Session 11 – Linear Models

Assignment – 1

1. Use the given link and locate the bank marketing dataset. Data Set Link

Perform the below operations:

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

View(bank)

> View(bank)

> dim(bank) # 4119 observations with 21 attributes

[1] 4521 17

> str(bank)

Classes ‘tbl\_df’, ‘tbl’ and '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" ...

- attr(\*, "spec")=List of 2

..$ cols :List of 17

.. ..$ age : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ job : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ marital : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ education: list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ default : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ balance : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ housing : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ loan : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ contact : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ day : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ month : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ duration : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ campaign : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ pdays : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ previous : list()

.. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"

.. ..$ poutcome : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

.. ..$ y : list()

.. .. ..- attr(\*, "class")= chr "collector\_character" "collector"

..$ default: list()

.. ..- attr(\*, "class")= chr "collector\_guess" "collector"

..- attr(\*, "class")= chr "col\_spec"

> install.packages("psych")

library(psych)

> describe(bank)

vars n mean sd median trimmed mad min max range skew kurtosis

age 1 4521 41.17 10.58 39 40.48 10.38 19 87 68 0.70 0.35

job\* 2 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

marital\* 3 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

education\* 4 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

default\* 5 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

balance 6 4521 1422.66 3009.64 444 802.41 658.27 -3313 71188 74501 6.59 88.25

housing\* 7 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

loan\* 8 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

contact\* 9 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

day 10 4521 15.92 8.25 16 15.80 10.38 1 31 30 0.09 -1.04

month\* 11 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

duration 12 4521 263.96 259.86 185 216.44 143.81 4 3025 3021 2.77 12.51

campaign 13 4521 2.79 3.11 2 2.14 1.48 1 50 49 4.74 37.11

pdays 14 4521 39.77 100.12 -1 11.56 0.00 -1 871 872 2.72 7.94

previous 15 4521 0.54 1.69 0 0.12 0.00 0 25 25 5.87 51.91

poutcome\* 16 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

y\* 17 4521 NaN NA NA NaN NA Inf -Inf -Inf NA NA

se

age 0.16

job\* NA

marital\* NA

education\* NA

default\* NA

balance 44.76

housing\* NA

loan\* NA

contact\* NA

day 0.12

month\* NA

duration 3.86

campaign 0.05

pdays 1.49

previous 0.03

poutcome\* NA

y\* NA

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"))

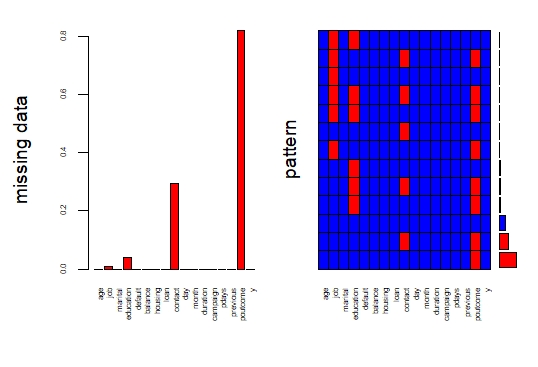
> sapply(missing, function(x) sum(is.na(x)))

age job marital education default balance housing loan contact

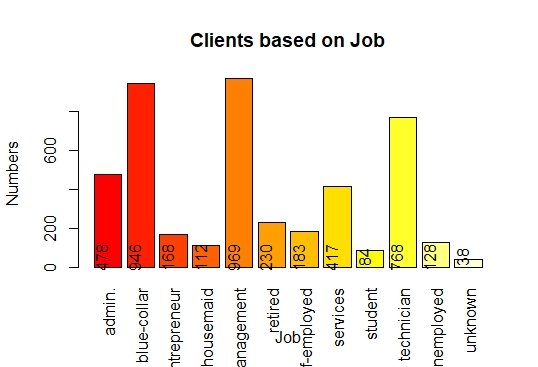
0 38 0 187 0 0 0 0 1324

day month duration campaign pdays previous poutcome y

0 0 0 0 0 0 3705 0



1. Show a distribution of clients based on a Job.
2. t <- table(bank$job)
3. > # distribution in tabular form
4. > t
5. admin. blue-collar entrepreneur housemaid management retired
6. 478 946 168 112 969 230
7. self-employed services student technician unemployed unknown
8. 183 417 84 768 128 38
9. > # distribution in graphical form
10. > title <- barplot(t, xlab = "Job", ylab = "Numbers", main = "Clients based on Job",
11. + col = heat.colors(12), las=3)
12. > text(title, 0, t, pos = 3, srt = 90)



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

Ho : There is NO association between Job and Marital Status

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

# 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, are omitted

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

Pearson's Chi-squared test

data: missing$job and missing$marital

X-squared = 367.09, df = 20, p-value < 2.2e-16

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

Ho : There is NO association between Job and Education.

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

Pearson's Chi-squared test

data: missing$job and missing$education

X-squared = 2638.8, df = 20, p-value < 2.2e-16

# 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, are omitted