Delivery 1 - Data Quality Report

En aquesta entrega es fa un analisis del conjunt de dades del fitxer cresco.txt. El conjunt de dades representa un conjunt de persones. Disposem d'una serie de caracteristiques d'aquestes persones, com per exemple, el tipus de feina, la vivenda, l'estat civil...

A partir d'aquesta informació hem de decidir si els donem un credit o els hi deneguem.

Les dades

A continuació veiem un resum de les dades que disposem

```
setwd("/Users/guillem/Google Drive/FIB/ADEI/R/R-working-directory")
base = read.table("dades.txt",header=T,sep='\t',na.string='99999999')
cresco<-base
summary(cresco)
```

```
##  dictamen      anys.feina        vivenda           plan
##  \032:   1   Min.   : 0.000   Min.   :0.000   Min.   : 6.00
##  0   :   1   1st Qu.: 2.000   1st Qu.:2.000   1st Qu.:36.00
##  1   :3200   Median : 5.000   Median :2.000   Median :48.00
##  2   :1254   Mean   : 7.987   Mean   :2.657   Mean   :46.44
##              3rd Qu.:12.000   3rd Qu.:4.000   3rd Qu.:60.00
##              Max.   :48.000   Max.   :6.000   Max.   :72.00
##              NA's   :1        NA's   :1       NA's   :1
##       edat        estat.civil      registres      tipus.feina
##  Min.   :18.00   Min.   :0.000   Min.   :1.000   Min.   :0.000
##  1st Qu.:28.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :36.00   Median :2.000   Median :1.000   Median :1.000
##  Mean   :37.08   Mean   :1.879   Mean   :1.174   Mean   :1.676
##  3rd Qu.:45.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:3.000
##  Max.   :68.00   Max.   :5.000   Max.   :2.000   Max.   :4.000
##  NA's   :1       NA's   :1       NA's   :1       NA's   :1
##     despeses        ingressos       patrimoni       carrecs.patr
##  Min.   : 35.00   Min.   :  0.0   Min.   :     0   Min.   :    0.0
##  1st Qu.: 35.00   1st Qu.: 80.0   1st Qu.:     0   1st Qu.:    0.0
##  Median : 51.00   Median :120.0   Median :  3000   Median :    0.0
##  Mean   : 55.57   Mean   :130.6   Mean   :  5403   Mean   :  342.9
##  3rd Qu.: 72.00   3rd Qu.:165.0   3rd Qu.:  6000   3rd Qu.:    0.0
##  Max.   :180.00   Max.   :959.0   Max.   :300000   Max.   :30000.0
##  NA's   :1        NA's   :35      NA's   :48       NA's   :19
##    import.sol      preu.be
##  Min.   : 100   Min.   :  105
##  1st Qu.: 700   1st Qu.: 1118
##  Median :1000   Median : 1400
##  Mean   :1039   Mean   : 1463
##  3rd Qu.:1300   3rd Qu.: 1692
##  Max.   :5000   Max.   :11140
##  NA's   :1      NA's   :1
```

Anàlisis de les dades

Per tal de poder tractar amb les dades hem de fer un primer anàlisis d'aquestes per tal de trobar aquelles que son errates o que tenen un valor molt poc coherent amb la resta de variables que disposem individuos.

La primera operació que fem sobre el dataset es un head i un tail

```
head(base, 2)
```

```
##   dictamen anys.feina vivenda plan edat estat.civil registres tipus.feina
## 1        1          9       1   60   30           2         1           3
## 2        1         17       1   60   58           3         1           1
##   despeses ingressos patrimoni carrecs.patr import.sol preu.be
## 1       73       129         0            0        800     846
## 2       48       131         0            0       1000    1658
```

```
tail(base,2)
```

```
##      dictamen anys.feina vivenda plan edat estat.civil registres
## 4455        1          5       2   60   32           2         1
## 4456     \032         NA      NA   NA   NA          NA        NA
##      tipus.feina despeses ingressos patrimoni carrecs.patr import.sol
## 4455           3       60       140      4000         1000       1350
## 4456          NA       NA        NA        NA           NA         NA
##      preu.be
## 4455    1650
## 4456      NA
```

Al fer el tail() ens donem compte que el ultim individuo es erroni totalment i decidim borrar-lo del dataset.

```
badInd = cresco[4456,]
badInd
```

```
##      dictamen anys.feina vivenda plan edat estat.civil registres
## 4456     \032         NA      NA   NA   NA          NA        NA
##      tipus.feina despeses ingressos patrimoni carrecs.patr import.sol
## 4456          NA       NA        NA        NA           NA         NA
##      preu.be
## 4456      NA
```

```
cresco <- cresco[-4456,]
```

Per començar mirarem el nombre de variables que tenen algun individu sense emplenar, es a dir, que es un NaN.

```
numberNArow <- function(dataFrame, output) {
  output = sum(is.na(dataFrame))
}

missXrow = apply(base, 1, numberNArow, output = 'outputfile')
missXcol = apply(base, 2, numberNArow, output = 'outputfile')
```

Despres d'executar aquest codi tenim el nombre de NaN que hi ha per fila dins la següent variable (no mostrem el resultat perque un vector d'uns i zeros amb totes les files que te el nostre seet de dades):

```
missXrow
```

I per columna:

```
missXcol
```

```
##     dictamen   anys.feina      vivenda         plan         edat
##            0            1            1            1            1
##  estat.civil    registres  tipus.feina     despeses    ingressos
##            1            1            1            1           35
##    patrimoni carrecs.patr   import.sol      preu.be
##           48           19            1            1
```

Si mirem el total de NaN que tenim en el nostre set de dades tenim que:

```
sum(missXcol)
```

```
## [1] 112
```

Que ha de ser igual a la suma de valors per fila:

```
sum(missXrow)
```

```
## [1] 112
```

Per tal d'afegir aquestes dades en el nostre set de dades utilitzam les següents comandes:

```
rbind(cresco,missXcol)
cresco$missValues <- missXrow
```

Anàlisis dels atributs

Cada individu del conjunt de dades te els següents atributs: dictamen ,anys.feina ,vivenda ,plan ,edat ,estat.civil ,registres ,tipus.feina ,despeses ,ingressos ,patrimoni i carrecs.patr

Alguns d'aquets atributs son variables continues i altres categoriques, començarem fent el analisis de les categoriques.

El que farem serà mirar quins son els posibles valor que poden prendre i mirar quins individuos no compleixen aquets requesits

Analisis de les variables categoriques

Les variables catagoriques de les que disposem en el nostre dataset són les següents: vivenda, estat.civil i tipus.feina.

Analisis de la vivenda

```
v <- cresco$vivenda
llista <-which(v != 1 & v != 2 & v != 3 & v != 4 & v != 5 & v != 6)
llista
```

```
## [1]   30  240 1060 1677 2389 2996
```

Tenim 6 individuos que no compleixen els valors definits al enunciat.

