Xavier Gonzalez – Exercici Sessió 2 - Profiling

SESIÓN 2: PROFILING

Una de las mayores preocupaciones en todas las empresas es evitar la fuga de clientes, este es el caso del sector bancario. Para poder tener un conocimiento fiable y poder adoptar políticas de prevención, se han listado las bajas ocurridas en un periodo de tiempo determinado. Para ellas se ha recogido la información que se dispone del ex cliente: sus características y posición en el año anterior a la baja y el cambio ocurrido hasta antes de tres meses de la baja. La información se completa con una muestra aleatoria de clientes que no han sido baja en el periodo considerado, para los que se recoge la misma información. Con los datos recogidos se ha formado el fichero “churn.txt”.

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

```
> setwd("/media/xabee/XABEE_USB/BIG_DATA/Analytics/SESSION 2_cursBigData_Profiling/exer2")
> df <- read.table("churn.txt", header = TRUE)
```

'sexo' 'No informado' se remplazo por 'MUJER'. Se observan diversos campos que contienen el nombre de la cabecera de la columna, que se elimina, excepto en la cabecera:

"edatcat " se sustituye "" 2000 veces
"Pension " se sustituye "" 1994 veces
"Nomina " sustituye "" 1996 veces
...

```
> summary(df)
```

```
Baja       edatcat        sexo          antig         Nomina     Pension     Debito_normal Debito_aff    VISA      VISA_aff   MCard
NO:1000   16-17: 25   HOMBRE:1134   Min.   : 3.00   NO  :1393   NO  :1262   NO:1912       NO  :1485   NO  :1543   NO:1965   NO  :1955
SI:1000   18-25:128   MUJER : 866   1st Qu.:13.00   SI  : 603   SI  : 732   SI:  88       SI  : 513   SI  : 456   SI:  35   SI  :  44
          26-35:345                 Median :18.00   NA's:   4   NA's:   6                 NA's:   2   NA's:   1              NA's:   1
          36-45:343                 Mean   :17.38
          46-55:306                 3rd Qu.:21.00
          56-65:261                 Max.   :99.00
          66.. :592
```

```
Amex       Total_activo    Total_Plazo      Total_Inversion   Total_Seguros    Total_Vista       dif_resid oper_caj_Libreta
NO:1987   Min.   :    0   Min.   :    0.0   Min.   :    0.0   Min.   :    0   Min.   :    0.00   NO:1982   Min.   :-1157.500
SI:  13   1st Qu.:    0   1st Qu.:    0.0   1st Qu.:    0.0   1st Qu.:    0   1st Qu.:   51.75   SI:  18   1st Qu.:    0.000
          Median :    0   Median :    0.0   Median :    0.0   Median :    0   Median :  206.00             Median :    0.000
          Mean   :  618   Mean   : 1332.7   Mean   :  853.1   Mean   :  279   Mean   :  569.17             Mean   :   -7.404
          3rd Qu.:    0   3rd Qu.:  472.2   3rd Qu.:    0.0   3rd Qu.:    0   3rd Qu.:  657.00             3rd Qu.:    5.000
          Max.   :32772   Max.   :43400.0   Max.   :62017.0   Max.   :45455   Max.   :12738.00             Max.   :  774.750
```

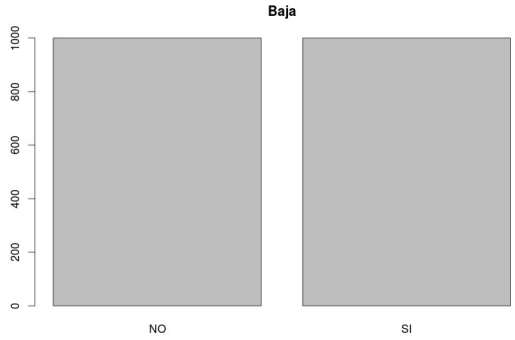
```
oper_ven_Libreta       dif_CC          dif_Libreta          dif_Plazo         dif_Ahorro        dif_Largo_plazo
Min.   :-6378.260   Min.   :-3312.54   Min.   :-11811.900   Min.   :-15000.0   Min.   :-24208.000   Min.   :-15913.04
1st Qu.:   -6.750   1st Qu.:    0.00   1st Qu.:   -56.910   1st Qu.:     0.0   1st Qu.:     0.000   1st Qu.:     0.00
Median :    0.000   Median :    0.00   Median :     1.765   Median :     0.0   Median :     0.000   Median :     0.00
Mean   :    2.541   Mean   :   26.93   Mean   :   -41.937   Mean   :   114.9   Mean   :     7.051   Mean   :    26.11
3rd Qu.:   51.562   3rd Qu.:    0.00   3rd Qu.:    98.000   3rd Qu.:     0.0   3rd Qu.:     0.000   3rd Qu.:     0.00
Max.   : 5038.670   Max.   : 9715.28   Max.   : 12737.000   Max.   : 27000.0   Max.   :  4008.000   Max.   : 10071.00
```

