Visualització de dades: Pràctica 2

Autor: Anna Mayoral Hernando

Gener 2023

Contents

Introducció

Presentació

Aquest exercici correspont a la practica 2 de l'Assignatura de visualització de dades del màster universitaria de ciència de dades de la UOC.

Objectius

- Desenvolupar una visualització de dades mitjançant l'ús de diferents eines i tècniques, basades en el conjunt de dades d'una entitat sense de lucre (ONG) i validat a la primera part de la pràctica. Per realitzar la visualització el més útil possible realitzaré un anàlisis previ de la base dades per detectar necessitats (valors perduts, extrems, homegeïtzar dades, extracció de característiques etc)

Anàlisi exploratòria

El primer que farem serà un anàlisis exploratori de les dades del dataset, per això carregarem el fitxer de dades:

```
# Carreguem les llibreries necessàries
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if(!require('Rmisc')) install.packages('Rmisc'); library('Rmisc')
if(!require('dplyr')) install.packages('dplyr'); library('dplyr')
if(!require('xfun')) install.packages('xfun'); library('xfun')
if(!require('readxl')) install.packages('xfun'); library('readxl')

# Càrrega de l'arxiu
path = 'colaboradores.xlsx'
sociosData <-  read_xlsx(path)

# Mostrem les primeres dades
head(sociosData)
```

```
## # A tibble: 6 x 17
##      ID Estado Origen Sexo  Fecha_Alta          Fecha_Baja          Motivo_Baja
##   <dbl> <chr>  <chr>  <chr> <dttm>              <dttm>              <chr>
## 1     3 Baja   TLMK   H     2017-07-12 00:00:00 2019-04-24 00:00:00 Defunción
## 2    16 Activa TLMK   H     2017-09-07 00:00:00 NA                  <NA>
## 3    24 Activa ONLINE H     2019-09-02 00:00:00 NA                  <NA>
## 4    42 Activa SEDE   H     2020-03-02 00:00:00 NA                  <NA>
## 5    48 Activa TLMK   H     2017-03-05 00:00:00 NA                  <NA>
## 6    60 Activa TLMK   H     2021-05-27 00:00:00 NA                  <NA>
## # ... with 10 more variables: Cuota_Anualizada <dbl>,
## #   Num_Devoluciones_Total <dbl>, Edad <dbl>, Country <chr>, Provincia <chr>,
## #   CCAA <chr>, CP <chr>, Aceptar_recibir_info <dbl>, Idioma <chr>,
## #   Profesion <chr>
```

