Entrega-1

2023-02-20

Posada apunt i descarrega de paquets necessàris per a importar les funcions necessàries.

R Markdown

Obtenim les dades:

```
df <- read.csv2("./bank-additional-full.csv")
```

Agafem una mostra de 5000 persones de forma aleatoria per poder fer el nostre estudi. L'única condició que posarem serà que hi hagi un rang de y = yes

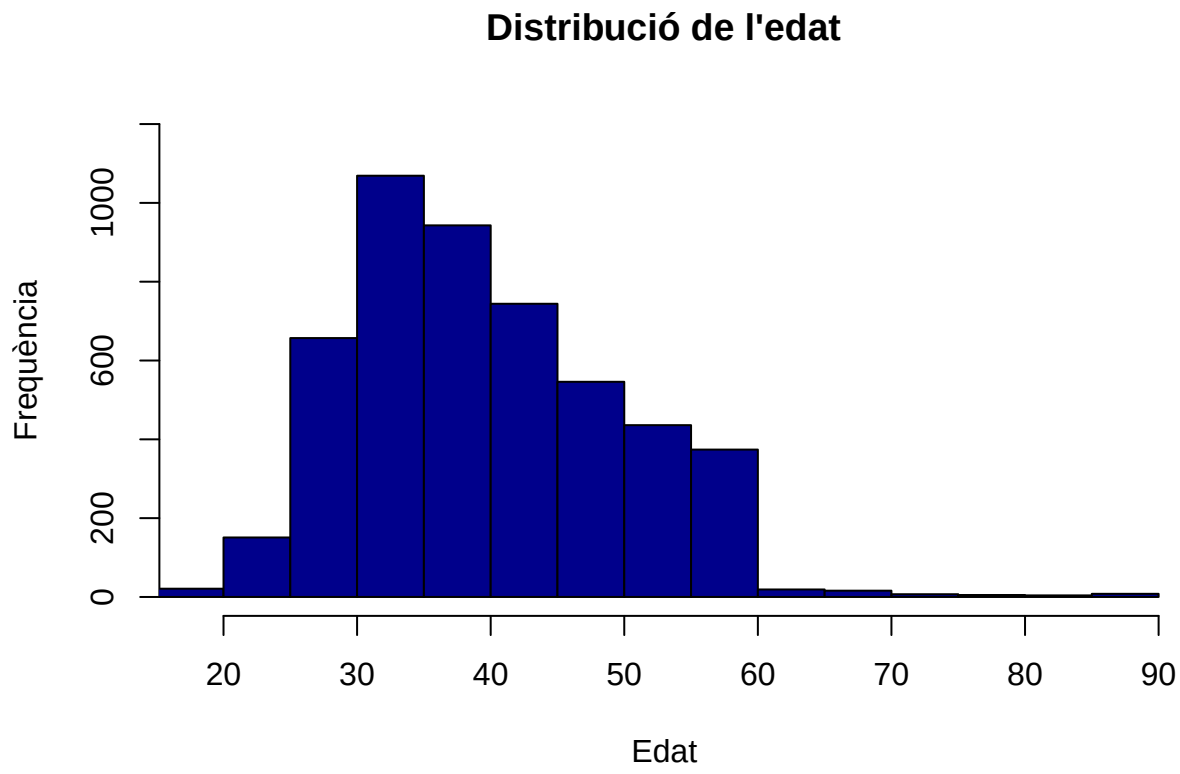
```
set.seed(49643638)
n <- 5000
number_of_trues = as.integer(runif(1, min = 2400, max=2600))

df_yes = df[df$y=="yes",]
df_yes = df_yes[sample(1:number_of_trues), ]

df_no = df[df$y=="no",]
df_no = df_no[sample(1:(n-number_of_trues)),]
df = rbind(df_yes, df_no)
```

```
df$age <- as.numeric(df$age)

df$job <- as.factor(df$job)
df$marital <- as.factor(df$marital)
df$education <- as.factor(df$education)
df$default <- as.factor(df$default)
df$housing <- as.factor(df$housing)
df$loan <- as.factor(df$loan)
df$contact <- as.factor(df$contact)
df$month <- as.factor(df$month)
df$day_of_week <- as.factor(df$day_of_week)

df$duration <- as.numeric(df$duration)
df$campaign <- as.numeric(df$campaign)
df$previous <- as.numeric(df$previous)

df$poutcome <- as.factor(df$poutcome)

df$emp.var.rate <- as.numeric(df$emp.var.rate)
df$cons.price.idx <- as.numeric(df$cons.price.idx)
```

```r
df$cons.conf.idx <- as.numeric(df$cons.conf.idx)
df$euribor3m <- as.numeric(df$euribor3m)
df$nr.employed <- as.numeric(df$nr.employed)

df$y <- as.factor(df$y)
```

Exploració de les dades

```r
hist(df$age,
     col = "blue4",
     xlim = c(min(df$age),
              max(df$age)),
              ylim = c(0, 1200),
              main = "Distribució de l'edat",
              xlab = "Edat",
              ylab = "Frequència")
```



Age

Agruparem en quatre noves categories: Jove[0,25], Jove-Adult[26,45], Adult[46,65], Gran[+66].

Realitzem aquesta distinció pels següents motius:

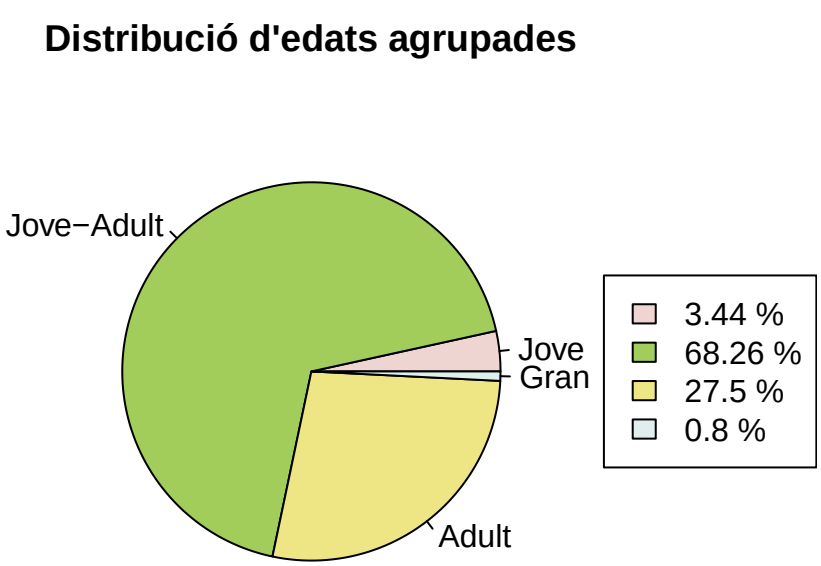
- Jove: No solen tenir gaire poder adquisitiu propi

- Jove-Adult: És quan s'acostumen a fer més plans de futur i a tenir més capacitat econòmica
- Adult: Solen ser persones amb una vida estable i sense gaires canvis econòmics grans
- Gran: Persones amb la vida feta, sense canvis econòmics (Com que l'edat més gran registrada és 88 anys, entendrem que no tenim outliers)

La variable passarà a ser categòrica.

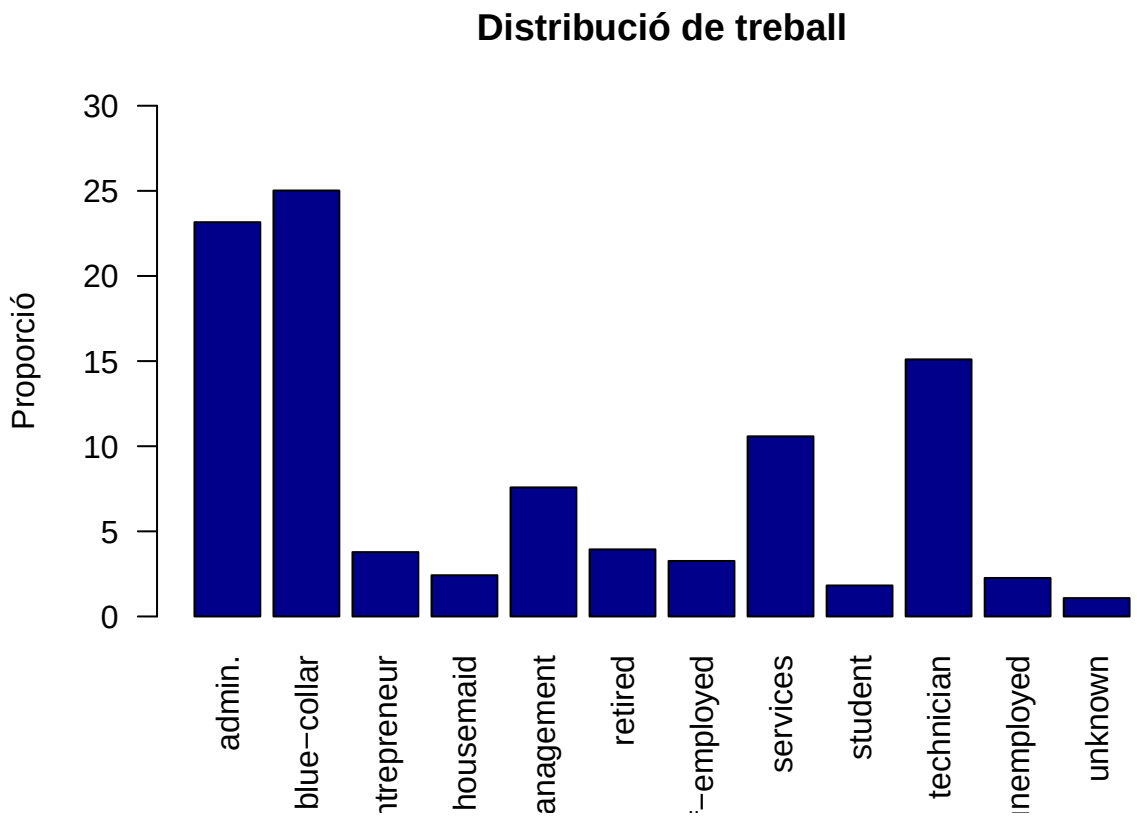
```
df$age_num <- df$age
df$age <- cut(df$age,
              breaks = c(0, 25, 45, 65, max(df$age)),
              labels = c("Jove", "Jove-Adult", "Adult", "Gran"))
df$age <- as.factor(df$age)
```

```
pie(table(df$age),
        col = c("mistyrose2", "darkolivegreen3", "khaki2", "azure2"),
        main = "Distribució d'edats agrupades")
legend("right", fill = c("mistyrose2", "darkolivegreen3", "khaki2", "azure2") , legend = paste(100*prop
```




```
barplot(100*prop.table(table(df$job)),
        ylim = c(0, 30),
        col = "blue4",
```

```
        main = "Distribució de treball",
        ylab = "Proporció",
        las = 2)
```

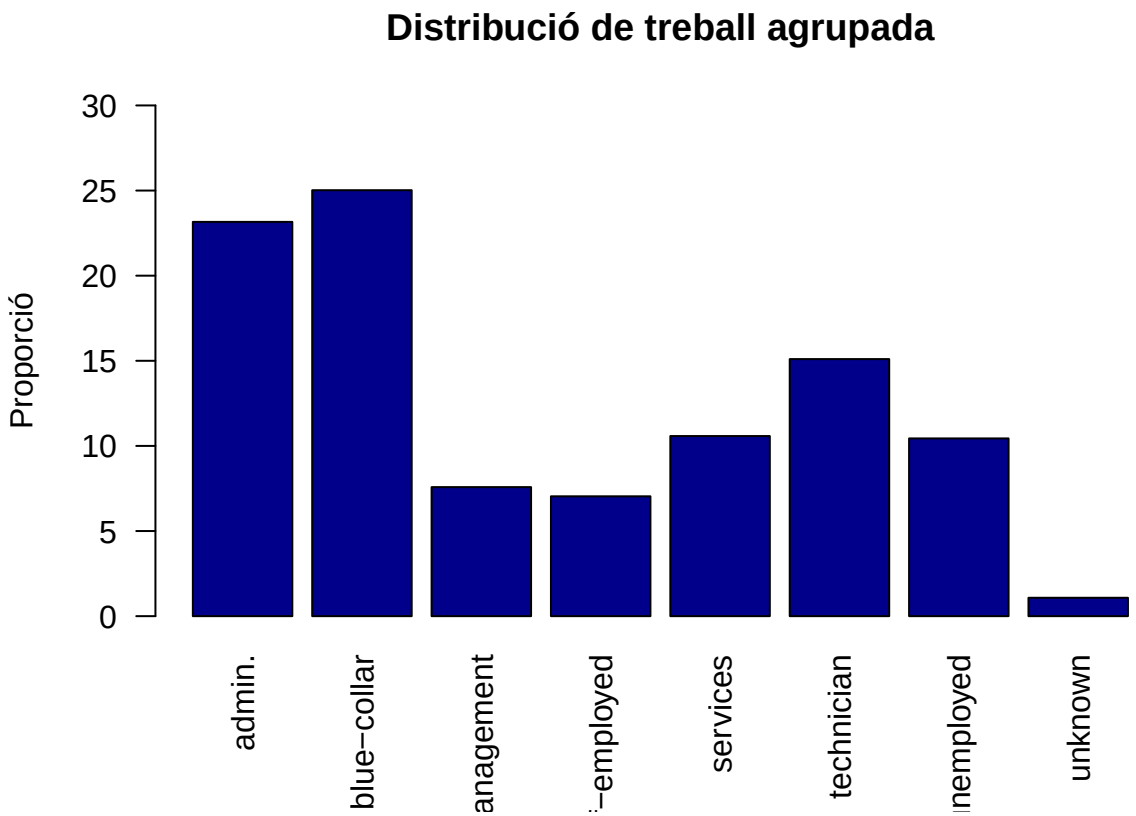


Les categories “retired”, “unemployed”, “student” i “housemaid” passaran a ser “unemployed”, ja que tenim en compte que són persones que no cotitzen.

