Exercises Session 18/12/19: Profiling

Daniel Zanchetta and Lais Zanchetta

24/12/2019

1. Lea este fichero y efectúe un “summary” de los datos. ¿Detecta algún error o inconsistencia?. Si es así, corríjalo.

```
churn <- read.delim(file = file.choose(),header = TRUE,sep = "")
summary(churn)
```

```
##       Baja                edatcat              sexo          antig
##  Baja NO:1000   edatcat 16-17: 25   HOMBRE      :1134   Min.   : 3.00
##  Baja SI:1000   edatcat 18-25:128   No informado: 866   1st Qu.:13.00
##                 edatcat 26-35:345                       Median :18.00
##                 edatcat 36-45:343                       Mean   :17.38
##                 edatcat 46-55:306                       3rd Qu.:21.00
##                 edatcat 56-65:261                       Max.   :99.00
##                 edatcat 66.. :592
##        Nomina           Pension             Debito_normal
##  Nomina NO:1393   Pension NO:1262   Debito normal NO:1912
##  Nomina SI: 603   Pension SI: 732   Debito normal SI:  88
##  NA's     :   4   NA's      :   6
##
##
##
##
##           Debito_aff        VISA              VISA_aff         MCard
##  Debito aff. NO:1485   VISA NO:1543   VISA aff. NO:1965   MCard
NO:1955
##  Debito aff. SI: 513   VISA SI: 456   VISA aff. SI:  35   MCard SI:
44
##  NA's          :   2   NA's   :   1                       NA's    :
1
##
##
##
##
##       Amex       Total_activo    Total_Plazo      Total_Inversion
##  Amex NO:1987   Min.   :    0   Min.   :    0.0   Min.   :    0.0
##  Amex SI:  13   1st Qu.:    0   1st Qu.:    0.0   1st Qu.:    0.0
##                 Median :    0   Median :    0.0   Median :    0.0
##                 Mean   :  618   Mean   : 1332.7   Mean   :  853.1
##                 3rd Qu.:    0   3rd Qu.:  472.2   3rd Qu.:    0.0
##                 Max.   :32772   Max.   :43400.0   Max.   :62017.0
##
##  Total_Seguros    Total_Vista                  dif_resid
```

```
##  Min.   :    0   Min.   :    0.00   cambio resid. NO:1982
##  1st Qu.:    0   1st Qu.:   51.75   cambio resid. SI:  18
##  Median :    0   Median :  206.00
##  Mean   :  279   Mean   :  569.17
##  3rd Qu.:    0   3rd Qu.:  657.00
##  Max.   :45455   Max.   :12738.00
##
##  oper_caj_Libreta    oper_ven_Libreta        dif_CC
##  Min.   :-1157.500   Min.   :-6378.260   Min.   :-3312.54
##  1st Qu.:    0.000   1st Qu.:   -6.750   1st Qu.:    0.00
##  Median :    0.000   Median :    0.000   Median :    0.00
##  Mean   :   -7.404   Mean   :    2.541   Mean   :   26.93
##  3rd Qu.:    5.000   3rd Qu.:   51.562   3rd Qu.:    0.00
##  Max.   :  774.750   Max.   : 5038.670   Max.   : 9715.28
##
##   dif_Libreta           dif_Plazo          dif_Ahorro
##  Min.   :-11811.900   Min.   :-15000.0   Min.   :-24208.000
##  1st Qu.:   -56.910   1st Qu.:     0.0   1st Qu.:     0.000
##  Median :     1.765   Median :     0.0   Median :     0.000
##  Mean   :   -41.937   Mean   :   114.9   Mean   :     7.051
##  3rd Qu.:    98.000   3rd Qu.:     0.0   3rd Qu.:     0.000
##  Max.   : 12737.000   Max.   : 27000.0   Max.   :  4008.000
##
##  dif_Largo_plazo     dif_Fondos_inv     dif_Seguros
##  Min.   :-15913.04   Min.   :-7746.1   Min.   :-3905.05
##  1st Qu.:     0.00   1st Qu.:    0.0   1st Qu.:    0.00
##  Median :     0.00   Median :    0.0   Median :    0.00
##  Mean   :    26.11   Mean   :  261.8   Mean   :   17.82
##  3rd Qu.:     0.00   3rd Qu.:    0.0   3rd Qu.:    0.00
##  Max.   : 10071.00   Max.   :62017.0   Max.   :19461.00
##
##  dif_Planes_pension  dif_Hipoteca       dif_Prest_personales
##  Min.   :-8246.55   Min.   :-26654.00   Min.   :-8676.00
##  1st Qu.:    0.00   1st Qu.:     0.00   1st Qu.:    0.00
##  Median :    0.00   Median :     0.00   Median :    0.00
##  Mean   :  -39.86   Mean   :    13.28   Mean   :   17.51
##  3rd Qu.:    0.00   3rd Qu.:     0.00   3rd Qu.:    0.00
##  Max.   :    0.00   Max.   : 32772.00   Max.   : 6741.00
##
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)

churn_tidy <- churn %>% drop_na() %>%
  separate(Baja, into = c("Baja_Rem","Baja"), sep = " ", extra = "merge",
fill = "left") %>%
  separate(edatcat, into = c("edatcat_Rem",
"edatcat","edatcat_Rem2","edatcat_Rem3"), sep = "([\\ \\.\\.])", extra =
"merge", fill = "right") %>%
  separate(Nomina, into = c("Nomina_Rem", "Nomina"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(Pension, into = c("Pension_Rem", "Pension"), sep = " ", extra
= "merge", fill = "left") %>%
  separate(Debito_normal, into =
c("Debito_normal_Rem","Debito_normal_Rem2", "Debito_normal"), sep = "([\\
\\ ])", extra = "merge", fill = "left") %>%
  separate(Debito_aff, into = c("Debito_aff_Rem","Debito_aff_Rem2",
"Debito_aff"), sep = "([\\ \\. ])", extra = "merge", fill = "left") %>%
  separate(VISA, into = c("VISA_Rem", "VISA"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(VISA_aff, into = c("VISA_aff_Rem","VISA_aff_Rem2",
"VISA_aff"), sep = "([\\ \\. ])", extra = "merge", fill = "left") %>%
  separate(MCard, into = c("MCard_Rem", "MCard"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(Amex, into = c("Amex_Rem", "Amex"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(dif_resid, into = c("dif_resid_Rem","dif_resid_Rem2",
"dif_resid"), sep = "([\\ \\. ])", extra = "merge", fill = "left") %>%
  select(-c("Baja_Rem",
"edatcat_Rem","edatcat_Rem2","edatcat_Rem3","Nomina_Rem","Pension_Rem","D
ebito_normal_Rem","Debito_normal_Rem2","Debito_aff_Rem","Debito_aff_Rem2"
,"VISA_Rem","VISA_aff_Rem","VISA_aff_Rem2","MCard_Rem","Amex_Rem","dif_re
sid_Rem","dif_resid_Rem2"))

head(churn_tidy)

##   Baja edatcat   sexo antig Nomina Pension Debito_normal Debito_aff
VISA
## 1   NO   46-55 HOMBRE    15     NO      NO            NO         NO
NO
## 2   NO   46-55 HOMBRE    19     NO      NO            SI         NO
NO
## 3   NO   36-45 HOMBRE    15     SI      NO            NO         NO
SI
## 4   NO   36-45 HOMBRE    18     SI      NO            NO         NO
SI
## 5   NO   36-45 HOMBRE    21     SI      NO            NO         SI
```

```
NO
## 6   NO      66 HOMBRE    11     NO      SI             NO          NO
NO
##   VISA_aff MCard Amex Total_activo Total_Plazo Total_Inversion
## 1       NO    NO   NO         1025           0               0
## 2       NO    NO   NO            0           0               0
## 3       NO    NO   NO        11101       10000               0
## 4       NO    NO   NO           57           0               0
## 5       NO    NO   NO            0           0              40
## 6       NO    NO   NO            0           0               0
##   Total_Seguros Total_Vista dif_resid oper_caj_Libreta
oper_ven_Libreta
## 1             0          10        NO             0.00
0.00
## 2           850           0        NO             0.00          -
65.00
## 3             0        1492        SI           351.50
0.00
## 4             0          54        NO           -21.75          -
79.75
## 5             0          82        NO           346.16
132.00
## 6             0           0        NO             0.00
0.00
##    dif_CC dif_Libreta dif_Plazo dif_Ahorro dif_Largo_plazo
dif_Fondos_inv
## 1    0.00      -12.62         0          0               0
0
## 2    0.00      -21.69         0          0               0
0
## 3 1548.22        0.00     10000          0               0
0
## 4    0.00      -34.19         0          0               0
0
## 5   -0.01     -225.54         0          0               0
0
## 6    0.00        0.00         0          0               0
0
##   dif_Seguros dif_Planes_pension dif_Hipoteca dif_Prest_personales
## 1        0.00                  0          101                    0
## 2       -0.41                  0            0                    0
## 3        0.00                  0        11101                    0
## 4        0.00                  0            0                    0
## 5        0.00                  0            0                    0
## 6        0.00                  0            0                    0

#Summary with the tidy dataframe
summary(churn_tidy)
```

```
##      Baja              edatcat                    sexo          antig
##  Length:1993        Length:1993        HOMBRE      :1130   Min.   :
3.00
##  Class :character   Class :character   No informado: 863   1st
Qu.:13.00
##  Mode  :character   Mode  :character                       Median
:18.00
##                                                           Mean
:17.38
##                                                           3rd
Qu.:21.00
##                                                           Max.
:99.00
##     Nomina            Pension          Debito_normal
##  Length:1993        Length:1993        Length:1993
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Debito_aff            VISA             VISA_aff
##  Length:1993        Length:1993        Length:1993
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     MCard               Amex            Total_activo      Total_Plazo
##  Length:1993        Length:1993        Min.   :    0.0   Min.   :    0
##  Class :character   Class :character   1st Qu.:    0.0   1st Qu.:    0
##  Mode  :character   Mode  :character   Median :    0.0   Median :    0
##                                        Mean   :  616.8   Mean   : 1322
##                                        3rd Qu.:    0.0   3rd Qu.:  431
##                                        Max.   :32772.0   Max.   :43400
##  Total_Inversion   Total_Seguros    Total_Vista       dif_resid
##  Min.   :    0.0   Min.   :    0   Min.   :    0.0   Length:1993
##  1st Qu.:    0.0   1st Qu.:    0   1st Qu.:   51.0   Class :character
##  Median :    0.0   Median :    0   Median :  206.0   Mode  :character
##  Mean   :  853.3   Mean   :  280   Mean   :  565.4
##  3rd Qu.:    0.0   3rd Qu.:    0   3rd Qu.:  655.0
##  Max.   :62017.0   Max.   :45455   Max.   :12738.0
##  oper_caj_Libreta    oper_ven_Libreta        dif_CC
##  Min.   :-1157.500   Min.   :-6378.260   Min.   :-3312.54
##  1st Qu.:    0.000   1st Qu.:   -6.750   1st Qu.:    0.00
##  Median :    0.000   Median :    0.000   Median :    0.00
##  Mean   :   -7.765   Mean   :    2.691   Mean   :   26.87
##  3rd Qu.:    5.000   3rd Qu.:   52.500   3rd Qu.:    0.00
##  Max.   :  774.750   Max.   : 5038.670   Max.   : 9715.28
##   dif_Libreta          dif_Plazo          dif_Ahorro
##  Min.   :-11811.90   Min.   :-15000.0   Min.   :-24208.000
```

```
##  1st Qu.:   -57.25   1st Qu.:     0.0   1st Qu.:     0.000
##  Median :     1.56   Median :     0.0   Median :     0.000
##  Mean   :   -47.32   Mean   :   102.2   Mean   :     7.076
##  3rd Qu.:    95.61   3rd Qu.:     0.0   3rd Qu.:     0.000
##  Max.   : 12737.00   Max.   : 27000.0   Max.   :  4008.000
##  dif_Largo_plazo     dif_Fondos_inv   dif_Seguros
dif_Planes_pension
##  Min.   :-15913.04   Min.   :-7746   Min.   :-3905.05   Min.   :-8247
##  1st Qu.:     0.00   1st Qu.:    0   1st Qu.:    0.00   1st Qu.:    0
##  Median :     0.00   Median :    0   Median :    0.00   Median :    0
##  Mean   :    26.15   Mean   :  260   Mean   :   17.88   Mean   :  -40
##  3rd Qu.:     0.00   3rd Qu.:    0   3rd Qu.:    0.00   3rd Qu.:    0
##  Max.   : 10071.00   Max.   :62017   Max.   :19461.00   Max.   :    0
##   dif_Hipoteca       dif_Prest_personales
##  Min.   :-26654.00   Min.   :-8676.00
##  1st Qu.:     0.00   1st Qu.:    0.00
##  Median :     0.00   Median :    0.00
##  Mean   :     9.94   Mean   :   17.57
##  3rd Qu.:     0.00   3rd Qu.:    0.00
##  Max.   : 32772.00   Max.   : 6741.00
```

