Assignment No. 11.1

This assignment is based on the dataset named bank, which has the following characteristics

> str(bank)

'data.frame': 4521 obs. of 17 variables:

$ age : int 30 33 35 30 59 35 36 39 41 43 ...

$ job : chr "unemployed" "services" "management" "management" ...

$ marital : chr "married" "married" "single" "married" ...

$ education: chr "primary" "secondary" "tertiary" "tertiary" ...

$ default : chr "no" "no" "no" "no" ...

$ balance : int 1787 4789 1350 1476 0 747 307 147 221 -88 ...

$ housing : chr "no" "yes" "yes" "yes" ...

$ loan : chr "no" "yes" "no" "yes" ...

$ contact : chr "cellular" "cellular" "cellular" "unknown" ...

$ day : int 19 11 16 3 5 23 14 6 14 17 ...

$ month : chr "oct" "may" "apr" "jun" ...

$ duration : int 79 220 185 199 226 141 341 151 57 313 ...

$ campaign : int 1 1 1 4 1 2 1 2 2 1 ...

$ pdays : int -1 339 330 -1 -1 176 330 -1 -1 147 ...

$ previous : int 0 4 1 0 0 3 2 0 0 2 ...

$ poutcome : chr "unknown" "failure" "failure" "unknown" ...

$ y : chr "no" "no" "no" "no" ...

The following is the data definitions of each of the variables

Input variables:  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

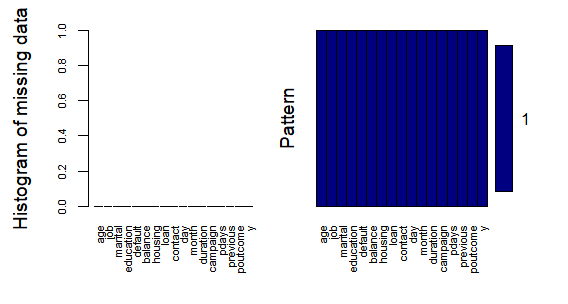
The following questions were asked to be performed in the assignment

1. Create a visual for representing missing values in the dataset.
2. Show a distribution of clients based on Job
3. Check whether there is any relationship between Job and Marital Status
4. Check whether there is any association between Job and Association

MISSING DATA VISUALISATION

> library(VIM)

> missdataplot <- aggr(bank, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))



> library(mice)

> md.pattern(bank)

/\ /\

{ `---' }

{ O O }

==> V <== No need for mice. This data set is completely observed.

\ \|/ /

`-----'

age job marital education default balance housing loan contact day

4521 1 1 1 1 1 1 1 1 1 1

0 0 0 0 0 0 0 0 0 0

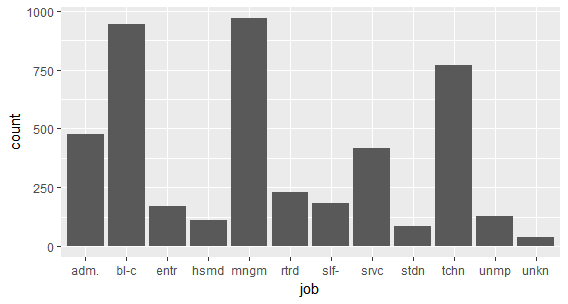
month duration campaign pdays previous poutcome y

4521 1 1 1 1 1 1 1 0

0 0 0 0 0 0 0 0

VISUALISATION OF CLIENTS BASED ON THEIR JOB

|  |
| --- |
| > library(ggplot2)  > ggplot(bank,aes(job)) + geom\_bar() + scale\_x\_discrete(labels = abbreviate) |
|  |
| |  | | --- | |  | |



RELATIONSHIP BETWEEN JOB AND MARITAL STATUS

> JobMarry <- table(bank$job, bank$marital)

> addmargins(JobMarry)

divorced married single Sum

admin. 69 266 143 478

blue-collar 79 693 174 946

entrepreneur 16 132 20 168

housemaid 13 84 15 112

management 119 557 293 969

retired 43 176 11 230

self-employed 15 127 41 183

services 62 236 119 417

student 0 10 74 84

technician 89 411 268 768

unemployed 22 75 31 128

unknown 1 30 7 38

Sum 528 2797 1196 4521

> chisq.test(JobMarry)