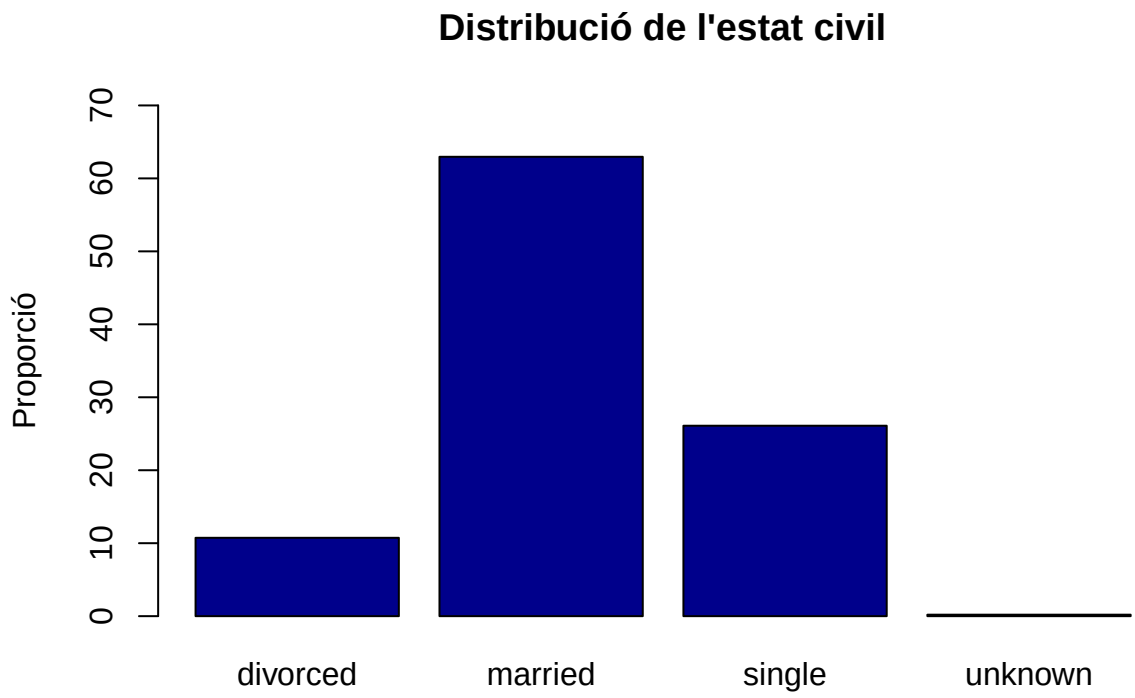
Els “enterpeneur” passaran a ser “self-employed”. vector indicating the indexes of the quantitative supplementary variables

```
df$job <- as.character(df$job)
df$job <- ifelse(df$job %in% c("retired", "unemployed", "student", "housemaid"), "unemployed", df$job)
df$job <- ifelse(df$job == "entrepreneur", "self-employed", df$job)
df$job <- as.factor(df$job)
```

```
barplot(100*prop.table(table(df$job)),
        ylim = c(0, 30),
        col = "blue4",
        main = "Distribució de treball agrupada",
        ylab = "Proporció",
        las = 2)
```

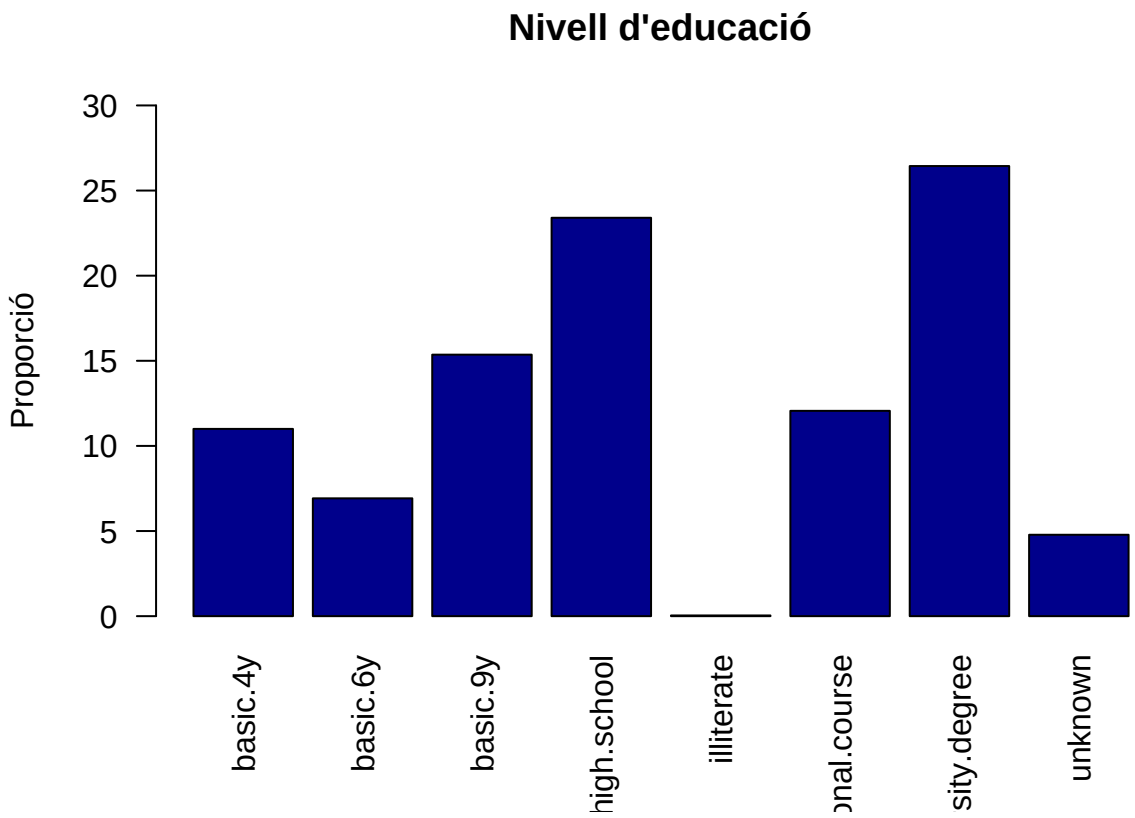



```
barplot(100*prop.table(table(df$marital)),
        ylim = c(0, 70),
        main = "Distribució de l'estat civil",
        ylab = "Proporció",
        col = "blue4")
```



Marital

```
barplot(100*prop.table(table(df$education)),
        ylim = c(0, 30),
        col = "blue4",
        main = "Nivell d'educació",
        ylab = "Proporció",
        las = 2)
```

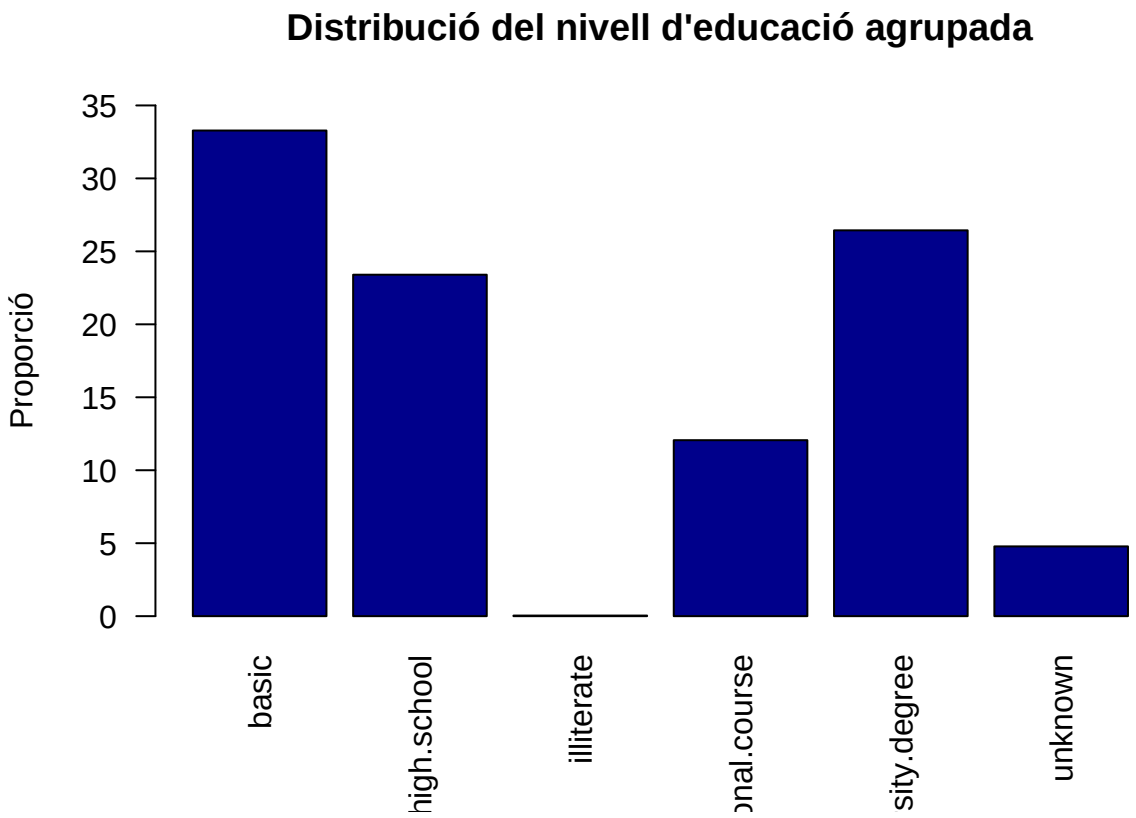


Education

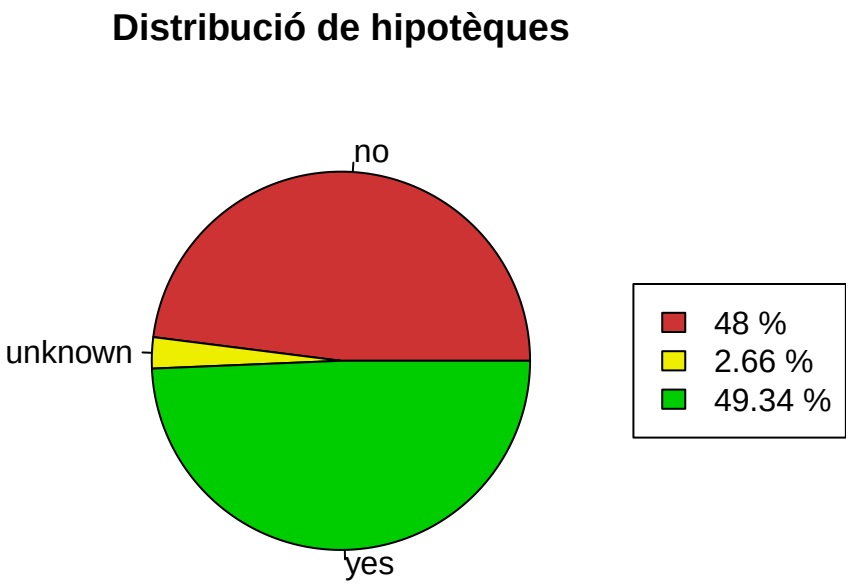
Les categories "basic.4y", "basic.6y", "basic.9y" passaran a ser "basic", ja que no aporta informació saber quin nivell de "basic" tenen.

```
df$education <- as.character(df$education)
df$education <- ifelse(df$education %in% c("basic.4y", "basic.6y", "basic.9y"), "basic", df$education)
df$education <- as.factor(df$education)
```

```
barplot(100*prop.table(table(df$education)),
        ylim = c(0, 35),
        col = "blue4",
        main = "Distribució del nivell d'educació agrupada",
        ylab = "Proporció",
        las = 2)
```