Analisis de la estat civil

```
v <- cresco$estat.civil
llista <-which(v != 1 & v != 2 & v != 3 & v != 4 & v != 5)
llista
```

```
## [1] 3320
```

Tenim un element que no compleix els valors definits al enunciat

Analisis de la tipus feina

```
v <- cresco$tipus.feina
llista <-which(v != 1 & v != 2 & v != 3 & v != 4)
llista
```

```
## [1]  30 912
```

Tenim 2 elements que no compleixen els valors definits al enunciat

Analisis de les variables continues

Les variables continues les estudiarem d'una manera diferent a les categoriques. Les variables continues del nostre set de dades son les seguents: anys.feina, plan, edat, despeses, ingressos, patrimoni, carrecs.patrimoni, import.sol i preu.be.

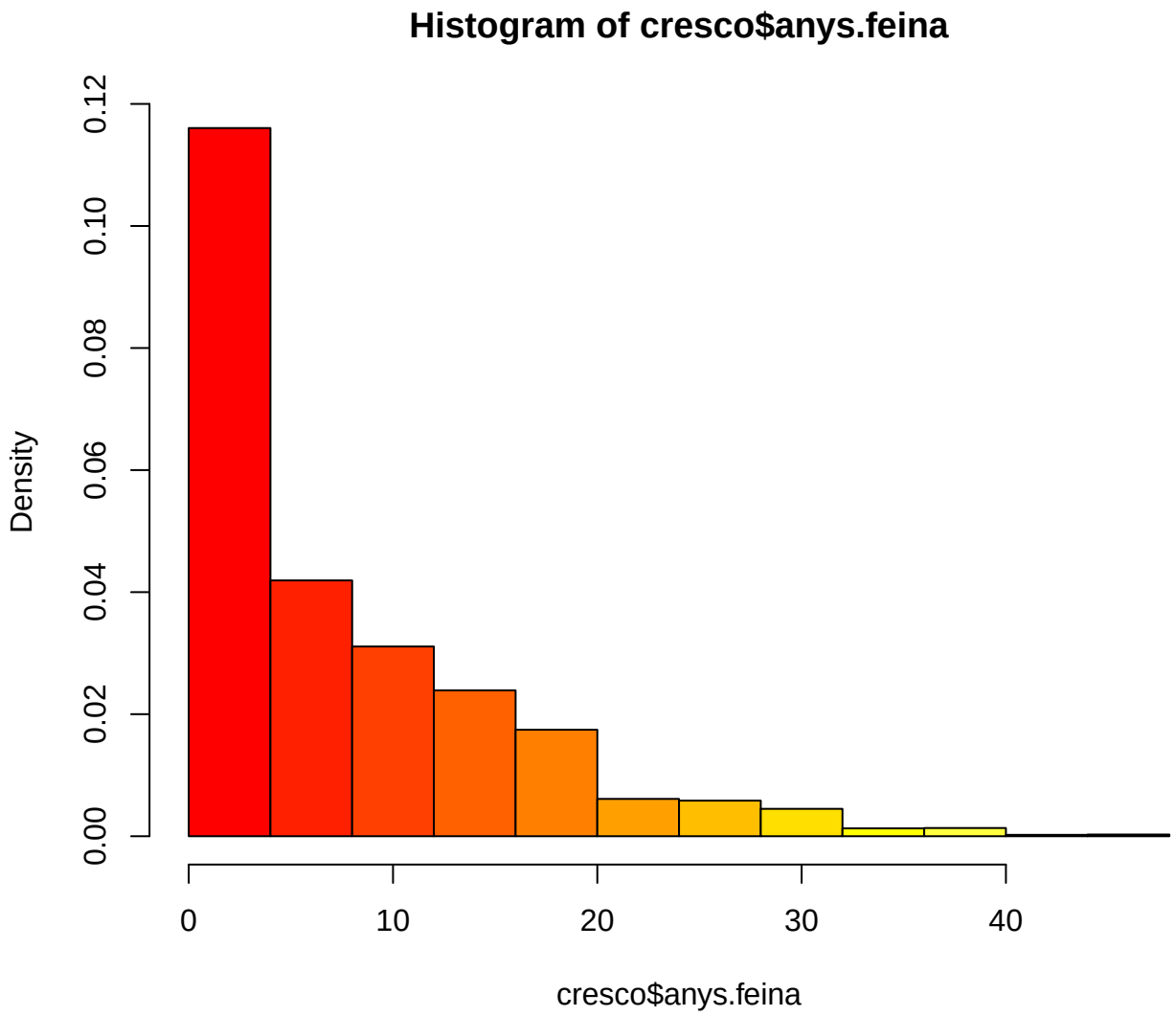
Per començar a analitzar les variables continues mirarem quens dels individuos són outliers per cada atribut. Només buscarem outliers extrems i la formula que seguirem és la següent: (Q1-3×IQR, Q3+3×IQR)

A a partir d'aquí, utilitzarem la funció escrita a continuació per a trobar el llindar a partir del cual, una variable es un outlier.

```
calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, souts=s.x[5]+3*iqr )
  }
```

Anàlisis de la variable anys.feina

Comencem mirant un histograma de la variable:



Podem comprovar que aquesta variable no segueix una distribució normal i que en el cas de que tinguem algun outlier aquest tindrà un valor semblant al 40 o major.

Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$anys.feina)
```

```
## $souti
## 1st Qu.
##     -28
##
## $souts
## 3rd Qu.
##      42
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 42 o menor que -28 (com que no te una distribució normal no en tenim cap) son outliers extrems.

