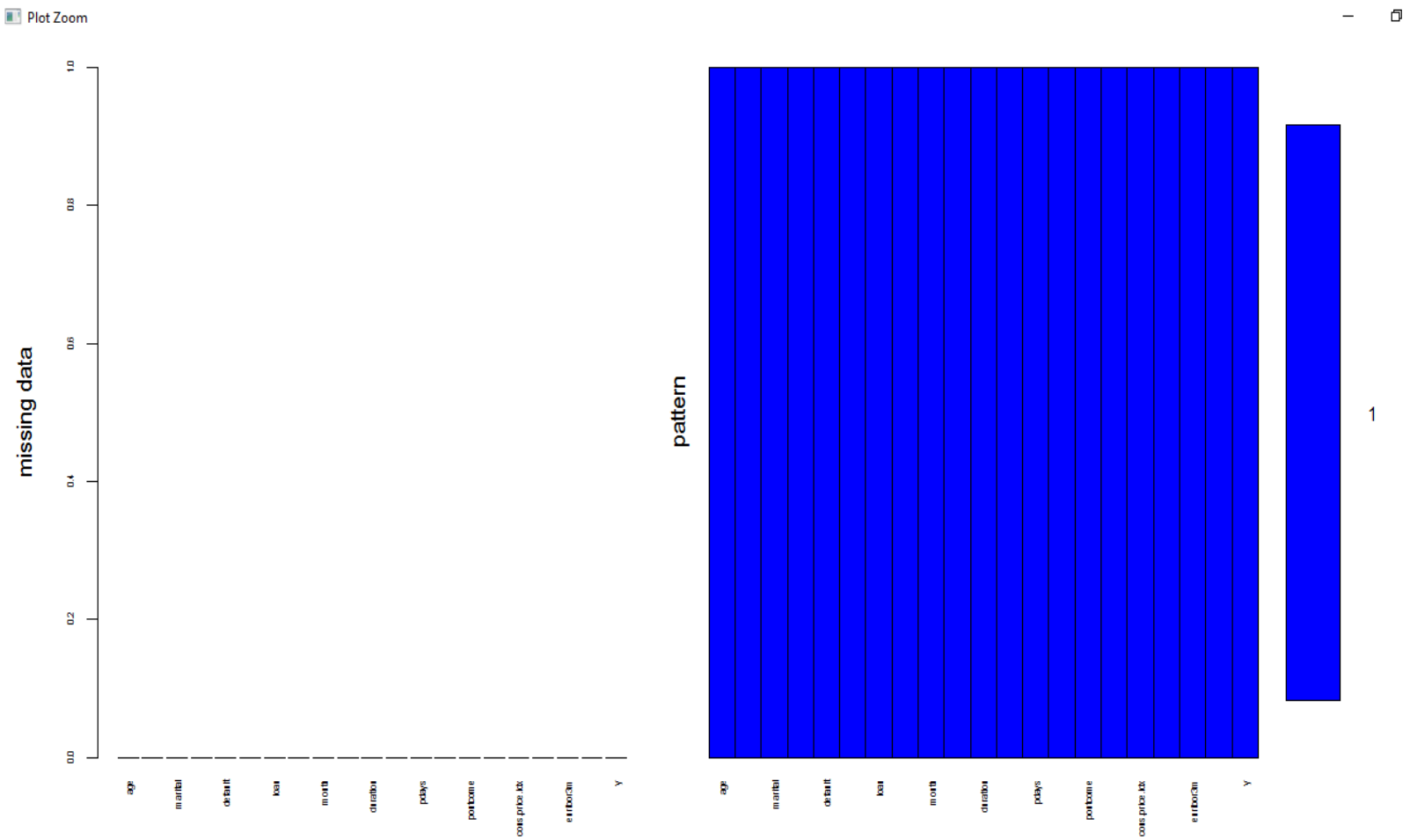
Console




```
> # a. Create a visual for representing missing values in the dataset.
> library(psych)
> psych::describe(bank)
              vars     n    mean     sd  median trimmed    mad     min     max   range  skew kurtosis   se
age              1 41188   40.02  10.42   38.00   39.30  10.38   17.00   98.00   81.00  0.78     0.79 0.05
job*             2 41188    5.63   3.56    4.00    5.39   2.97    1.00   12.00   11.00  0.46    -1.40 0.02
marital*         3 41188    3.16   0.61    3.00    3.21   0.00    1.00    4.00    3.00 -0.16    -0.23 0.00
education*       4 41188    5.41   2.23    5.00    5.57   2.97    1.00    8.00    7.00 -0.29    -1.12 0.01
default*         5 41188    1.79   0.41    2.00    1.86   0.00    1.00    3.00    2.00 -1.43     0.06 0.00
housing*         6 41188    2.50   0.55    3.00    2.53   0.00    1.00    3.00    2.00 -0.44    -0.94 0.00
loan*            7 41188    2.13   0.40    2.00    2.06   0.00    1.00    3.00    2.00  1.01     1.99 0.00
contact*         8 41188    1.37   0.48    1.00    1.33   0.00    1.00    2.00    1.00  0.56    -1.69 0.00
month*           9 41188    5.23   2.32    5.00    5.31   2.97    1.00   10.00    9.00 -0.31    -1.03 0.01
day_of_week*    10 41188    3.00   1.40    3.00    3.01   1.48    1.00    5.00    4.00  0.01    -1.27 0.01
duration        11 41188  258.29 259.28  180.00  210.61 139.36    0.00 4918.00 4918.00  3.26    20.24 1.28
campaign        12 41188    2.57   2.77    2.00    1.99   1.48    1.00   56.00   55.00  4.76    36.97 0.01
pdays           13 41188  962.48 186.91  999.00  999.00   0.00    0.00  999.00  999.00 -4.92    22.23 0.92