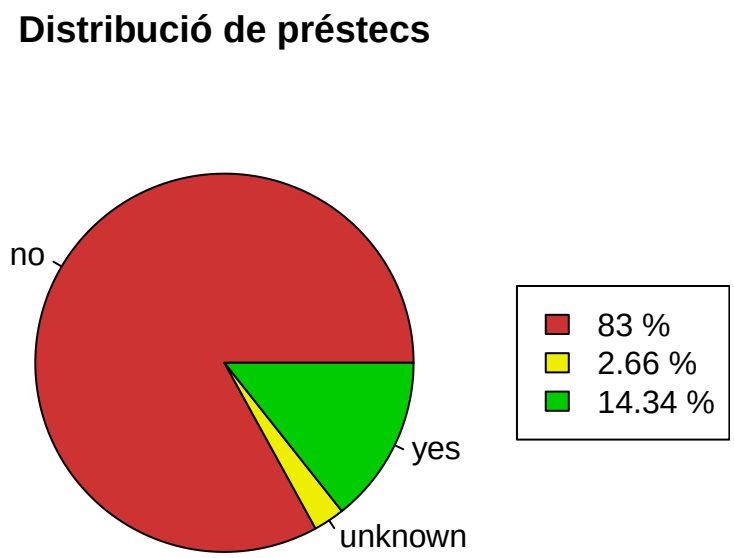



```
pie(prop.table(table(df$housing)),
        col = c("brown3", "yellow2", "green3"),
        main = "Distribució de hipotèques")
legend("right", fill = c("brown3", "yellow2", "green3") , legend = paste(100*prop.table(table(df$housing
```



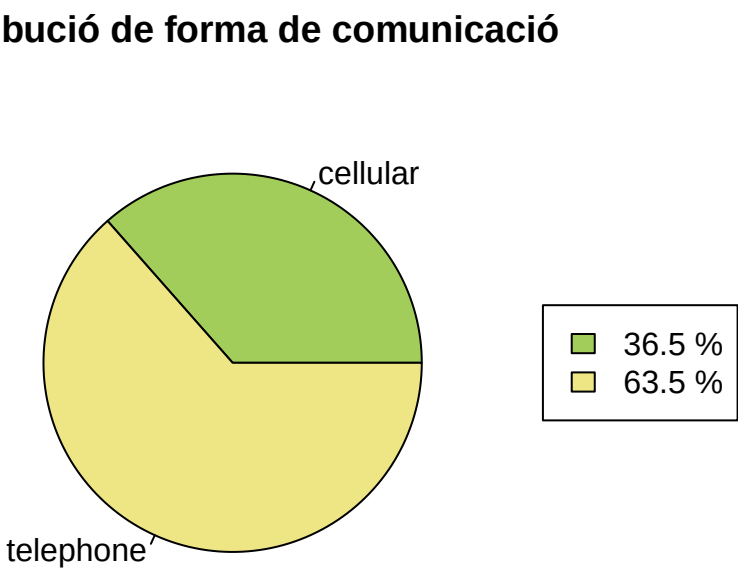
Housing

```
pie(prop.table(table(df$loan)),
        col = c("brown3", "yellow2", "green3"),
        main = "Distribució de préstecs")
legend("right", fill = c("brown3", "yellow2", "green3") , legend = paste(100*prop.table(table(df$loan))
```



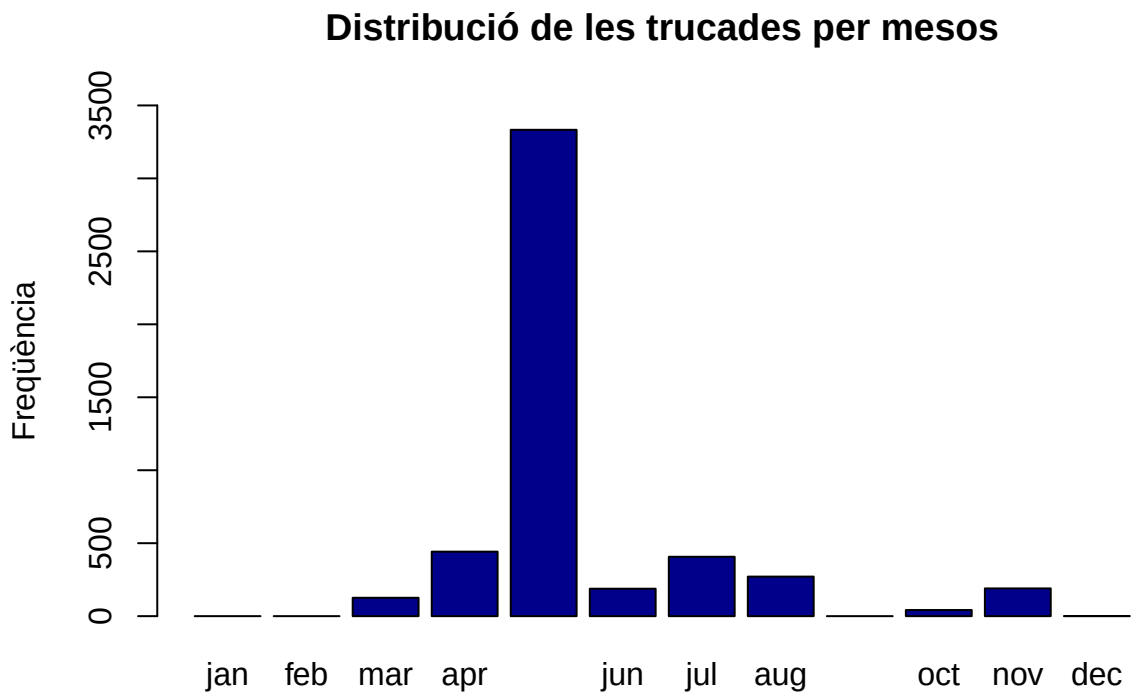
Loan

```
pie(table(df$contact),
        col = c("darkolivegreen3", "khaki2"),
        main = "Distribució de forma de comunicació")
legend("right", fill = c("darkolivegreen3", "khaki2") , legend = paste(100*prop.table(table(df$contact)),
```



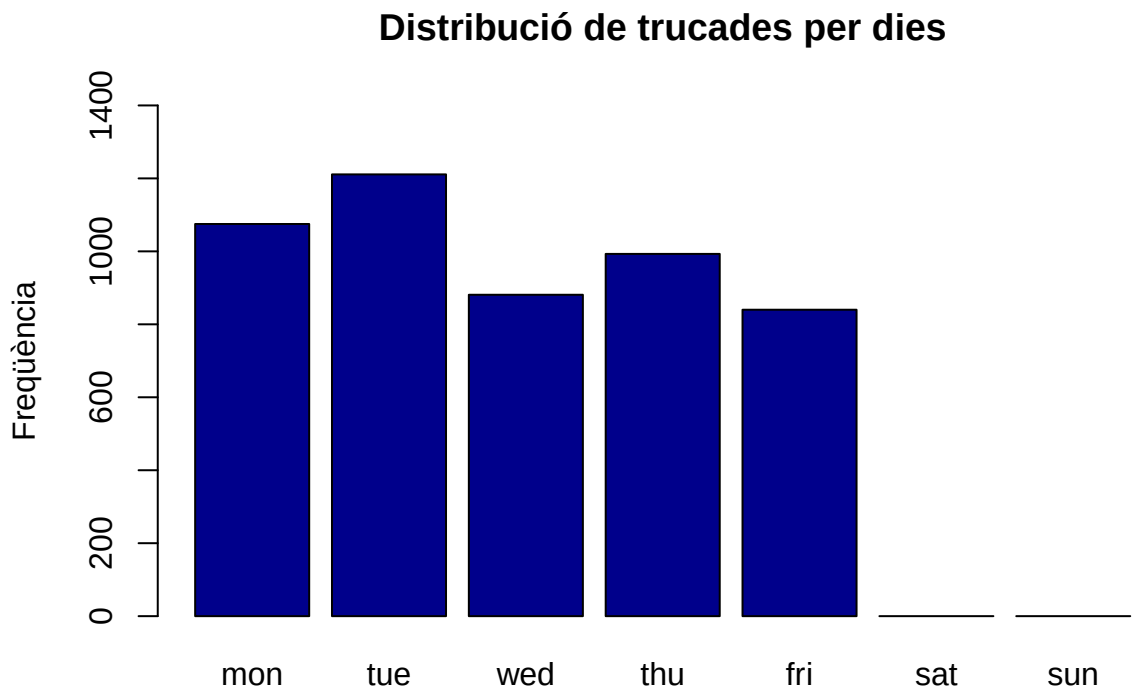
Contact

```
mesos <- c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec")

df$month <- factor(df$month, levels = mesos)

barplot(table(df$month),
        ylim = c(0, 3500),
        col = "blue4",
        ylab = "Freqüència",
        main = "Distribució de les trucades per mesos")
```



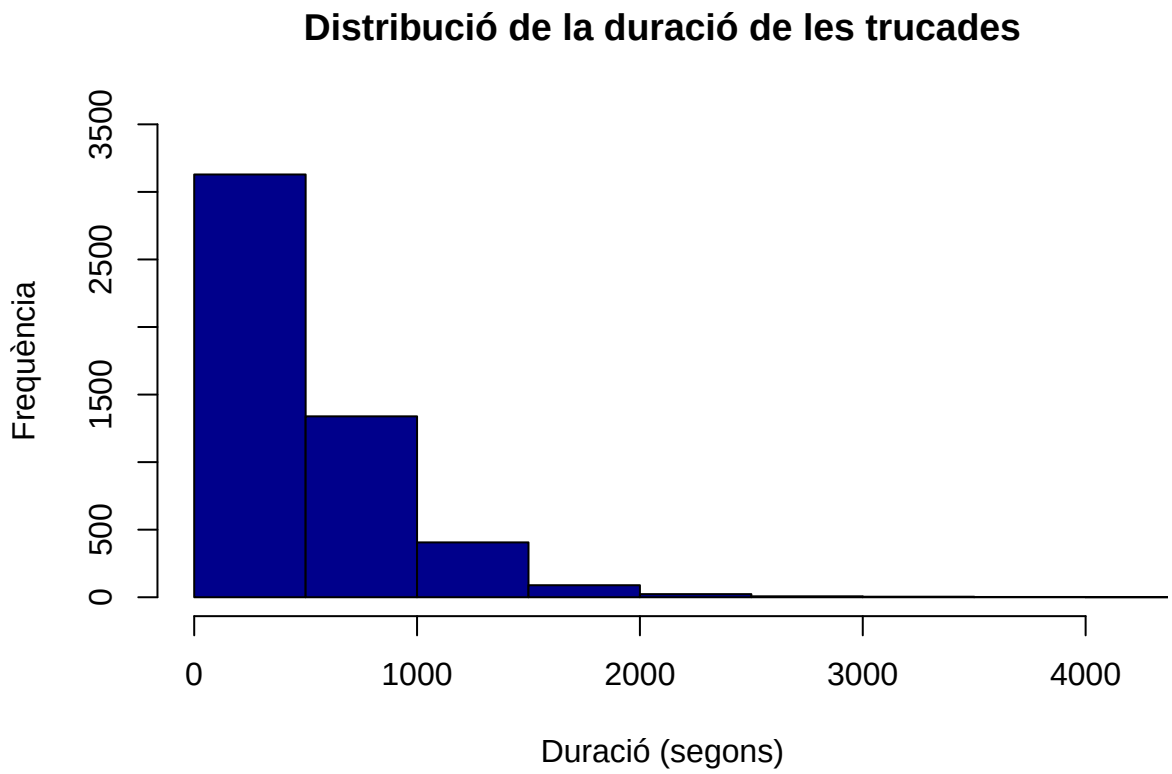
Month

```
days <- c("mon", "tue", "wed", "thu", "fri", "sat", "sun")

df$day_of_week <- factor(df$day_of_week, levels = days)

barplot(table(df$day_of_week),
        ylim = c(0, 1400),
        col = "blue4",
        ylab = "Freqüència",
        main = "Distribució de trucades per dies")
```