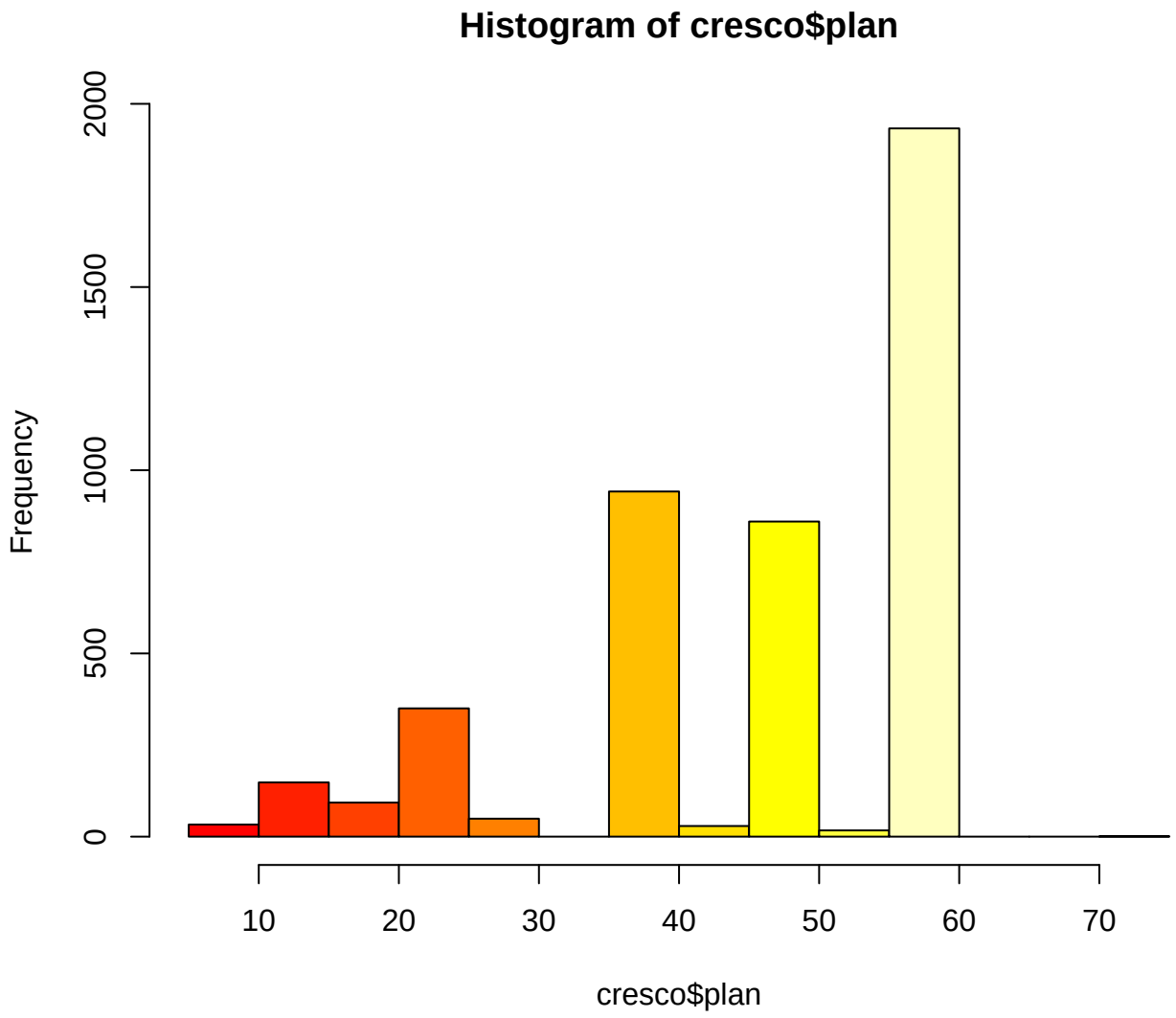
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which( cresco$anys.feina > 42);
llista
```

```
## [1] 1036 2586 3461 3524 3532 4155 4255
```

Anàlisis de la variable plan

Comencem mirant un histograma de la variable:



Podem comprovar que aquesta variable no segueix una distribució normal i que en el cas de que tinguem algun outlier aquest tindrà un valor semblant al 40 o major.

Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$plan)
```

```
## $souti
## 1st Qu.
##     -36
##
## $souts
## 3rd Qu.
##     132
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 132 o menor que -36 (com que no te una distribució normal no en tenim cap) son outliers extrems.

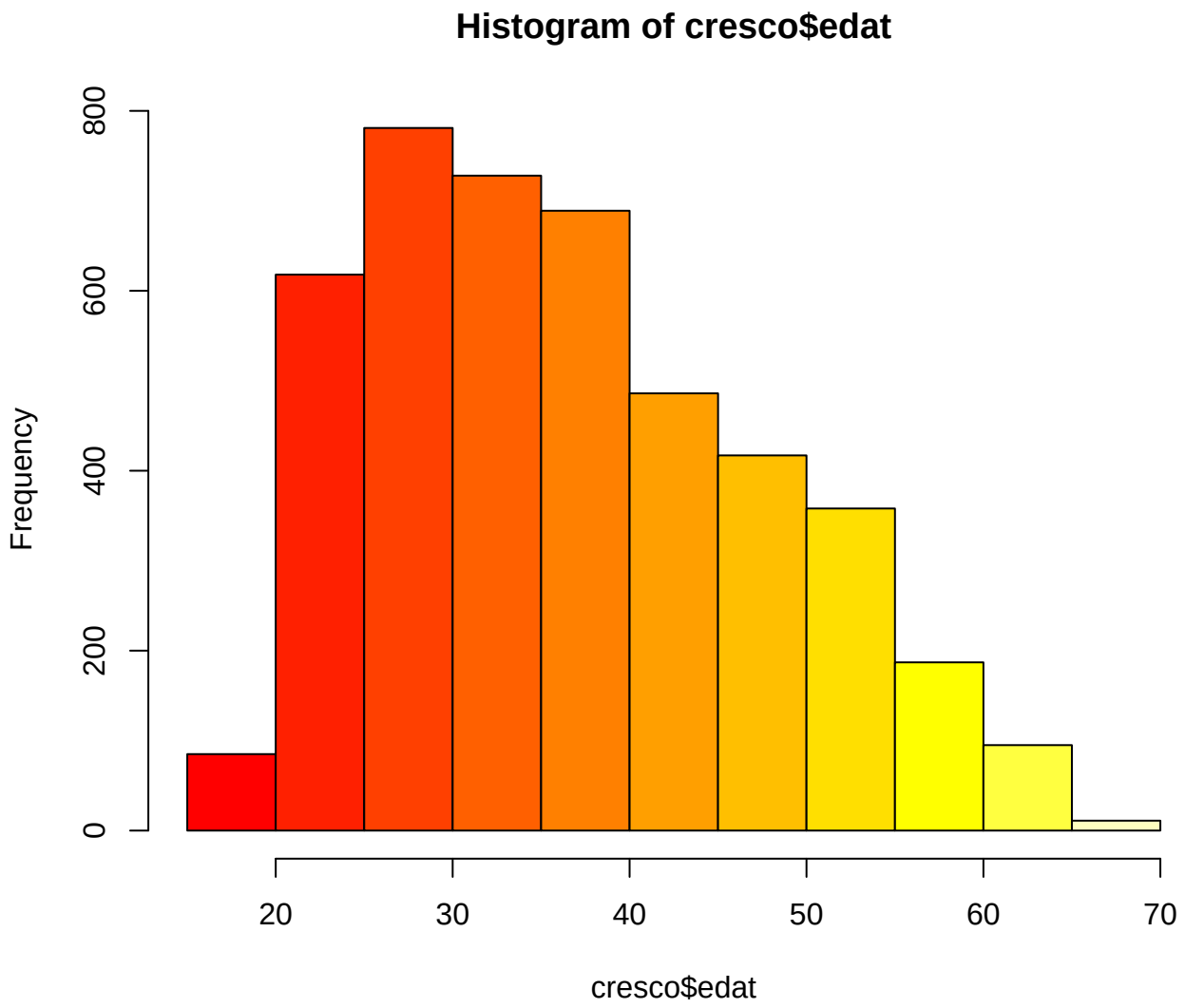
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which( cresco$plan >= 132);
llista
```

```
## integer(0)
```

Anàlisis de la variable edat

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$edat)
```

```
## $souti
## 1st Qu.
##     -23
##
## $souts
## 3rd Qu.
##      96
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 96 o menor que -23 (com que no te una distribució normal no en tenim cap) son outliers extrems.

