



ACADGILD

SESSION 11: Linear Models

Assignment 1

Submitted by: Munmun Ghosal

Login Id: munmun55@gmail.com

(M):+91-8007178659

Data Analytics

Table of Contents

1. Problem Statement.....	3
2. Solution.....	3

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?

2. Solution

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

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

```
# Import BankMArketiing 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:

```

> 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)
[1] 4119    21
> str(bank)
'data.frame':   4119 obs. of  21 variables:
 $ age          : int  30 39 25 38 47 32 32 41 31 35 ...
 $ job          : Factor w/ 12 levels "admin.", "blue-collar",...: 2 8 8 8 1 8
1 3 8 2 ...
 $ marital      : Factor w/ 4 levels "divorced", "married",...: 2 3 2 2 2 3 3
2 1 2 ...
 $ education    : Factor w/ 8 levels "basic.4y", "basic.6y",...: 3 4 4 3 7 7 7
7 6 3 ...
 $ default      : Factor w/ 3 levels "no", "unknown",...: 1 1 1 1 1 1 1 2 1 2
...
 $ housing      : Factor w/ 3 levels "no", "unknown",...: 3 1 3 2 3 1 3 3 1 1
...
 $ loan         : Factor w/ 3 levels "no", "unknown",...: 1 1 1 2 1 1 1 1 1 1
...
 $ contact      : Factor w/ 2 levels "cellular", "telephone": 1 2 2 2 1 1 1 1 1
1 2 ...
 $ month        : Factor w/ 10 levels "apr", "aug", "dec",...: 7 7 5 5 8 10 10
8 8 7 ...
 $ day_of_week  : Factor w/ 5 levels "fri", "mon", "thu",...: 1 1 5 1 2 3 2 2 4
3 ...
 $ duration     : int  487 346 227 17 58 128 290 44 68 170 ...
 $ campaign     : int   2  4  1  3  1  3  4  2  1  1 ...
 $ pdays       : int  999 999 999 999 999 999 999 999 999 999 ...
 $ previous     : int   0  0  0  0  0  2  0  0  1  0 ...
 $ poutcome     : Factor w/ 3 levels "failure", "nonexistent",...: 2 2 2 2 2 2 1
2 2 1 2 ...
 $ emp.var.rate : num  -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...
 $ cons.price.idx: num  92.9 94 94.5 94.5 93.2 ...
 $ cons.conf.idx : num  -46.2 -36.4 -41.8 -41.8 -42 -37.5 -37.5 -42 -42 -36.4
...
 $ euribor3m    : num   1.31 4.86 4.96 4.96 4.19 ...
 $ nr.employed  : num  5099 5191 5228 5228 5196 ...
 $ y            : Factor w/ 2 levels "no", "yes": 1 1 1 1 1 1 1 1 1 1 ...

```

```

> library(psych)
> psych::describe(bank)

```

	vars	n	mean	sd	median	trimmed	mad	min
age	1	4119	40.11	10.31	38.00	39.44	10.38	18.00
job*	2	4119	4.82	3.61	4.00	4.60	4.45	1.00
marital*	3	4119	2.18	0.61	2.00	2.22	0.00	1.00
education*	4	4119	4.78	2.15	4.00	4.92	2.97	1.00
default*	5	4119	1.20	0.40	1.00	1.12	0.00	1.00
housing*	6	4119	2.08	0.98	3.00	2.10	0.00	1.00
loan*	7	4119	1.35	0.74	1.00	1.19	0.00	1.00
contact*	8	4119	1.36	0.48	1.00	1.32	0.00	1.00
month*	9	4119	5.29	2.31	5.00	5.37	2.97	1.00
day_of_week*	10	4119	3.01	1.39	3.00	3.01	1.48	1.00
duration	11	4119	256.79	254.70	181.00	210.52	136.40	0.00
campaign	12	4119	2.54	2.57	2.00	1.99	1.48	1.00
pdays	13	4119	960.42	191.92	999.00	999.00	0.00	0.00
previous	14	4119	0.19	0.54	0.00	0.06	0.00	0.00
poutcome*	15	4119	1.92	0.37	2.00	1.99	0.00	1.00
emp.var.rate	16	4119	0.08	1.56	1.10	0.27	0.44	-3.40
cons.price.idx	17	4119	93.58	0.58	93.75	93.58	0.56	92.20
cons.conf.idx	18	4119	-40.50	4.59	-41.80	-40.59	6.52	-50.80
euribor3m	19	4119	3.62	1.73	4.86	3.81	0.16	0.64
nr.employed	20	4119	5166.48	73.67	5191.00	5178.54	55.00	4963.60