2. Especifique cuál es la variable de respuesta y cuáles son las explicativas y el tipo de todas ellas.

Resp.: La variable de respuesta es la "Baja". Todas las demás variables del dataset son calificadas como explicativas. Abajo son todas las variables con sus respectivos tipos:

Variable Respuesta:

- "Baja" - Categórica

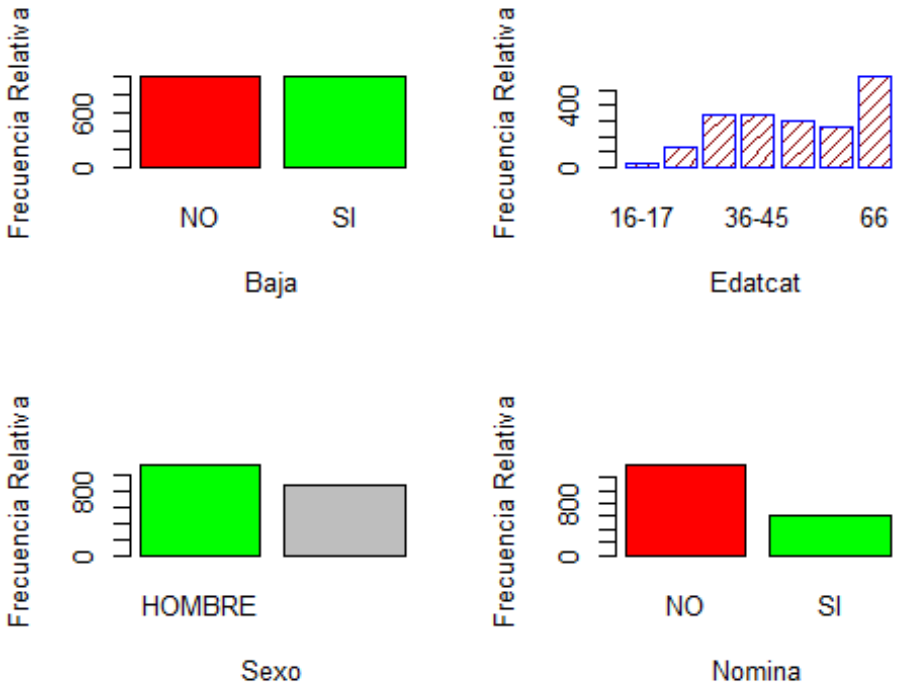
Variables explicativas:

- "edatcat" – Categórica
- "sexo" - Categórica
- "antig" - Continua
- "Nomina" - Categórica
- "Pension" - Categórica
- "Debito_normal" - Categórica
- "Debito_aff" - Categórica
- "VISA" - Categórica
- "VISA_aff" – Categórica
- "MCard" - Categórica
- "Amex" - Categórica
- "Total_activo" - Continua
- "Total_Plazo" - Continua
- "Total_Inversion" - Continua
- "Total_Seguros" - Continua
- "Total_Vista" - Continua
- "dif_resid" - Categórica

