Bank Marketting\_univariate Analysis

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Importing the dataset

bank <- read.csv("bank.csv",header = TRUE, sep = ",")

summary(bank)

## age job marital education   
## Min. :19.00 management :969 divorced: 528 primary : 678   
## 1st Qu.:33.00 blue-collar:946 married :2797 secondary:2306   
## Median :39.00 technician :768 single :1196 tertiary :1350   
## Mean :41.17 admin. :478 unknown : 187   
## 3rd Qu.:49.00 services :417   
## Max. :87.00 retired :230   
## (Other) :713   
## default balance housing loan contact   
## no :4445 Min. :-3313 no :1962 no :3830 cellular :2896   
## yes: 76 1st Qu.: 69 yes:2559 yes: 691 telephone: 301   
## Median : 444 unknown :1324   
## Mean : 1423   
## 3rd Qu.: 1480   
## Max. :71188   
##   
## day month duration campaign   
## Min. : 1.00 may :1398 Min. : 4 Min. : 1.000   
## 1st Qu.: 9.00 jul : 706 1st Qu.: 104 1st Qu.: 1.000   
## Median :16.00 aug : 633 Median : 185 Median : 2.000   
## Mean :15.92 jun : 531 Mean : 264 Mean : 2.794   
## 3rd Qu.:21.00 nov : 389 3rd Qu.: 329 3rd Qu.: 3.000   
## Max. :31.00 apr : 293 Max. :3025 Max. :50.000   
## (Other): 571   
## pdays previous poutcome y   
## Min. : -1.00 Min. : 0.0000 failure: 490 no :4000   
## 1st Qu.: -1.00 1st Qu.: 0.0000 other : 197 yes: 521   
## Median : -1.00 Median : 0.0000 success: 129   
## Mean : 39.77 Mean : 0.5426 unknown:3705   
## 3rd Qu.: -1.00 3rd Qu.: 0.0000   
## Max. :871.00 Max. :25.0000   
##

str(bank)

## 'data.frame': 4521 obs. of 17 variables:  
## $ age : int 30 33 35 30 59 35 36 39 41 43 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 11 8 5 5 2 5 7 10 3 8 ...  
## $ marital : Factor w/ 3 levels "divorced","married",..: 2 2 3 2 2 3 2 2 2 2 ...  
## $ education: Factor w/ 4 levels "primary","secondary",..: 1 2 3 3 2 3 3 2 3 1 ...  
## $ default : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ balance : int 1787 4789 1350 1476 0 747 307 147 221 -88 ...  
## $ housing : Factor w/ 2 levels "no","yes": 1 2 2 2 2 1 2 2 2 2 ...  
## $ loan : Factor w/ 2 levels "no","yes": 1 2 1 2 1 1 1 1 1 2 ...  
## $ contact : Factor w/ 3 levels "cellular","telephone",..: 1 1 1 3 3 1 1 1 3 1 ...  
## $ day : int 19 11 16 3 5 23 14 6 14 17 ...  
## $ month : Factor w/ 12 levels "apr","aug","dec",..: 11 9 1 7 9 4 9 9 9 1 ...  
## $ 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 : Factor w/ 4 levels "failure","other",..: 4 1 1 4 4 1 2 4 4 1 ...  
## $ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

Here we have 4521 observations and 17 variables in the data set.

Let us check for NA’s and blank spaces

colSums(is.na(bank))

## age job marital education default balance housing   
## 0 0 0 0 0 0 0   
## loan contact day month duration campaign pdays   
## 0 0 0 0 0 0 0   
## previous poutcome y   
## 0 0 0

colSums(bank =="")

## age job marital education default balance housing   
## 0 0 0 0 0 0 0   
## loan contact day month duration campaign pdays   
## 0 0 0 0 0 0 0   
## previous poutcome y   
## 0 0 0

There is no missing values in the data set

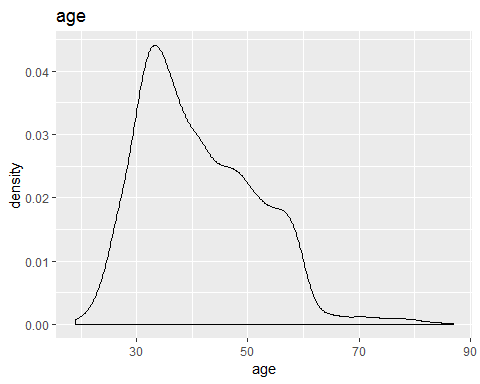
1. age(age of the customer)

summary(bank$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 33.00 39.00 41.17 49.00 87.00

minimum of age is 19 maximum is 87 and mean age is 41.17 Now visualize the data

library(ggplot2)  
ggplot(bank, aes(x=age)) +labs(title="age") +  
 geom\_density(aes(fill = age), alpha = 0.5)