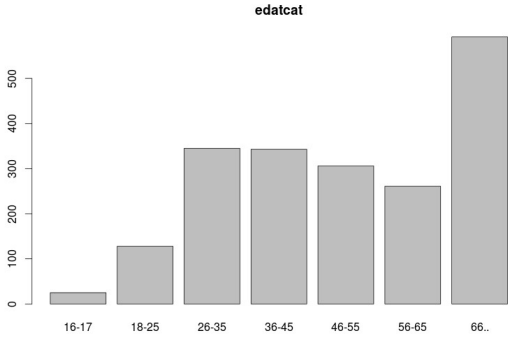
```
dif_Fondos_inv     dif_Seguros       dif_Planes_pension  dif_Hipoteca       dif_Prest_personales
Min.   :-7746.1   Min.   :-3905.05   Min.   :-8246.55   Min.   :-26654.00   Min.   :-8676.00
1st Qu.:    0.0   1st Qu.:    0.00   1st Qu.:    0.00   1st Qu.:     0.00   1st Qu.:    0.00
Median :    0.0   Median :    0.00   Median :    0.00   Median :     0.00   Median :    0.00
Mean   :  261.8   Mean   :   17.82   Mean   :  -39.86   Mean   :    13.28   Mean   :   17.51
3rd Qu.:    0.0   3rd Qu.:    0.00   3rd Qu.:    0.00   3rd Qu.:     0.00   3rd Qu.:    0.00
Max.   :62017.0   Max.   :19461.00   Max.   :    0.00   Max.   : 32772.00   Max.   : 6741.00
```

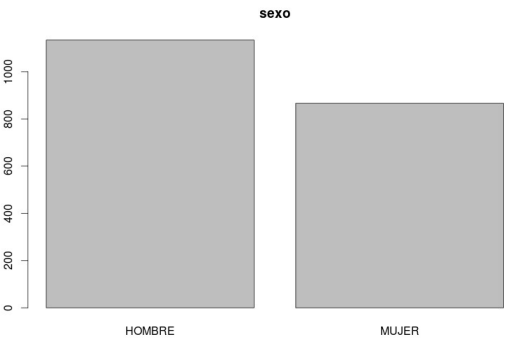
2. Especifique cuál es la variable de respuesta y cuáles son las explicativas y el tipo de todas ellas.

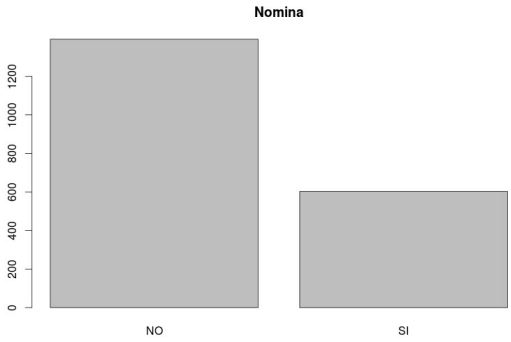
La variable de respuesta es “Baja” y las explicativas serían las demás, algunas cualitativas ('edatcat', 'sexo', 'Nomina', 'Pension', 'Debito_normal', 'Debito_aff', 'VISA', 'VISA_aff', 'MCard', 'Amex', 'dif_resid') y el resto cuantitativas.

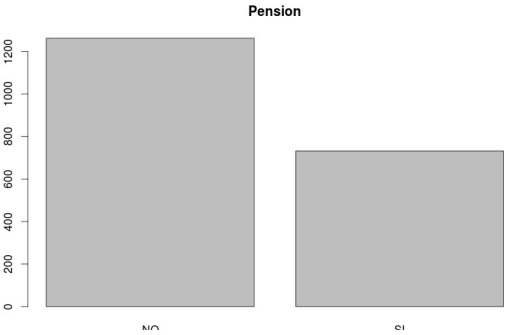
3. Efectúe una gráfica de los datos; un diagrama de barras para las variables categóricas y un histograma para las variables continuas.