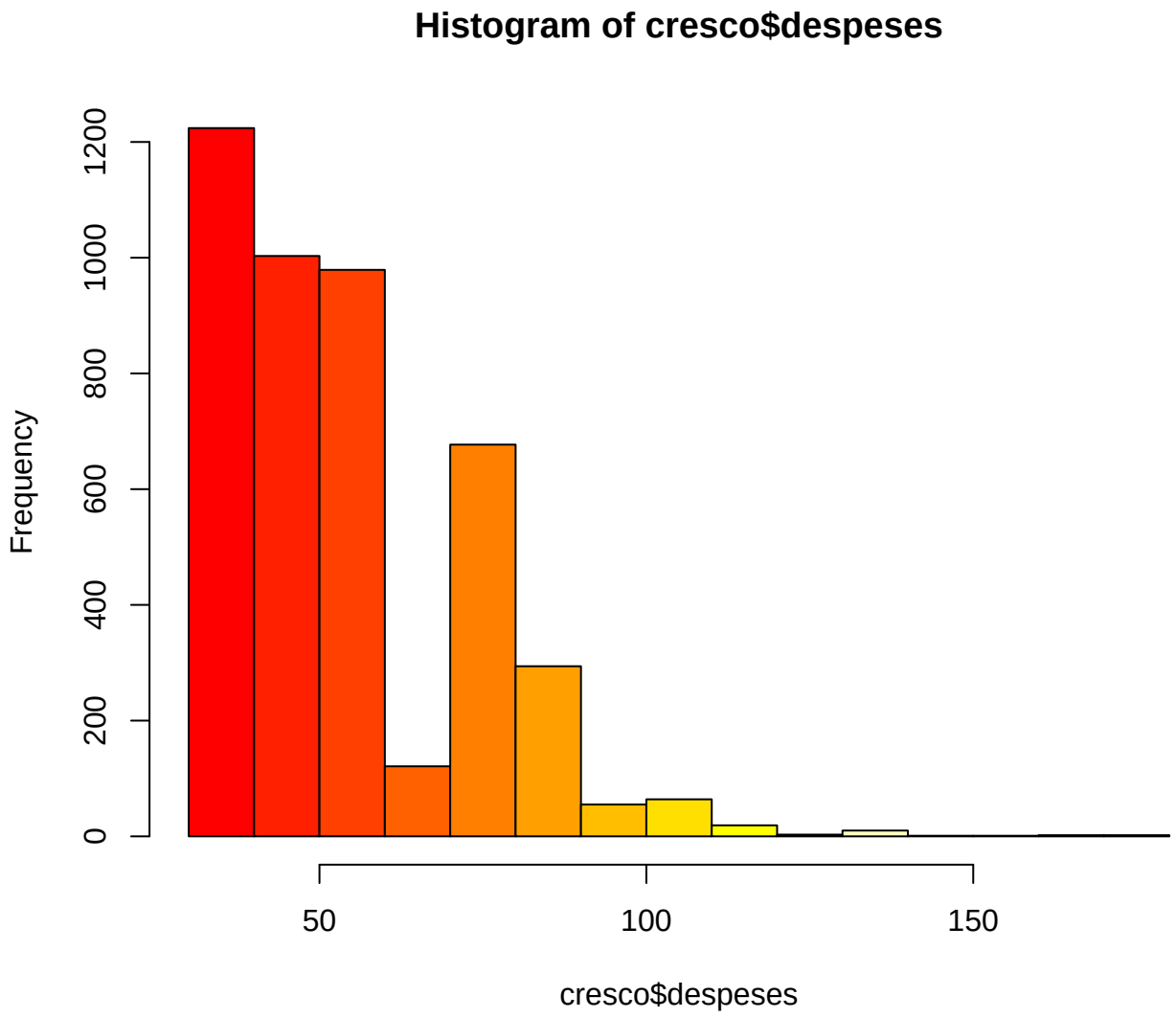
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which( cresco$edat >= 96);
llista
```

```
## integer(0)
```

Anàlisis de la variable despeses

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$despeses)
```

```
## $souti
## 1st Qu.
##     -76
##
## $souts
## 3rd Qu.
##     183
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 183 o menor que -76 (com que no te una distribució normal no en tenim cap) son outliers extrems.

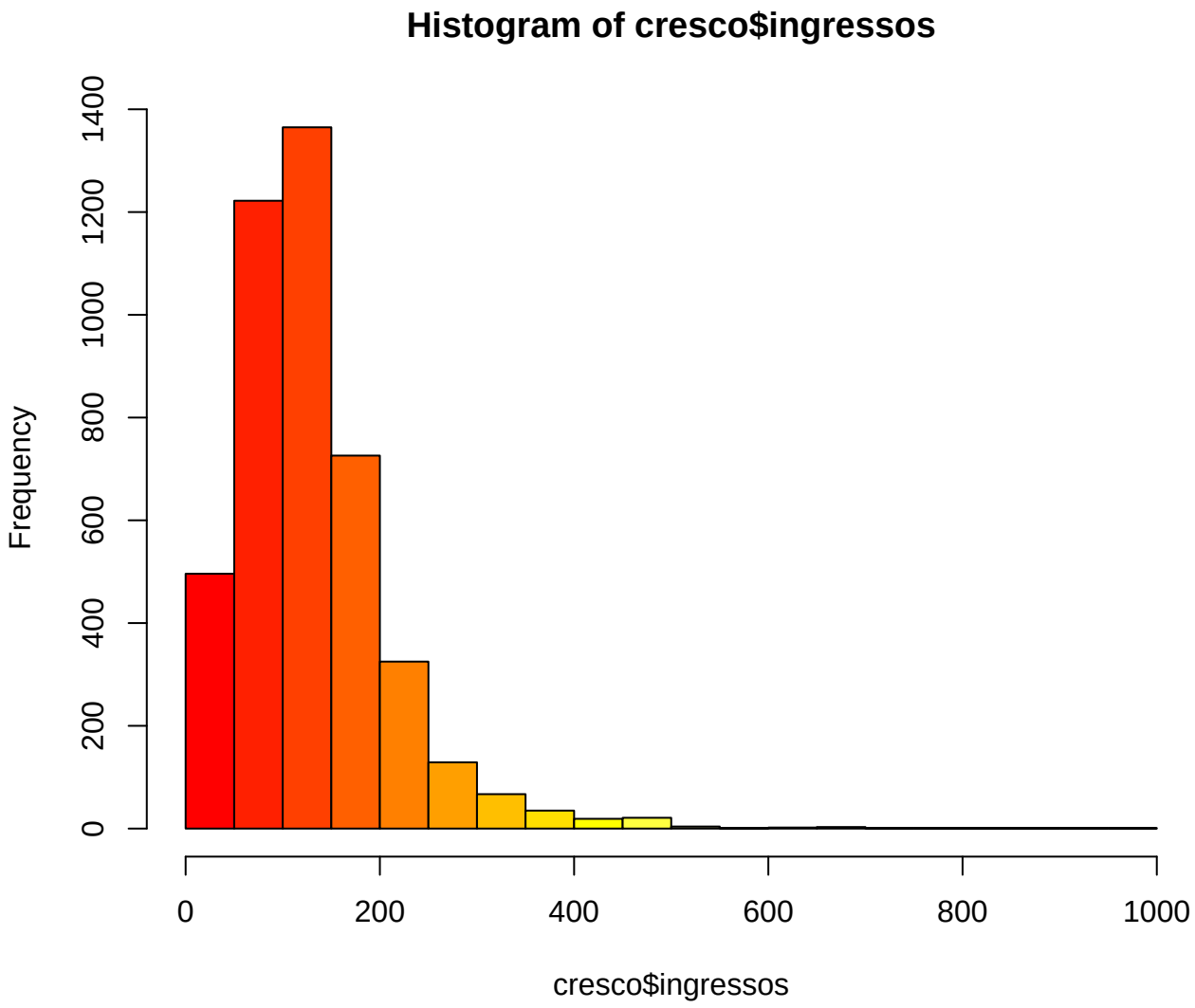
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which(cresco$despeses >= 183);
llista
```

```
## integer(0)
```