 Most of the customers are from 30 to 40

2.job(Type of job). There are 12 type of job catagories.

table(bank$job)

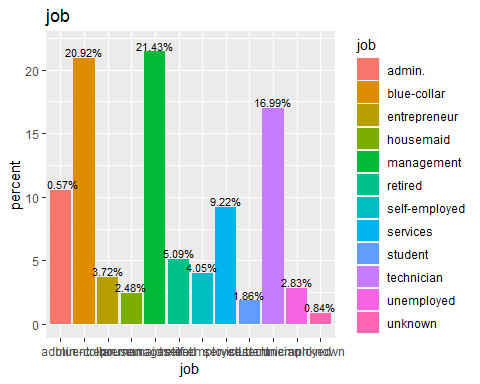
##   
## admin. blue-collar entrepreneur housemaid management   
## 478 946 168 112 969   
## retired self-employed services student technician   
## 230 183 417 84 768   
## unemployed unknown   
## 128 38

prop.table(table(bank$job))

##   
## admin. blue-collar entrepreneur housemaid management   
## 0.10572882 0.20924574 0.03715992 0.02477328 0.21433311   
## retired self-employed services student technician   
## 0.05087370 0.04047777 0.09223623 0.01857996 0.16987392   
## unemployed unknown   
## 0.02831232 0.00840522

Now visualize it

ggplot(bank, aes(x = job, fill = job)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'job', y = 'percent', fill = 'job', title = 'job')

 Highest number of customers having management job (21.43%) and next is blue-collar job (20.92%)

3.marital(marital status).

table(bank$marital)

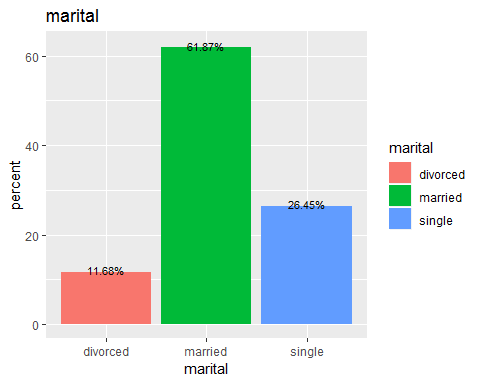
##   
## divorced married single   
## 528 2797 1196

prop.table(table(bank$marital))

##   
## divorced married single   
## 0.1167883 0.6186684 0.2645432

Now visualize it

ggplot(bank, aes(x = marital, fill = marital)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'marital', y = 'percent', fill = 'marital', title = 'marital')

 62% customers are married.

4.education(education of the customer)

table(bank$education)

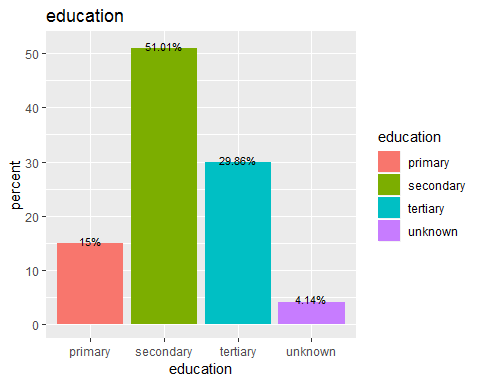
##   
## primary secondary tertiary unknown   
## 678 2306 1350 187

prop.table(table(bank$education))

##   
## primary secondary tertiary unknown   
## 0.14996682 0.51006415 0.29860650 0.04136253

visualize it

ggplot(bank, aes(x = education, fill = education)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'education', y = 'percent', fill = 'education', title = 'education')

 Half of the customers are having secondary education. 5. default(whether the customer has credit in default or not)

table(bank$default)

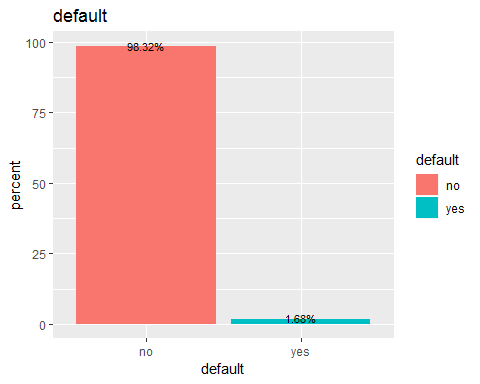
##   
## no yes   
## 4445 76

prop.table(table(bank$default))