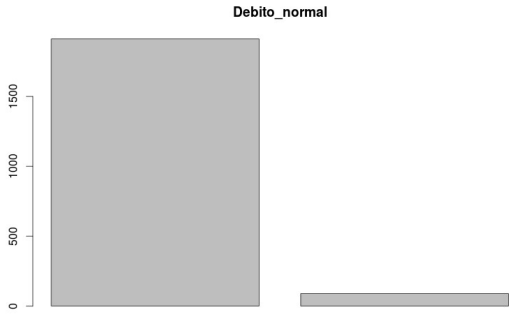


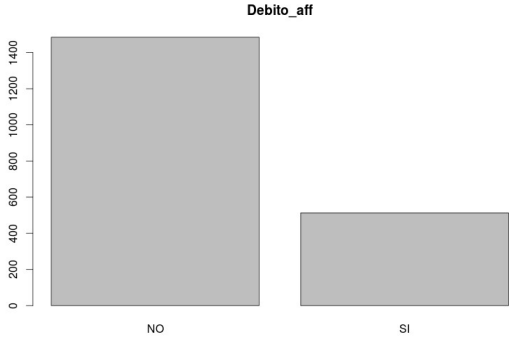


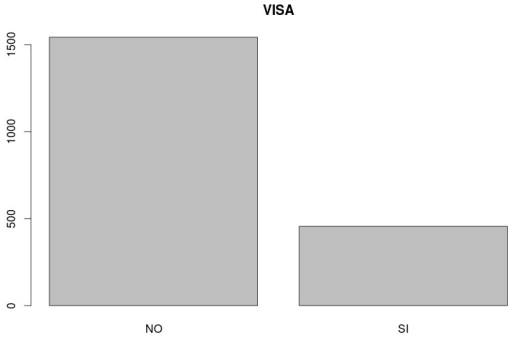


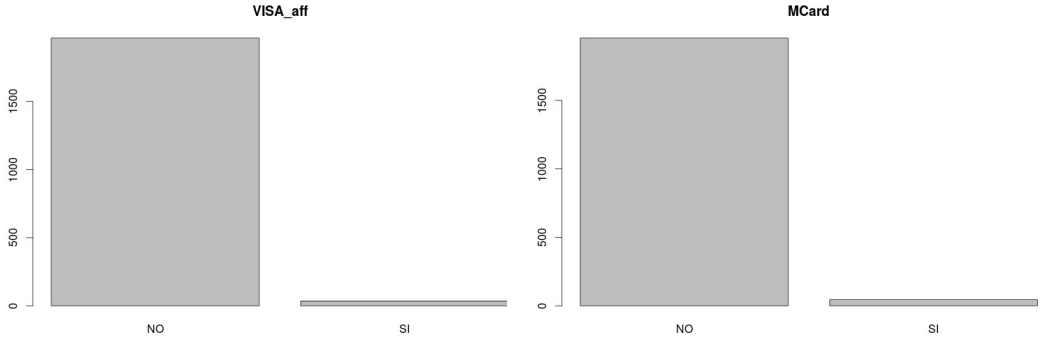




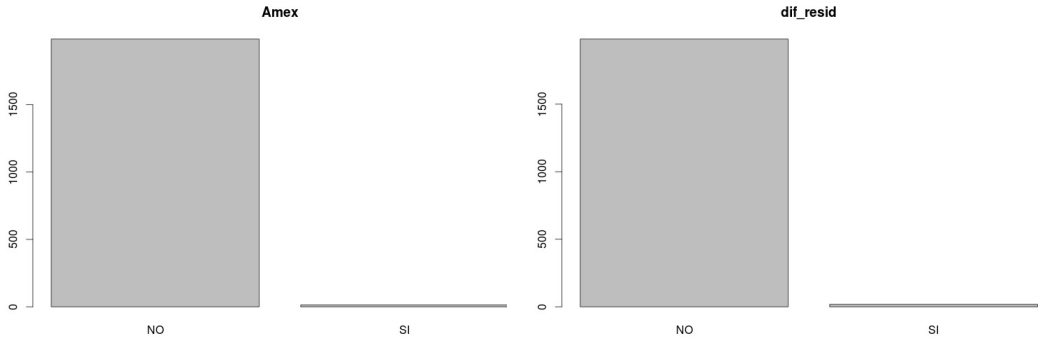




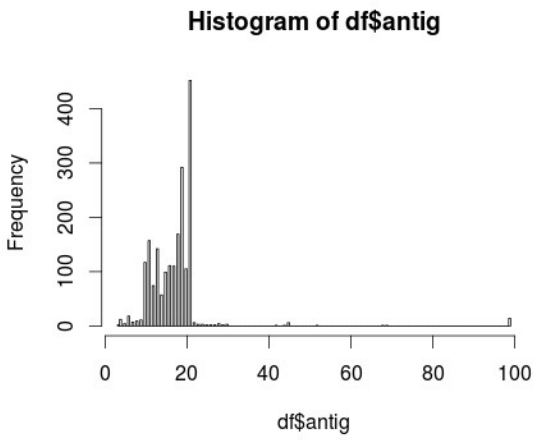


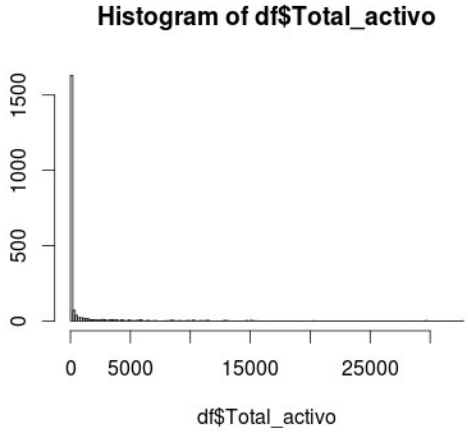

VISA_aff
0
500
1000
1500
NO
SI

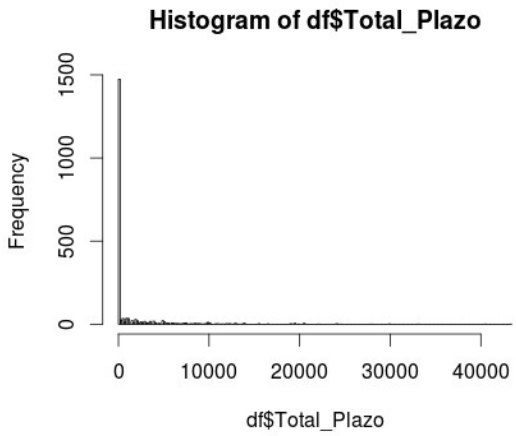
MCard
0
500
1000
1500
NO
SI

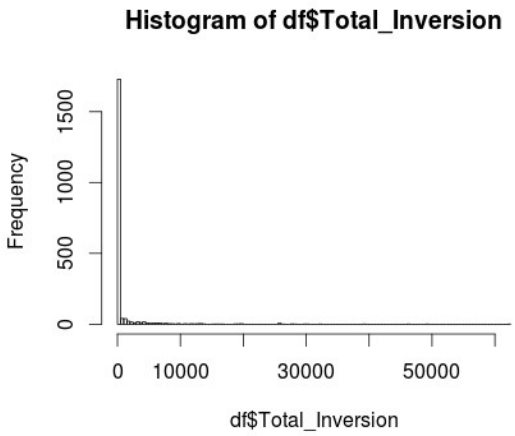

Amex
0
500
1000
1500
NO
SI