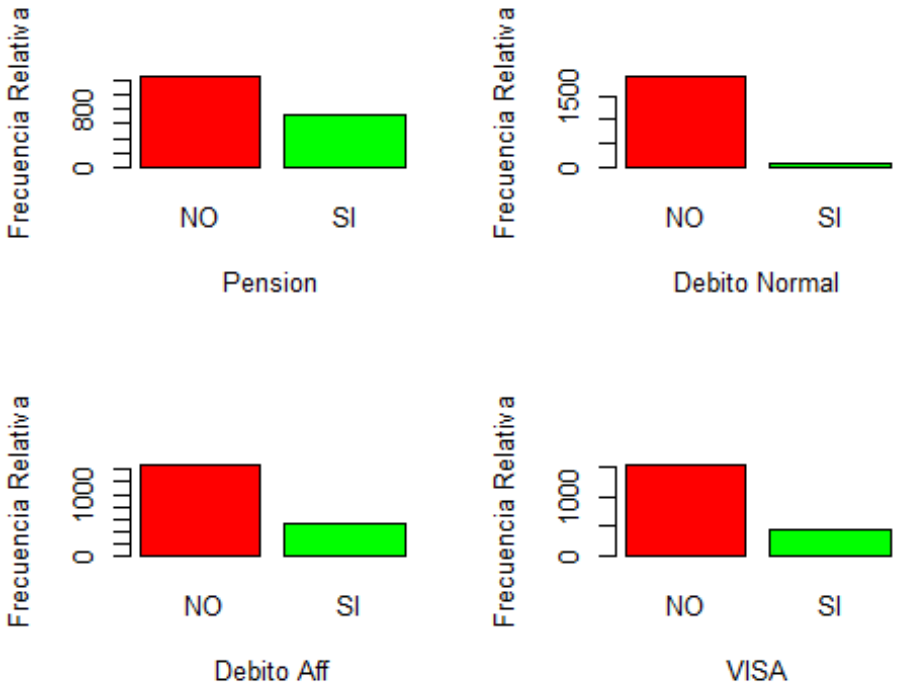
- “oper_caj_Libreta” - Continua
- “oper_ven_Libreta” - Continua
- “dif_CC” - Continua
- “dif_Libreta” - Continua
- “dif_Plazo” - Continua
- “dif_Ahorro” - Continua
- “dif_Largo_plazo” - Continua
- “dif_Fondos_inv” - Continua
- “dif_Seguros” - Continua
- “dif_Planes_pension” - Continua
- “dif_Hipoteca” - Continua
- “dif_Prest_personales” - Continua

3. Efectúe una gráfica de los datos; un diagrama de barras para las variables categóricas y un histograma para las variables continuas.

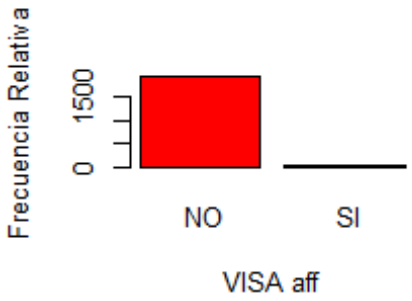
```
#Grafica para variables tipo categórica
par(mfrow=c(2, 2))
baja <- table(churn_tidy$Baja)
barplot(baja, ylab = "Frecuencia Relativa", xlab = "Baja", col =
c("red","green"))

freq.edat <- table(churn_tidy$edatcat)
barplot(freq.edat, ylab = "Frecuencia Relativa", xlab = "Edatcat", border
= "blue" ,col = "darkred", density = 20)

freq.sexo <- table(churn_tidy$sexo)
barplot(freq.sexo, ylab = "Frecuencia Relativa", xlab = "Sexo", col =
c("green","grey"))

freq.nomina <- table(churn_tidy$Nomina)
barplot(freq.nomina, ylab = "Frecuencia Relativa", xlab = "Nomina", col =
c("red","green"))
```

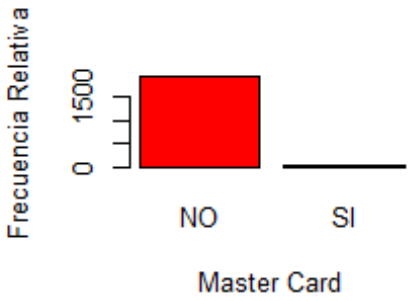


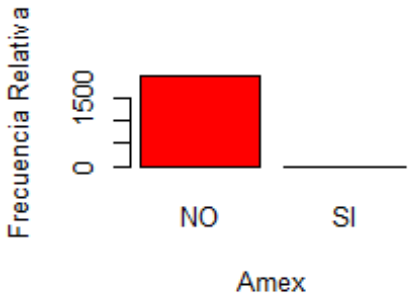

```r
freq.pension <- table(churn_tidy$Pension)
barplot(freq.pension, ylab = "Frecuencia Relativa", xlab = "Pension", col
= c("red","green"))

freq.deb <- table(churn_tidy$Debito_normal)
barplot(freq.deb, ylab = "Frecuencia Relativa", xlab = "Debito Normal",
col = c("red","green"))

freq.deb.aff <- table(churn_tidy$Debito_aff)
barplot(freq.deb.aff, ylab = "Frecuencia Relativa", xlab = "Debito Aff",
col = c("red","green"))

freq.visa <- table(churn_tidy$VISA)
barplot(freq.visa, ylab = "Frecuencia Relativa", xlab = "VISA", col =
c("red","green"))
```

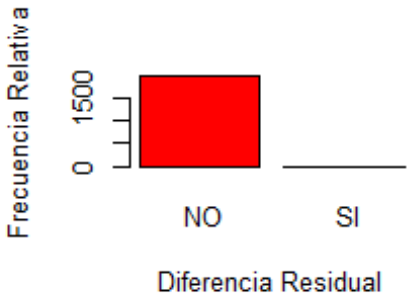



```r
freq.visa.aff <- table(churn_tidy$VISA_aff)
barplot(freq.visa.aff, ylab = "Frecuencia Relativa", xlab = "VISA aff",
col = c("red","green"))

freq.mcard <- table(churn_tidy$MCard)
barplot(freq.mcard, ylab = "Frecuencia Relativa", xlab = "Master Card",
col = c("red","green"))

freq.amex <- table(churn_tidy$Amex)
barplot(freq.amex, ylab = "Frecuencia Relativa", xlab = "Amex", col =
c("red","green"))

freq.difresid <- table(churn_tidy$dif_resid)
barplot(freq.difresid, ylab = "Frecuencia Relativa", xlab = "Diferencia
Residual", col = c("red","green"))
```

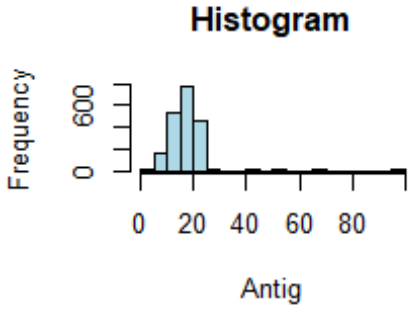


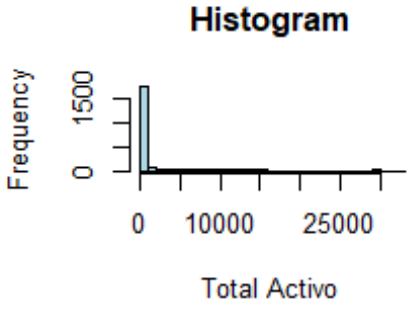


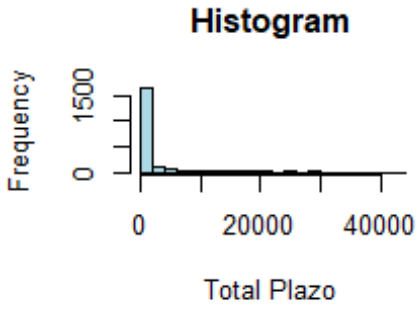


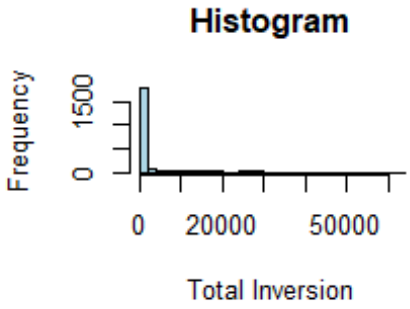



```r
#Grafica para variables tipo continuas
par(mfrow=c(2, 2))
hist(churn_tidy$antig,breaks = 30,col = "lightblue",xlab = "Antig", main
= "Histogram")
hist(churn_tidy$Total_activo,breaks = 30,col = "lightblue",xlab = "Total
Activo", main = "Histogram")
hist(churn_tidy$Total_Plazo,breaks = 30,col = "lightblue",xlab = "Total
Plazo", main = "Histogram")
hist(churn_tidy$Total_Inversion,breaks = 30,col = "lightblue",xlab =
"Total Inversion", main = "Histogram")
```