previous        14 41188    0.17   0.49    0.00    0.05   0.00    0.00    7.00    7.00  3.83    20.11 0.00
poutcome*       15 41188    1.93   0.36    2.00    2.00   0.00    1.00    3.00    2.00 -0.88     3.98 0.00
emp.var.rate    16 41188    0.08   1.57    1.10    0.27   0.44   -3.40    1.40    4.80 -0.72    -1.06 0.01
cons.price.idx  17 41188   93.58   0.58   93.75   93.58   0.56   92.20   94.77    2.57 -0.23    -0.83 0.00
cons.conf.idx   18 41188  -40.50   4.63  -41.80  -40.60   6.52  -50.80  -26.90   23.90  0.30    -0.36 0.02
euribor3m       19 41188    3.62   1.73    4.86    3.81   0.16    0.63    5.04    4.41 -0.71    -1.41 0.01
nr.employed     20 41188 5167.04  72.25 5191.00 5178.43  55.00 4963.60 5228.10  264.50 -1.04     0.00 0.36
y*              21 41188    1.11   0.32    1.00    1.02   0.00    1.00    2.00    1.00  2.45     4.00 0.00
> library(VIM)
> missing <- bank
> missing[missing == "unknown"] <- NA
> aggr(missing, col=c('blue', 'red'),
+      numbers=TRUE, sortvars= TRUE,
+      labels=names(missing), cex.axis=0.5,
+      gap=3, ylab=c("missing data","pattern"))
> sapply(missing, function(x) sum(is.na(x)))
           age            job        marital      education        default        housing           loan
             0              0              0              0              0              0              0
       contact          month    day_of_week       duration       campaign          pdays       previous
             0              0              0              0              0              0              0
     poutcome   emp.var.rate cons.price.idx  cons.conf.idx      euribor3m    nr.employed              y
             0              0              0              0              0              0              0
```