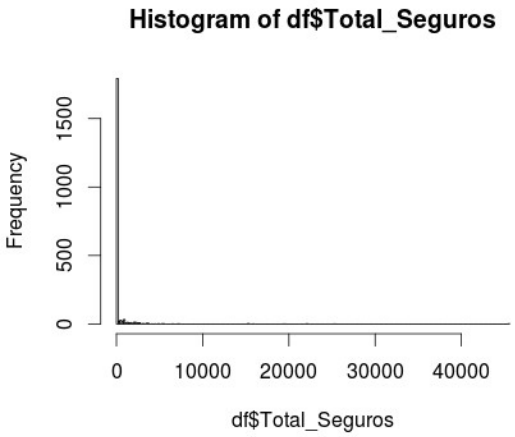
dif_resid
0
500
1000
1500
NO
SI

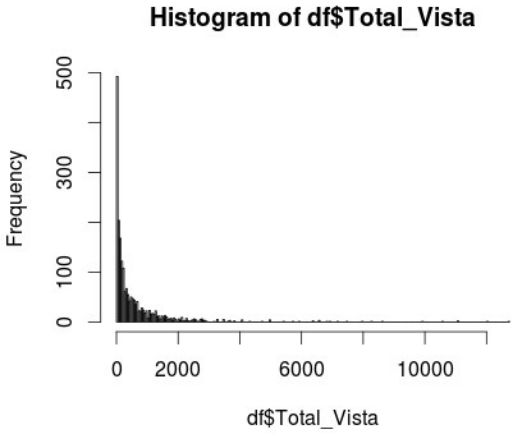

Histogram of df$antig
Frequency
0
100
200
300
400
0
20
40
60
80
100
df$antig

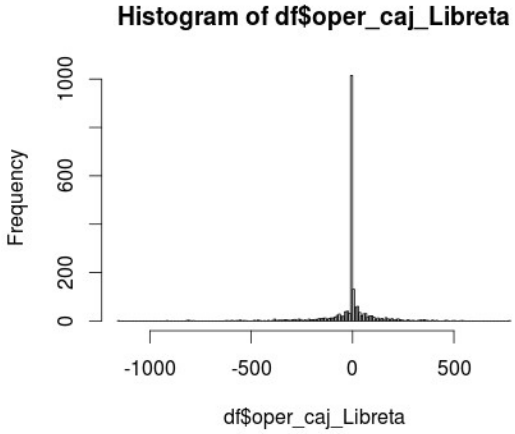

Histogram of df$Total_activo
0
500
1000
1500
0
5000
15000
25000
df$Total_activo

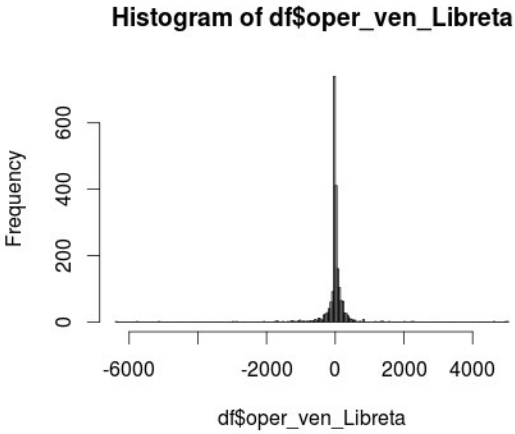

Histogram of df$Total_Plazo
Frequency
0
500
1000
1500
0
10000
20000
30000
40000
df$Total_Plazo

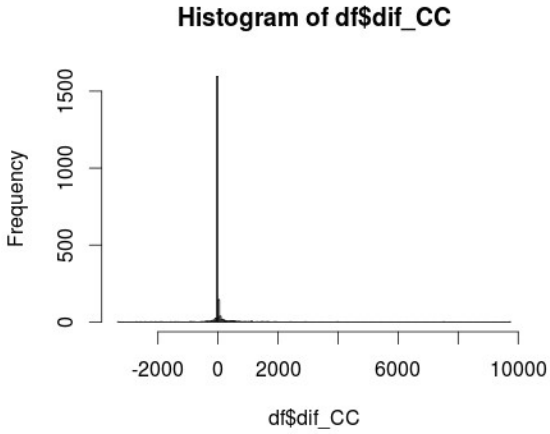