Verifiquem l'estructura del joc de dades principal. Veiem el nombre de columnes que tenim i exemples dels continguts de les files.

```
# Resum
str(sociosData)
```

```
## tibble [84,121 x 17] (S3: tbl_df/tbl/data.frame)
##  $ ID                    : num [1:84121] 3 16 24 42 48 60 73 76 77 103 ...
##  $ Estado                : chr [1:84121] "Baja" "Activa" "Activa" "Activa" ...
##  $ Origen                : chr [1:84121] "TLMK" "TLMK" "ONLINE" "SEDE" ...
##  $ Sexo                  : chr [1:84121] "H" "H" "H" "H" ...
##  $ Fecha_Alta            : POSIXct[1:84121], format: "2017-07-12" "2017-09-07" ...
##  $ Fecha_Baja            : POSIXct[1:84121], format: "2019-04-24" NA ...
##  $ Motivo_Baja           : chr [1:84121] "Defunción" NA NA NA ...
##  $ Cuota_Anualizada      : num [1:84121] 1008 120 240 600 360 ...
##  $ Num_Devoluciones_Total: num [1:84121] 1 0 0 0 0 0 3 0 0 0 ...
##  $ Edad                  : num [1:84121] 83 84 82 83 83 67 85 NA 93 66 ...
```

```
##  $ Country              : chr [1:84121] "Spain" "Spain" "Spain" "Spain" ...
##  $ Provincia            : chr [1:84121] "Barcelona" "Barcelona" "Barcelona" "Barcelona" ...
##  $ CCAA                 : chr [1:84121] "Catalunya" "Catalunya" "Catalunya" "Catalunya" ...
##  $ CP                   : chr [1:84121] "08035" "08720" "08017" "08006" ...
##  $ Aceptar_recibir_info : num [1:84121] 0 1 1 1 1 1 0 1 1 1 ...
##  $ Idioma               : chr [1:84121] "CAT" "CAT" "ESP" "CAT" ...
##  $ Profesion            : chr [1:84121] NA "FINAN" "PENS" NA ...
```

```r
# resum estadístic
summary(sociosData)
```

```
##        ID             Estado             Origen             Sexo
##  Min.   :      3   Length:84121       Length:84121       Length:84121
##  1st Qu.: 110535   Class :character   Class :character   Class :character
##  Median : 319648   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 433714
##  3rd Qu.: 792347
##  Max.   :1099931
##
##    Fecha_Alta                    Fecha_Baja                   Motivo_Baja
##  Min.   :2013-01-03 00:00:00   Min.   :2013-12-20 00:00:00   Length:84121
##  1st Qu.:2018-04-10 00:00:00   1st Qu.:2019-04-12 00:00:00   Class :character
##  Median :2019-12-30 00:00:00   Median :2020-09-14 00:00:00   Mode  :character
##  Mean   :2019-10-28 00:57:05   Mean   :2020-07-08 14:19:12
##  3rd Qu.:2021-09-03 00:00:00   3rd Qu.:2022-01-07 00:00:00
##  Max.   :2022-12-31 00:00:00   Max.   :2022-12-31 00:00:00
##                                NA's   :57468
##  Cuota_Anualizada Num_Devoluciones_Total      Edad          Country
##  Min.   :   5.0   Min.   : 0.0000        Min.   : 18.00   Length:84121
##  1st Qu.: 120.0   1st Qu.: 0.0000        1st Qu.: 52.00   Class :character
##  Median : 144.0   Median : 0.0000        Median : 63.00   Mode  :character
##  Mean   : 156.7   Mean   : 0.9845        Mean   : 61.76
##  3rd Qu.: 180.0   3rd Qu.: 1.0000        3rd Qu.: 73.00
##  Max.   :2400.0   Max.   :76.0000        Max.   :106.00
##                                          NA's   :6783
##   Provincia             CCAA                CP            Aceptar_recibir_info
##  Length:84121       Length:84121       Length:84121       Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:1.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.0000
##                                                           Mean   :0.9714
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
##
##     Idioma            Profesion
##  Length:84121       Length:84121
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Realitzem la primera modificació, i canviem el nom de la columna Sexe per Gènere que és més correcte

```
# Sex->gender
colnames(sociosData)[4]<-"Genero"
```

Veiem que tenim **17** variables i **84.121** registres

Revisem la descripció de les variables contingudes al fitxer i els tipus de variables es correspon al que hem carregat. Les organitzem lògicament per donar-los sentit i construïm un petit diccionari de dades.

- **ID** (Numèric) identificador del col · laborador

FETS A ESTUDIAR

- **Estado** (String) Estat de la col · laboració (Activo o baja)
- **Origen** (String) Canal de captació del col · laborador
- **Género** (String) Gènere del col · laborador (H = Hombre / M = Mujer)
- **Motivo_Baja** (String) Motiu de baixa de la col · laboració
- **Numero Devoluciones Total** (Numèric) Número de devolucions totals durant la col · laboració

DIMENSIÓ GEOGRÀFICA

- **Provincia** (String) província del col · laborador (N/A per alguns col · laboradors)
- **CP** (String) Codi postal del col · laborador (N/A per alguns col · laboradors)
- **CCAA** (String) Comunitat autònoma del col · laborador (N/A per alguns col · laboradors)
- **Country** (String) País del col · laborador

DIMENSIÓ TEMPORAL

- **Fecha Alta** (date) Data d'inici de la col · laboració
- **Fecha Baja** (date) Data de finalització de la col · laboració

DIMENSIÓ COL · LABORACIÓ

- **Cuota Anualizada** (Numeric) Quota anual de col · laboració

DIMENSIÓ INFORMACIÓ

- **Aceptar recibir info** (Boolean) El col · laborador accepta que li enviem comunicacions (0 = No, 1= Si)

ALTRES

- **Idioma** (String) idioma de preferència de comunicació (Cat / Esp)
- **Profesion** (string) Professió del col · laborador (N/A per alguns col · laboradors)
- **Edad** (Numèric) Edat del col · laborador

El següent pas serà la neteja de dades, mirant si hi ha valors nulls.