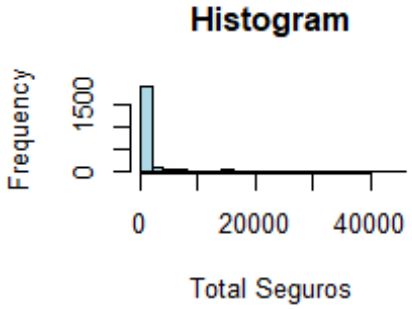


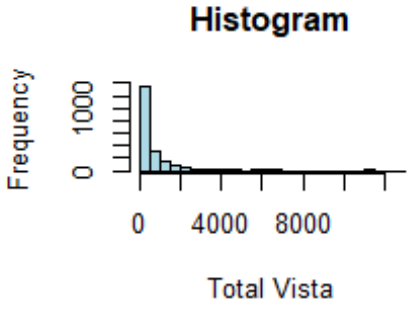


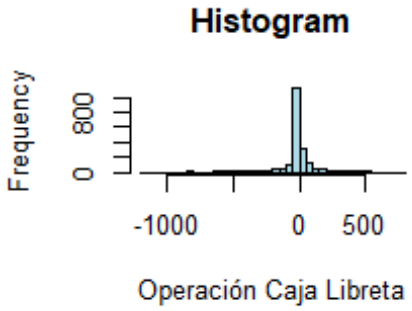


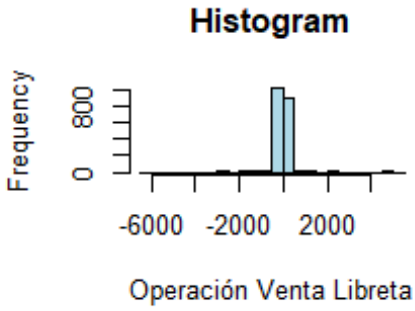



```
hist(churn_tidy$Total_Seguros,breaks = 30,col = "lightblue",xlab = "Total
Seguros", main = "Histogram")
hist(churn_tidy$Total_Vista,breaks = 30,col = "lightblue",xlab = "Total
Vista", main = "Histogram")
hist(churn_tidy$oper_caj_Libreta,breaks = 30,col = "lightblue",xlab =
"Operación Caja Libreta", main = "Histogram")
hist(churn_tidy$oper_ven_Libreta,breaks = 30,col = "lightblue",xlab =
"Operación Venta Libreta", main = "Histogram")
```

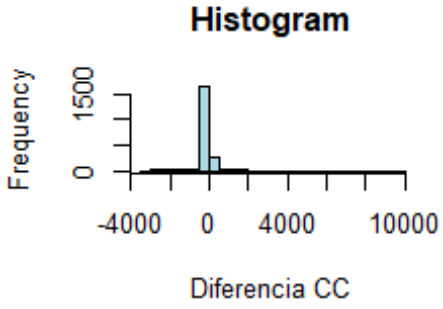


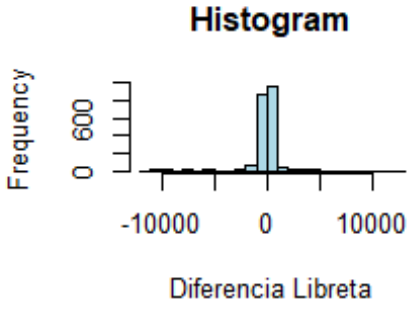


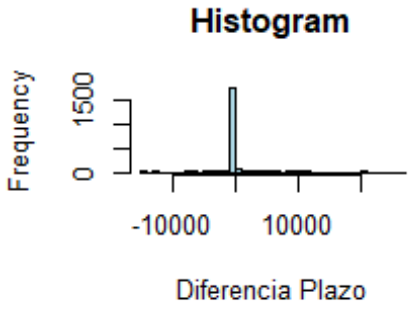


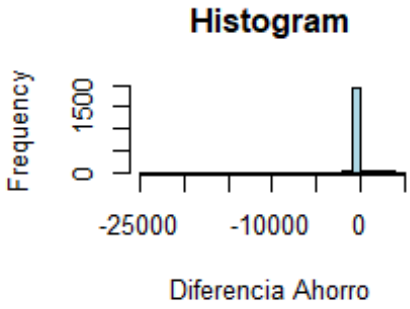



```r
hist(churn_tidy$dif_CC,breaks = 30,col = "lightblue",xlab = "Diferencia CC", main = "Histogram")
hist(churn_tidy$dif_Libreta,breaks = 30,col = "lightblue",xlab = "Diferencia Libreta", main = "Histogram")
hist(churn_tidy$dif_Plazo,breaks = 30,col = "lightblue",xlab = "Diferencia Plazo", main = "Histogram")
hist(churn_tidy$dif_Ahorro,breaks = 30,col = "lightblue",xlab = "Diferencia Ahorro", main = "Histogram")
```