Histogram of df$Total_Inversion
Frequency
0
500
1000
1500
0
10000
30000
50000
df$Total_Inversion

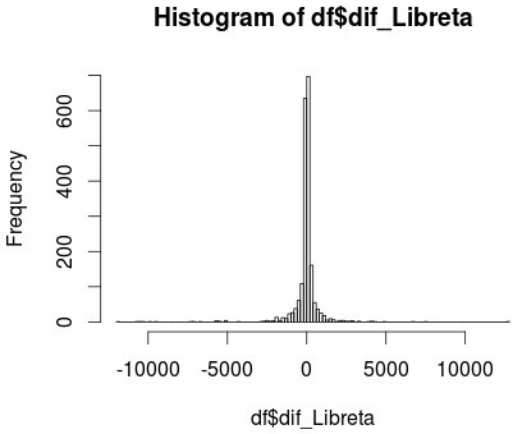

Histogram of df$Total_Seguros
Frequency
0
500
1000
1500
0
10000
20000
30000
40000
df$Total_Seguros

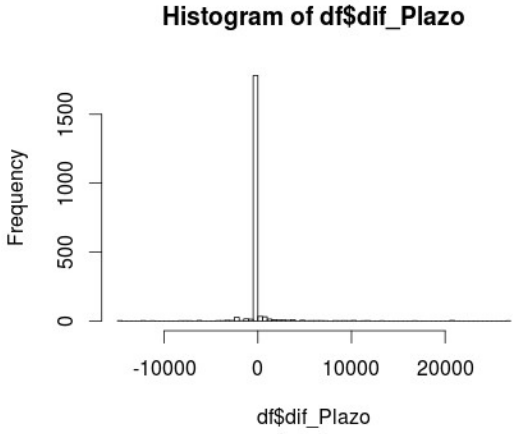

Histogram of df$Total_Vista
Frequency
0
100
300
500
0
2000
6000
10000
df$Total_Vista

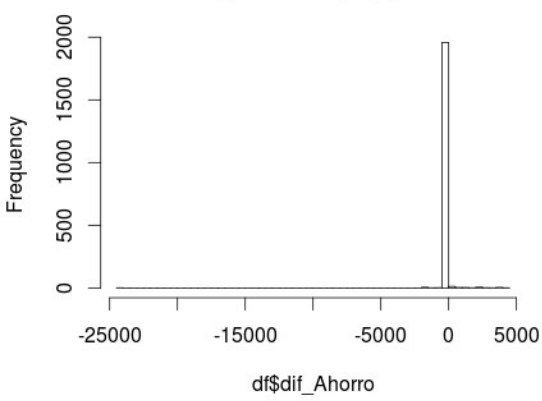

Histogram of df$oper_caj_Libreta
Frequency
0
200
600
1000
-1000
-500
0
500
df$oper_caj_Libreta