```
# Revisió dels N/A
colSums(is.na(sociosData))
```

```
##                     ID                 Estado                 Origen
##                      0                      0                      0
##                 Genero             Fecha_Alta             Fecha_Baja
##                      0                      0                  57468
##            Motivo_Baja       Cuota_Anualizada Num_Devoluciones_Total
##                  57468                      0                      0
##                   Edad                Country              Provincia
##                   6783                      0                   6028
##                   CCAA                     CP   Aceptar_recibir_info
##                   6028                   6028                      0
##                 Idioma              Profesion
##                      0                  12923
```

Observem que hi han valors nuls a les dades. Primer de tot el que observem és que Fecha Baja i Motivo Baja tenen els mateixos valors nuls (n=57.468), això ens indica que totes les persones que estan de baixa tenen data de finalització i motiu de baixa (no hi ha discrepància de valors). Finalment observem com professió, província, CP, CCAA i edat si que tenen valors nulls.

El primer que farem serà assignar el valor "Desconegut" per als valors buits de la variable Província, CCAA i Professió. També assignarem el valor 60000 als CP buits.

```
# NA->Desconegut
sociosData$Provincia[is.na(sociosData$Provincia)] <- "Desconegut"
sociosData$CCAA[is.na(sociosData$CCAA)] <- "Desconegut"
sociosData$CP[is.na(sociosData$CP)] <- "60000"
sociosData$Profesion[is.na(sociosData$Profesion)] <- "Desconegut"
```

També assignarem, la mitjana de les edats als valors buits de la variable "Edat".

```
# NA edat -> mean
sociosData$Edad[is.na(sociosData$Edad)] <- mean(sociosData$Edad,na.rm=T)
```

```
summary(sociosData[,"Edad"])
```

```
##       Edad
##  Min.   : 18.00
##  1st Qu.: 53.00
##  Median : 61.76
##  Mean   : 61.76
##  3rd Qu.: 72.00
##  Max.   :106.00
```

De la informació mostrada destaquem que el col · laborador més jove té 18 anys i el més gran 106 anys. La mitjana d'edat la teníem gairebé en 62 anys. Com a primera conclusió extraurem que és una base de dades amb mitjanes de col · laboració força adulta.

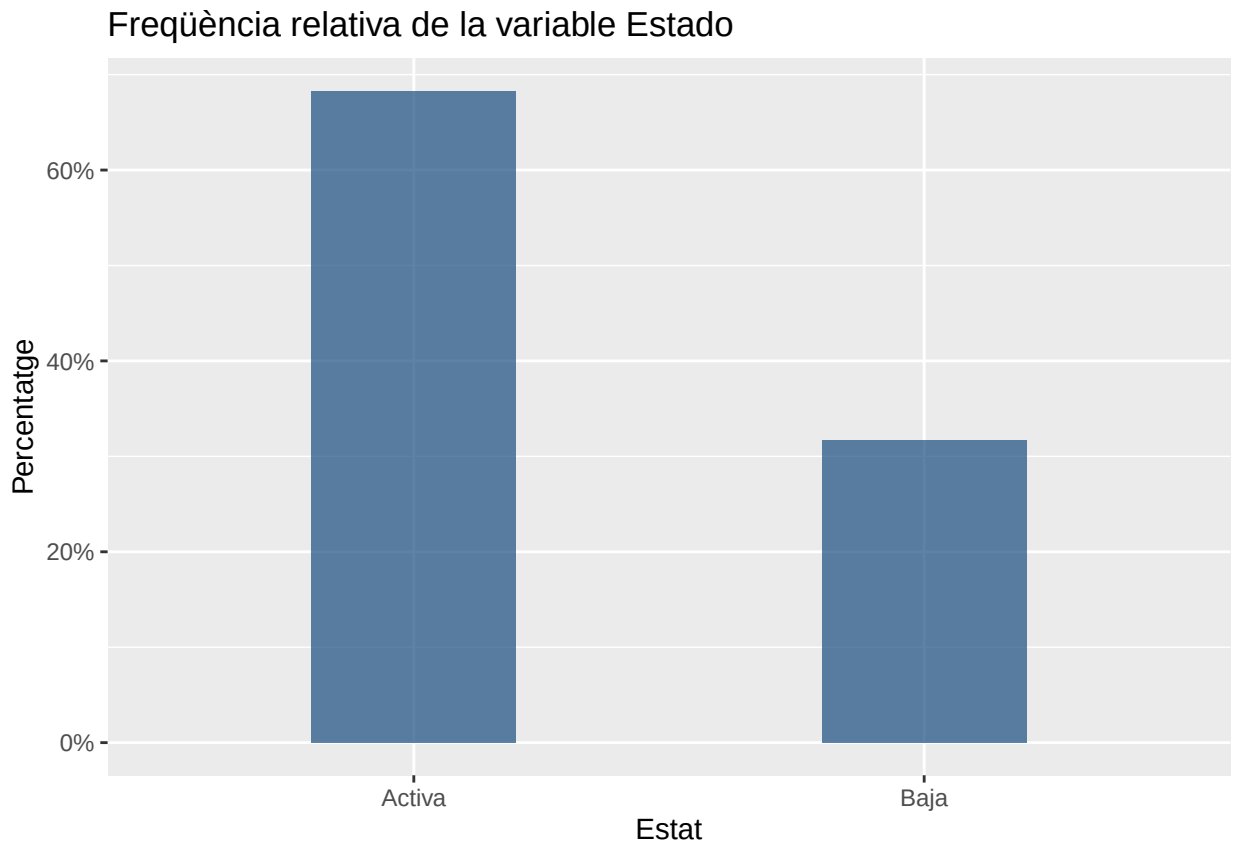
Si observem els NA (valors nuls) veiem que les dades estan prou bé. Decidim substituir el valor NA de Província, CP, CCAA i Professió per Desconegut per una major llegibilitat. També proposem substituir els NA de l'edat per la mitjana.

Ara crearem una seria d'histogrames i descriure els valors per veure les dades en general i així fer una primera aproximació a les dades