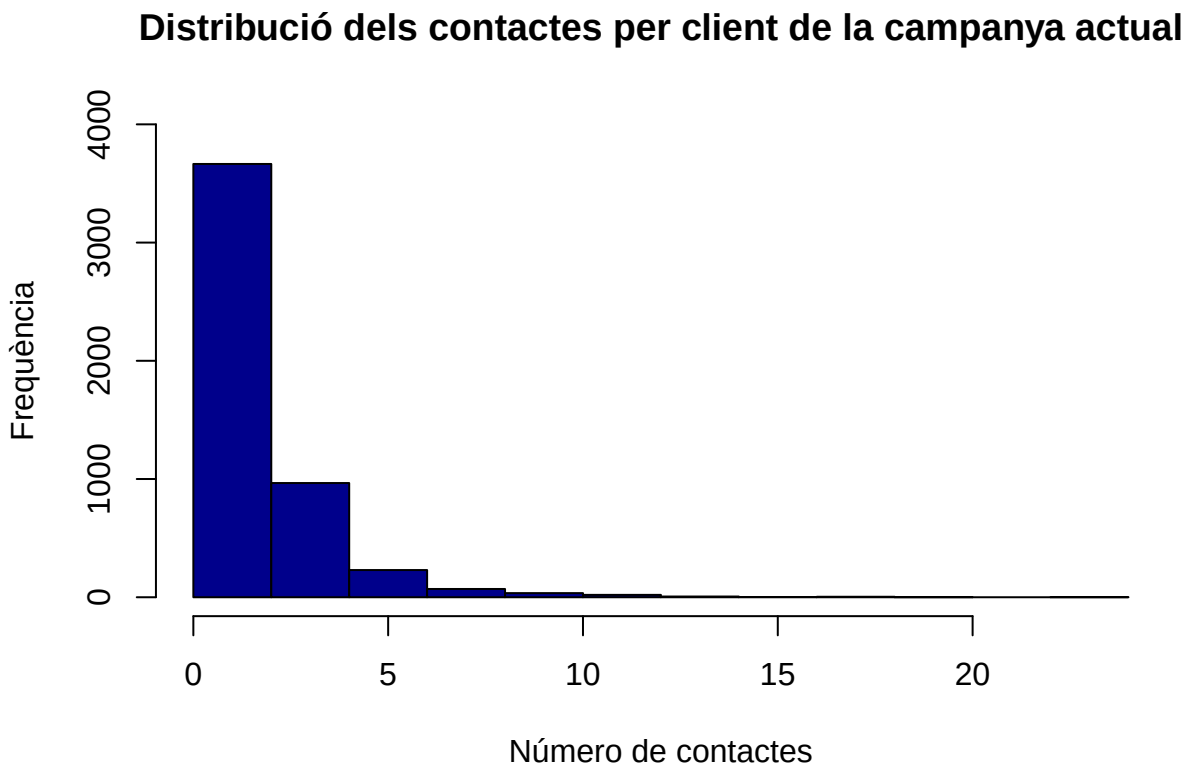
Day of the week

```
hist(df$duration,
     col = "blue4",
     xlim = c(min(df$duration),
              max(df$duration)),
              ylim = c(0, 3500),
              main = "Distribució de la duració de les trucades",
              xlab = "Duració (segons)",
              ylab = "Frequència")
```



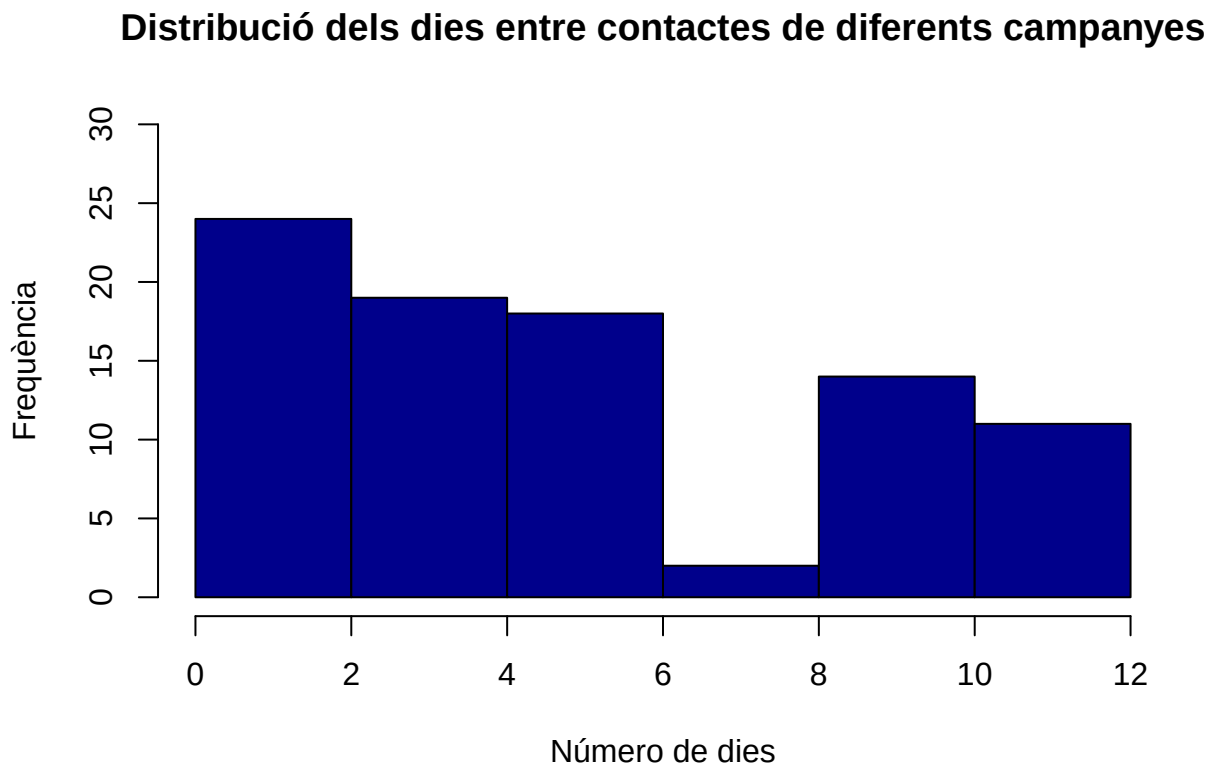
Duration

```
hist(df$campaign,
     col = "blue4",
               ylim = c(0, 4000),
               main = "Distribució dels contactes per client de la campanya actual",
               xlab = "Número de contactes",
               ylab = "Frequència")
```



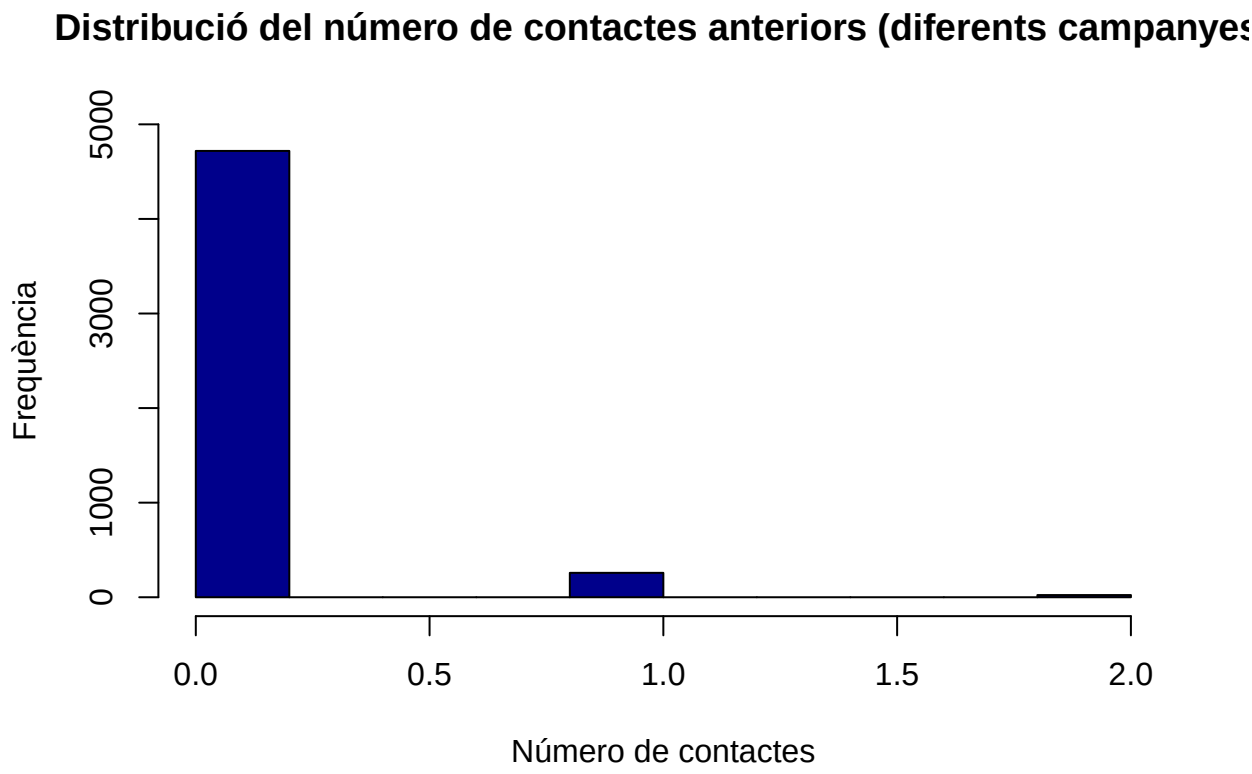
Campaign

```
pdays_aux <- subset(df$pdays, df$pdays != 999)
hist(pdays_aux,
     col = "blue4",
               ylim = c(0, 30),
               main = "Distribució dels dies entre contactes de diferents campanyes",
               xlab = "Número de dies",
               ylab = "Frequència")
```



Pdays

```
hist(df$previous,
     col = "blue4",
              ylim = c(0, 5000),
              main = "Distribució del número de contactes anteriors (diferents campanyes)",
              xlab = "Número de contactes",
              ylab = "Frequència")
```



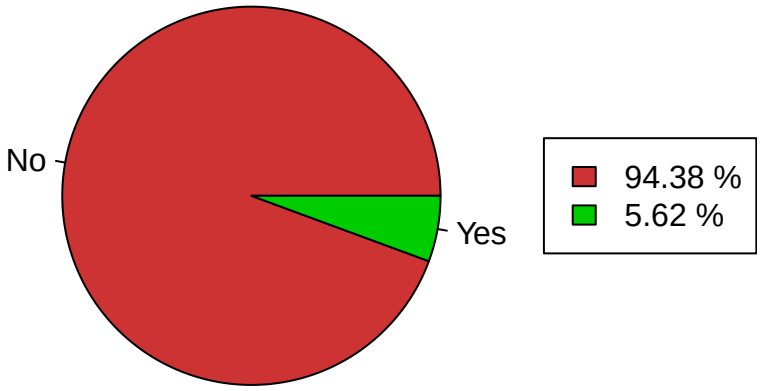
Previous

Ho passem a dos valors possibles: No[0] Yes[+1].

