Trabajo módulo n°3 Data Science

Cristian Riquelme Fernández

Presentación del caso

EL presente trabajo toma a modo de ejercicio práctico, un caso extraído desde la plataforma online https://www.kaggle.com. En la descripción de la problemática se expone lo siguiente:

"Los datos están relacionados con campañas de marketing directo de una institución bancaria portuguesa. Sus clientes no estaban invirtiendo lo suficiente para depósitos a largo plazo y esto se estaba plasmando en la disminución de sus ingresos.Las campañas de marketing se basaron en llamadas telefónicas. A menudo, se requería más de un contacto con el mismo cliente, para poder acceder si el producto (depósito a plazo bancario) estaría suscrito ('sí') o no ('no') suscrito."

Extraído desde https://www.kaggle.com/datasets/rashmiranu/banking-dataset-classification?select=new_test.csv

La finalidad del presente análisis es determinar una caracterización de los clientes que suscribieron el depósito a largo plazo.

Análisis exploratorio y preparación del contenido de base de datos

```
dat=read.csv(file.choose(),sep=",",dec=".",header=T)

library(agricolae)

#Revizando la cantidad de variables

names(dat)
```

```
##  [1] "age"         "job"         "marital"     "education"   "default"
##  [6] "housing"     "loan"        "contact"     "month"       "day_of_week"
## [11] "duration"    "campaign"    "pdays"       "previous"    "poutcome"
## [16] "y"
```

```
#Revizando tipo de variables

str(dat)
```

```
## 'data.frame':    32950 obs. of  16 variables:
##  $ age        : int  49 37 78 36 59 29 26 30 50 33 ...
##  $ job        : chr  "blue-collar" "entrepreneur" "retired" "admin." ...
##  $ marital    : chr  "married" "married" "married" "married" ...
##  $ education  : chr  "basic.9y" "university.degree" "basic.4y" "university.degree" ...
##  $ default    : chr  "unknown" "no" "no" "no" ...
```

```
##  $ housing    : chr  "no" "no" "no" "yes" ...
##  $ loan       : chr  "no" "no" "no" "no" ...
##  $ contact    : chr  "cellular" "telephone" "cellular" "telephone" ...
##  $ month      : chr  "nov" "nov" "jul" "may" ...
##  $ day_of_week: chr  "wed" "wed" "mon" "mon" ...
##  $ duration   : int  227 202 1148 120 368 256 449 126 574 498 ...
##  $ campaign   : int  4 2 1 2 2 2 1 2 1 5 ...
##  $ pdays      : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ poutcome   : chr  "nonexistent" "failure" "nonexistent" "nonexistent" ...
##  $ y          : chr  "no" "no" "yes" "no" ...
```

Se Cambian tipos de variables "character" por "factor", para que los lea de forma cualitativa nominal

```
dat$marital = as.factor(dat$marital)
dat$job = as.factor(dat$job)
dat$marital = as.factor(dat$marital)
dat$education = as.factor(dat$education)
dat$default = as.factor(dat$default)
dat$housing = as.factor(dat$housing)
dat$loan = as.factor(dat$loan)
dat$contact = as.factor(dat$contact)
dat$month = as.factor(dat$month)
dat$day_of_week = as.factor (dat$day_of_week)
dat$poutcome = as.factor (dat$poutcome)
dat$y = as.factor(dat$y)

#Sumario a la data

summary(dat)
```

```
##       age                 job           marital                    education
##  Min.   :17.00   admin.     :8314   divorced: 3675   university.degree  :9736
##  1st Qu.:32.00   blue-collar:7441   married :19953   high.school        :7596
##  Median :38.00   technician :5400   single  : 9257   basic.9y           :4826
##  Mean   :40.01   services   :3196   unknown :   65   professional.course:4192
##  3rd Qu.:47.00   management :2345                    basic.4y           :3322
##  Max.   :98.00   retired    :1366                    basic.6y           :1865
##                  (Other)    :4888                    (Other)            :1413
##     default         housing           loan            contact
##  no     :26007   no     :14900   no     :27131   cellular :20908
##  unknown: 6940   unknown:  796   unknown:  796   telephone:12042
##  yes    :    3   yes    :17254   yes    : 5023
##
##
##
##
##      month       day_of_week    duration         campaign          pdays
##  may    :11011   fri:6322    Min.   :   0.0   Min.   : 1.000   Min.   :  0.0
##  jul    : 5763   mon:6812    1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.:999.0
##  aug    : 4948   thu:6857    Median : 180.0   Median : 2.000   Median :999.0
##  jun    : 4247   tue:6444    Mean   : 258.1   Mean   : 2.561   Mean   :962.1
##  nov    : 3266   wed:6515    3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0
```

```
##  apr    : 2085               Max.   :4918.0   Max.   :56.000   Max.   :999.0
##  (Other): 1630
##     previous             poutcome       y
##  Min.   :0.0000   failure    : 3429   no :29238
##  1st Qu.:0.0000   nonexistent:28416   yes: 3712
##  Median :0.0000   success    : 1105
##  Mean   :0.1747
##  3rd Qu.:0.0000
##  Max.   :7.0000
##
```