```r
# Taula de freqüència per conèixer el % de socis actius i de baixa
situacion = as.table(table(x=sociosData$Estado))
prop.table(situacion)
```

```
## x
##    Activa      Baja
## 0.6831588 0.3168412
```

```r
# Gràfic il·lustratiu dels socis de baixa i actius
ggplot(data=sociosData, aes(x = Estado)) +
  geom_bar(width = 0.4, fill=rgb(0.1,0.3,0.5,0.7), aes(y = (..count..)/sum(..count..))) +
  scale_x_discrete("Estat") +
  scale_y_continuous("Percentatge",labels=scales::percent) +
  labs(title = "Freqüència relativa de la variable Estado")
```



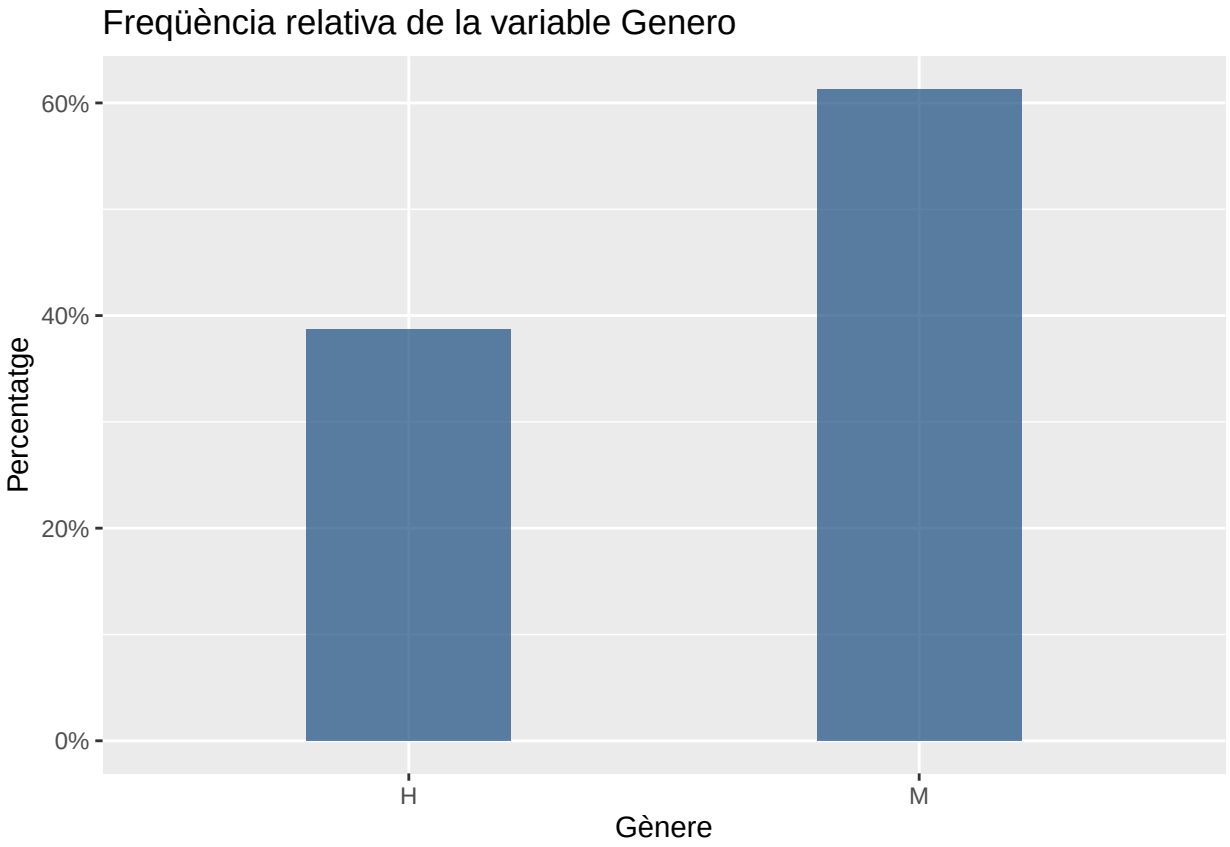
El primer que observem és que més del 68% dels col · laboradors estan actualment actius i per tant col · laboren econòmicament mentre que gairebé el 32% del col · laboradors ja no col · laboren econòmicament.