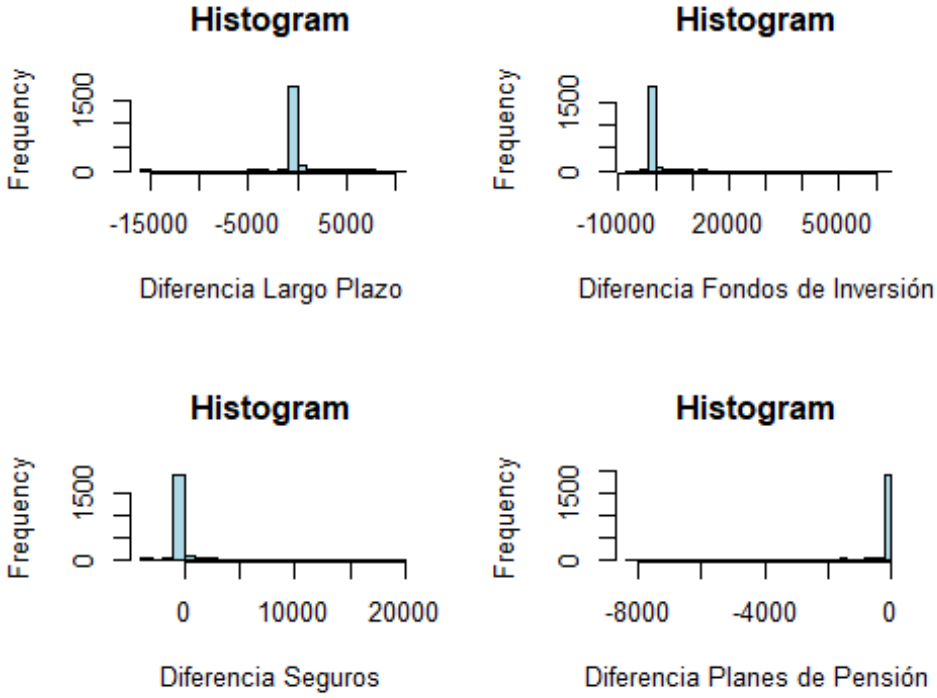




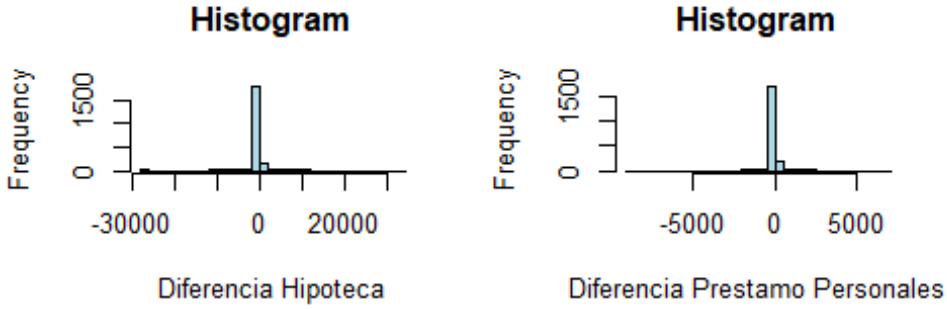





```r
hist(churn_tidy$dif_Largo_plazo,breaks = 30,col = "lightblue",xlab =
"Diferencia Largo Plazo", main = "Histogram")
hist(churn_tidy$dif_Fondos_inv,breaks = 30,col = "lightblue",xlab =
"Diferencia Fondos de Inversión", main = "Histogram")
hist(churn_tidy$dif_Seguros,breaks = 30,col = "lightblue",xlab =
"Diferencia Seguros", main = "Histogram")
hist(churn_tidy$dif_Planes_pension,breaks = 30,col = "lightblue",xlab =
"Diferencia Planes de Pensión", main = "Histogram")
```



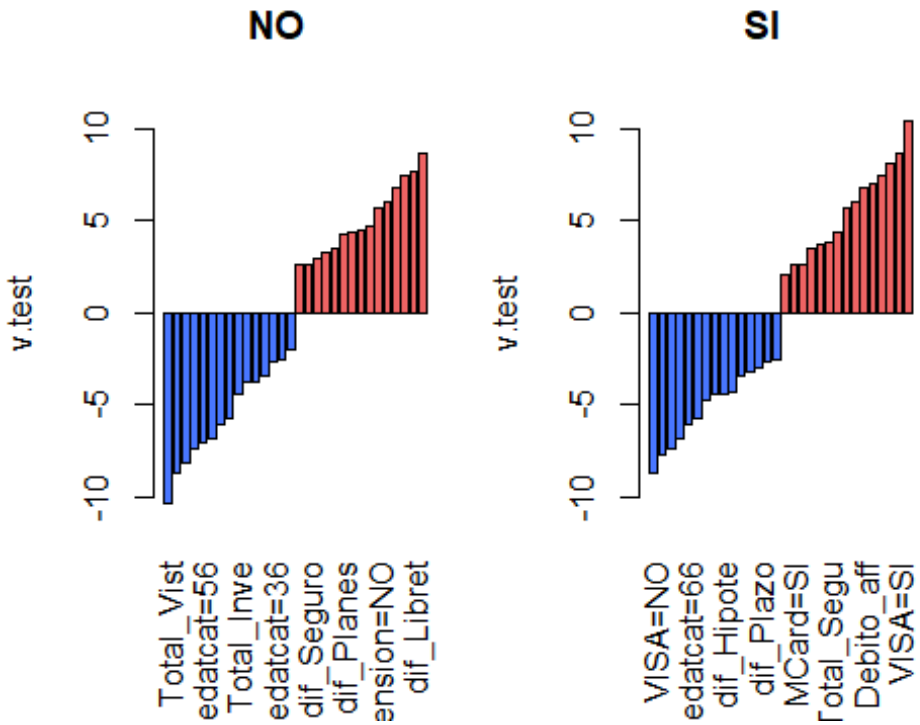

```
hist(churn_tidy$dif_Hipoteca,breaks = 30,col = "lightblue",xlab =
"Diferencia Hipoteca", main = "Histogram")
hist(churn_tidy$dif_Prest_personales,breaks = 30,col = "lightblue",xlab =
"Diferencia Prestamo Personales", main = "Histogram")
```



4. Efectúe el "profiling" de las bajas (con la función catdes de la librería "FactoMineR"). Interprete el resultado.

```
library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 3.6.2

descBaja <- catdes(churn_tidy, num.var=1)
descBaja$quanti$SI

##                          v.test Mean in category Overall mean
## Total_Vista          10.417630        811.50000   565.386854
## Total_Plazo           8.154336       1991.57600  1321.739087
## Total_activo          6.078187        930.03300   616.766683
## Total_Inversion       4.386655       1222.43300   853.314099
## Total_Seguros         3.801778        447.48300   279.960361
## antig                 2.044712         17.76700    17.381836
## dif_Seguros          -2.664971        -22.78806    17.884390
## dif_Plazo            -2.973505         -8.72000   102.248836
## dif_Prest_personales -3.247204         -6.62700    17.568991
## dif_Planes_pension   -4.310552        -79.68979   -39.998419
## dif_Hipoteca         -4.421666       -156.41700     9.941295
## oper_caj_Libreta     -4.769913        -20.75696    -7.765429
## dif_Libreta          -7.688177       -214.85129   -47.318440
##                      sd in category  Overall sd      p.value
## Total_Vista             1194.850299 1058.122407 2.060223e-25
## Total_Plazo             4535.166258 3679.175482 3.511027e-16
```

```
## Total_activo             2755.986574 2308.392562 1.215487e-09
## Total_Inversion          3989.995560 3768.797959 1.151072e-05
## Total_Seguros            2546.855669 1973.589325 1.436614e-04
## antig                       8.041313    8.436904 4.088326e-02
## dif_Seguros               622.215940  683.561855 7.699486e-03
## dif_Plazo                1842.742846 1671.484775 2.944198e-03
## dif_Prest_personales      435.744101  333.736676 1.165450e-03
## dif_Planes_pension        579.497275  412.413830 1.628473e-05
## dif_Hipoteca             1831.209112 1685.111379 9.794297e-06
## oper_caj_Libreta          133.879118  121.988791 1.843056e-06
## dif_Libreta              1146.230653  975.992902 1.492465e-14

descBaja$category$SI

##                  Cla/Mod Mod/Cla    Global      p.value    v.test
## VISA=SI         68.00000    30.6 22.579027 4.744719e-18  8.659348
## Nomina=SI       62.83333    37.7 30.105369 1.009594e-13  7.439641
## edatcat=56-65   70.38462    18.3 13.045660 1.636159e-12  7.062430
## Debito_aff= SI  63.18898    32.1 25.489212 9.232302e-12  6.817990
## Pension=SI      58.54993    42.8 36.678374 1.234848e-08  5.694844
## edatcat=46-55   60.00000    18.3 15.303562 1.901998e-04  3.731690
## VISA_aff= SI    82.14286     2.3  1.404917 5.373258e-04  3.461424
## edatcat=36-45   56.59824    19.3 17.109885 9.225116e-03  2.603597
## MCard=SI        71.05263     2.7  1.906673 9.405861e-03  2.596939
## MCard=NO        49.76982    97.3 98.093327 9.405861e-03 -2.596939
## VISA_aff= NO    49.72010    97.7 98.595083 5.373258e-04 -3.461424
## edatcat=18-25   31.25000     4.0  6.422479 8.377620e-06 -4.455299
## Pension=NO      45.32488    57.2 63.321626 1.234848e-08 -5.694844
## edatcat=66      39.66102    23.4 29.603613 1.083682e-09 -6.096570
## Debito_aff= NO  45.72391    67.9 74.510788 9.232302e-12 -6.817990
## Nomina=NO       44.72362    62.3 69.894631 1.009594e-13 -7.439641
## VISA=NO         44.97732    69.4 77.420973 4.744719e-18 -8.659348

plot(descBaja,barplot = T)
```



Resp.: Para realizar la interpretación hemos analisado los resultados obtenidos através de la función catdes, donde “quanti” representa la descripción de cada categoria por cada variable continua, y “category” representa la descripción de cada categoria por cada categoria entre todas variables categoricas. Ha sido considerada

solo la variable Baja (num.var=1). Con esto, los analisis se centraron en Overall mean, Mean in Category, v.test para "quanti", y en Global, Mod/Cla y v.test para "category". La interpretación que hacemos es que las variables que más caracterizan la baja son de aquellos que tienen más en Total a Vista, total Plazo y en totales de activo y de inversión. Aun, estan en el mismo perfil de baja aquellos que tienen VISA, nomina, que tienen edad entre 56 y 65 años y pensión. Para comentarlo tambien, hay una una diferencia negativa en v.test sobre, por ejemplo, la diferencia de Libreta en los ultimos 3 meses antes de su baja. Por lo tanto, se podría llegar a pensar que el "Profile" para las bajas es de personas mayores, probablemente recien jubiladas, y que tienen condiciones financieras estables.