Pearson's Chi-squared test

data: JobMarry

X-squared = 373.18, df = 22, p-value < 2.2e-16

Warning message:

In chisq.test(JobMarry) : Chi-squared approximation may be incorrect

The above result REJECTS THE null hypothesis “There is no relationship between job and marital status”. The p-value is very less than 0, so the ALTERNATE HYPOTHESIS MAY BE TRUE that THERE IS RELATIONSHIP BETWEEN JOB AND MARITAL STATUS. The incorrectness of Hypothesis comes due to Student and Unknown job status rows, which less than 5 observations in the column have divorced. Chi Square test requires that atleast 5 observations should be recorded under each row column pivot in the expected contingency table, and we are assuming that the observed table would be nearer to the expected.

We correct the data by removing the student and unknown rows and then again test for association between job and marital status. The following is the result

> JobMarry2 <- JobMarry[-c(9,12),]

> JobMarry2

divorced married single

admin. 69 266 143

blue-collar 79 693 174

entrepreneur 16 132 20

housemaid 13 84 15

management 119 557 293

retired 43 176 11

self-employed 15 127 41

services 62 236 119

technician 89 411 268

unemployed 22 75 31

> chisq.test(JobMarry2)

Pearson's Chi-squared test

data: JobMarry2

X-squared = 204.02, df = 18, p-value < 2.2e-16

The warning message is gone and the P-value is quite less than 0 and it REJECTS the Null Hypothesis that “there is no significant association between job type and marital status.

ASSOCIATION BETWEEN JOB AND EDUCATION

|  |
| --- |
| > JobEdu <- table(bank$job, bank$education)  > addmargins(JobEdu)    primary secondary tertiary unknown Sum  admin. 17 393 51 17 478  blue-collar 369 524 12 41 946  entrepreneur 26 58 73 11 168  housemaid 57 28 22 5 112  management 39 116 787 27 969  retired 80 105 31 14 230  self-employed 15 76 88 4 183  services 25 363 16 13 417  student 2 47 19 16 84  technician 15 520 211 22 768  unemployed 26 68 32 2 128  unknown 7 8 8 15 38  Sum 678 2306 1350 187 4521    > chisq.test(JobEdu)  Pearson's Chi-squared test  data: JobEdu  X-squared = 2840, df = 33, p-value < 2.2e-16  Warning message:  In chisq.test(JobEdu) : Chi-squared approximation may be incorrect |
|  |
| |  | | --- | |  | |

The above result REJECTS THE null hypothesis “There is no relationship between job and education”. The p-value is very less than 0, so the ALTERNATE HYPOTHESIS MAY BE TRUE that THERE IS RELATIONSHIP BETWEEN JOB AND EDUCATION. The incorrectness of Hypothesis comes due to student, self-employed, housemaid, and unemployed, and unknown, job status rows, which less than 5 observations in the column have divorced. Chi Square test requires that atleast 5 observations should be recorded under each row column pivot in the expected contingency table, and we are assuming that the observed table would be nearer to the expected.

We correct the data by removing the student and unknown rows and then again test for association between job and marital status. The following is the result

> JobEdu2 <- JobEdu[-c(4,7,9,11,12),]

> JobEdu2

primary secondary tertiary unknown

admin. 17 393 51 17

blue-collar 369 524 12 41

entrepreneur 26 58 73 11

management 39 116 787 27

retired 80 105 31 14

services 25 363 16 13

technician 15 520 211 22

> chisq.test(JobEdu2)

Pearson's Chi-squared test

data: JobEdu2

X-squared = 2529.1, df = 18, p-value < 2.2e-16

The above result again REJECTS THE NULL hypothesis that “There is no significant association between job status and educational background, due to the very low p-value, hence the ALTERNATE hypothesis MAY BE TRUE, that there is significant association between educational background and job type.