Assignment - 1

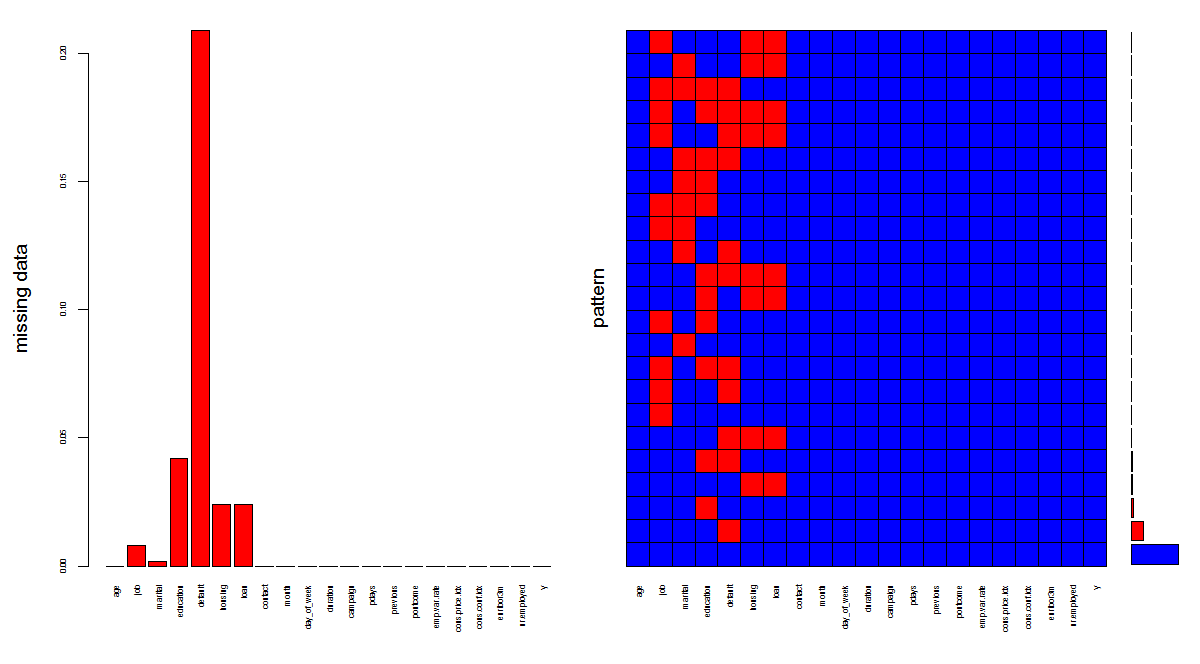
Session 11 – Linear Models

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

Perform the below operations:

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

|  |
| --- |
| > bank.additional.full <- read.csv("E:/ACADGILD-DATA ANALYTICS/bank-additional-full.csv", sep=";")  > View(bank.additional.full)  > View(bank.additional.full)  > dim(bank.additional.full) # 4119 observations with 21 attributes  [1] 41188 21  > str(bank.additional.full) # All have the correct class  '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",..: 4 8 8 1 8 8 1 2 10 8 ...  $ marital : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...  $ education : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...  $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...  $ housing : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...  $ loan : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...  $ 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 ...  > psych::describe(bank.additional.full)  vars n mean sd median trimmed mad min max range skew  age 1 41188 40.02 10.42 38.00 39.30 10.38 17.00 98.00 81.00 0.78  job\* 2 41188 4.72 3.59 3.00 4.48 2.97 1.00 12.00 11.00 0.45  marital\* 3 41188 2.17 0.61 2.00 2.21 0.00 1.00 4.00 3.00 -0.06  education\* 4 41188 4.75 2.14 4.00 4.88 2.97 1.00 8.00 7.00 -0.24  default\* 5 41188 1.21 0.41 1.00 1.14 0.00 1.00 3.00 2.00 1.44  housing\* 6 41188 2.07 0.99 3.00 2.09 0.00 1.00 3.00 2.00 -0.14  loan\* 7 41188 1.33 0.72 1.00 1.16 0.00 1.00 3.00 2.00 1.82  contact\* 8 41188 1.37 0.48 1.00 1.33 0.00 1.00 2.00 1.00 0.56  month\* 9 41188 5.23 2.32 5.00 5.31 2.97 1.00 10.00 9.00 -0.31  day\_of\_week\* 10 41188 3.00 1.40 3.00 3.01 1.48 1.00 5.00 4.00 0.01  duration 11 41188 258.29 259.28 180.00 210.61 139.36 0.00 4918.00 4918.00 3.26  campaign 12 41188 2.57 2.77 2.00 1.99 1.48 1.00 56.00 55.00 4.76  pdays 13 41188 962.48 186.91 999.00 999.00 0.00 0.00 999.00 999.00 -4.92  previous 14 41188 0.17 0.49 0.00 0.05 0.00 0.00 7.00 7.00 3.83  poutcome\* 15 41188 1.93 0.36 2.00 2.00 0.00 1.00 3.00 2.00 -0.88  emp.var.rate 16 41188 0.08 1.57 1.10 0.27 0.44 -3.40 1.40 4.80 -0.72  cons.price.idx 17 41188 93.58 0.58 93.75 93.58 0.56 92.20 94.77 2.57 -0.23  cons.conf.idx 18 41188 -40.50 4.63 -41.80 -40.60 6.52 -50.80 -26.90 23.90 0.30  euribor3m 19 41188 3.62 1.73 4.86 3.81 0.16 0.63 5.04 4.41 -0.71  nr.employed 20 41188 5167.04 72.25 5191.00 5178.43 55.00 4963.60 5228.10 264.50 -1.04  y\* 21 41188 1.11 0.32 1.00 1.02 0.00 1.00 2.00 1.00 2.45  kurtosis se  age 0.79 0.05  job\* -1.39 0.02  marital\* -0.34 0.00  education\* -1.21 0.01  default\* 0.07 0.00  housing\* -1.95 0.00  loan\* 1.38 0.00  contact\* -1.69 0.00  month\* -1.03 0.01  day\_of\_week\* -1.27 0.01  duration 20.24 1.28  campaign 36.97 0.01  pdays 22.23 0.92  previous 20.11 0.00  poutcome\* 3.98 0.00  emp.var.rate -1.06 0.01  cons.price.idx -0.83 0.00  cons.conf.idx -0.36 0.02  euribor3m -1.41 0.01  nr.employed 0.00 0.36  y\* 4.00 0.00  > library(VIM)  > missing <- bank.additional.full  > 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  0 330 80 1731 8597 990  loan contact month day\_of\_week duration campaign  990 0 0 0 0 0  pdays previous poutcome emp.var.rate cons.price.idx cons.conf.idx  0 0 0 0 0 0  euribor3m nr.employed y  0 0 0 |
|  |
| |  | | --- | | > | |