5. Represente visualmente la relación de las variables explicativas con la variable de respuesta; para ello discretize las variables continuas (esto es, recodifíquelas según un cierto número de intervalos; tenga en cuenta el significado especial del valor 0 a la hora de establecer los intervalos de recodificación) y represente mediante barplots el porcentaje de baja de las modalidades de las variables categóricas (tanto las categóricas originales como las continuas recodificadas).

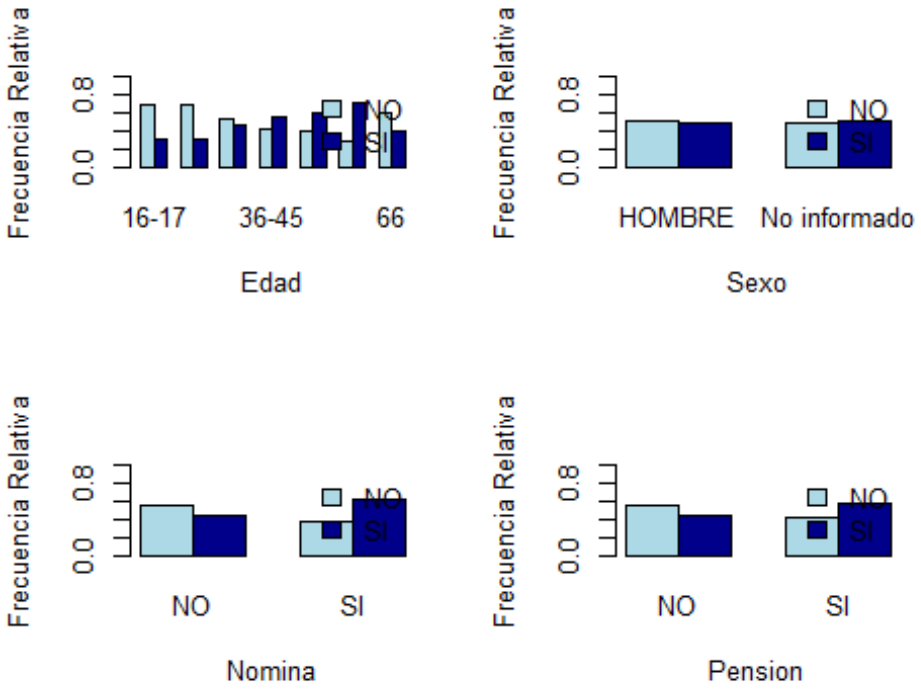
```
churn_tidy2 = churn_tidy

churn_tidy2$Rec_tot_activo = cut(churn_tidy2$Total_activo,
breaks=c(0,0.0001,150,400,1000,3000,99000),include.lowest=T)

churn_tidy2$Rec_tot_plazo = cut(churn_tidy2$Total_Plazo,
breaks=c(0,0.0001,700,2000,4000,8000,99000),include.lowest=T)

churn_tidy2$Rec_tot_inversion = cut(churn_tidy2$Total_Inversion,
breaks=c(0,0.0001,700,2000,4000,8000,99000),include.lowest=T)

churn_tidy2$Rec_tot_seguros = cut(churn_tidy2$Total_Seguros,
breaks=c(0,0.0001,150,400,1000,3000,99000),include.lowest=T)

churn_tidy2$Rec_tot_vista = cut(churn_tidy2$Total_Vista,
breaks=c(0,0.0001,50,150,400,1000,99000),include.lowest=T)

churn_tidy2$Rec_oper_caj_Libreta = cut(churn_tidy2$oper_caj_Libreta,
breaks=c(-9000,-100,-20,-0.0001,0,20,100,9000))

churn_tidy2$Rec_oper_ven_Libreta = cut(churn_tidy2$oper_ven_Libreta,
breaks=c(-9000,-100,-20,-0.0001,0,20,100,9000))

churn_tidy2$Rec_dif_CC= cut(churn_tidy2$dif_CC, breaks=c(-99000,-100,-
0.0001,0,20,200,1000,99000))
```

```
churn_tidy2$Rec_dif_Libreta= cut(churn_tidy2$dif_Libreta, breaks=c(-
99000,-100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Plazo= cut(churn_tidy2$dif_Plazo, breaks=c(-99000,-
100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Ahorro= cut(churn_tidy2$dif_Ahorro, breaks=c(-99000,-
100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Largo_plazo= cut(churn_tidy2$dif_Largo_plazo,
breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Fondos_inv= cut(churn_tidy2$dif_Fondos_inv,
breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Seguros= cut(churn_tidy2$dif_Seguros, breaks=c(-
99000,-100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Planes_pension= cut(churn_tidy2$dif_Planes_pension,
breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Hipoteca= cut(churn_tidy2$dif_Hipoteca, breaks=c(-
99000,-100,-0.0001,0,20,200,1000,99000))

churn_tidy2$Rec_dif_Prest_personales=
cut(churn_tidy2$dif_Prest_personales, breaks=c(-99000,-100,-
0.0001,0,20,200,1000,99000))
```

Grafica de la variable Baja con las categóricas originales:

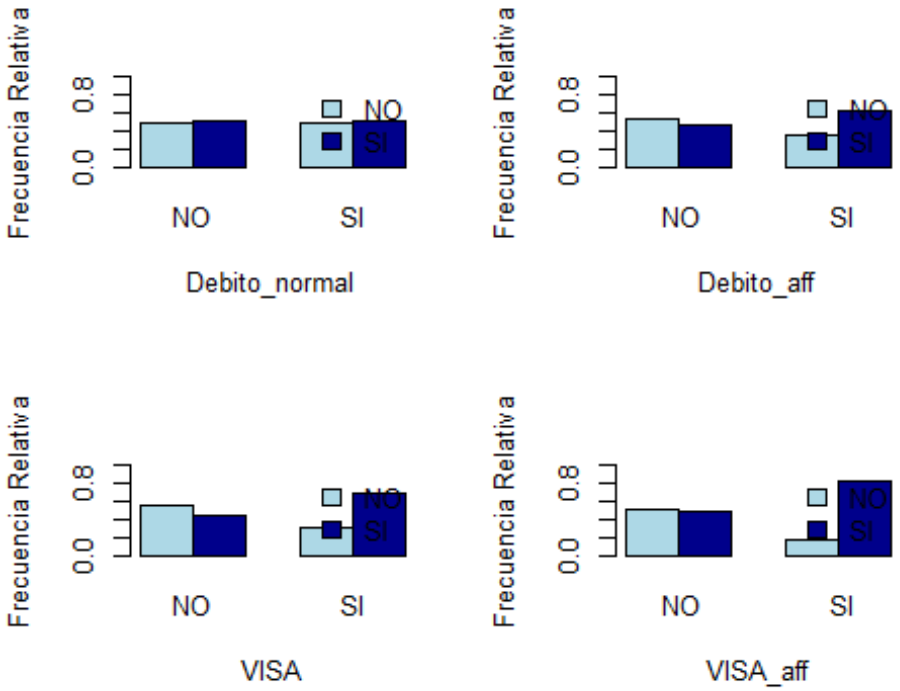
```
par(mfrow=c(2,2))
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$edatcat) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Edad",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$sexo) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Sexo",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Nomina) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Nomina",
```

```
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Pension) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Pension",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```

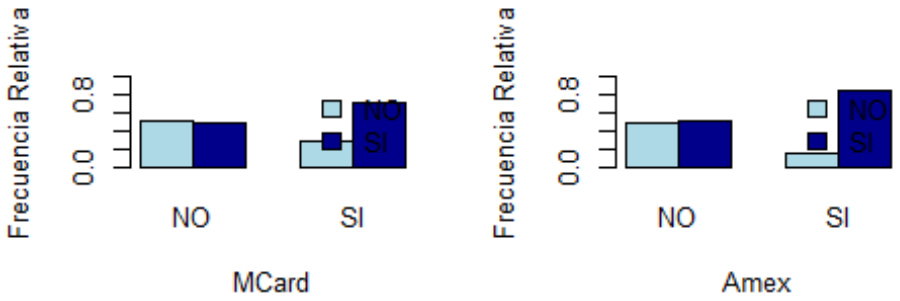