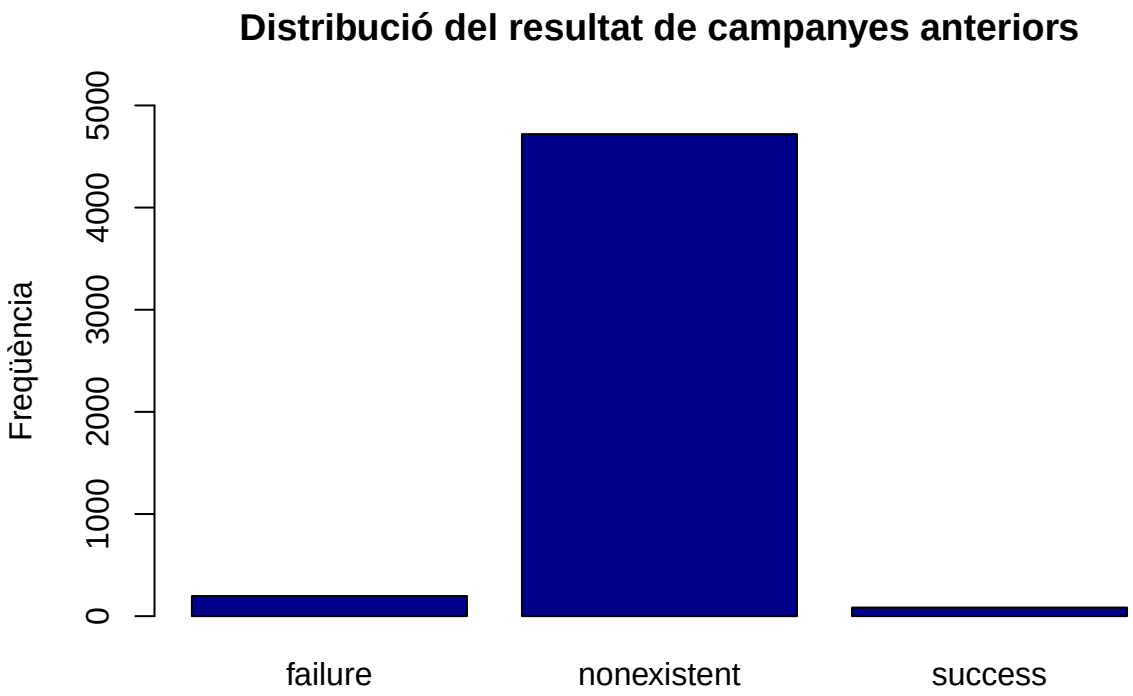
La variable passarà a ser categòrica.

```
df <- df %>% mutate(previous = ifelse(previous >= 1, "Yes" , previous))
df <- df %>% mutate(previous = ifelse(previous == 0, "No" , previous))
df$previous <- as.factor(df$previous)
```

```
pie(table(df$previous), main = "Distribució unificada de contactes anteriors (diferents campanyes)",
    col = c("brown3", "green3"))
legend("right", fill = c("brown3", "green3") , legend = paste(100*prop.table(table(df$previous)), "%"))
```

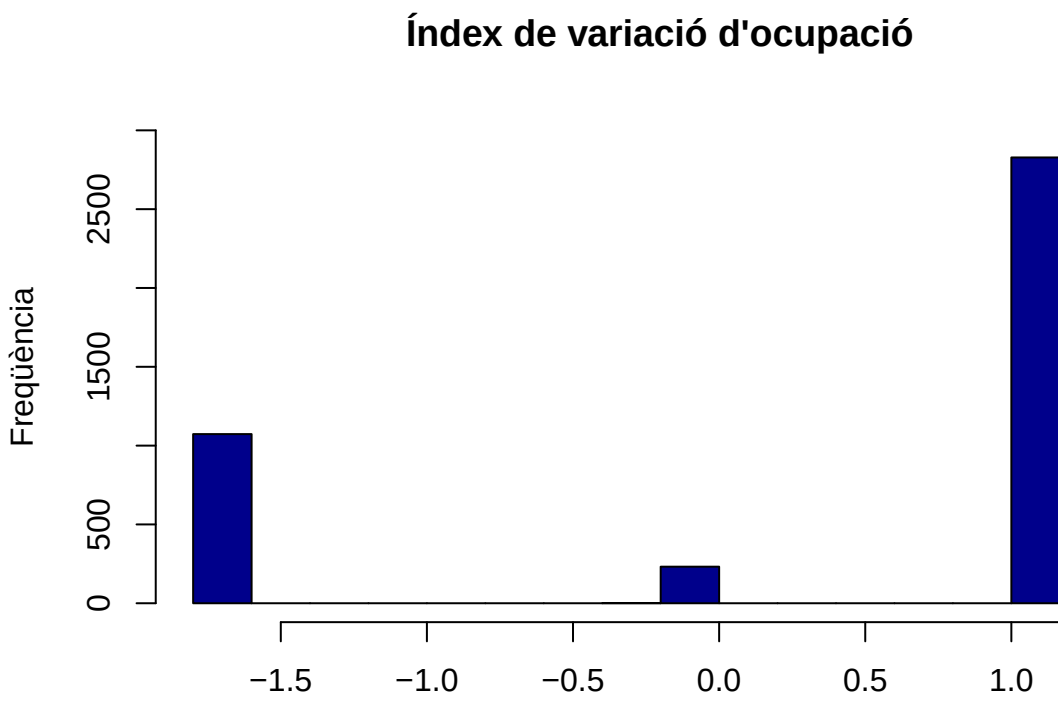



```
barplot(table(df$poutcome),
        ylim = c(0, 5000),
        col = "blue4",
        ylab = "Freqüència",
        main = "Distribució del resultat de campanyes anteriors")
```



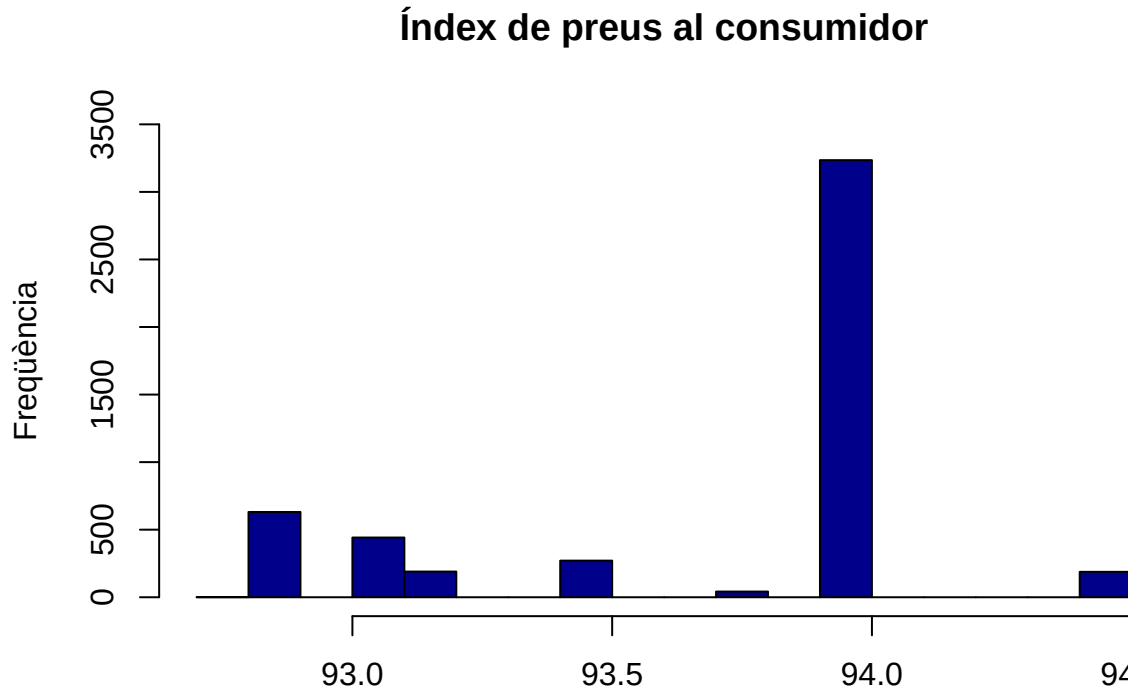
Poutcome

```
hist(df$emp.var.rate,
     ylim = c(0, 3000),
        col = "blue4",
        xlab = "",
        ylab = "Freqüència",
        main = "Índex de variació d'ocupació")
```



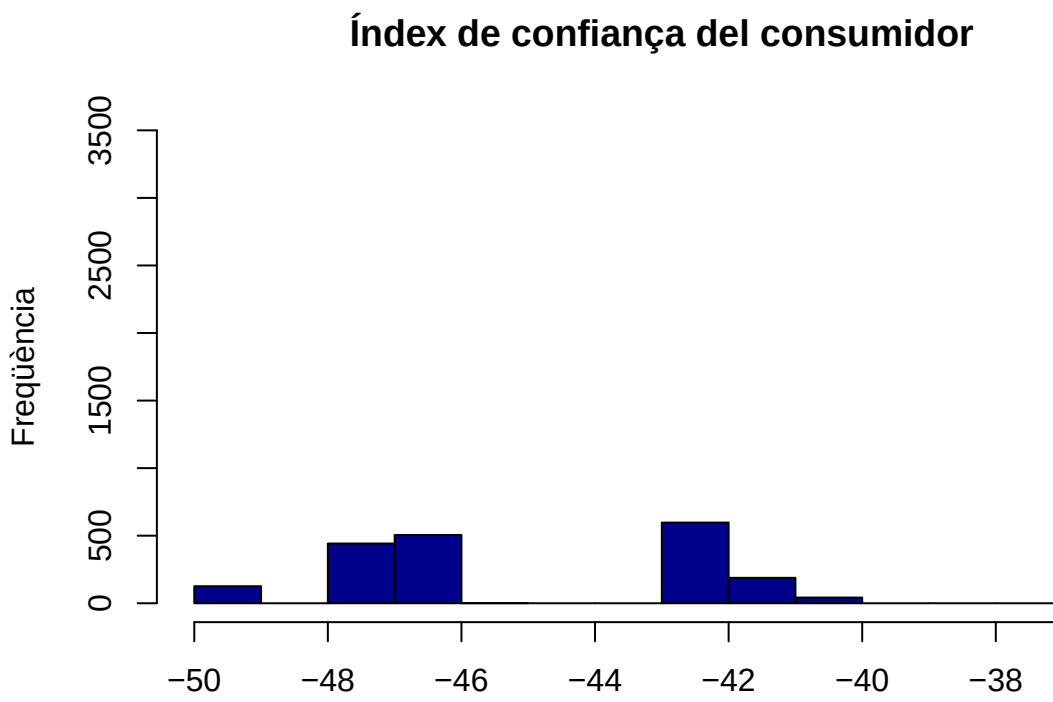
Employment variation rate

```
hist(df$cons.price.idx,
     ylim = c(0, 3500),
        col = "blue4",
        xlab = "",
        ylab = "Freqüència",
        main = "Índex de preus al consumidor")
```



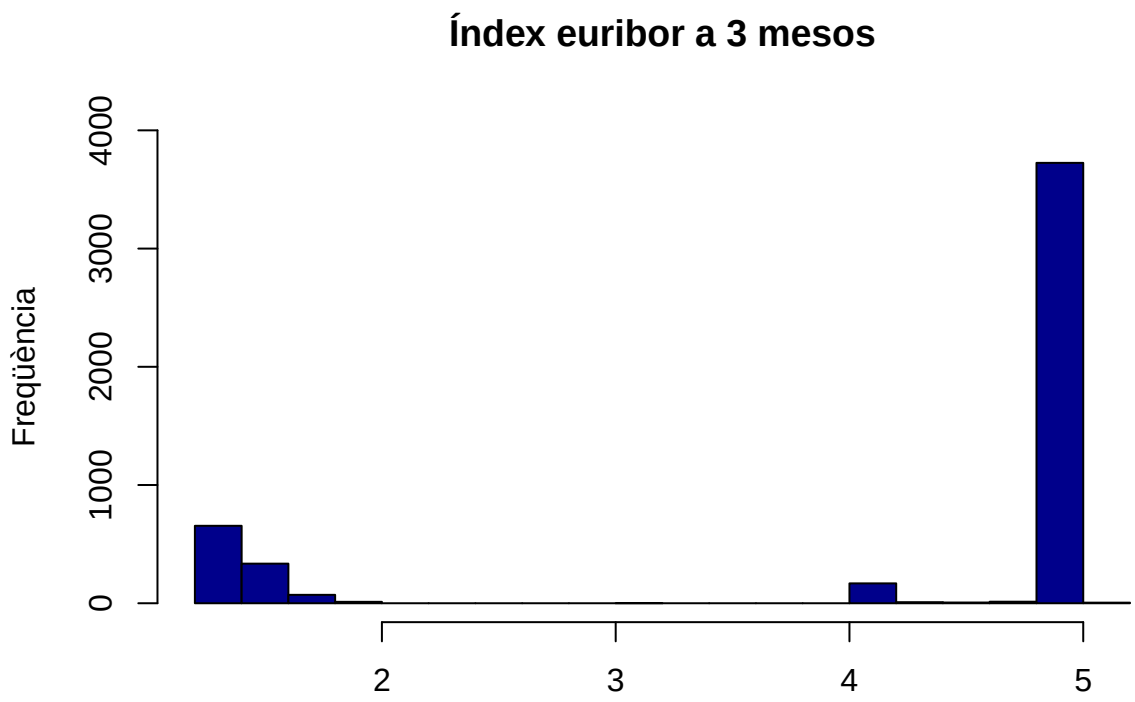
Consumer price index

```
hist(df$cons.conf.idx,
     ylim = c(0, 3500),
        col = "blue4",
        xlab = "",
        ylab = "Freqüència",
        main = "Índex de confiança del consumidor")
```