Análisis descriptivo general

1.Edad

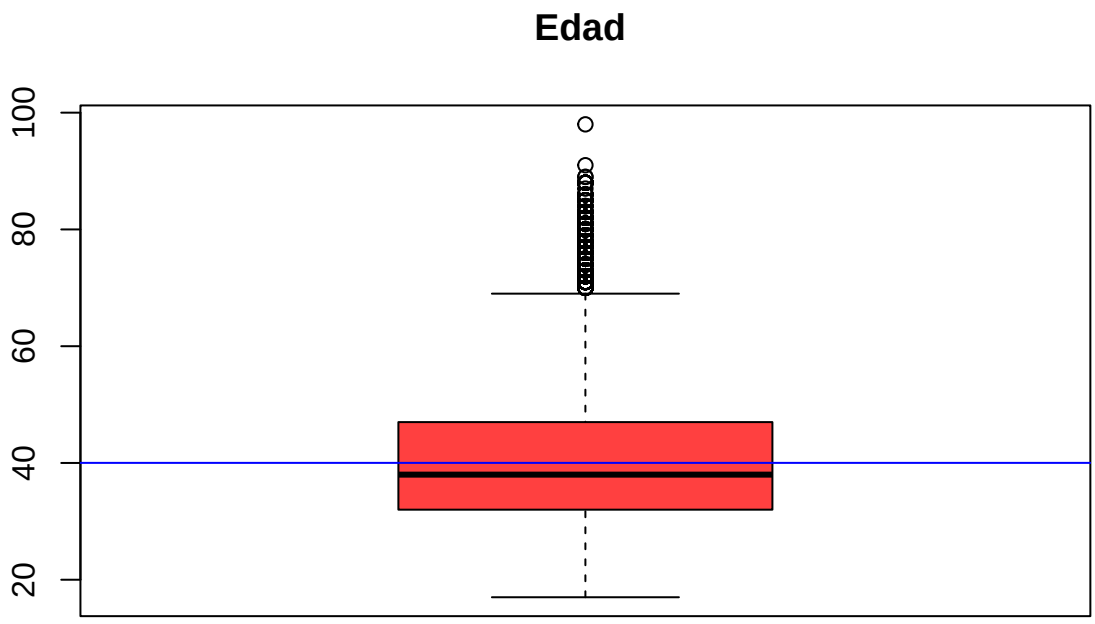
```
summary(dat$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   40.01   47.00   98.00
```

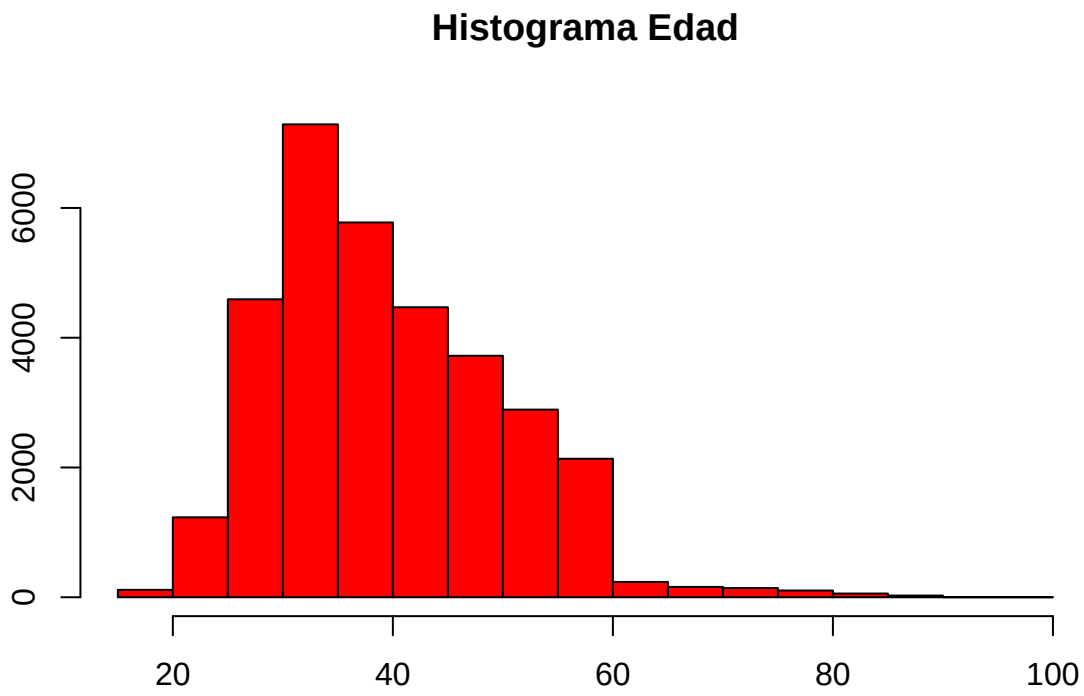
```
sd(dat$age, na.rm = TRUE)
```

```
## [1] 10.40364
```

```
boxplot(dat$age, col="brown1", main="Edad ")
abline(h= mean(dat$age), col="blue")
```




```
HistAge= hist(dat$age, main="Histograma Edad", col="red", xlab="", ylab="")
```




```
tabla = table.freq(HistAge)
tabla
```

```
##    Lower Upper Main Frequency Percentage    CF   CPF
## 1     15    20 17.5       114        0.3   114   0.3
## 2     20    25 22.5      1232        3.7  1346   4.1
## 3     25    30 27.5      4593       13.9  5939  18.0
## 4     30    35 32.5      7289       22.1 13228  40.1
## 5     35    40 37.5      5777       17.5 19005  57.7
## 6     40    45 42.5      4470       13.6 23475  71.2
## 7     45    50 47.5      3722       11.3 27197  82.5
## 8     50    55 52.5      2892        8.8 30089  91.3
## 9     55    60 57.5      2135        6.5 32224  97.8
## 10    60    65 62.5       236        0.7 32460  98.5
## 11    65    70 67.5       159        0.5 32619  99.0
## 12    70    75 72.5       142        0.4 32761  99.4
## 13    75    80 77.5       104        0.3 32865  99.7
## 14    80    85 82.5        57        0.2 32922  99.9
## 15    85    90 87.5        26        0.1 32948 100.0
## 16    90    95 92.5         1        0.0 32949 100.0
## 17    95   100 97.5         1        0.0 32950 100.0
```