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

> t <- table(bank.additional.full$job)

> # distribution in tabular form

> t

admin. blue-collar entrepreneur housemaid management retired

10422 9254 1456 1060 2924 1720

self-employed services student technician unemployed unknown

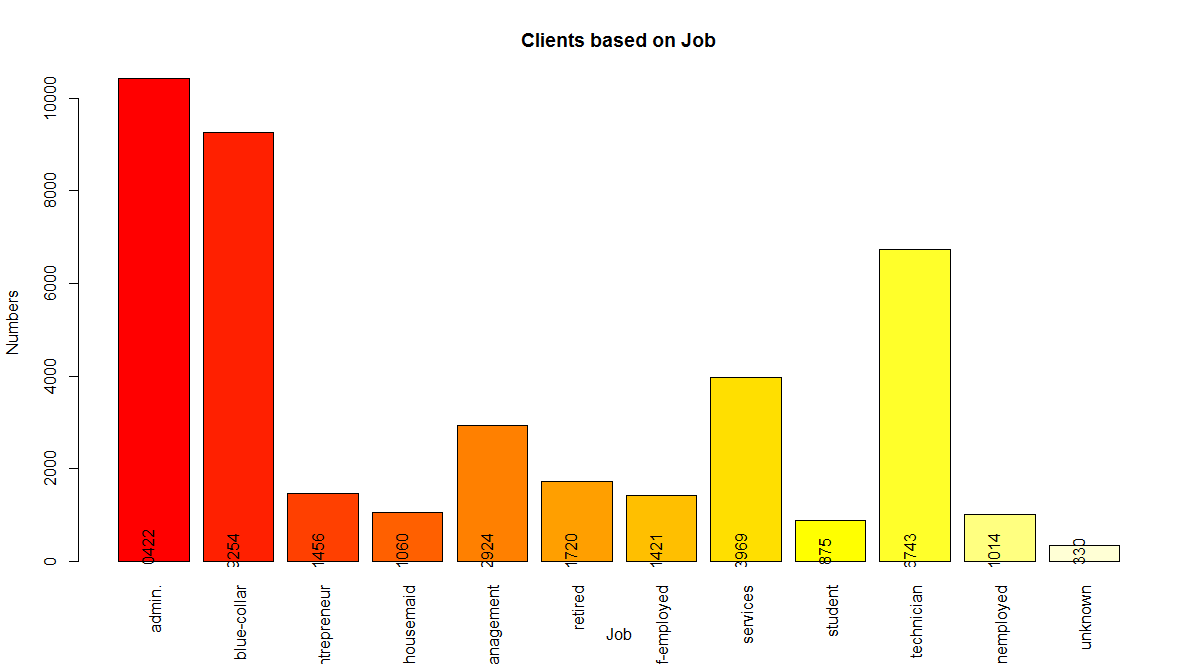
1421 3969 875 6743 1014 330

> # 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)



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

|  |
| --- |
| > chisq.test(missing$job, missing$marital)  Pearson's Chi-squared test  data: missing$job and missing$marital  X-squared = 4045.1, df = 20, p-value < 2.2e-16  > # there is association between Job and Marital status at 95% confidence level  > # Since NA values are very less, are omitted |
|  |
| |  | | --- | | > | |

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

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

Pearson's Chi-squared test

data: missing$job and missing$education

X-squared = 35560, df = 60, p-value < 2.2e-16