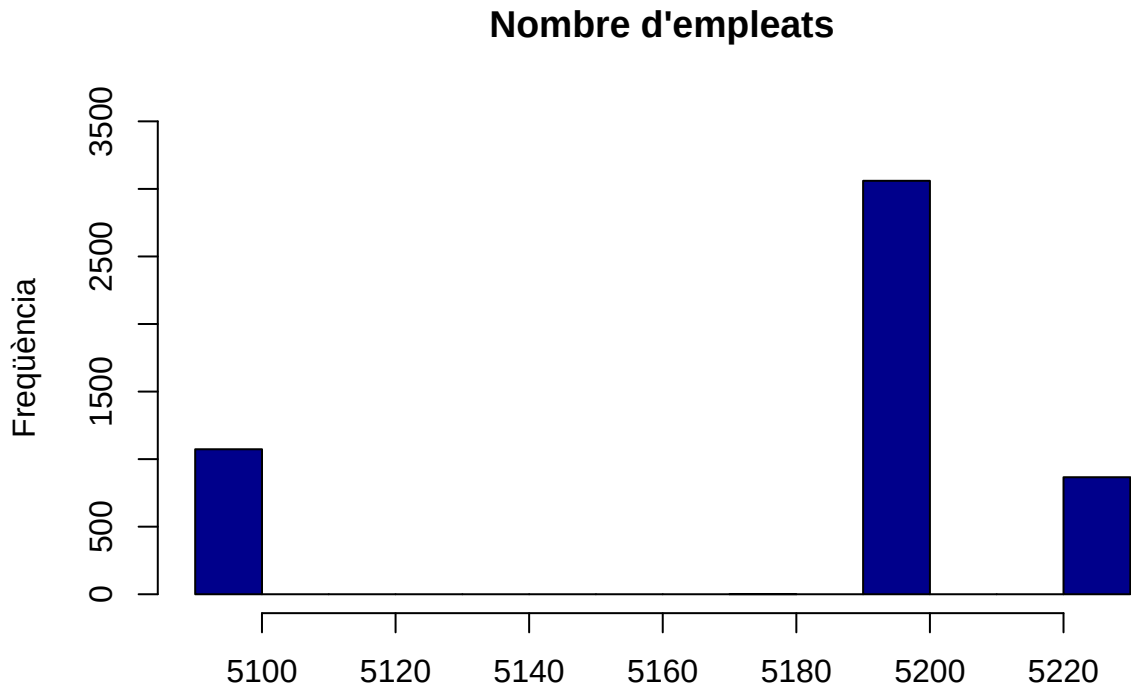
Consumer confidence index

```
hist(df$euribor3m,
     ylim = c(0, 4000),
        col = "blue4",
        xlab = "",
        ylab = "Freqüència",
        main = "Índex euribor a 3 mesos")
```



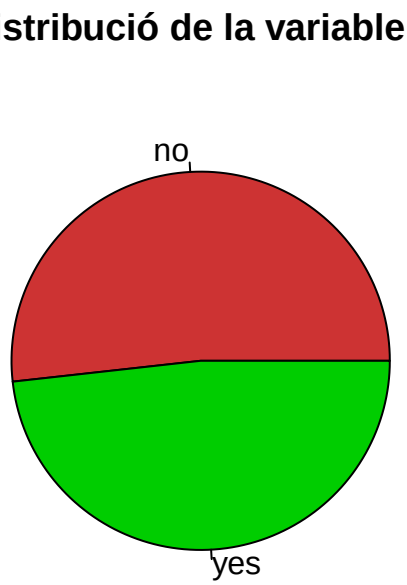
Euribor 3 month rate

```
hist(df$nr.employed,
     ylim = c(0, 3500),
        col = "blue4",
        xlab = "",
        ylab = "Freqüència",
        main = "Nombre d'empleats")
```



Number of employees

```
pie(prop.table(table(df$y)),
        col = c("brown3", "green3"),
        main = "Distribució de la variable y")
```



Subscribed deposit

Qualitat de les dades

Per variable

Nombre de missings Passem tots els valors "unknown" a NA's per tractar-los com a missings i contem el total de NA's per variable.

```
df[df == "unknown"] <- NA

#Mostrem el nombre de missings per variable
colSums(is.na(df))
```

```
##            age            job        marital      education        default
##              0             54             10            239           1244
##        housing           loan        contact          month    day_of_week
##            133            133              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        age_num
##              0              0
```

Nombre d'errors 1- Que una persona tingui pdays de 999 i que previous "yes" és una contradicció

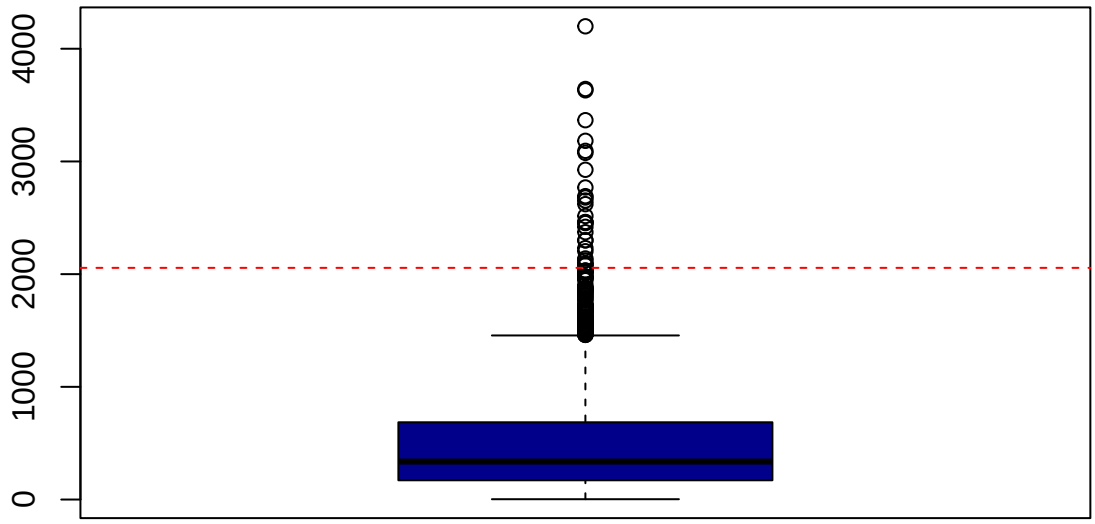
```
err_ind <- which(df$pdays == 999 & df$previous == "Yes")
length(err_ind)
```

```
## [1] 193
```

```
df$previous[err_ind]<-NA
df$pdays[err_ind]<-NA
```

Nombre d'outliers Busquem els outliers de les variables utilitzant la tècnica del tercer quantil

```
boxplot(df$duration, col = "blue4")
treshold <- quantile(df$duration, 0.75)*3
abline(h = treshold, col = "red", lty = "dashed")
```

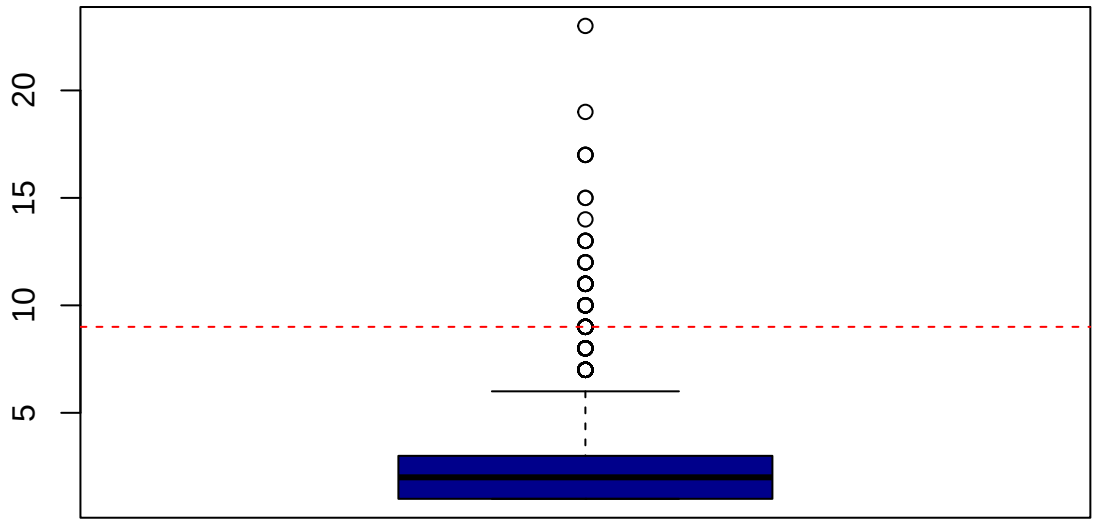


Trobem el nombre d'outliers

```
out_dur <- which(df$duration > treshold)
length(out_dur)
```

```
## [1] 30
```

```
boxplot(df$campaign, col = "blue4")
treshold <- quantile(df$campaign, 0.75)*3
abline(h = treshold, col = "red", lty = "dashed")
```



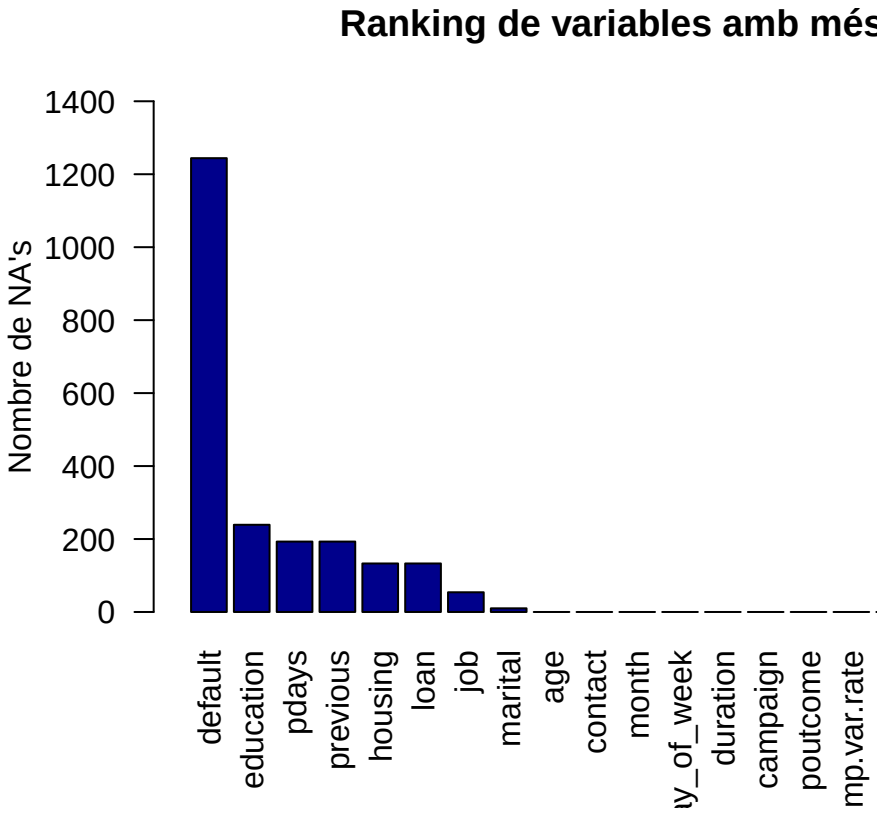
Trobem el nombre d'outliers

```
out_camp <- which(df$campaign > treshold)
length(out_camp)
```

```
## [1] 45
```

```
nas_per_var <- data.frame(num_nas = colSums(is.na(df)),
                          variable = names(df))

# Ordenem el data frame per nombre de NA's descendent
nas_per_var <- nas_per_var[order(-nas_per_var$num_nas),]

# Fem el gràfic de barres
barplot(nas_per_var$num_nas,
        names.arg = nas_per_var$variable,
        ylim = c(0, 1400),
        col = "blue4",
```

```
        ylab = "Nombre de NA's",
        main = "Ranking de variables amb més NA's",
        las = 2)
```



Ranking de variables amb més valors NA

Per individus

Tornem a recuperar els valors de pdays i previous per no contar-los com a missings, sino com a errors.

```
df$pdays[err_ind]<-999
df$previous[err_ind]<-"Yes"
```

```
n_missings <- rowSums(is.na(df))
table(n_missings)
```

Nombre de missings

```
## n_missings
##    0    1    2    3    4
## 3506 1250  174   65    5
```

Nombre d'errors Com que només hem detectat un error, que pdays sigui 999 i previous sigui "yes", sabem que tenim 193 combinacions de valors que compleixen l'error. Per tant, tindrem que 193 individus tenen 2 errors (1 a pdays i 1 a previous) i la resta (4807) no tenen cap error.