```

y*          21 4119      1.11  0.31      1.00      1.01  0.00      1.00
      max range skew kurtosis se
age      88.00 70.00 0.72  0.43 0.16
job*     12.00 11.00 0.41 -1.42 0.06
marital*  4.00  3.00 -0.03 -0.29 0.01
education* 8.00  7.00 -0.28 -1.21 0.03
default*  3.00  2.00  1.55  0.43 0.01
housing*  3.00  2.00 -0.16 -1.95 0.02
loan*     3.00  2.00  1.72  1.03 0.01
contact*  2.00  1.00  0.60 -1.64 0.01
month*    10.00 9.00 -0.31 -1.02 0.04
day_of_week* 5.00  4.00  0.00 -1.26 0.02
duration 3643.00 3643.00 3.29 20.72 3.97
campaign  35.00 34.00  4.00 25.24 0.04
pdays    999.00 999.00 -4.77 20.77 2.99
previous   6.00  6.00  4.02 22.08 0.01
poutcome*  3.00  2.00 -0.84  3.55 0.01
emp.var.rate 1.40  4.80 -0.73 -1.04 0.02
cons.price.idx 94.77  2.57 -0.22 -0.82 0.01
cons.conf.idx -26.90 23.90  0.29 -0.32 0.07
euribor3m    5.04  4.41 -0.71 -1.40 0.03
nr.employed 5228.10 264.50 -1.08  0.06 1.15
y*          2.00  1.00  2.50  4.25 0.00

```

```
> library(VIM)
```

```
Loading required package: colorspace
```

```
Loading required package: grid
```

```
Loading required package: data.table
```

```
data.table 1.11.8 Latest news: r-datatable.com
```

```
VIM is ready to use.
```

```
Since version 4.0.0 the GUI is in its own package VIMGUI.
```

Please use the package to use the new (and old) GUI.

Suggestions and bug-reports can be submitted at:

<https://github.com/alexxkova/VIM/issues>

Attaching package: 'VIM'

The following object is masked from 'package:datasets':

sleep

Warning message:

package 'VIM' was built under R version 3.5.2

```
>
```

```
> 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"))
```

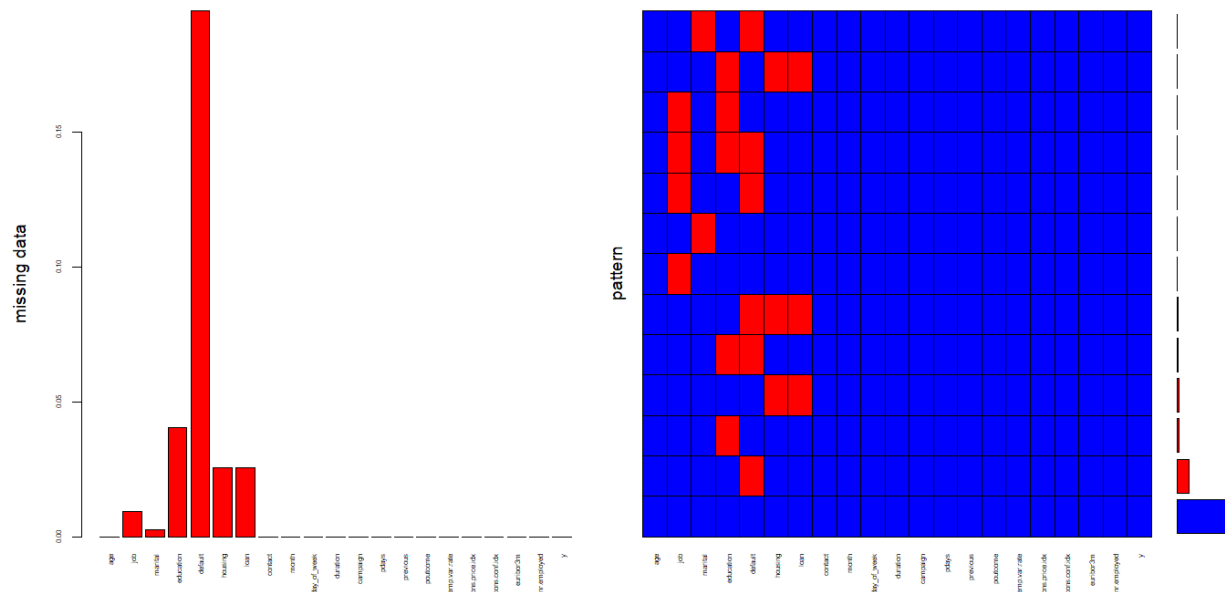
Warning message:

In plot.aggr(res, ...) : not enough horizontal space to display frequencies