##   
## no yes   
## 0.98318956 0.01681044

Visualize it

ggplot(bank, aes(x = default, fill = default)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'default', y = 'percent', fill = 'default', title = 'default')

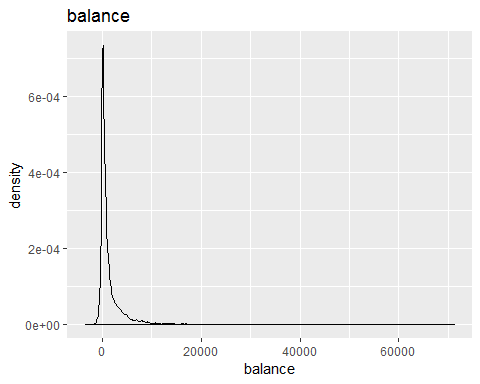
 98% customers has no credit in default 6.balance(Average yearly balance in Euros)

summary(bank$balance)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3313 69 444 1423 1480 71188

minimum balance is -3313, maximum is 71188 and mean baance is 1423. visualize it

ggplot(bank, aes(x=balance)) +labs(title="balance") +  
 geom\_density(aes(fill = balance), alpha = 0.5)



7.housing(whether the customer has housing loan or not)

table(bank$housing)

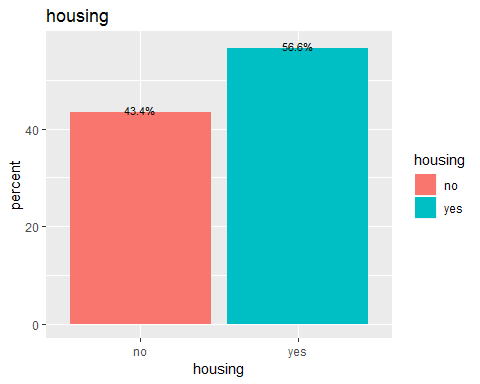
##   
## no yes   
## 1962 2559

prop.table(table(bank$housing))

##   
## no yes   
## 0.4339748 0.5660252

visualize it

ggplot(bank, aes(x = housing, fill = housing)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'housing', y = 'percent', fill = 'housing', title = 'housing')

 56.6% customers has housing loan.

8.loan(whether the customer has personal loan or not)

table(bank$loan)

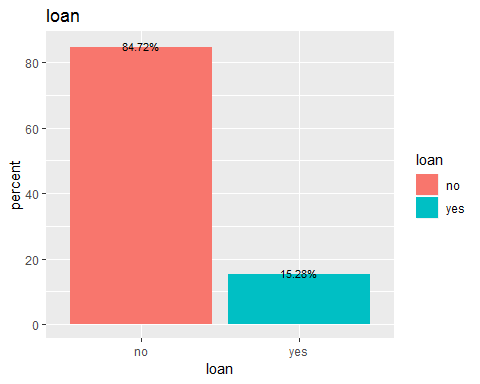
##   
## no yes   
## 3830 691

prop.table(table(bank$loan))

##   
## no yes   
## 0.8471577 0.1528423

visualize it

ggplot(bank, aes(x = loan, fill = loan)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'loan', y = 'percent', fill = 'loan', title = 'loan')

 84.7% has personal loan.

9.contact(how the last contact for marketing campaign has been made)

table(bank$contact)

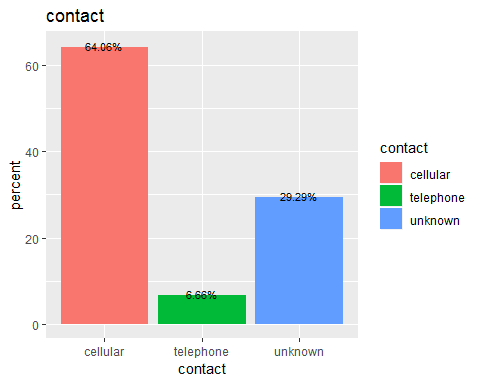
##   
## cellular telephone unknown   
## 2896 301 1324

prop.table(table(bank$contact))

##   
## cellular telephone unknown   
## 0.64056625 0.06657819 0.29285556

visualize it

ggplot(bank, aes(x = contact, fill = contact)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'contact', y = 'percent', fill = 'contact', title = 'contact')

 64% customers are contacted by cellular.