Anàlisis de la variable ingressos

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$ingressos)
```

```
## $souti
## 1st Qu.
##    -175
```

```
##
## $souts
## 3rd Qu.
##     420
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 420 o menor que -175 (com que no te una distribució normal no en tenim cap) son outliers extrems.

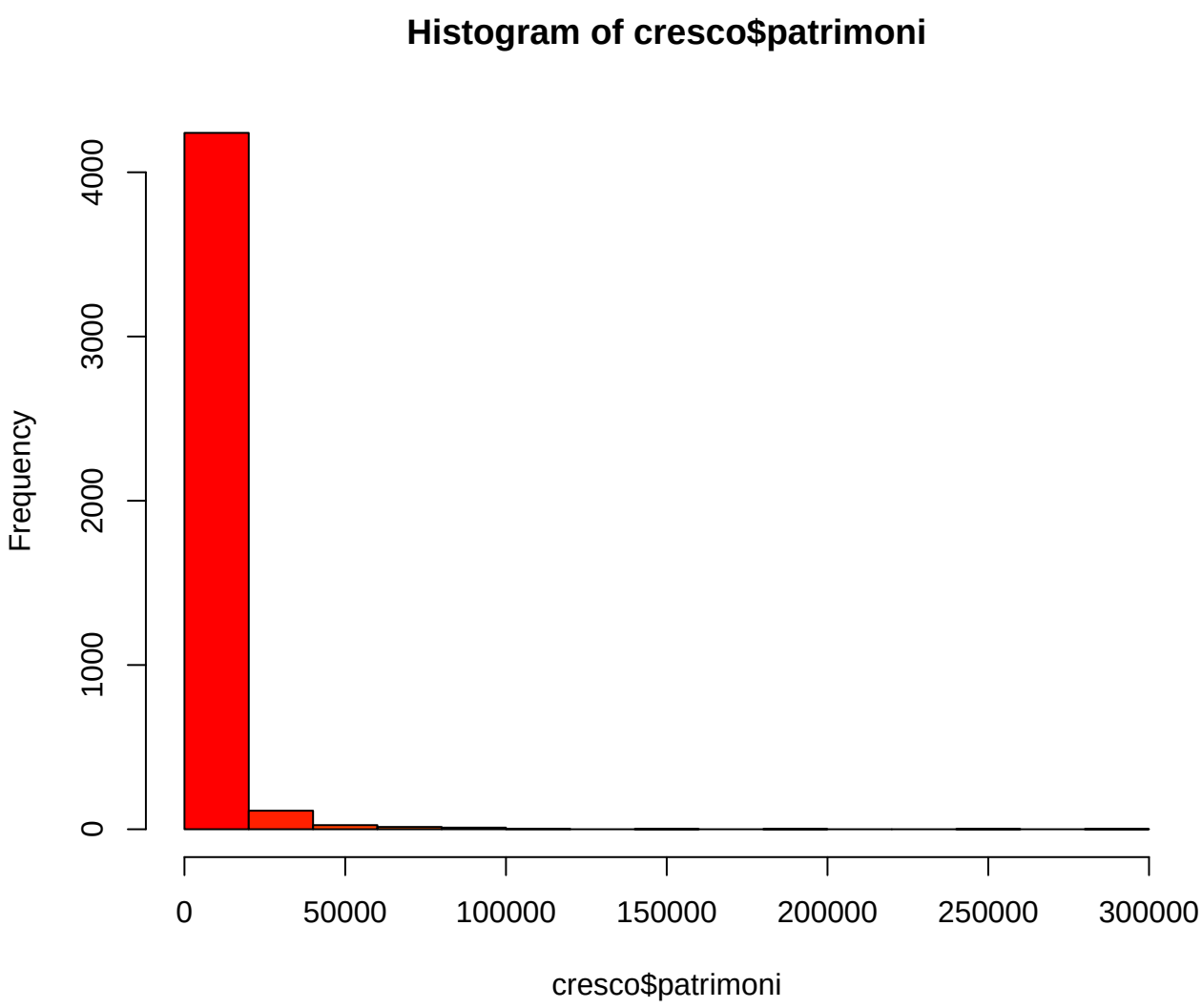
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which(cresco$ingressos >= 420);
llista
```