Esborrarem la variable pdays ja que no ens aporta cap informació adicional.

```
df <- subset(df, select = -pdays)
```

```
out_ind <- rowSums(cbind(as.numeric(df$duration %in% out_dur), as.numeric(df$campaign %in% out_camp)))
table(out_ind)
```

Nombre d'outliers

```
## out_ind
##    0    1
## 4934   66
```

Afegim variable que conta els NA Abans de contar els missings, passarem a NA's els index de les variables que hem trobat errors o outliers.

```
df$campaign[out_camp]<-NA
df$duration[out_dur]<-NA
df$previous[err_ind]<-NA
```

Creem la nova variable i contem els NA's

```
df$na_count <- apply(df, 1, function(x) sum(is.na(x)))
```

```
corr <- df[,c("duration", "campaign", "na_count")]
cor_mat <- cor(corr)

# Ordenar les correlacions per valor absolut
cor_ranking <- sort(abs(cor_mat), decreasing = TRUE)

# Mostrar el ranking
print(cor_ranking)
```

Correlació de les variables

```
## [1] 1 1 1
```

Relació entre grup d'edat i valors atípics

Imputació de missings

Variables categòriques Imputem tots els NAs que tenim en el conjunt de variables categòriques (var_cat).

Dades abans de imputar.

```
var_cat <- c("age", "job", "marital", "education", "housing", "loan", "contact", "month", "day_of_week"
summary(df[,var_cat])
```

```
##          age                job           marital
##  Jove      : 172   blue-collar:1251   divorced: 537
##  Jove-Adult:3413   admin.     :1158   married :3148
##  Adult     :1375   technician : 755   single  :1305
##  Gran      :  40   services   : 529   unknown :   0
##                    unemployed : 522   NA's    :  10
##                    (Other)    : 731
##                    NA's       :  54
##                education       housing          loan           contact
##  basic              :1664   no     :2400   no     :4150   cellular :1825
##  high.school        :1170   unknown:   0   unknown:   0   telephone:3175
##  illiterate         :   2   yes    :2467   yes    : 717
##  professional.course: 603   NA's   : 133   NA's   : 133
##  university.degree  :1322
##  unknown            :   0
##  NA's               : 239
##      month      day_of_week previous           poutcome      y
##  may    :3333   mon:1075    No  :4719   failure    : 197   no :2588
##  apr    : 442   tue:1211    Yes :  88   nonexistent:4719   yes:2412
##  jul    : 407   wed: 881    NA's: 193   success    :  84
##  aug    : 271   thu: 993
##  nov    : 190   fri: 840
##  jun    : 188   sat:   0
##  (Other): 169   sun:   0
```

```
res.immca<-imputeMCA(df[,var_cat],ncp = length(var_cat)-1)
summary(res.immca$completeObs)
```

```
##          age                   job           marital
##  Jove      : 172   admin.       :1169   divorced: 537
##  Jove-Adult:3413   blue-collar  :1292   married :3158
##  Adult     :1375   management   : 379   single  :1305
##  Gran      :  40   self-employed: 352
##                    services     : 530
##                    technician   : 756
##                    unemployed   : 522
##                education    housing     loan           contact
##  basic              :1767   no :2473   no :4283   cellular :1825
##  high.school        :1212   yes:2527   yes: 717   telephone:3175
##  illiterate         :   2
##  professional.course: 632
##  university.degree  :1387
##
```

```
##
##      month      day_of_week previous          poutcome      y
##  may    :3333   mon:1075    No :4912   failure    : 197   no :2588
##  apr    : 442   tue:1211    Yes:  88   nonexistent:4719   yes:2412
##  jul    : 407   wed: 881               success    :  84
##  aug    : 271   thu: 993
##  nov    : 190   fri: 840
##  jun    : 188
##  (Other): 169
```

```
df[,var_cat]<-res.immca$completeObs
```

Variables numèriques Imputem tots els NAs que tenim en el conjunt de variables numeriques (var_num).

Dades abans de imputar.

```
var_num <- c("duration", "campaign", "emp.var.rate", "cons.conf.idx", "cons.price.idx", "euribor3m", "n
summary(df[,var_num])
```

```
##     duration         campaign      emp.var.rate     cons.conf.idx
##  Min.   :   3.0   Min.   :1.000   Min.   :-1.8000   Min.   :-50.00
##  1st Qu.: 170.0   1st Qu.:1.000   1st Qu.:-0.1000   1st Qu.:-42.70
##  Median : 334.0   Median :2.000   Median : 1.1000   Median :-36.40
##  Mean   : 463.6   Mean   :2.065   Mean   : 0.4737   Mean   :-39.63
##  3rd Qu.: 676.8   3rd Qu.:3.000   3rd Qu.: 1.1000   3rd Qu.:-36.40
##  Max.   :2033.0   Max.   :9.000   Max.   : 1.4000   Max.   :-36.10
##  NA's   :30       NA's   :45
##  cons.price.idx    euribor3m      nr.employed
##  Min.   :92.76   Min.   :1.244   Min.   :5099
##  1st Qu.:93.20   1st Qu.:4.343   1st Qu.:5191
##  Median :93.99   Median :4.856   Median :5191
##  Mean   :93.72   Mean   :4.100   Mean   :5178
##  3rd Qu.:93.99   3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :94.47   Max.   :5.045   Max.   :5228
##
```

Imputació dels missings.

```
res.impca<-imputePCA(df[,var_num],ncp = length(var_num)-1)
summary(res.impca$completeObs)
```

```
##     duration         campaign      emp.var.rate     cons.conf.idx
##  Min.   :   3.0   Min.   :1.000   Min.   :-1.8000   Min.   :-50.00
##  1st Qu.: 171.0   1st Qu.:1.000   1st Qu.:-0.1000   1st Qu.:-42.70
##  Median : 336.0   Median :2.000   Median : 1.1000   Median :-36.40
##  Mean   : 464.8   Mean   :2.068   Mean   : 0.4737   Mean   :-39.63
##  3rd Qu.: 680.0   3rd Qu.:3.000   3rd Qu.: 1.1000   3rd Qu.:-36.40
##  Max.   :2033.0   Max.   :9.000   Max.   : 1.4000   Max.   :-36.10
##  cons.price.idx    euribor3m      nr.employed
##  Min.   :92.76   Min.   :1.244   Min.   :5099
##  1st Qu.:93.20   1st Qu.:4.343   1st Qu.:5191
##  Median :93.99   Median :4.856   Median :5191
```

```
##  Mean   :93.72   Mean   :4.100   Mean   :5178
##  3rd Qu.:93.99   3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :94.47   Max.   :5.045   Max.   :5228
```

```
df[,var_num ]<-res.impca$completeObs
```