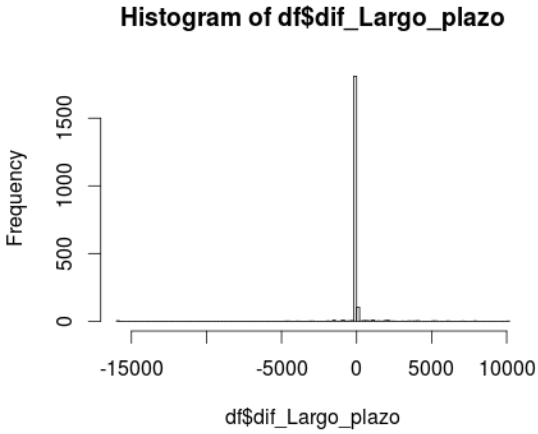

Histogram of df$oper_ven_Libreta
Frequency
0
200
400
600
-6000
-2000
0
2000
4000
df$oper_ven_Libreta

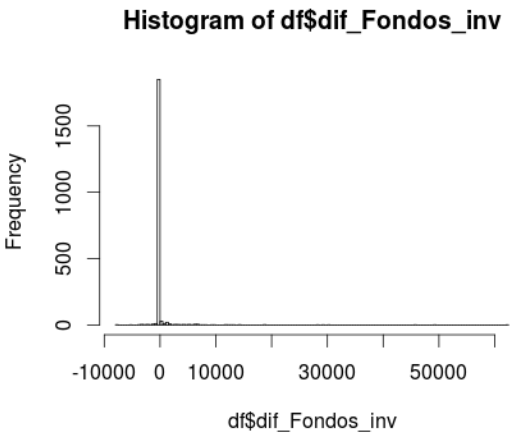

Histogram of df$dif_CC
Frequency
0
500
1000
1500
-2000
0
2000
6000
10000
df$dif_CC

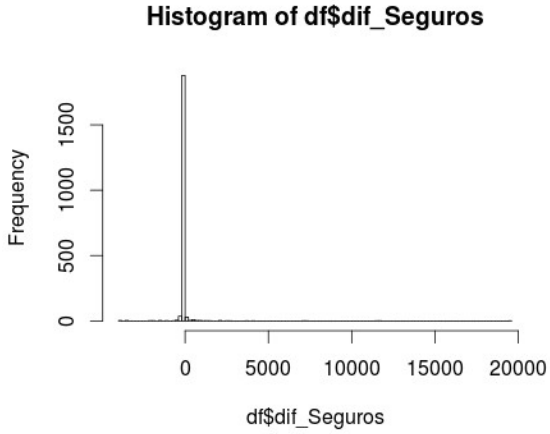

Histogram of df$dif_Libreta
Frequency
0
200
400
600
-10000
-5000
0
5000
10000
df$dif_Libreta

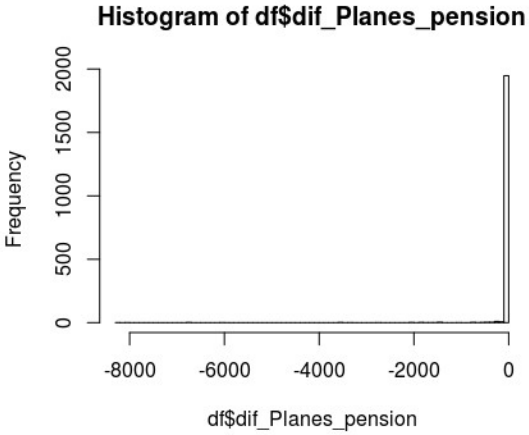

Histogram of df$dif_Plazo
Frequency
0
500
1000
1500
-10000
0
10000
20000
df$dif_Plazo

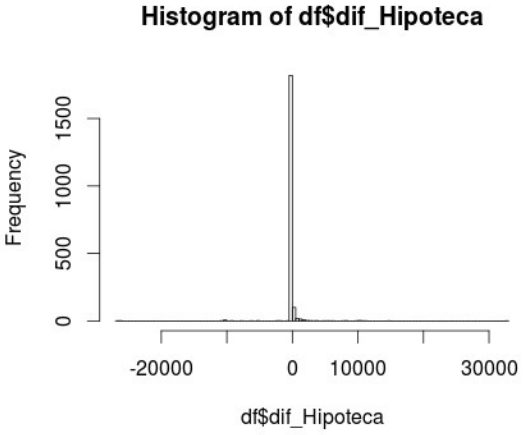

Histogram of df$dif_Ahorro
Frequency
0
500
1000
1500
2000
-25000
-15000
-5000
0
5000
df$dif_Ahorro