```
##  [1]   60   77   94  126  145  287  328  368  391  601  633  680  755  774
## [15]  807  813  839  920  921  964  976 1064 1073 1230 1303 1313 1368 1370
## [29] 1419 1428 1474 1620 1723 2008 2069 2390 2519 2576 2630 2655 2737 2830
## [43] 2865 2873 3247 3589 3730 4089 4266 4365
```

Anàlisis de la variable patrimoni

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$patrimoni)
```

```
## $souti
## 1st Qu.
##  -18000
##
## $souts
## 3rd Qu.
##   24000
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 24000 o menor que -18000 (com que no te una distribució normal no en tenim cap) son outliers extrems.

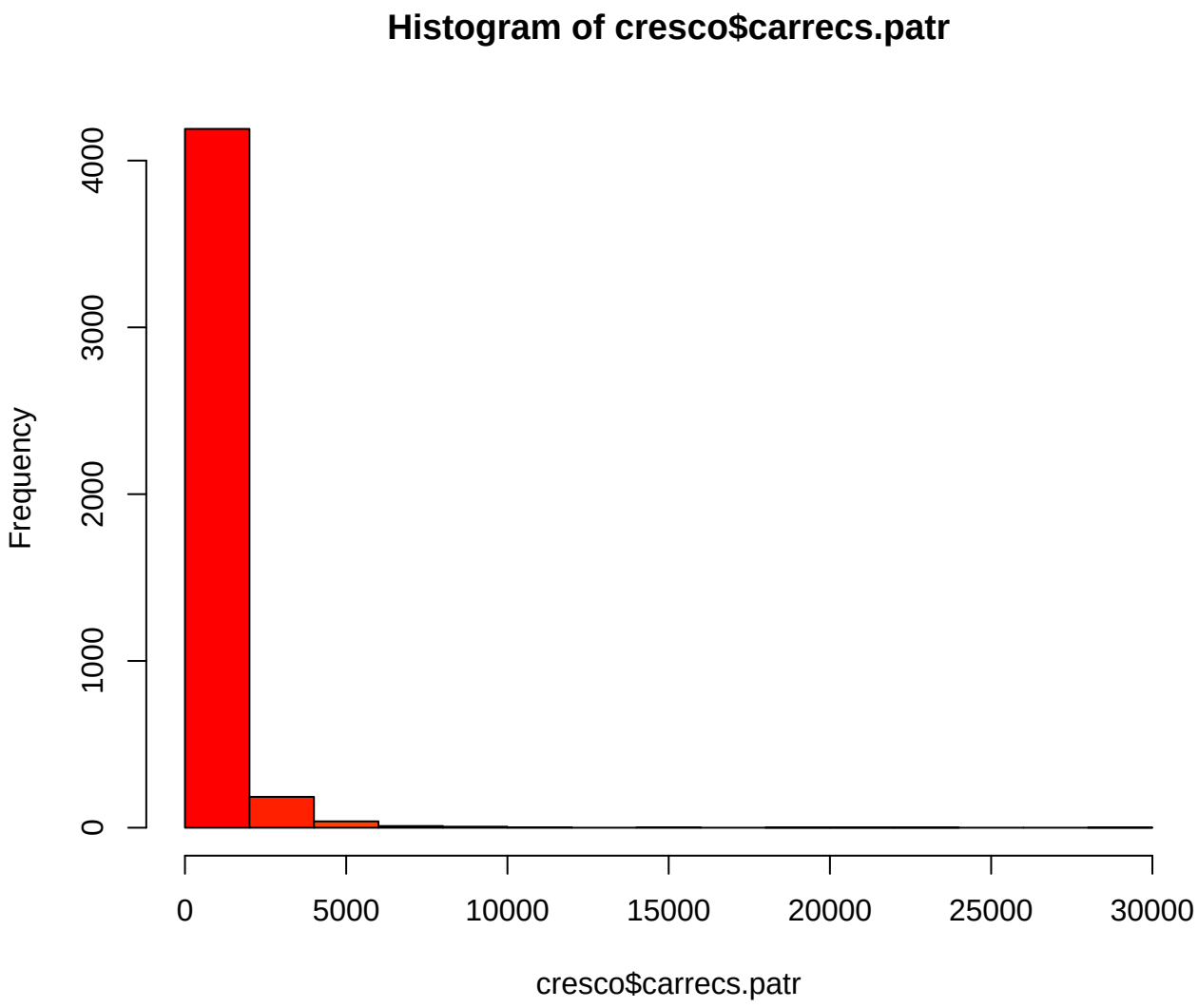
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which(cresco$patrimoni >= 24000);
llista
```

```
##   [1]   43   79   94  126  134  143  144  166  177  219  232  284  286  294
##  [15]  391  394  399  555  614  644  648  667  698  763  791  837  845  873
##  [29]  878  921  964 1042 1100 1208 1233 1277 1325 1346 1365 1370 1392 1502
##  [43] 1533 1548 1651 1676 1682 1684 1693 1702 1710 1769 1792 1802 1809 1817
##  [57] 1851 1960 1972 1973 2003 2008 2009 2024 2069 2071 2085 2089 2109 2128
##  [71] 2139 2255 2277 2288 2310 2331 2334 2390 2403 2464 2485 2494 2525 2568
##  [85] 2611 2619 2655 2657 2676 2708 2737 2778 2872 2873 2917 2924 2944 2954
##  [99] 2971 2979 3008 3040 3047 3049 3051 3052 3110 3121 3223 3227 3233 3277
## [113] 3304 3325 3327 3339 3348 3353 3415 3439 3467 3494 3524 3539 3568 3608
## [127] 3730 3732 3739 3745 3758 3770 3806 3892 3950 3998 4056 4086 4123 4130
## [141] 4133 4142 4199 4254 4290 4338 4377 4387 4404 4412 4421 4442
```

Anàlisis de la variable carrecs.patr

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$carrecs.patr)
```

```
## $souti
## 1st Qu.
##       0
##
## $souts
## 3rd Qu.
##       0
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 24000 o menor que -18000 (com que no te una distribució normal no en tenim cap) son outliers extrems.

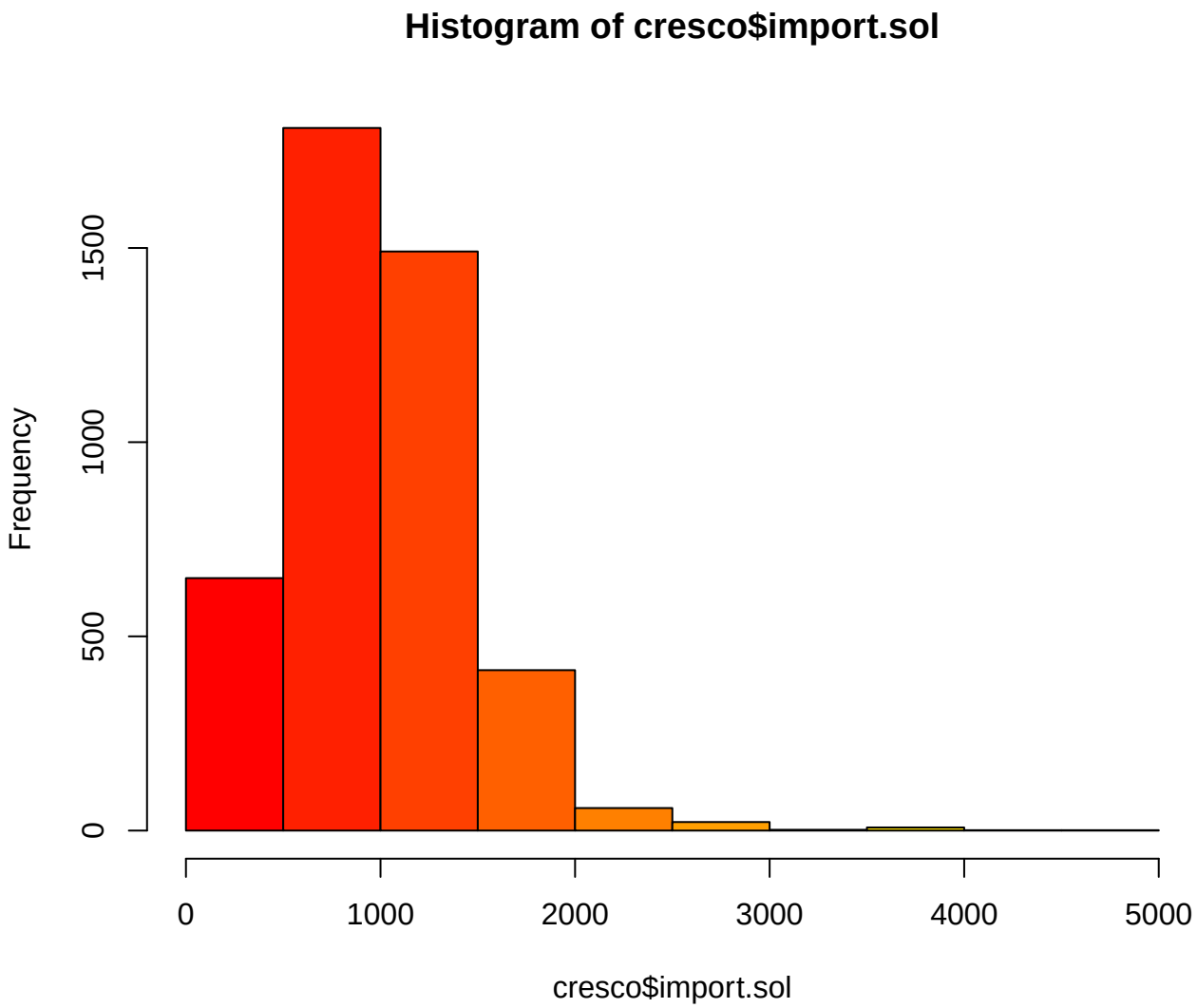
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which(cresco$carrecs.patr >= 24000);
llista
```