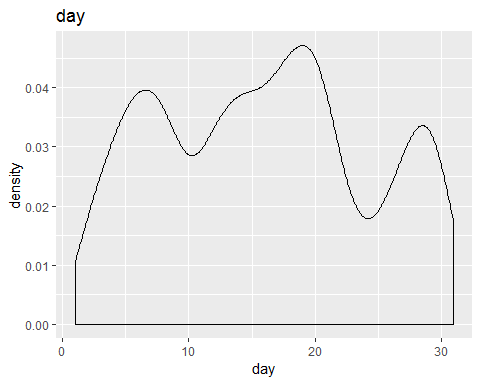
10.day(Day of the month last time customer was contacted)

summary(bank$day)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 9.00 16.00 15.92 21.00 31.00

visualize it

ggplot(bank, aes(x=day)) +labs(title="day") +  
 geom\_density(aes(fill = day), alpha = 0.5)

 Most customers contacted on 15 to 20

1. month(month of the year last time customer was contacted)

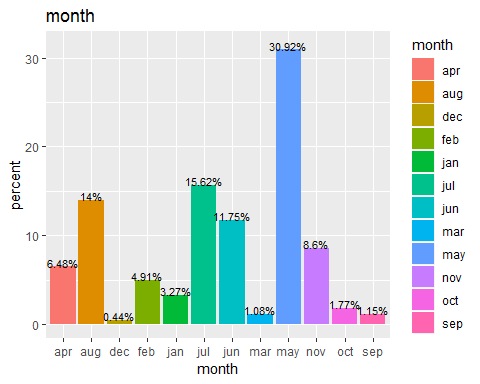
table(bank$month)

##   
## apr aug dec feb jan jul jun mar may nov oct sep   
## 293 633 20 222 148 706 531 49 1398 389 80 52

prop.table(table(bank$month))

##   
## apr aug dec feb jan jul   
## 0.06480867 0.14001327 0.00442380 0.04910418 0.03273612 0.15616014   
## jun mar may nov oct sep   
## 0.11745189 0.01083831 0.30922362 0.08604291 0.01769520 0.01150188

ggplot(bank, aes(x = month, fill = month)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'month', y = 'percent', fill = 'month', title = 'month')

 May has more customers contacted

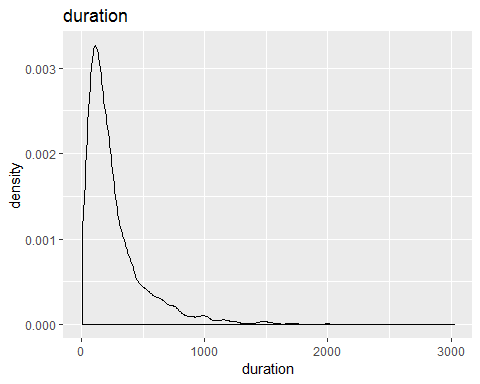
12.duration(last contact duration in seconds)

summary(bank$duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4 104 185 264 329 3025

visualize it

ggplot(bank, aes(x=duration)) +labs(title="duration") +  
 geom\_density(aes(fill = duration), alpha = 0.5)



Mean duration for last contact is 264 seconds

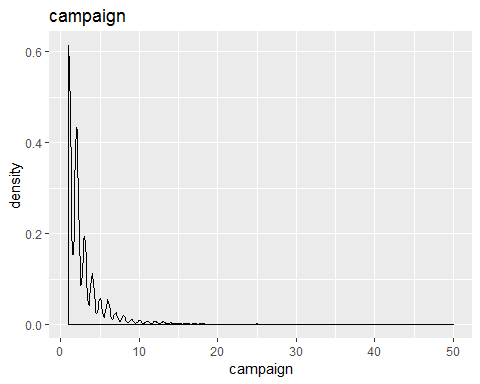
13.campaign(Number of contacts performed during the marketing campaign and for this customer)

summary(bank$campaign)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 2.794 3.000 50.000

Visualize

ggplot(bank, aes(x=campaign)) +labs(title="campaign") +  
 geom\_density(aes(fill = campaign), alpha = 0.5)

 Mean number of contacts performed a cuctomer is 2 times

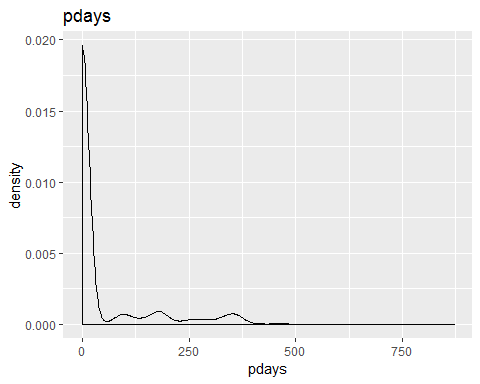
14.pdays(Number of days that passed by after the client was last contacted from a previous campaign)

summary(bank$pdays)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.00 -1.00 -1.00 39.77 -1.00 871.00

Visualize

ggplot(bank, aes(x=pdays)) +labs(title="pdays") +  
 geom\_density(aes(fill = pdays), alpha = 0.5)

 Mean number of days that passed by after the client has contacted from previous campaign is 39.