```
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Debito_normal) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Debito_normal", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Debito_aff) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Debito_aff",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$VISA) %>%
```

```r
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "VISA",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$VISA_aff) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "VISA_aff",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```



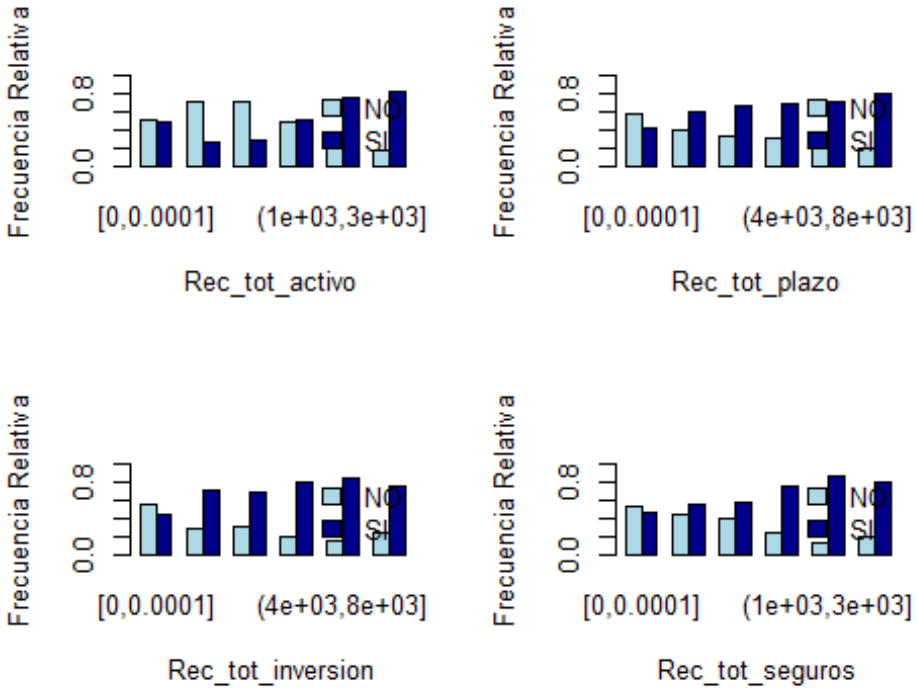

```r
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$MCard) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "MCard",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Amex) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Amex",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```



Grafica de la variable Baja con las continuas codificadas:

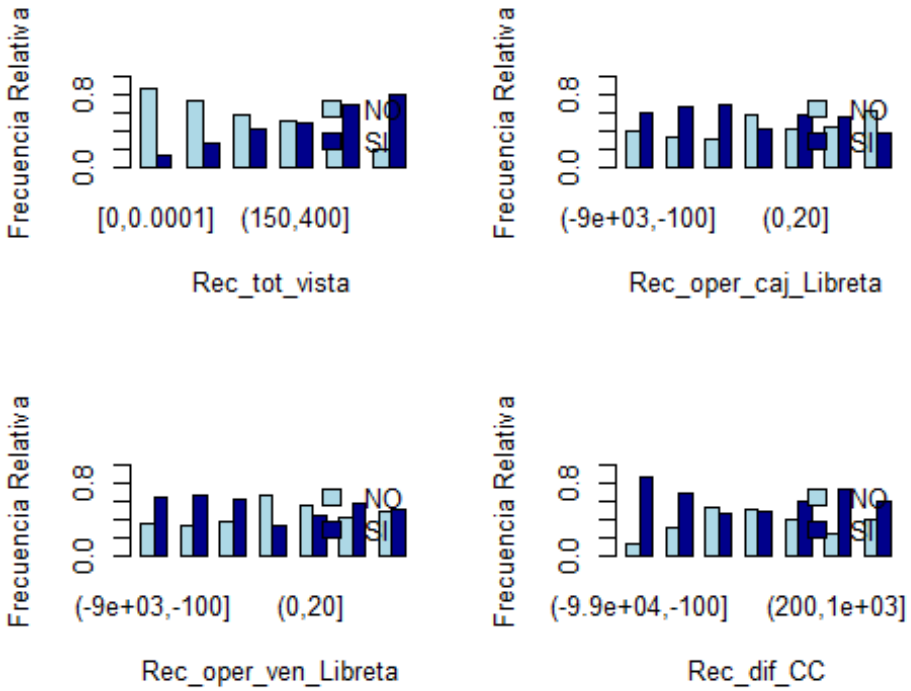
```
par(mfrow=c(2,2))
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_tot_activo) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_tot_activo", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_tot_plazo) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_tot_plazo", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_tot_inversion)
%>% prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_tot_inversion", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```

```
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_tot_seguros) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_tot_seguros", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```



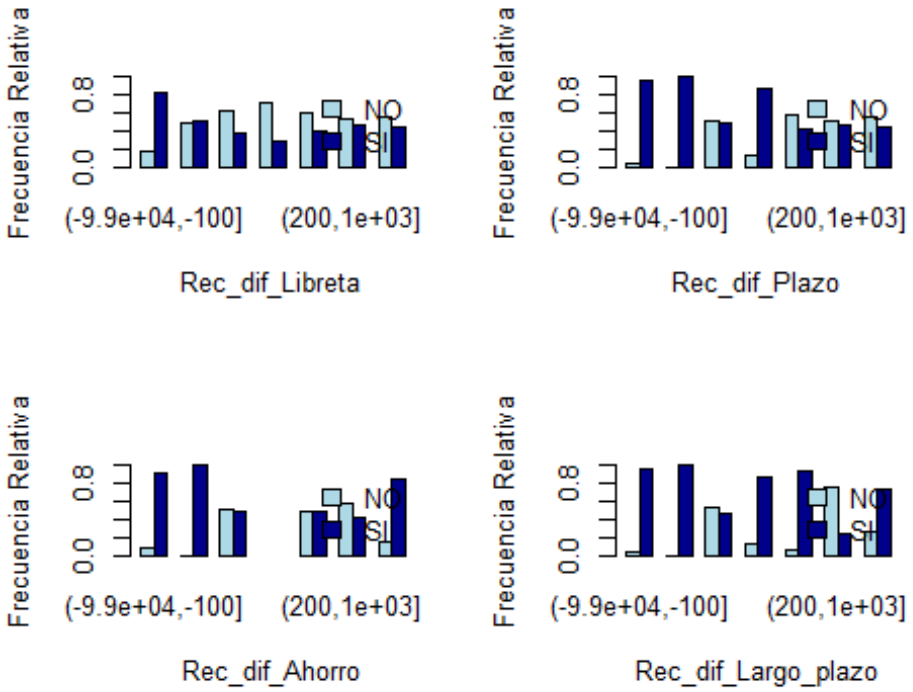

```
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_tot_vista) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_tot_vista", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_oper_caj_Libreta)
%>% prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_oper_caj_Libreta", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_oper_ven_Libreta)
%>% prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
```

```
"Rec_oper_ven_Libreta", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_CC) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab = "Rec_dif_CC",
beside = TRUE, ylim = c(0,1), col = c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```