```
## [1] 1392
```

Anàlisis de la variable import.sol

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$import.sol)
```

```
## $souti
## 1st Qu.
##   -1100
##
## $souts
## 3rd Qu.
##    3100
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 3100 o menor que -1100 (com que no te una distribució normal no en tenim cap) son outliers extrems.

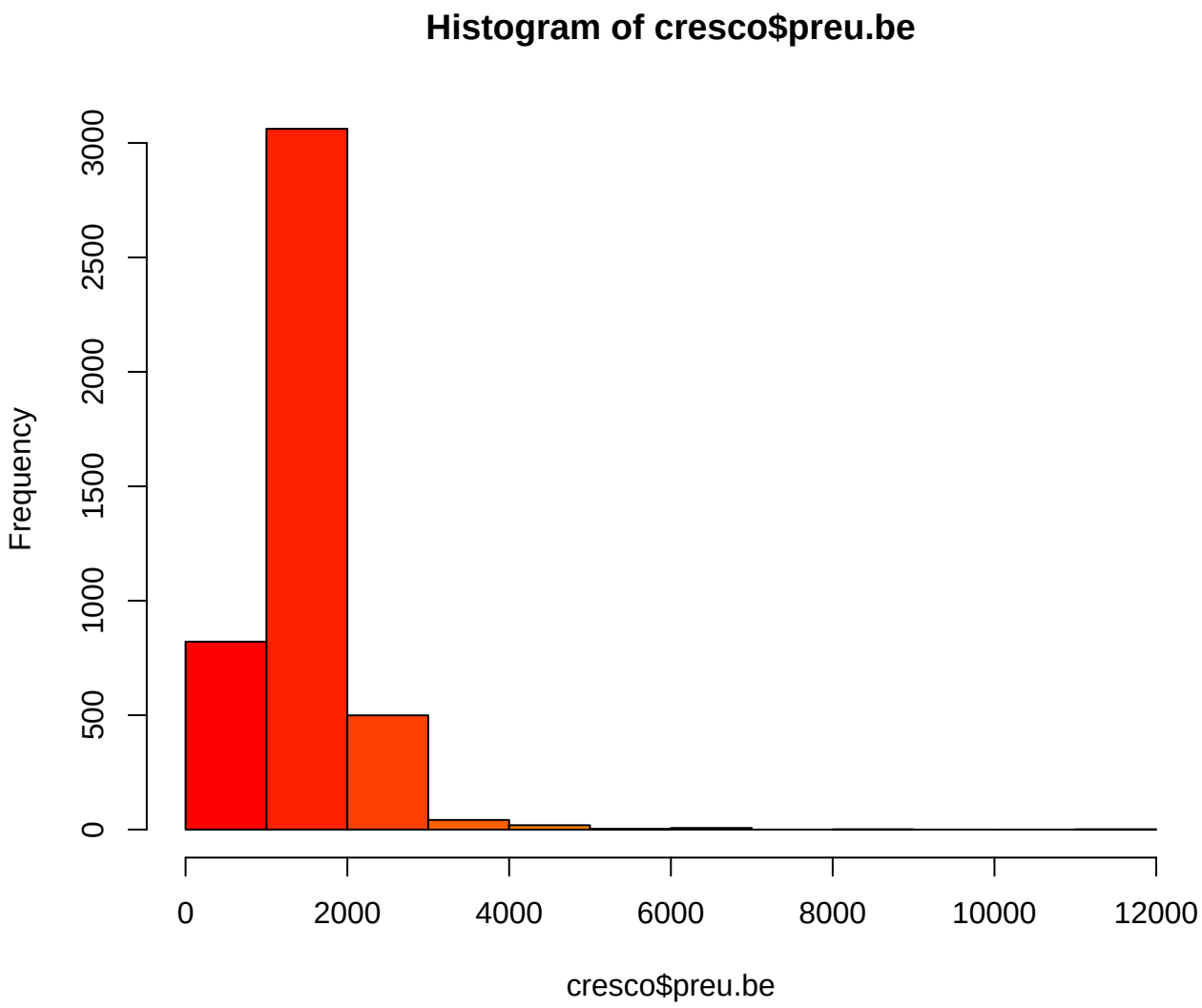
Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which(cresco$import.sol >= 3100);
llista
```

```
##  [1]  143  312 1312 1325 1893 2390 2476 2690 2951 3139 3929
```

Anàlisis de la variable preu.be

Comencem mirant un histograma de la variable:



Per tal de trobar els outliers apliquem la formula descrita anteriorment:

```
calcQ(cresco$preu.be)
```

```
## $souti
## 1st Qu.
##    -604
##
## $souts
## 3rd Qu.
##    3414
```

Al executar-la ens donem compte que les variables que tinguin un valor major de 3100 o menor que -604 (com que no te una distribució normal no en tenim cap) son outliers extrems.