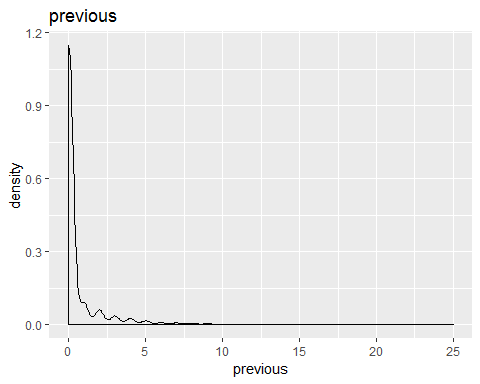
15.previous(Number of contacts performed before this campaign and for this client)

summary(bank$previous)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.5426 0.0000 25.0000

Visualize

ggplot(bank, aes(x=previous)) +labs(title="previous") +  
 geom\_density(aes(fill = previous), alpha = 0.5)

 Maximum number of contacts performed before this campaign is 25

16.poutcome(Outcome of the previous marketing campaign)

table(bank$poutcome)

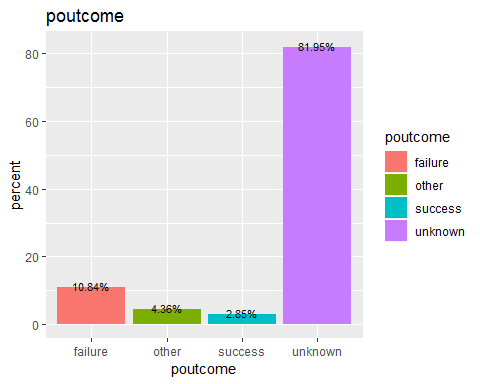
##   
## failure other success unknown   
## 490 197 129 3705

prop.table(table(bank$poutcome))

##   
## failure other success unknown   
## 0.10838310 0.04357443 0.02853351 0.81950896

Visualize

ggplot(bank, aes(x = poutcome, fill = poutcome)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'poutcome', y = 'percent', fill = 'poutcome', title = 'poutcome')

 Here 10% are failured, 2% success and 81% unknown.

17.y(Class attribute showing whether the client has subscribed a term deposit or not)

table(bank$y)

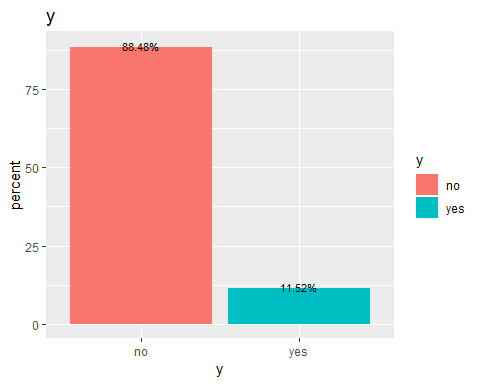
##   
## no yes   
## 4000 521

prop.table(table(bank$y))

##   
## no yes   
## 0.88476 0.11524

visualize

ggplot(bank, aes(x = y, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'y', y = 'percent', fill = 'y', title = 'y')

 There is class imbalance so 88% customers didn’t subscribed a deposit and 11% subscribed a deposit.

Correlation matrix

#install.packages("corrplot")  
library(corrplot)

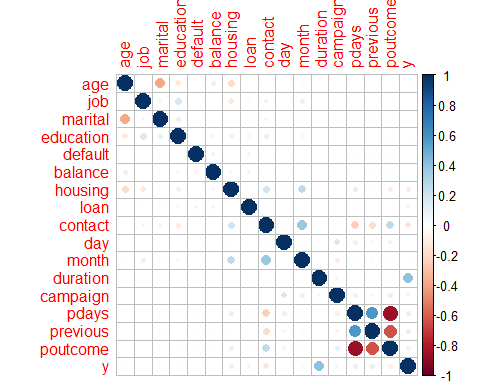
## corrplot 0.84 loaded

#install.packages("psych")  
library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

bank\_cor <- bank  
for(i in 1:ncol(bank\_cor)){  
   
 bank\_cor[,i]<- as.integer(bank\_cor[,i])  
}  
  
corrplot(cor(bank\_cor))

 pdays, previous and poutcome are highly correlated variables