```r
# Taula de freqüència per conèixer els % de socis que son Homes o dones
situacion= as.table(table(x=sociosData$Genero))
prop.table(situacion)
```

```
## x
```

```
##         H         M
## 0.3867168 0.6132832
```

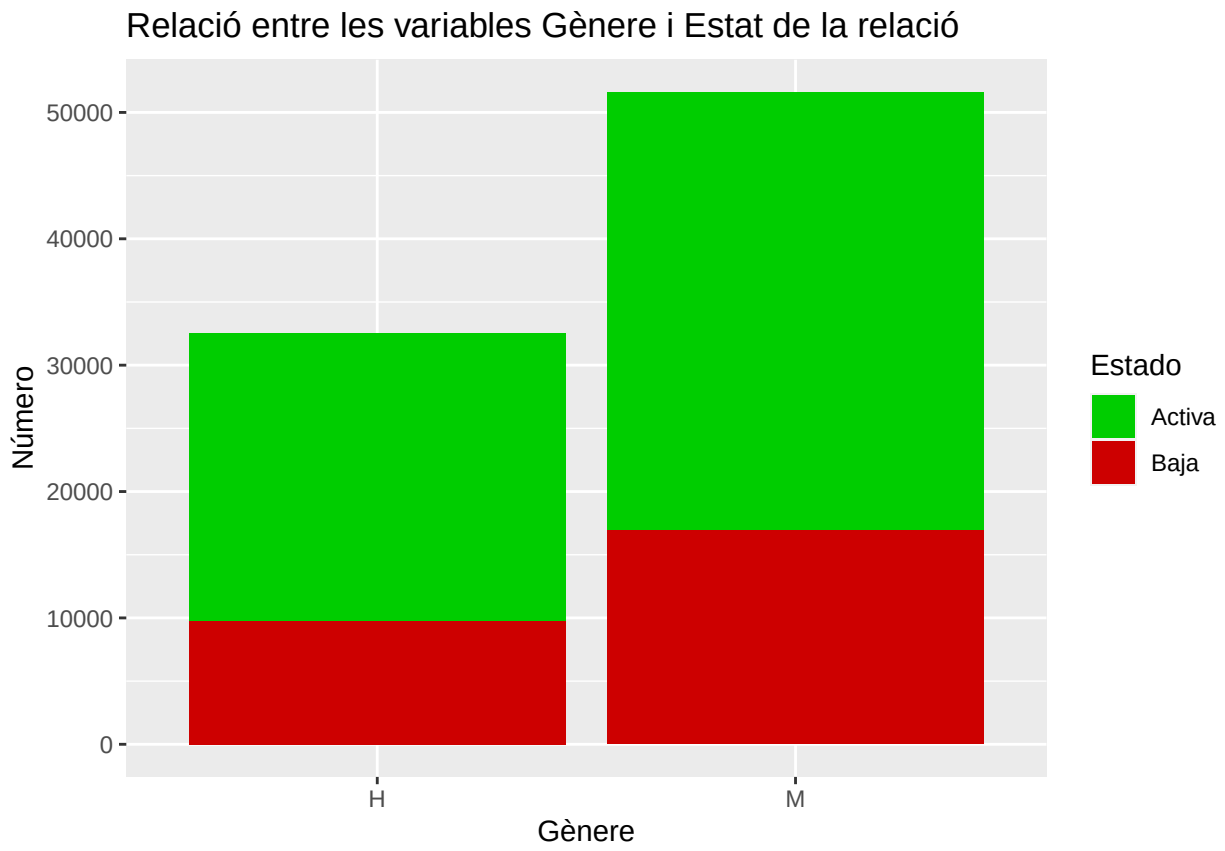
```
# Gràfic il·lustratiu dels col·laboradors (actius o no per Gènere)
ggplot(data=sociosData, aes(x = Genero)) +
  geom_bar(width = 0.4, fill=rgb(0.1,0.3,0.5,0.7), aes(y = (..count..)/sum(..count..))) +
  scale_x_discrete("Gènere") +