Se observa que el 50% de los casos posee una edad entre los 32 hasta 47 años. La mediana es de 38 y la media de 40.1. La distribución posee asimetria positiva, los datos atípicos sobre el límite superior desplazan la media hacia los datos extremos.Al obervar la tabla de frecuencia se identifica que el rango que va de los 30 a 35 años posee mayor cantidad de clientes con 7289 casos.

2.Estado civil

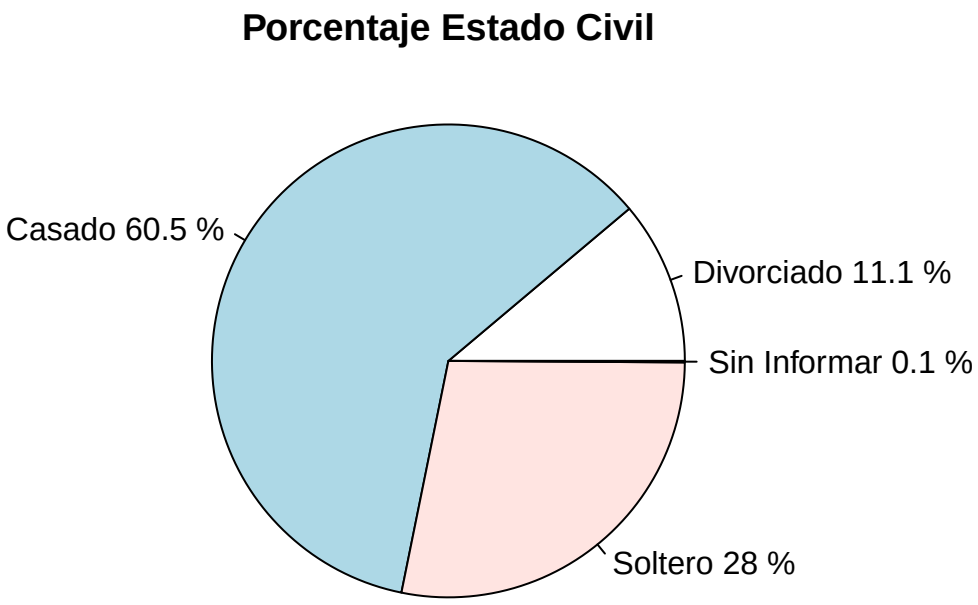
```
summary(dat$marital)
```

```
## divorced  married   single  unknown
##     3675    19953     9257       65
```

```
table(dat$marital)/32950
```

```
##
##    divorced     married      single     unknown
## 0.111532625 0.605553869 0.280940819 0.001972686
```

```
labels = c("Divorciado", "Casado", "Soltero", "Sin Informar")
porcentajes = c(0.111*100, 0.605*100, 0.280*100, 0.001*100)
tags =paste(labels, porcentajes, "%", sep = " ")
pie(porcentajes, labels  =tags, main = "Porcentaje Estado Civil", radius = 1)
```



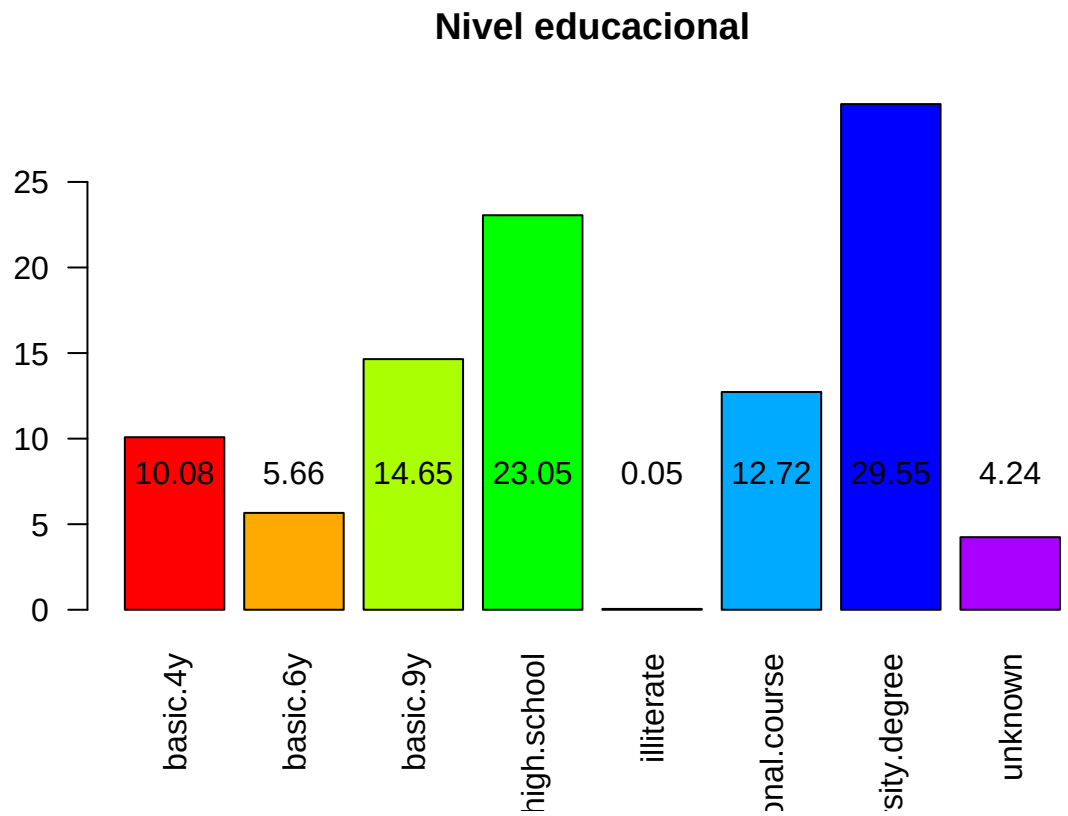
Se observa que la mayoria de la población considerada para el análisis esta casada al representar un 60,5% de los casos, los solteros son un 28%, divorciados 11,1% y sin informar un 0,1%.