Conclusion/Interpretation:

Visual for representing missing values in the dataset is created



b. Show a distribution of clients based on a job.

The R-script for the given problem is as follows:

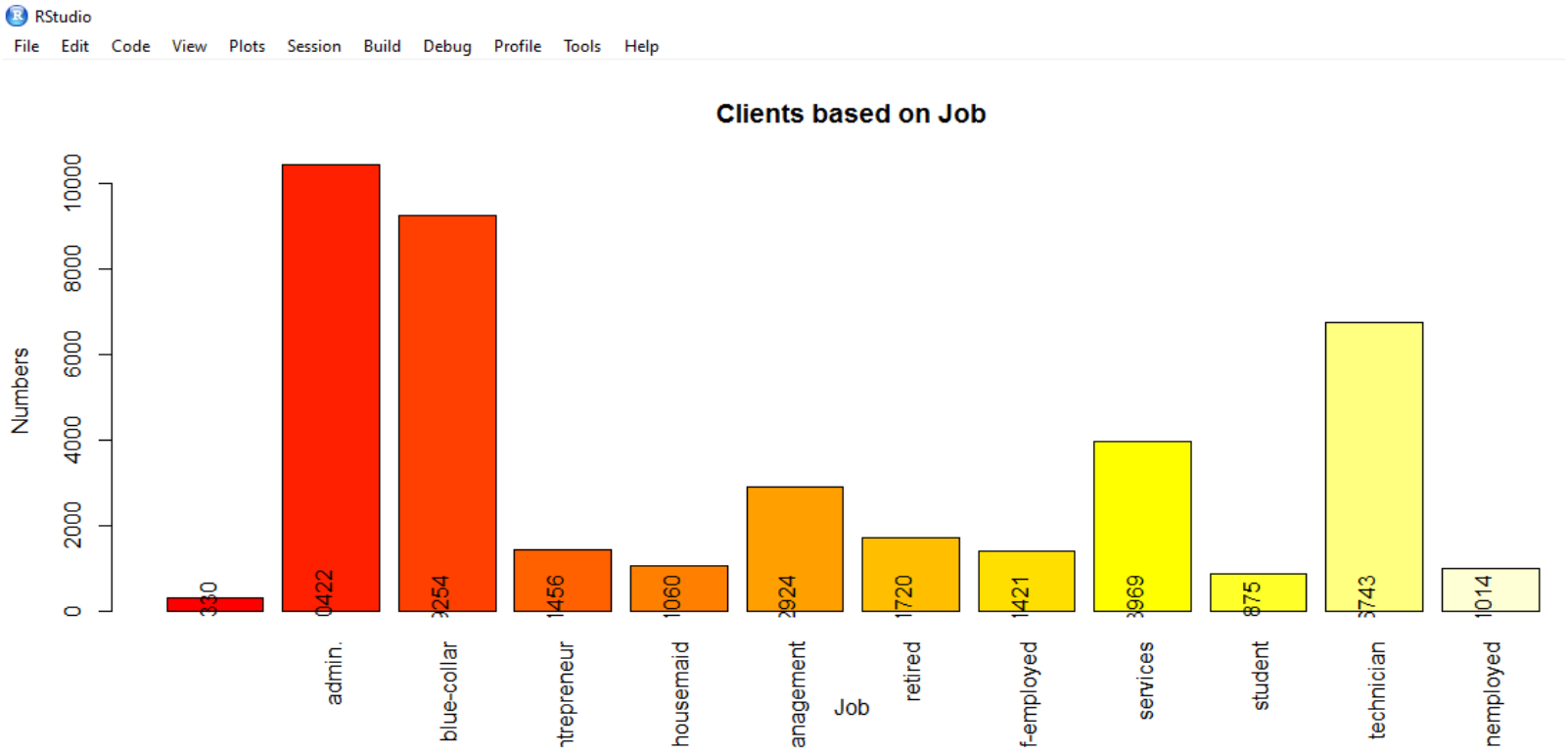
```
t <- table(bank$job)
  # distribution in tabular form
t
# distribution in graphical form
title <- barplot(t, xlab = "Job", ylab = "Numbers", main = "Clients based on Job", col =
            heat.colors(12), las=3)
text(title, 0, t, pos = 3, srt = 90)
```

The output of the R-Script (from Console window) is given as follows:

```
> t <- table(bank$job)
> # distribution in tabular form
> t

                    admin.    blue-collar  entrepreneur     housemaid    management
               330      10422         9254         1456         1060         2924
self-employed       services      student    technician   unemployed
              1421      3969          875          6743         1014
```

```
> # distribution in graphical form
> title <- barplot(t, xlab = "Job", ylab = "Numbers", main = "Clients based on Job",
+                  col = heat.colors(12), las=3)
> text(title, 0, t, pos = 3, srt = 90)
```



Conclusion/Interpretation:

Distribution of clients based on a job is obtained in tabular and graphical form.

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

The R-script for the given problem is as follows:

```
chisq.test(missing$job, missing$marital)
```

The output of the R-Script (from Console window) is given as follows:

```
> chisq.test(missing$job, missing$marital)

        Pearson's Chi-squared test

data:  missing$job and missing$marital
X-squared = 4197.5, df = 33, p-value < 2.2e-16
```

Conclusion/Interpretation:

Ho : There is NO association between Job and Marital Status

Since P Value is less than 0.05, there is association between Job and Marital status at 95% confidence level. Since NA values are very less, they are omitted.

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

The R-script for the given problem is as follows:

```
chisq.test(missing$job, missing$education)
```

The output of the R-Script (from Console window) is given as follows:

```
> chisq.test(missing$job, missing$education)

        Pearson's Chi-squared test

data:  missing$job and missing$education
X-squared = 37338, df = 77, p-value < 2.2e-16
```

Conclusion/Interpretation:

Ho : There is NO association between Job and Education.

Since the P value is less than 0.05, there is association between Job and Education at 95% confidence level. Since NA values are very less, they are omitted