  scale_y_continuous("Percentatge",labels=scales::percent) +
  labs(title = "Freqüència relativa de la variable Genero")
```



En aquest segons anàlisis visual, observem com les dones tendeixen a col · laborar més amb la causa que els homes, això pot ser degut a que les dones estan més sensibilitzades amb la causa. Observem com 6 de cada 10 col · laboradors son dones.

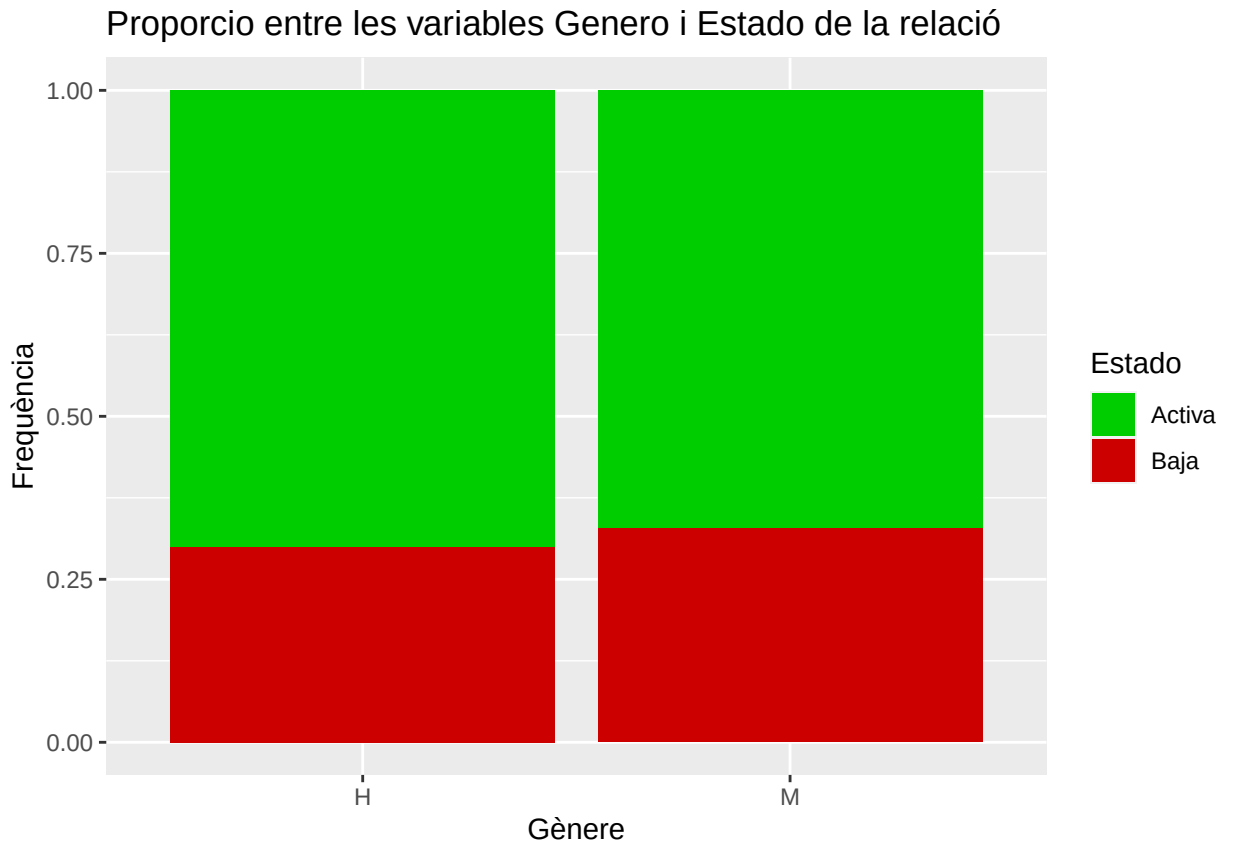
Ara analitzarem si hi ha més baixes que siguin dones o homes