3.Nivel educacional

```
educ_barr= table(dat$education)
educ_barr
```

```
##
##           basic.4y           basic.6y           basic.9y        high.school
##               3322               1865               4826               7596
##         illiterate professional.course   university.degree            unknown
##                 16               4192               9736               1397
```

```
porce =prop.table(educ_barr)*100
educ_bar =barplot(porce, las=2, main = "Nivel educacional", col= rainbow(9))
text(educ_bar, c(8), round(porce,2))
```



Se observa que la mayoría de la población considerada para el análisis posee estudios universitarios al representar un 29,5% de los casos, high school representan un 23,05%, enseñanza básica completa 14,65%

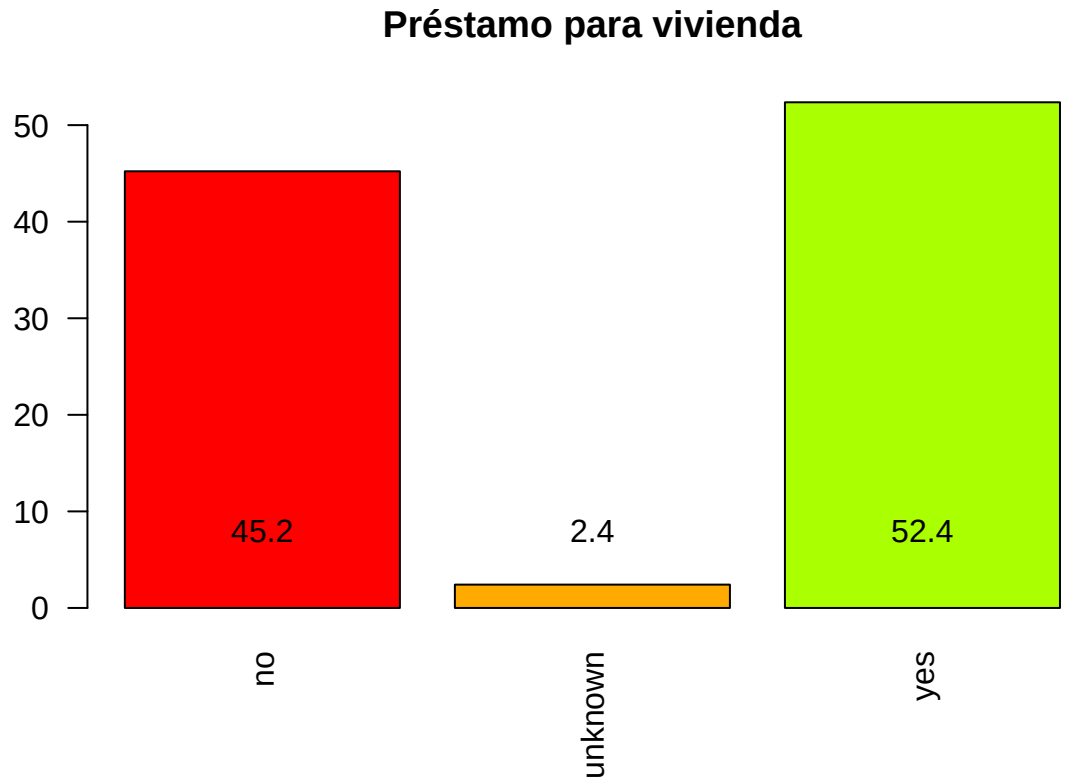
4.Clientes con algún prestamo para vivienda

```
hous_bar= table(dat$housing)
hous_bar
```

```
##
```

```
##      no unknown     yes
##   14900     796   17254
```

```
porce =prop.table(hous_bar)*100
hous_barr =barplot(porce, las=2, main = "Préstamo para vivienda", col= rainbow(9))
text(hous_barr, c(8), round(porce,1))
```