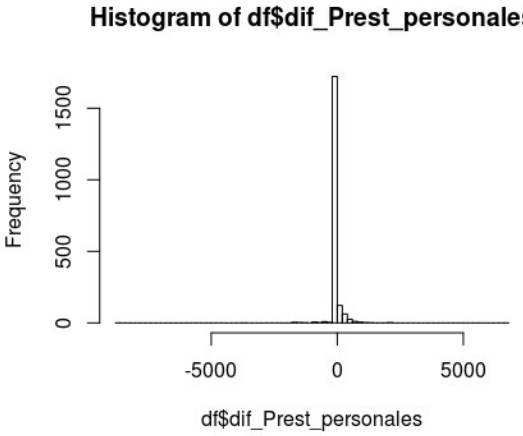

Histogram of df$dif_Largo_plazo
Frequency
0
500
1000
1500
-15000
-5000
0
5000
10000
df$dif_Largo_plazo


Histogram of df$dif_Fondos_inv
Frequency
0
500
1000
1500
-10000
0
10000
30000
50000
df$dif_Fondos_inv


Histogram of df$dif_Seguros
Frequency
0
500
1000
1500
0
5000
10000
15000
20000
df$dif_Seguros


Histogram of df$dif_Planes_pension
Frequency
0
500
1000
1500
2000
-8000
-6000
-4000
-2000
0
df$dif_Planes_pension


Histogram of df$dif_Hipoteca
Frequency
0
500
1000
1500
-20000
0
10000
30000
df$dif_Hipoteca


Histogram of df$dif_Prest_personales
Frequency
0
500
1000
1500
-5000
0
5000
df$dif_Prest_personales

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

Los resultados obtenidos para la variables categóricas, muestran las variables más ligadas a 'Baja':

```
> Dataset <- read.table("/media/xabee/XABEE_USB/BIG_DATA/Analytics/SESSION
2_cursBigData_Profiling/exer2/churn.txt",  header=TRUE, sep="", na.strings="NA",
dec=".", strip.white=TRUE)
[5] NOTE: The dataset Dataset has 2000 rows and 30 columns.
> res=catdes(Dataset, num.var = 1)
> res$test.chi2
```

```
                  p.value           df
edatcat     3.772902e-21  6
VISA        6.501235e-16  2
Nomina      3.564497e-13  2
Debito_aff 1.456320e-10  2
Pension     5.495301e-09  2
```

Si concretamos los casos 'Baja' SI/NO:

```
> res$category
```

```
$NO
               Cla/Mod Mod/Cla Global      p.value    v.test
VISA=NO        55.02268    84.9  77.15 9.794477e-17  8.307251
Nomina=NO      55.27638    77.0  69.65 7.657091e-13  7.167158
Debito_aff=NO  54.27609    80.6  74.25 7.443526e-11  6.511437
edatcat=66..   60.47297    35.8  29.60 1.177508e-09  6.083277
Pension=NO     54.67512    69.0  63.10 4.478523e-08  5.470862
edatcat=18-25  68.75000     8.8   6.40 1.013616e-05  4.414248
Pension=NA    100.00000     0.6   0.30 1.550793e-02  2.420295
edatcat=36-45  43.73178    15.0  17.15 1.081337e-02 -2.548673
edatcat=46-55  40.19608    12.3  15.30 1.920596e-04 -3.729238
Pension=SI     41.53005    30.4  36.60 8.411597e-09 -5.759996
Debito_aff=SI  37.42690    19.2  25.65 3.508650e-11 -6.623469
edatcat=56-65  29.88506     7.8  13.05 1.930610e-12 -7.039406
Nomina=SI      37.47927    22.6  30.15 1.607876e-13 -7.377915
VISA=SI        32.89474    15.0  22.80 5.835818e-17 -8.368496

$SI
               Cla/Mod Mod/Cla Global      p.value    v.test
VISA=SI       67.10526    30.6  22.80 5.835818e-17  8.368496
Nomina=SI     62.52073    37.7  30.15 1.607876e-13  7.377915
edatcat=56-65 70.11494    18.3  13.05 1.930610e-12  7.039406
Debito_aff=SI 62.57310    32.1  25.65 3.508650e-11  6.623469
Pension=SI    58.46995    42.8  36.60 8.411597e-09  5.759996
edatcat=46-55 59.80392    18.3  15.30 1.920596e-04  3.729238
edatcat=36-45 56.26822    19.3  17.15 1.081337e-02  2.548673
Pension=NA     0.00000     0.0   0.30 1.550793e-02 -2.420295
edatcat=18-25 31.25000     4.0   6.40 1.013616e-05 -4.414248
Pension=NO    45.32488    57.2  63.10 4.478523e-08 -5.470862
edatcat=66..  39.52703    23.4  29.60 1.177508e-09 -6.083277
Debito_aff=NO 45.72391    67.9  74.25 7.443526e-11 -6.511437
Nomina=NO     44.72362    62.3  69.65 7.657091e-13 -7.167158
VISA=NO       44.97732    69.4  77.15 9.794477e-17 -8.307251
```