Profiling

Variables numériques

Variables categòriques

```
#edat
res.catdes<-catdes(df,grep("^y$", colnames(df)), proba=0.05)
res.catdes$test.chi2 # relació entre les variables y la variable resposta
```

```
##                  p.value df
## contact     0.000000e+00  1
## month       0.000000e+00  8
## poutcome    4.272175e-70  2
## default     1.928184e-43  1
## day_of_week 3.199213e-31  4
## marital     1.423436e-23  2
## previous    1.085202e-22  1
## age         2.484262e-22  3
## education   3.990582e-22  4
## job         4.547525e-15  6
## housing     7.595941e-11  1
```

```
res.catdes$category
```

```
## $no
##                              Cla/Mod    Mod/Cla Global       p.value     v.test
## month=may                   77.64776 100.000000  66.66  0.000000e+00        Inf
## contact=telephone           81.51181 100.000000  63.50  0.000000e+00        Inf
## poutcome=nonexistent        54.84213 100.000000  94.38  1.396630e-93  20.521050
## default=NA                  68.72990  33.037094  24.88  2.743947e-44  13.959750
## previous=No                 52.68730 100.000000  98.24  5.977949e-29  11.166052
## day_of_week=tue             64.07927  29.984544  24.22  3.778666e-23   9.909677
## marital=married             56.71311  69.204019  63.16  4.208146e-20   9.182598
## education=basic             59.59253  40.687790  35.34  2.212776e-16   8.209952
## job=blue-collar             59.98452  29.945904  25.84  5.702943e-12   6.886880
## housing=no                  56.40922  53.902628  49.46  7.492004e-11   6.510462
## age=Adult                   57.67273  30.641422  27.50  2.478903e-07   5.159288
## day_of_week=mon             56.74419  23.570325  21.50  2.206025e-04   3.694174
## job=services                58.49057  11.978362  10.60  1.022175e-03   3.284351
## job=unemployed              46.55172   9.389490  10.44  1.196767e-02  -2.513096
## job=technician              47.08995  13.755796  15.12  5.330588e-03  -2.786346
## day_of_week=fri             47.14286  15.301391  16.80  3.355222e-03  -2.933168
## job=admin.                  45.16681  20.401855  23.38  2.586795e-07  -5.151305
```

```
## day_of_week=wed             43.47333  14.799073  17.62  5.911229e-08  -5.421466
## day_of_week=thu             42.59819  16.344668  19.86  1.088140e-10  -6.454169
## housing=yes                 47.21013  46.097372  50.54  7.492004e-11  -6.510462
## age=Jove                    27.32558   1.816074   3.44  4.108495e-11  -6.600111
## age=Gran                     0.00000   0.000000   0.80  1.832334e-13  -7.360492
## month=oct                    0.00000   0.000000   0.84  4.189623e-14  -7.554968
## education=university.degree 41.31218  22.140649  27.74  4.843918e-20  -9.167439
## marital=single              39.84674  20.092736  26.10  1.101467e-23 -10.032102
## poutcome=success             0.00000   0.000000   1.68  1.190303e-27 -10.897069
## previous=Yes                 0.00000   0.000000   1.76  5.977949e-29 -11.166052
## month=mar                    0.00000   0.000000   2.52  2.269537e-41 -13.472530
## default=no                  46.13951  66.962906  75.12  2.743947e-44 -13.959750
## month=jun                    0.00000   0.000000   3.76  5.976573e-62 -16.609220
## month=nov                    0.00000   0.000000   3.80  1.276337e-62 -16.701583
## poutcome=failure             0.00000   0.000000   3.94  5.700944e-65 -17.021382
## month=aug                    0.00000   0.000000   5.42  3.904937e-90 -20.131561
## month=jul                    0.00000   0.000000   8.14 5.437191e-138 -25.004681
## month=apr                    0.00000   0.000000   8.84 1.155522e-150 -26.143923
## contact=cellular             0.00000   0.000000  36.50  0.000000e+00       -Inf
##
## $yes
##                              Cla/Mod   Mod/Cla Global       p.value     v.test
## contact=cellular           100.00000 75.663350  36.50  0.000000e+00        Inf
## month=apr                  100.00000 18.325041   8.84 1.155522e-150  26.143923
## month=jul                  100.00000 16.873964   8.14 5.437191e-138  25.004681
## month=aug                  100.00000 11.235489   5.42  3.904937e-90  20.131561
## poutcome=failure           100.00000  8.167496   3.94  5.700944e-65  17.021382
## month=nov                  100.00000  7.877280   3.80  1.276337e-62  16.701583
## month=jun                  100.00000  7.794362   3.76  5.976573e-62  16.609220
## default=no                  53.86049 83.872305  75.12  2.743947e-44  13.959750
## month=mar                  100.00000  5.223881   2.52  2.269537e-41  13.472530
## previous=Yes               100.00000  3.648425   1.76  5.977949e-29  11.166052
## poutcome=success           100.00000  3.482587   1.68  1.190303e-27  10.897069
## marital=single              60.15326 32.545605  26.10  1.101467e-23  10.032102
## education=university.degree 58.68782 33.747927  27.74  4.843918e-20   9.167439
## month=oct                  100.00000  1.741294   0.84  4.189623e-14   7.554968
## age=Gran                   100.00000  1.658375   0.80  1.832334e-13   7.360492
## age=Jove                    72.67442  5.182421   3.44  4.108495e-11   6.600111
## housing=yes                 52.78987 55.306799  50.54  7.492004e-11   6.510462
## day_of_week=thu             57.40181 23.631841  19.86  1.088140e-10   6.454169
## day_of_week=wed             56.52667 20.646766  17.62  5.911229e-08   5.421466
## job=admin.                  54.83319 26.575456  23.38  2.586795e-07   5.151305
## day_of_week=fri             52.85714 18.407960  16.80  3.355222e-03   2.933168
## job=technician              52.91005 16.583748  15.12  5.330588e-03   2.786346
## job=unemployed              53.44828 11.567164  10.44  1.196767e-02   2.513096
## job=services                41.50943  9.121061  10.60  1.022175e-03  -3.284351
## day_of_week=mon             43.25581 19.278607  21.50  2.206025e-04  -3.694174
## age=Adult                   42.32727 24.129353  27.50  2.478903e-07  -5.159288
## housing=no                  43.59078 44.693201  49.46  7.492004e-11  -6.510462
## job=blue-collar             40.01548 21.434494  25.84  5.702943e-12  -6.886880
## education=basic             40.40747 29.601990  35.34  2.212776e-16  -8.209952
## marital=married             43.28689 56.674959  63.16  4.208146e-20  -9.182598
## day_of_week=tue             35.92073 18.034826  24.22  3.778666e-23  -9.909677
## previous=No                 47.31270 96.351575  98.24  5.977949e-29 -11.166052
```

```
## default=NA                   31.27010 16.127695  24.88  2.743947e-44 -13.959750
## poutcome=nonexistent         45.15787 88.349917  94.38  1.396630e-93 -20.521050
## month=may                    22.35224 30.887231  66.66  0.000000e+00       -Inf
## contact=telephone            18.48819 24.336650  63.50  0.000000e+00       -Inf
```

```
res.catdes$quanti  # Global association to numeric variables
```

```
## $no
##                    v.test Mean in category Overall mean sd in category
## cons.conf.idx   53.169902      -36.4000000   -39.626300   0.000000e+00
## cons.price.idx  43.816029       93.9940000    93.722201   0.000000e+00
## euribor3m       38.675184        4.8560696     4.100294   8.352458e-04
## emp.var.rate    37.512348        1.1000000     0.473680   0.000000e+00
## nr.employed     22.081559     5191.0000000  5177.923760   0.000000e+00
## age_num          6.658336       41.0641422    40.159000   8.885746e+00
## na_count         6.469584        0.4741113     0.416200   6.947767e-01
## campaign        -3.462143        2.0011383     2.067662   1.294232e+00
## duration       -42.854275      243.9352241   464.790464   2.007280e+02
##                 Overall sd       p.value
## cons.conf.idx    4.4439989  0.000000e+00
## cons.price.idx   0.4543068  0.000000e+00
## euribor3m        1.4311845  0.000000e+00
## emp.var.rate     1.2228047 5.794591e-308
## nr.employed     43.3698783 4.754022e-108
## age_num          9.9560293  2.769453e-11
## na_count         0.6555742  9.827295e-11
## campaign         1.4072408  5.358914e-04
## duration       377.4405914  0.000000e+00
##
## $yes
##                    v.test Mean in category Overall mean sd in category
## duration        42.854275      701.7611779   464.790464    378.9391840
## campaign         3.462143        2.1390405     2.067662      1.5159317
## na_count        -6.469584        0.3540630     0.416200      0.6045811
## age_num         -6.658336       39.1878109    40.159000     10.9058598
## nr.employed    -22.081559     5163.8933665  5177.923760     59.3196912
## emp.var.rate   -37.512348       -0.1983416     0.473680      1.4923455
## euribor3m      -38.675184        3.2893702     4.100294      1.7249824
## cons.price.idx -43.816029       93.4305701    93.722201      0.5133574
## cons.conf.idx  -53.169902      -43.0880182   -39.626300      4.2174971
##                 Overall sd       p.value
## duration       377.4405914  0.000000e+00
## campaign         1.4072408  5.358914e-04
## na_count         0.6555742  9.827295e-11
## age_num          9.9560293  2.769453e-11
## nr.employed     43.3698783 4.754022e-108
## emp.var.rate     1.2228047 5.794591e-308
## euribor3m        1.4311845  0.000000e+00
## cons.price.idx   0.4543068  0.000000e+00
## cons.conf.idx    4.4439989  0.000000e+00
```

```
res.catdes$quali # Global association to factors
```

```
## NULL
```

Euribor3m: El tipus d'interès a tres mesos (Euribor3m) és la variable més fortament relacionada amb la variable target "y" segons el nostre analisi i tractament de variables. Si l'Euribor3m és baix, és més probable que el client contracti el dipòsit a termini. Si aquest és alt, també és l'influenciador més gran en què el resultat acabi sent negatiu.

Poutcome: La variable Poutcome (resultat de la campanya de màrqueting anterior) també està fortament relacionada amb la variable target "y". Si el resultat de la campanya anterior va ser exitós, és més probable que el client contracti el dipòsit a termini.

Duration: També es veu altament relacionada amb el resultat. Això pot ser degut a que com més temps duri la trucada, és més probable que l'agent de vendes hagi tingut l'oportunitat de persuadir el client i fer-li una oferta més atractiva.

Job: El tipus de treball del client també sembla estar relacionat amb la variable target "y". Els estudiants i els jubilats tenen més probabilitats de contractar el dipòsit a termini, mentre que els treballadors autònoms i els desocupats tenen menys probabilitats.

Mes: El mes en què es va realitzar l'última campanya de màrqueting també sembla estar relacionat amb el resultat negatiu de la variable target "y". En particular, els mesos de maig i juny tenen una taxa de rebuig més alta que altres mesos, mentre que el març, septembre i octubre estan molt relacionats amb un resultat positiu.

Contact: La forma de contacte també està relacionada amb el resultat de la variable target "y". Els clients contactats per telèfon fix tenen més probabilitats de rebutjar el dipòsit a termini que aquells contactats per correu electrònic o per telèfon mòbil.

Age: En general, els clients més joves tenen més probabilitats de rebutjar el dipòsit a termini que els clients més grans.

Campaign: El nombre de contactes realitzats durant l'última campanya de màrqueting també està relacionat amb el resultat negatiu de la variable target "y". En general, com més contactes es realitzin, és més probable que el client rebutgi el dipòsit a termini.

```
#duration
res.condes<-condes(df,grep("^duration$", colnames(df)), proba=0.05)
res.condes$test.chi2 # relació entre les variables y la variable resposta
```

```
## NULL
```

```
res.condes$category
```

```
##                              Estimate       p.value
## y=yes                      228.91298  0.000000e+00
## contact=cellular           165.16324 2.093855e-214
## month=jul                  333.85272 3.965288e-118
## month=aug                  280.72756  5.068452e-57
## month=jun                  341.05236  1.938499e-53
## month=nov                  208.13506  2.463898e-25
## poutcome=failure           106.55349  7.581864e-13
## day_of_week=wed             50.66271  1.397880e-06
## default=no                  21.12415  6.195630e-04
## marital=single              20.58647  2.151955e-03
## day_of_week=thu             26.20140  3.632670e-03
## housing=yes                 11.24417  3.517575e-02
## day_of_week=fri             19.96409  3.575222e-02
## age=Jove-Adult              57.19128  3.944233e-02
```

```
## housing=no                    -11.24417  3.517575e-02
## education=university.degree   -16.52057  1.044162e-02
## month=oct                    -245.16912  2.484858e-03
## marital=married               -19.32388  2.432148e-03
## day_of_week=mon               -36.34532  2.098132e-03
## job=unemployed                -47.45417  1.484963e-03
## age=Gran                     -146.97090  9.341034e-04
## default=NA                    -21.12415  6.195630e-04
## day_of_week=tue               -60.48287  3.926094e-09
## month=mar                    -265.09390  3.802976e-09
## poutcome=nonexistent          -90.99622  3.314190e-12
## month=may                    -178.18275 5.706054e-198
## contact=telephone            -165.16324 2.093855e-214
## y=no                         -228.91298  0.000000e+00
```

```
res.condes$quanti  # Global association to numeric variables
```

```
##                correlation      p.value
## campaign        0.09357119 3.371081e-11
## nr.employed     0.05776931 4.364235e-05
## age_num        -0.05293015 1.808138e-04
## emp.var.rate   -0.08878741 3.193708e-10
## euribor3m      -0.12017447 1.509610e-17
## cons.price.idx -0.18889226 2.199788e-41
## cons.conf.idx  -0.29184702 9.343997e-99
```

```
res.condes$quali # Global association to factors
```

```
##                       R2       p.value
## month       0.2687566917  0.000000e+00
## y           0.3673712587  0.000000e+00
## contact     0.1775236255 2.093855e-214
## day_of_week 0.0126405296  5.186401e-13
## poutcome    0.0108893480  1.316225e-12
## default     0.0023416784  6.195630e-04
## age         0.0027889411  2.967867e-03
## marital     0.0020923915  5.335796e-03
## housing     0.0008873747  3.517575e-02
## job         0.0026763395  3.729645e-02
```

Contacte: La forma de contacte utilitzada en l'última campanya de màrqueting té una alta correlació amb la durada de la trucada. En particular, els clients contactats per telèfon mòbil tendeixen a tenir trucades més curtes que aquells contactats per telèfon. La relació negativa entre "Duration" i "Contact" podria explicar-se pel fet que el correu electrònic i el telèfon mòbil són formes de contacte més breus i concises que una trucada telefònica.

Pdays: La variable "Pdays" representa el número de dies que han passat des que el client va ser contactat per última vegada per a una campanya de màrqueting anterior. Els clients que han estat contactats recentment (és a dir, menor valor en Pdays) tendeixen a tenir trucades més curtes en l'última campanya.

Previous: La variable "Previous" representa el número de contactes realitzats abans de l'última campanya de màrqueting. En general, com més gran sigui el número de contactes, menor serà la durada de l'última trucada. La relació negativa entre "Duration" i "Previous" podria explicar-se per la possibilitat que l'agent de vendes hagi hagut de repetir informació prèviament proporcionada en trucades anteriors.

Job: El tipus de treball del client també pot estar relacionat amb la durada de la trucada. En particular, els clients desocupats i els estudiants tendeixen a tenir trucades més curtes que altres tipus de treballadors. La relació negativa entre "Duration" i "Job" podria explicar-se pel fet que els clients desocupats i els estudiants poden tenir menys ingressos i, per tant, estar menys interessats en contractar un dipòsit a termini fix, el que es reflectiria en trucades més curtes.

```
# select only the factor variables from the dataset
cat_vars <- c("age", "job", "marital", "education", "default", "housing", "loan",
              "contact", "month", "day_of_week", "poutcome", "y")
df_cat <- df[, cat_vars]

# perform multivariate outlier analysis
#outliers <- mvoutlier(df_cat)

# create a new column to identify multivariate outliers
#df_cat$multivar_outlier <- ifelse(rownames(df_cat) %in% outliers, 1, 0)
```