Ara mirem quens calors d'aquesta variable tenen un balor superior a aquest valor:

```
llista<-which(cresco$preu.be >= 3414);
llista
```

```
##  [1]  126  143  156  269  284  312  693  763  864 1312 1325 1357 1614 1738
## [15] 1788 1883 1893 1979 2140 2156 2241 2390 2399 2476 2563 2645 2690 2705
## [29] 2778 2816 2951 3116 3139 3277 3306 3337 3339 3494 3516 3543 3674 3730
## [43] 3774 3857 3929 3952 4069 4276 4359 4370 4387
```

Anàlisis de la relació que hi ha entre les dades

Per tal de mirar la relació qui hi ha entre cada atribut de les dades que disposem amb la resta utilitzarem la llibreria anomenada: FactoMineR

```
library("FactoMineR")
```

Ara farem un estudi per cada variable categorica amb la resta de variables que disposem

Comparativa entre el atribut anys.feina i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="anys.feina"))
```

```
## $quanti
##             correlation       p.value
## edat          0.50578704 4.610031e-288
## estat.civil   0.16343290  4.806745e-28
## ingressos     0.13303017  6.551692e-19
## patrimoni     0.12686583  2.796466e-17
## despeses      0.12579829  3.526830e-17
## preu.be       0.04092199  6.299933e-03
## tipus.feina -0.10933905  2.520297e-13
## vivenda     -0.14587753  1.288383e-22
##
## $quali
##                   R2      p.value
## dictamen 0.06781714 1.286707e-68
##
## $category
##    Estimate      p.value
## 1  1.350401 7.197726e-70
## 2 -3.381740 5.707432e-70
```

Comparativa entre el atribut plan i la resta d'atributs

```r
condes(cresco,num.var=which(colnames(cresco)=="plan"))
```

```
## $quanti
##              correlation       p.value
## import.sol    0.43105036 4.788363e-201
## preu.be       0.12979560  3.386997e-18
## carrecs.patr  0.05768629  1.207249e-04
## estat.civil   0.04997839  8.468908e-04
## edat         -0.05196556  5.207427e-04
## patrimoni    -0.08487761  1.662907e-08
## tipus.feina  -0.13845433  1.641680e-20
##
## $quali
##                   R2      p.value
## dictamen 0.01031611 9.445277e-11
##
## $category
##    Estimate      p.value
## 2 -2.642369 1.792931e-11
## 1 -5.921003 1.464556e-11
```

Comparativa entre el atribut edat i la resta d'atributs

```r
condes(cresco,num.var=which(colnames(cresco)=="edat"))
```

```
## $quanti
##              correlation       p.value
## anys.feina    0.50578704 4.610031e-288
## estat.civil   0.32565230 1.457308e-110
## despeses      0.24825422  1.486017e-63
## patrimoni     0.18495833  3.214327e-35
## tipus.feina   0.17944667  1.483697e-33
## ingressos     0.11493086  1.787311e-14
## registres     0.06017052  5.853426e-05
## preu.be       0.04873927  1.137274e-03
## carrecs.patr -0.04581010  2.271781e-03
## plan         -0.05196556  5.207427e-04
## vivenda      -0.27078407  1.020309e-75
##
## $quali
##                    R2      p.value
## dictamen 0.009351241 8.265163e-10
##
## $category
##   Estimate      p.value
## 1 5.020986 1.580092e-10
## 2 2.693654 1.989910e-10
```

Comparativa entre el atribut despeses i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="despeses"))
```

```
## $quanti
##              correlation       p.value
## edat          0.24825422  1.486017e-63
## ingressos     0.23983127  7.129094e-59
## estat.civil   0.21031493  1.016421e-45
## anys.feina    0.12579829  3.526830e-17
## registres     0.05732897  1.289044e-04
## import.sol    0.04895832  1.080028e-03
## preu.be       0.04016226  7.340450e-03
## vivenda      -0.33409246 1.289858e-116
```

Comparativa entre el atribut ingressos i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="ingressos"))
```

```
## $quanti
##               correlation      p.value
## despeses       0.23983127 7.129094e-59
## preu.be        0.15932792 1.583446e-26
## patrimoni      0.14687224 1.540814e-22
## import.sol     0.13306912 6.398886e-19
## anys.feina     0.13303017 6.551692e-19
## edat           0.11493086 1.787311e-14
## carrecs.patr   0.10561696 2.090841e-12
## estat.civil    0.06893984 4.474818e-06
## vivenda       -0.11970917 1.393703e-15
## tipus.feina   -0.13631964 8.711412e-20
##
## $quali
##                   R2      p.value
## dictamen 0.04384669 9.667659e-44
##
## $category
##    Estimate      p.value
## 1  11.37614 6.277882e-45
## 2 -28.93402 5.406576e-45
```

Comparativa entre el atribut patrimoni i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="patrimoni"))
```

```
## $quanti
##              correlation      p.value
## tipus.feina   0.20861279 1.542092e-44
## preu.be       0.19955167 7.934644e-41
## carrecs.patr  0.19070397 2.256113e-37
## edat          0.18495833 3.214327e-35
## import.sol    0.14728068 8.424857e-23
## ingressos     0.14687224 1.540814e-22
## anys.feina    0.12686583 2.796466e-17
## estat.civil   0.06517727 1.486513e-05
## vivenda      -0.07819857 2.011100e-07
## plan         -0.08487761 1.662907e-08
##
## $quali
##                    R2      p.value
## dictamen 0.009430023 8.651017e-10
##
## $category
##    Estimate      p.value
## 1 1868.7213 1.043419e-10
## 2 -632.9676 1.081375e-10
```

Comparativa entre el atribut carrecs.patr i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="carrecs.patr"))
```

```
## $quanti
##             correlation      p.value
## patrimoni    0.19070397 2.256113e-37
## ingressos    0.10561696 2.090841e-12
## plan         0.05768629 1.207249e-04
## import.sol   0.05244815 4.739995e-04
## preu.be      0.04557643 2.392584e-03
## estat.civil  0.03685990 1.407239e-02
## tipus.feina  0.03115404 3.797588e-02
## edat        -0.04581010 2.271781e-03
## vivenda     -0.08789108 4.513105e-09
```

Comparativa entre el atribut import.sol i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="import.sol"))
```

```
## $quanti
##              correlation       p.value
## preu.be       0.72503979  0.000000e+00
## plan          0.43105036 4.788363e-201
## patrimoni     0.14728068  8.424857e-23
## ingressos     0.13306912  6.398886e-19
```

```
## registres     0.11034947  1.514704e-13
## tipus.feina   0.05583264  1.926564e-04
## estat.civil   0.05413042  3.007954e-04
## carrecs.patr  0.05244815  4.739995e-04
## despeses      0.04895832  1.080028e-03
##
## $quali
##                  R2     p.value
## dictamen 0.02409233 2.65362e-24
##
## $category
##    Estimate      p.value
## 2  -60.2954 3.609229e-25
## 1 -223.3462 2.625021e-25
```

Comparativa entre el atribut preu.be i la resta d'atributs

```
condes(cresco,num.var=which(colnames(cresco)=="preu.be"))
```

```
## $quanti
##              correlation      p.value
## import.sol    0.72503979 0.000000e+00
## patrimoni     0.19955167 7.934644e-41
## ingressos     0.15932792 1.583446e-26
## plan          0.12979560 3.386997e-18
## registres     0.08514332 1.257803e-08
## estat.civil   0.06253675 2.954022e-05
## tipus.feina   0.05649706 1.613637e-04
## edat          0.04873927 1.137274e-03
## carrecs.patr  0.04557643 2.392584e-03
## anys.feina    0.04092199 6.299933e-03
## despeses      0.04016226 7.340450e-03
```

Mirant les taules de comparatives anteriors podem veure el seguent: podem comprovar que hi ha molta relació entre la variable anys.feina i la variable edat, la qual cosa te molt de sentit ja que una persona amb més edat pot tenir més anys treballats. També hi ha molta relació entre ingressos i despeses, i entre patrimoni i tipus feina.