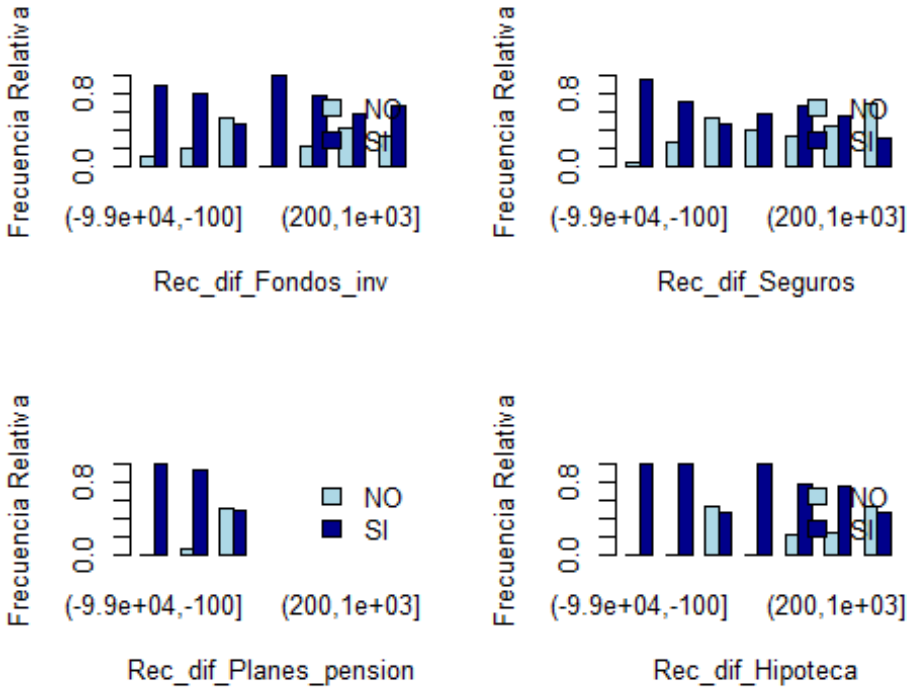

```
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Libreta) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Libreta", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Plazo) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Plazo", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```

```r
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Ahorro) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Ahorro", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Largo_plazo)
%>% prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Largo_plazo", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```



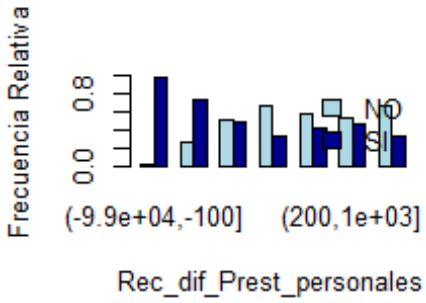

```r
freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Fondos_inv)
%>% prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Fondos_inv", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Seguros) %>%
prop.table(margin = 2)
```

```r
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Seguros", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <-
table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Planes_pension) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Planes_pension", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))

freq.rel.baja <- table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Hipoteca) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
"Rec_dif_Hipoteca", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```




```r
freq.rel.baja <-
table(churn_tidy2$Baja,churn_tidy2$Rec_dif_Prest_personales) %>%
prop.table(margin = 2)
barplot(freq.rel.baja, ylab = "Frecuencia Relativa", xlab =
```

```
"Rec_dif_Prest_personales", beside = TRUE, ylim = c(0,1), col =
c("lightblue","darkblue"))
legend('topright',legend=rownames(freq.rel.baja),bty ='n', fill =
c("lightblue","darkblue"))
```



6. Suponga que quiere analizar la compra de un producto a partir del barrio de residencia (alto o bajo) (indicador del poder adquisitivo del cliente). En un primer análisis de obtiene la siguiente tabla:

```
          Compra SI Compra NO   Total
```

Clase alta 20 373 393 Clase baja 6 316 322 ### En su opinión, ¿el poder adquisitivo del cliente, tiene alguna influencia sobre la compra o no del producto? (Responda sólo calculando las probabilidades, sin realizar la prueba de hipótesis de igualdad entre ambas probabilidades).

```
datoscompra <- data.frame("Clase" = c("Clase alta","Clase baja"),"Compra
SI" = c(20,6), "Compra NO" = c(373,316),"Total"=c(393,322))
datoscompra

##        Clase Compra.SI Compra.NO Total
## 1 Clase alta        20       373   393
## 2 Clase baja         6       316   322
```

```
perccompraalta <- datoscompra$Compra.SI[datoscompra$Clase == "Clase
alta"]/datoscompra$Total[datoscompra$Clase == "Clase alta"] * 100
paste("Porcentaje de compra en la clase alta: ",perccompraalta)
```

```
## [1] "Porcentaje de compra en la clase alta:  5.08905852417303"
```

```
percomprabaja <- datoscompra$Compra.SI[datoscompra$Clase == "Clase
baja"]/datoscompra$Total[datoscompra$Clase == "Clase baja"] * 100
paste("Porcentaje de compra en la clase baja: ",percomprabaja)
```

```
## [1] "Porcentaje de compra en la clase baja:  1.86335403726708"
```

Resp.: Solo calculando las probabilidades seria posible decir que el poder adquisitivo del cliente si que tiene influencia sobre la compra. Esto pues la probabilidad de alguien de la clase alta comprar el producto se presenta mas grande en relación a clase baja.

Un empleado senior de la compañía nos sugiere profundizar más en el análisis y tener en cuenta la edad de los clientes. Cruzando por edad (adulto o joven) los dos tipos de barrio mencionados, obtenemos las siguientes tablas:

ADULTOS Compra SI Compra NO Total Clase alta 3 176 179 Clase baja 4 293 297 JOVENES Compra SI Compra NO Total Clase alta 17 197 214 Clase baja 2 23 25 ### ¿Tenía razón el empleado de que era conveniente tener en cuenta la edad?. ¿Cuál de los dos factores, el barrio de residencia o la edad, es el determinante en la compra del producto en cuestión?

```
datosadulto <- data.frame("Clase" = c("Clase alta","Clase baja"),"Compra
SI" = c(3,4), "Compra NO" = c(176,293),"Total"=c(179,297))
datosadulto
```

```
##        Clase Compra.SI Compra.NO Total
## 1 Clase alta         3       176   179
## 2 Clase baja         4       293   297
```

```
datosjovenes <- data.frame("Clase" = c("Clase alta","Clase baja"),"Compra
SI" = c(17,2), "Compra NO" = c(197,23),"Total"=c(214,25))
datosjovenes
```

```
##        Clase Compra.SI Compra.NO Total
## 1 Clase alta        17       197   214
## 2 Clase baja         2        23    25
```

```
#Percentaje de jovenes por el total general
sum(datosjovenes$Total) / sum(datosadulto$Total,datosjovenes$Total) * 100
```

```
## [1] 33.42657
```

```
#Percentaje de adultos por el total general
sum(datosadulto$Total) / sum(datosadulto$Total,datosjovenes$Total) * 100
```

```
## [1] 66.57343

#Percentaje de compra por jovenes, en relacion al total del grupo
sum(datosjovenes$Compra.SI) / sum(datosjovenes$Total) * 100

## [1] 7.949791

#Percentaje de compra por adultos, en relacion al total del grupo
sum(datosadulto$Compra.SI) / sum(datosadulto$Total) * 100

## [1] 1.470588

#Percentaje de compra por clase alta, en relacion al total grupo
datosjovenes$Compra.SI[datosjovenes$Clase == "Clase alta"] /
datosjovenes$Total[datosjovenes$Clase == "Clase alta"] * 100

## [1] 7.943925

datosjovenes$Compra.SI[datosjovenes$Clase == "Clase baja"] /
datosjovenes$Total[datosjovenes$Clase == "Clase baja"] * 100

## [1] 8

#Percentaje de compra por clase baja, en relacion al total grupo
datosadulto$Compra.SI[datosadulto$Clase == "Clase alta"] /
datosadulto$Total[datosadulto$Clase == "Clase alta"] * 100

## [1] 1.675978

datosadulto$Compra.SI[datosadulto$Clase == "Clase baja"] /
datosadulto$Total[datosadulto$Clase == "Clase baja"] * 100

## [1] 1.346801
```

Resp.: Para realizar estos calculos, hemos intentado basarnos en el Paradojo de Simpson, en que demuestra que las relaciones estadisticas observadas en una población puede ser invertida/contradicha por sus subgrupos que forman esta población. Teniendo esto en cuenta, podemos decir que el empleado tenia razón en considerar la edad, porque esta variable nos ha parecido más determinante que el bario/clase para la compra del producto.