```
# Visualització gràfica entre Gènere i estat de la col·laboració
ggplot(data=sociosData,aes(x=Genero,fill=Estado))+
  geom_bar()+
  labs(title="Relació entre les variables Gènere i Estat de la relació",
       x= "Gènere",
       y="Número") +
  scale_fill_manual(values=c("green3","red3"))
```



Sembla que hi han més dones que homes que es donen de baixa, però aquests últim gràfic no ens serveix del tot, ja que en el dataset com hem dit abans hi ha més dones que homes per tant procedim a analitzar les dades però proporcionades:

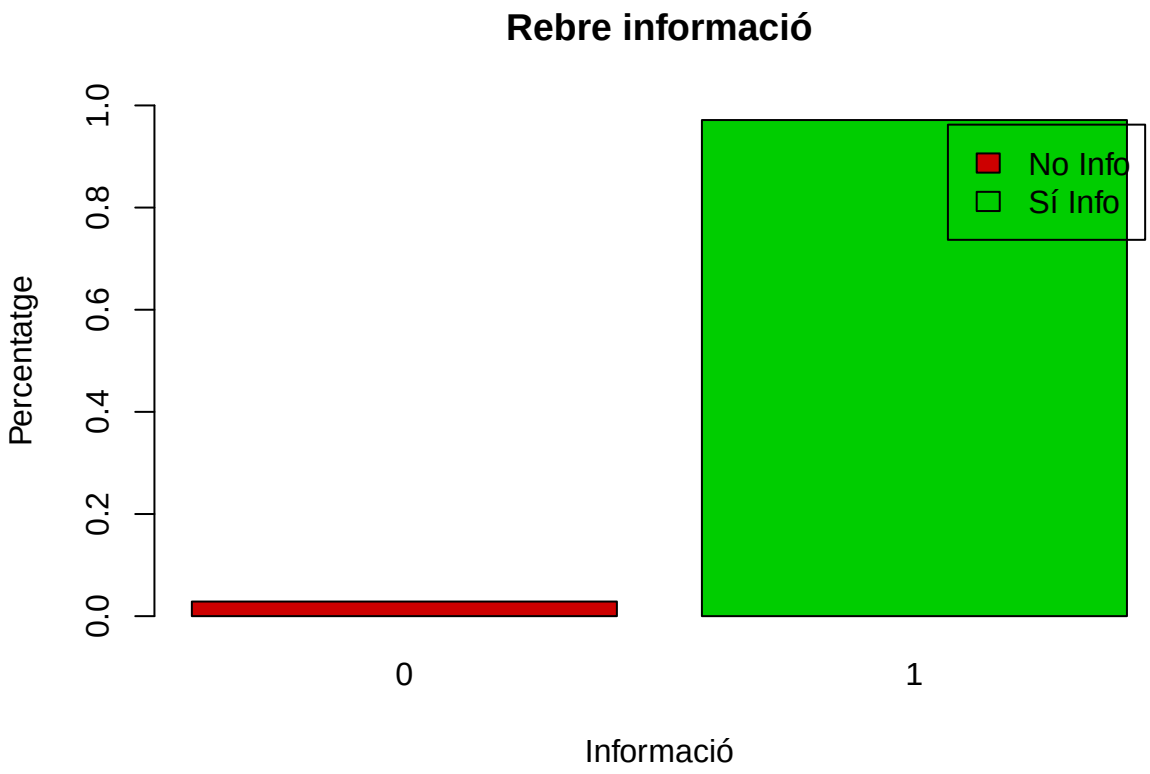
```
# Visualització gràfica entre Gènere i estat de la col·laboració
ggplot(sociosData, aes(Genero, fill=Estado)) +
  geom_bar(position = "fill") +
  labs(title="Proporcio entre les variables Genero i Estado de la relació",
       x= "Gènere",
       y="Frequència") +
  scale_fill_manual(values=c("green3","red3"))
```



Si analitzem el percentatge de baixes respecte les altes, observem com és pràcticament similar. Per tant la variable "Gènere" no és determinant a l'hora de donar-se de baixa.