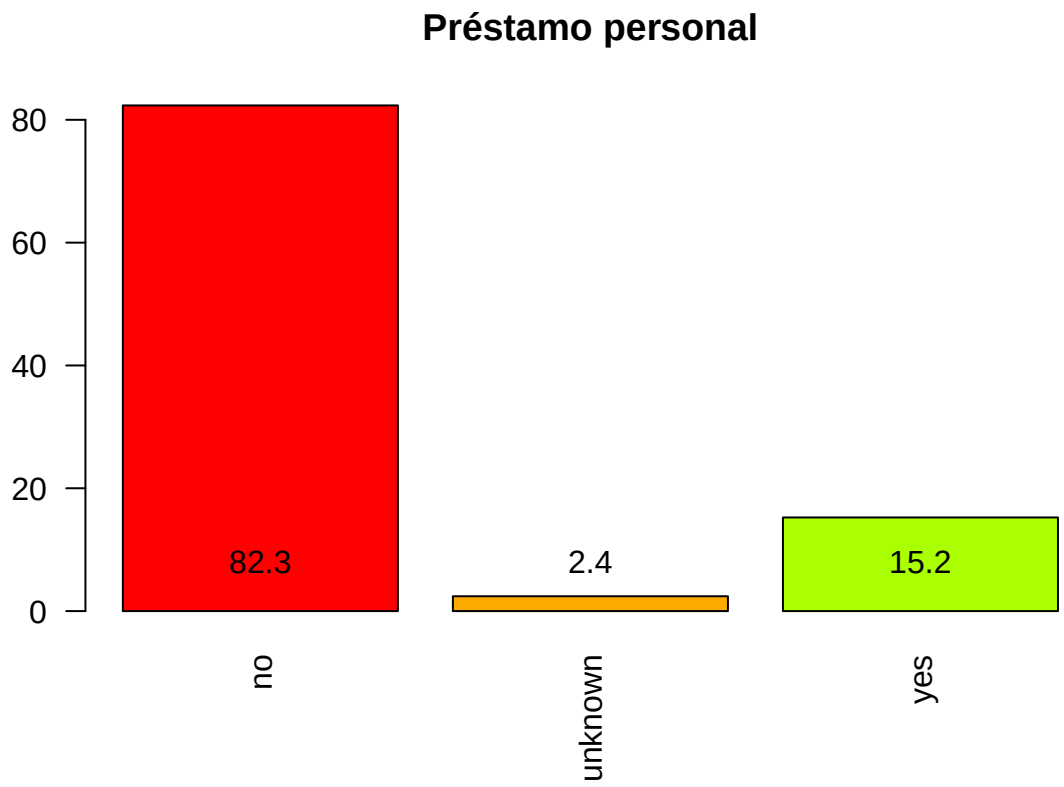
Al consultar por si los clientes del banco poseen algún prestamo para vivienda, se observa que; 52,4% señala sí tener y un 45,2% señala no tener.

5.Clientes con algún préstamo personal

```
loan_bar= table(dat$loan)
loan_bar
```

```
##
##      no unknown     yes
##   27131     796    5023
```

```
porce =prop.table(loan_bar)*100
loan_barr =barplot(porce, las=2, main = "Préstamo personal", col= rainbow(9))
text(loan_barr, c(8), round(porce,1))
```



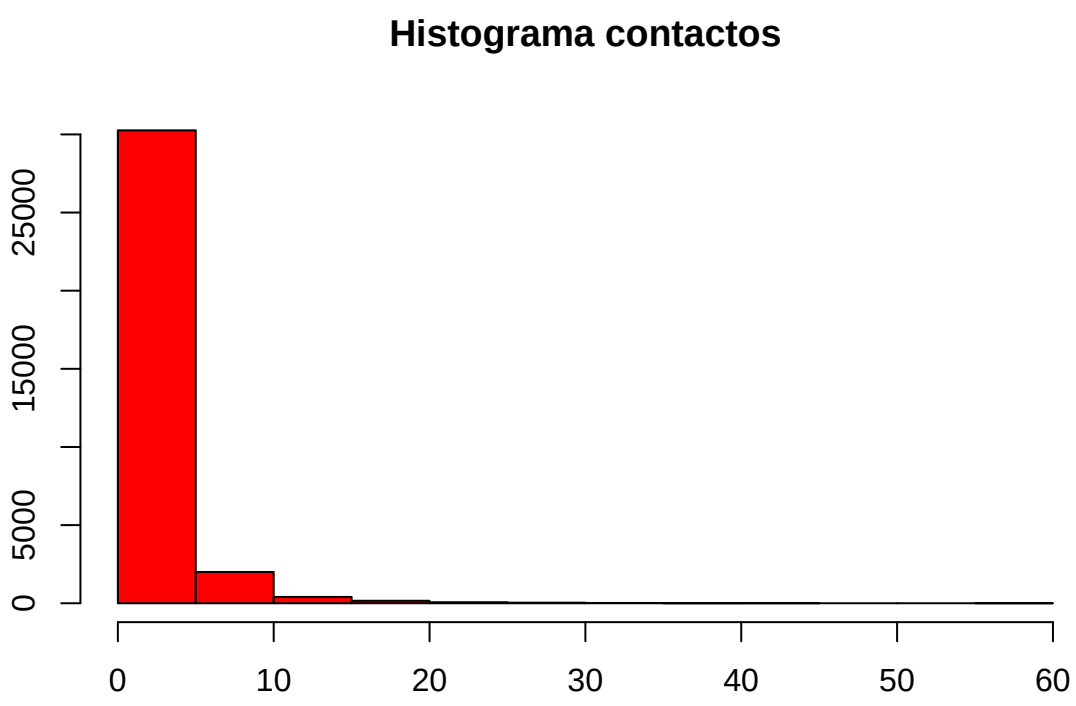
Al consultar por si los clientes del banco poseen algún prestamo personal, se observa que; 82,3% señala no tener y un 15,2% señala sí tener.