Finalmente podemos ver la influencia de las variables continuas:

```
> res$quanti
```

```
$NO
                       v.test Mean in category Overall mean sd in category
dif_Libreta          7.812057        130.97747    -41.93691    764.9084504
oper_caj_Libreta     4.887006          5.94846     -7.40425    107.5474789
dif_Hipoteca         4.499247        182.97200     13.27750   1508.7356300
dif_Planes_pension   4.325662         -0.02706    -39.85842      0.8552844
dif_Prest_personales 3.238918         41.64200     17.50750    175.9117956
dif_Plazo            3.208897        238.50193    114.89096   1583.1429939
dif_Seguros          2.660853         58.43165     17.82179    735.3870111
antig               -2.043187         16.99700     17.38200      8.7746790
```

```
Total_Seguros          -3.823860        110.47800    278.98050   1104.5954026
Total_Inversion        -4.386919        483.79900    853.11600   3484.5482618
Total_activo           -6.048264        305.92500    617.97900   1688.8493904
Total_Plazo            -7.965293        673.75000   1332.66300   2433.6459454
Total_Vista           -10.139952        326.84500    569.17250    859.2195616
                       Overall sd      p.value
dif_Libreta            989.628317 5.626217e-15
oper_caj_Libreta       122.161099 1.023809e-06
dif_Hipoteca          1686.298085 6.819465e-06
dif_Planes_pension     411.698250 1.520746e-05
dif_Prest_personales   333.153741 1.199839e-03
dif_Plazo             1722.295311 1.332452e-03
dif_Seguros            682.365391 7.794303e-03
antig                    8.424789 4.103395e-02
Total_Seguros         1970.201891 1.313784e-04
Total_Inversion       3763.969398 1.149676e-05
Total_activo          2306.775852 1.464147e-09
Total_Plazo           3698.560413 1.648327e-15
Total_Vista           1068.496777 3.672815e-24

$SI
                        v.test Mean in category Overall mean sd in category
Total_Vista           10.139952        811.50000    569.17250    1194.850299
Total_Plazo            7.965293       1991.57600   1332.66300    4535.166258
Total_activo           6.048264        930.03300    617.97900    2755.986574
Total_Inversion        4.386919       1222.43300    853.11600    3989.995560
Total_Seguros          3.823860        447.48300    278.98050    2546.855669
antig                  2.043187         17.76700     17.38200       8.041313
dif_Seguros           -2.660853        -22.78806     17.82179     622.215940
dif_Plazo             -3.208897         -8.72000    114.89096    1842.742846
dif_Prest_personales  -3.238918         -6.62700     17.50750     435.744101
dif_Planes_pension    -4.325662        -79.68979    -39.85842     579.497275
dif_Hipoteca          -4.499247       -156.41700     13.27750    1831.209112
oper_caj_Libreta      -4.887006        -20.75696     -7.40425     133.879118
dif_Libreta           -7.812057       -214.85129    -41.93691    1146.230653
                       Overall sd      p.value
Total_Vista           1068.496777 3.672815e-24
Total_Plazo           3698.560413 1.648327e-15
Total_activo          2306.775852 1.464147e-09
Total_Inversion       3763.969398 1.149676e-05
Total_Seguros         1970.201891 1.313784e-04
antig                    8.424789 4.103395e-02
dif_Seguros            682.365391 7.794303e-03
dif_Plazo             1722.295311 1.332452e-03
dif_Prest_personales   333.153741 1.199839e-03
dif_Planes_pension     411.698250 1.520746e-05
dif_Hipoteca          1686.298085 6.819465e-06
oper_caj_Libreta       122.161099 1.023809e-06
dif_Libreta            989.628317 5.626217e-15
```

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).

XXX

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 (5.09%)	373 (94.91%)	393
Clase baja	6 (1.86%)	316 (98.14%)	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).

Si, la clase alta tiene un 0.05% de probabilidades de hacer la compra frente a un 0.018% de la clase baja; todo según este primer análisis que posteriormente es rebatido.

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 (1.68%)	176 (98.32%)	179
Clase baja	4 (1.35%)	293 (98.65%)	297
JOVENES	Compra SI	Compra NO	Total
Clase alta	17 (7.94%)	197 (92.06%)	214
Clase baja	2 (8%)	23 (92%)	25

¿Tenía razón el empleado de que era conveniente tener en cuenta la edad?. ¿Cuál de los dos factores es el determinante en la compra del producto en cuestión?

Si, el empleado tenía razón, hay que tener en cuenta la edad, de hecho ser joven es el factor determinante puesto que el grupo 'joven' tiene mucha más probabilidad de hacer la compra que el grupo 'adulto', incluso el 'joven' de clase 'baja' resulta ser quien más compraría, a pesar de tener una muestra pequeña.