```
>
```

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

age	job	marital	education	default
0	39	11	167	803
housing	loan	contact	month	day_of_week
105	105	0	0	0
duration	campaign	pdays	previous	poutcome
0	0	0	0	0
emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
0	0	0	0	0
y				
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:

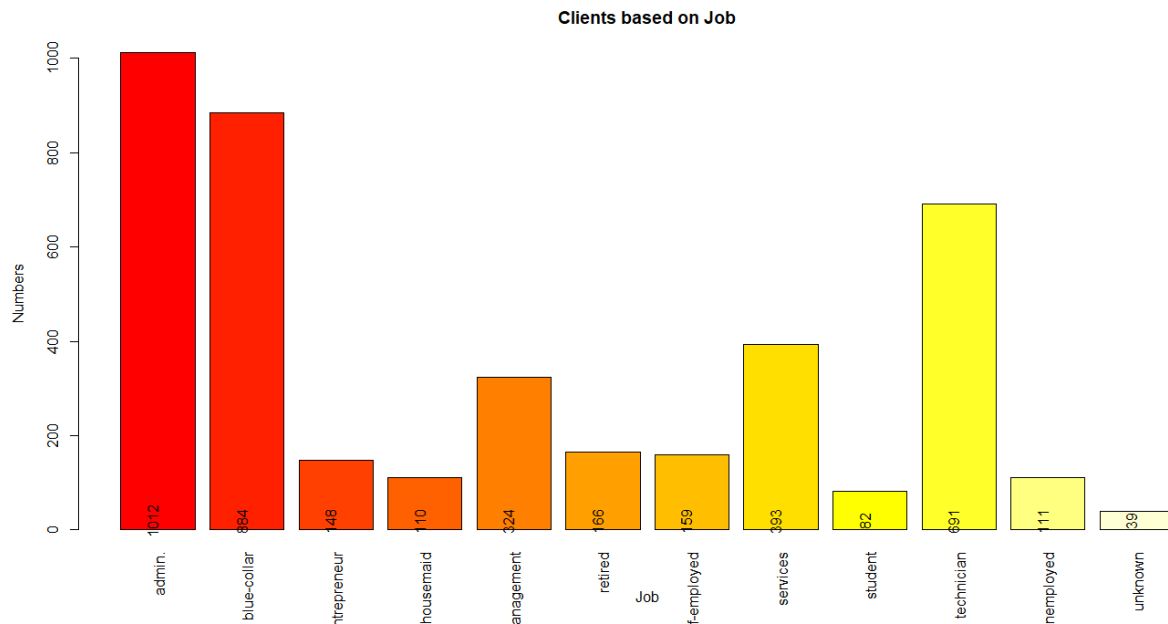
```
t <- table(bank$job)
# distribution in tabular form
t
# distribution in graphical form
title <- barplot(t, xlab = "Job", ylab = "Numbers", main = "Clients based on Job",
  col = heat.colors(12), las=3)
text(title, 0, t, pos = 3, srt = 90)
```

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
1012	884	148	110	324
retired	self-employed	services	student	technician
166	159	393	82	691
unemployed	unknown			
111	39			

```
> 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 = 395.28, df = 20, 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 = 3429, df = 60, 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