6.Cantidad de contactos durante la campaña (con histograma para ver tabla de frecuencia)

```
summary(dat$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.561   3.000  56.000
```

```
Histcon= hist(dat$campaign, main="Histograma contactos", col="red", xlab="", ylab="")
```




```
tabla = table.freq(Histcon)
tabla
```

```
##    Lower Upper Main Frequency Percentage    CF   CPF
## 1      0     5  2.5     30261       91.8 30261  91.8
## 2      5    10  7.5      2002        6.1 32263  97.9
## 3     10    15 12.5       405        1.2 32668  99.1
## 4     15    20 17.5       158        0.5 32826  99.6
## 5     20    25 22.5        65        0.2 32891  99.8
## 6     25    30 27.5        35        0.1 32926  99.9
## 7     30    35 32.5        17        0.1 32943 100.0
## 8     35    40 37.5         2        0.0 32945 100.0
## 9     40    45 42.5         4        0.0 32949 100.0
## 10    45    50 47.5         0        0.0 32949 100.0
## 11    50    55 52.5         0        0.0 32949 100.0
## 12    55    60 57.5         1        0.0 32950 100.0
```

Se observa que el 50% de los clientes fueron contactados durante la campaña entre 1 a 3 veces. La mediana es de 2 contactos y la media de 2.56.

7.Medio de contacto

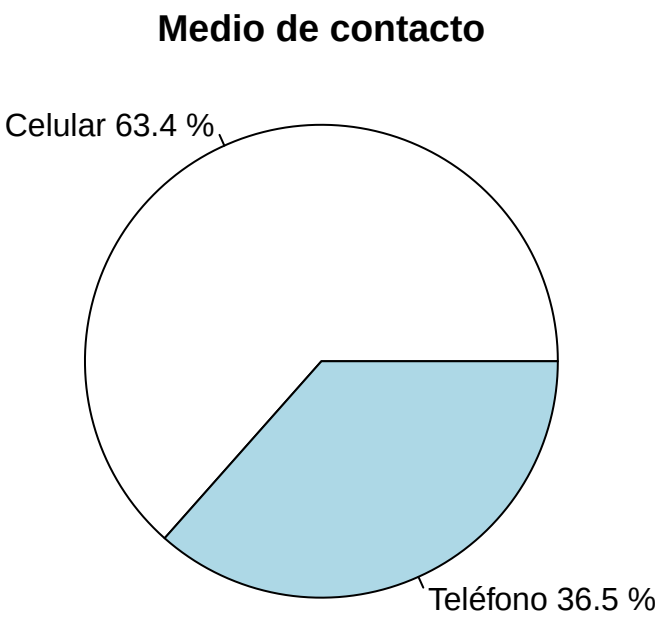
```
contact_pie= table(dat$contact)
contact_pie
```

```
##
##  cellular telephone
##     20908     12042
```

```
table(dat$contact)/32950
```

```
##
##  cellular telephone
## 0.6345372 0.3654628
```

```
labels = c("Celular", "Teléfono")
porcentajes = c(0.634*100, 0.365*100)
tags =paste(labels, porcentajes, "%", sep = " ")
pie(porcentajes, labels  =tags, main = "Medio de contacto", radius = 1)
```



Sobre el medio de contacto que se utilizó para realizar la campaña, un 63,4% fue contactado por celular y un 36,5% a través de teléfono fijo.

8.Mes del último contacto

```
summary(dat$month)
```

```
##   apr   aug   dec   jul   jun   mar   may   nov   oct   sep
##  2085  4948   143  5763  4247   436 11011  3266   587   464
```

9.Día del último contacto

```
summary(dat$day_of_week)
```

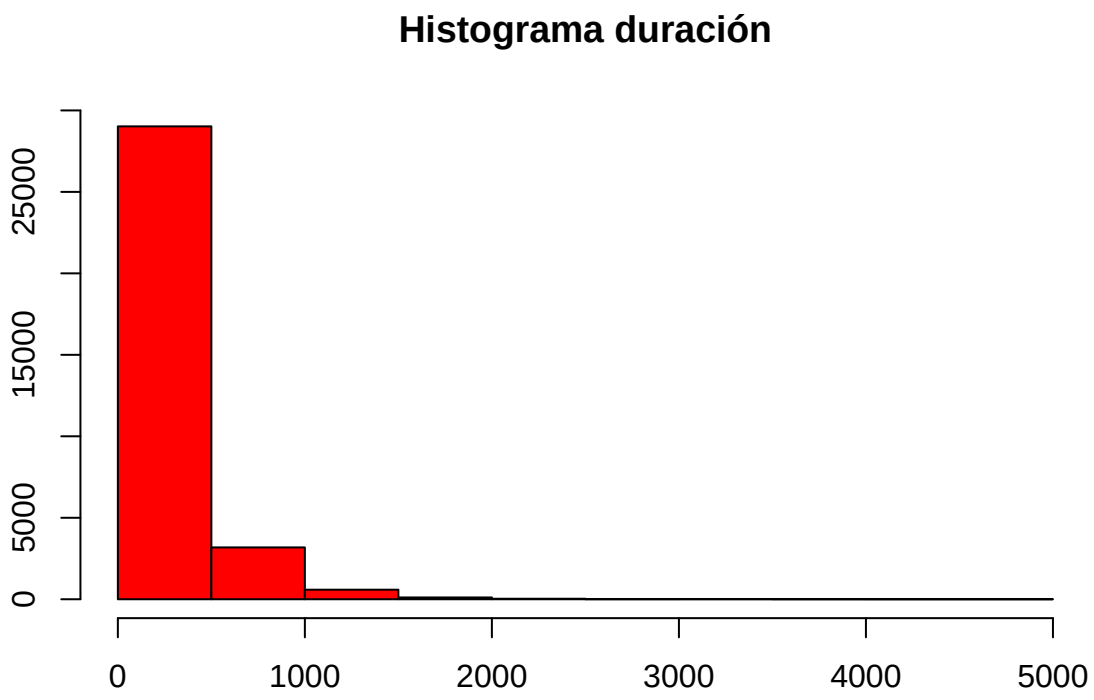
```
##  fri  mon  thu  tue  wed
## 6322 6812 6857 6444 6515
```

10.Duración de la llamada en segundos (con histograma para ver tabla de frecuencia)

```
summary(dat$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   103.0   180.0   258.1   319.0  4918.0
```

```
Histdurt= hist(dat$duration, main="Histograma duración", col="red", xlab="", ylab="")
```



```
tabla = table.freq(Histdurt)
tabla
```

```
##    Lower Upper Main Frequency Percentage    CF   CPF
## 1      0   500  250     29022       88.1 29022  88.1
## 2    500  1000  750      3181        9.7 32203  97.7
## 3   1000  1500 1250       586        1.8 32789  99.5
## 4   1500  2000 1750       110        0.3 32899  99.8
## 5   2000  2500 2250        31        0.1 32930  99.9
## 6   2500  3000 2750         7        0.0 32937 100.0
## 7   3000  3500 3250         8        0.0 32945 100.0
## 8   3500  4000 3750         3        0.0 32948 100.0
## 9   4000  4500 4250         1        0.0 32949 100.0
## 10  4500  5000 4750         1        0.0 32950 100.0
```

Se observa que el 50% de las llamadas registró una duración entre los 103 hasta 319 segundos (entre 1,7 a 5,3 minutos de duración). Al observar la tabla de frecuencia se obtiene que al menos 5 llamadas duraron más de una hora(4000 segundos), y hubo una gran cantidad de personas que no contestaron (0 segundos) o tuvieron una llamada de corta duración.

11.Resultado de la campaña de marketing anterior

```
summary(dat$poutcome)
```

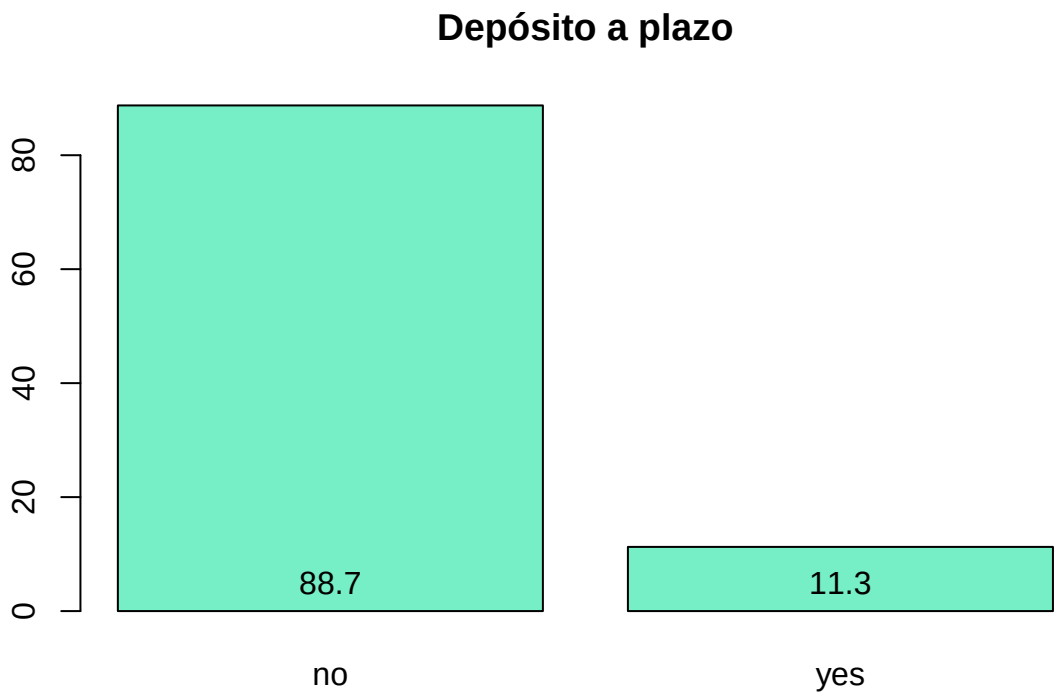
```
##     failure nonexistent     success
##        3429       28416        1105
```

12.¿El cliente ha suscrito un depósito a plazo? ('sí No')

```
dep_pla= table(dat$y)
dep_pla
```

```
##
##    no   yes
## 29238  3712
```

```
porce =prop.table(dep_pla)*100
dep_bar =barplot(porce, main = "Depósito a plazo", col=  "aquamarine2")
text(dep_bar, c(5,5), round(porce,1))
```



Finalmente, del total de clientes contactados durante el periodo de campaña un 11,3% suscribió el producto de interés para el banco y un 88,7% lo rechazó.

Se generan bases de datos separadas para breve análisis estratificado por quienes suscribieron o no el depósito a plazo de la campaña

```
dat_depyes= dat[dat$y=="yes",]
dat_depno= dat[dat$y=="no",]
summary(dat_depyes)
```

```
##       age                job          marital                   education
##  Min.   :17.00   admin.     :1070   divorced: 371   university.degree  :1345
##  1st Qu.:31.00   technician : 585   married :2020   high.school        : 815
##  Median :37.00   blue-collar: 515   single  :1310   professional.course: 473
##  Mean   :40.85   retired    : 348   unknown :  11   basic.9y           : 369
##  3rd Qu.:50.00   management : 269                   basic.4y           : 344
##  Max.   :98.00   services   : 254                   unknown            : 207
##                  (Other)    : 671                   (Other)            : 159
##     default        housing          loan           contact         month
##  no     :3351   no     :1628   no     :3058   cellular :3074   may    :699
##  unknown: 361   unknown:  90   unknown:  90   telephone: 638   jul    :532
##  yes    :   0   yes    :1994   yes    : 564                   aug    :531
##                                                              jun    :441
##                                                              apr    :427
```

```
##                                                                 nov    :332
##                                                                 (Other):750
##  day_of_week    duration         campaign          pdays          previous
##  fri:690     Min.   :  63.0   Min.   : 1.000   Min.   :  0.0   Min.   :0.0000
##  mon:698     1st Qu.: 252.0   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000
##  thu:823     Median : 448.0   Median : 2.000   Median :999.0   Median :0.0000
##  tue:748     Mean   : 549.4   Mean   : 2.055   Mean   :790.3   Mean   :0.4965
##  wed:753     3rd Qu.: 737.0   3rd Qu.: 2.000   3rd Qu.:999.0   3rd Qu.:1.0000
##              Max.   :4199.0   Max.   :23.000   Max.   :999.0   Max.   :6.0000
##
##        poutcome      y
##  failure    : 494   no :   0
##  nonexistent:2501   yes:3712
##  success    : 717
##
##
##
##
```

```
summary(dat_depno)
```

```
##       age                   job           marital
##  Min.   :17.00   admin.      :7244   divorced: 3304
##  1st Qu.:32.00   blue-collar :6926   married :17933
##  Median :38.00   technician  :4815   single  : 7947
##  Mean   :39.91   services    :2942   unknown :   54
##  3rd Qu.:47.00   management  :2076
##  Max.   :91.00   entrepreneur:1060
##                  (Other)     :4175
##                education       default         housing           loan
##  university.degree  :8391   no     :22656   no     :13272   no     :24073
##  high.school        :6781   unknown: 6579   unknown:  706   unknown:  706
##  basic.9y           :4457   yes    :    3   yes    :15260   yes    : 4459
##  professional.course:3719
##  basic.4y           :2978
##  basic.6y           :1709
##  (Other)            :1203
##       contact          month       day_of_week    duration
##  cellular :17834   may    :10312   fri:5632    Min.   :   0.0
##  telephone:11404   jul    : 5231   mon:6114    1st Qu.:  95.0
##                    aug    : 4417   thu:6034    Median : 164.0
##                    jun    : 3806   tue:5696    Mean   : 221.1
##                    nov    : 2934   wed:5762    3rd Qu.: 279.0
##                    apr    : 1658               Max.   :4918.0
##                    (Other):  880
##     campaign          pdays          previous            poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   failure    : 2935
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:25915
##  Median : 2.000   Median :999.0   Median :0.0000   success    :  388
##  Mean   : 2.625   Mean   :983.9   Mean   :0.1339
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :56.000   Max.   :999.0   Max.   :7.0000
##
##    y
```

```
##  no :29238
##  yes:    0
##
##
##
##
##
```

Edad por estrato

```
summary(dat_depyes$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   31.00   37.00   40.85   50.00   98.00
```

```
summary(dat_depno$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   39.91   47.00   91.00
```

Estado civil por estrato

```
summary(dat_depyes$marital)
```

```
## divorced  married   single  unknown
##      371     2020     1310       11
```

```
summary(dat_depno$marital)
```

```
## divorced  married   single  unknown
##     3304    17933     7947       54
```

Nivel educacional por estrato

```
summary(dat_depyes$education)
```

```
##            basic.4y            basic.6y            basic.9y         high.school
##                 344                 156                 369                 815
##          illiterate professional.course   university.degree             unknown
##                   3                 473                1345                 207
```

```
summary(dat_depno$education)
```

```
##            basic.4y            basic.6y            basic.9y         high.school
##                2978                1709                4457                6781
##          illiterate professional.course   university.degree             unknown
##                  13                3719                8391                1190
```

Tipo de trabajo por estrato

```
summary(dat_depyes$job)
```

```
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1070           515           100            86           269
##       retired self-employed      services       student    technician
##           348           119           254           217           585
##    unemployed       unknown
##           116            33
```

```
summary(dat_depno$job)
```

```
##        admin.   blue-collar  entrepreneur     housemaid    management
##          7244          6926          1060           769          2076
##       retired self-employed      services       student    technician
##          1018           980          2942           494          4815
##    unemployed       unknown
##           682           232
```

Duración de llamadas en segundos por estrato

```
summary(dat_depyes$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    63.0   252.0   448.0   549.4   737.0  4199.0
```

```
summary(dat_depno$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    95.0   164.0   221.1   279.0  4918.0
```

Conclusiones

De la clientela del banco contactada durante el periodo de campaña que sí acepto el producto ofrecido, se pueden extraer ciertas características que ayudarían a enfocar las estratégias de marketing a futuro. La media de la edad es de 41 años, la mayoría indica por estado civíl estar casado/a con 2020 casos, poseen estudios universitarios con 1345 casos, laboralmente pertenecen al sector administrativo con 1070. Respecto a las caracteristicas de la campaña, en cuanto a la duración de las llamadas, en este grupo hubo una media de 549 segundos (9,1 minutos) y un valor máximo de 419 segundos (1 hora con 9 minutos). Esta información podría porporcionar de manera básica una caracterización del tipo de cliente esperado por el banco.