Una altre variable important pot ser saber si tenim o no permisos per enviar informació i comunicar-nos. Crec que és important aquesta variable ja que la falta d'informació por fer que una persona es desvinculi.

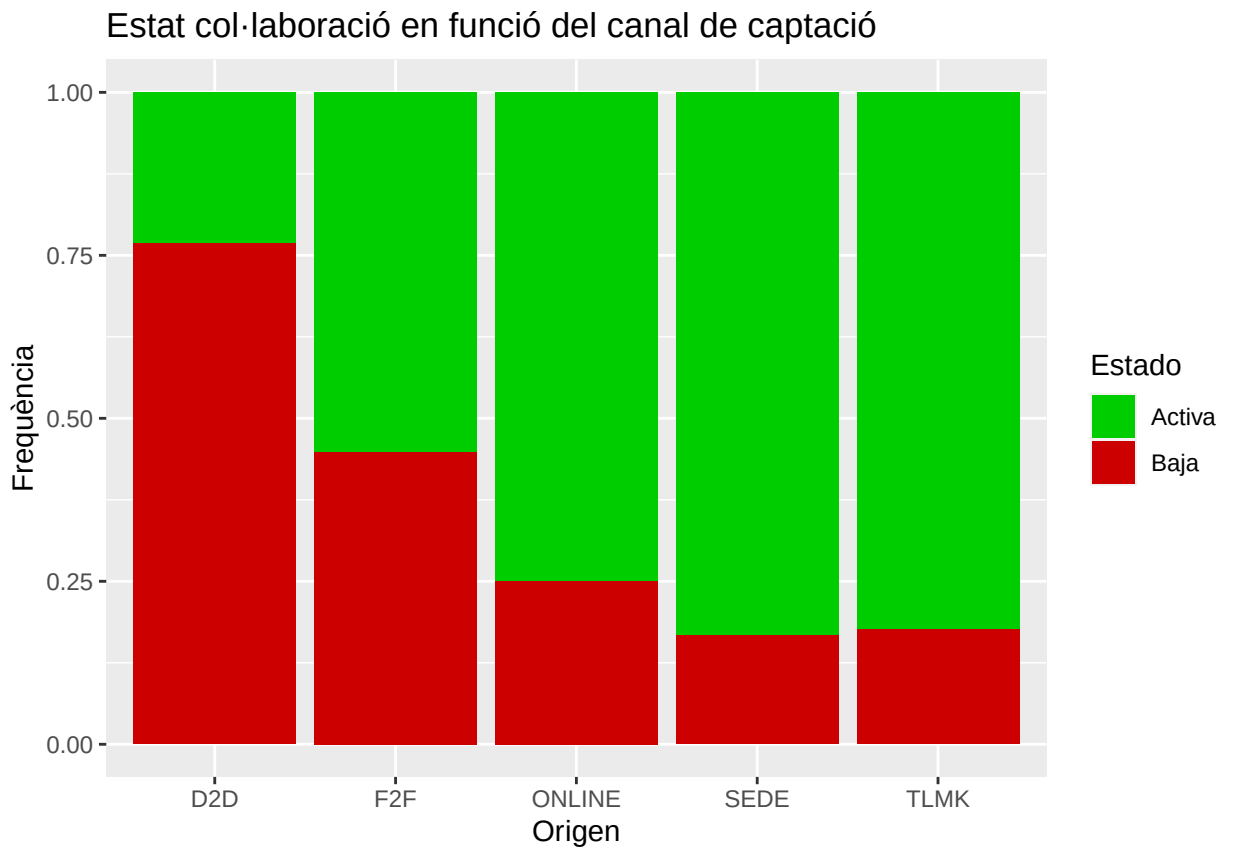
```
# Visualització gràfica dels permisos de comunicació
sociosData$Aceptar_recibir_info <- ifelse(sociosData$Aceptar_recibir_info %in% c(0), 0, 1)
counts <- table(sociosData$Aceptar_recibir_info)
barplot(prop.table(counts),col=c("red3","green3"), main="Rebre informació", legend.text=c("No Info","Sí
```



S'observa gràficament com gairebé el 100% dels socis, ens permeten que els hi enviem informació i per tant no és una variable determinant a l'hora de donar-se baixa.

Un altre variable que si que pot ser determinant en les baixes és l'origen de captació, ja que hi han canals de captació que poden donar un alt volum d'altes però poden ser més agressius i per tant tenir més baixes. Anem a fer un anàlisis exploratori dels canals de captació:

```
# Visualització gràfica dels canals d'entrada del socis i estat de la relació
ggplot(data = sociosData,aes(x=Origen,fill=Estado))+
  geom_bar(position="fill")+
  ylab("Frequència")+
  labs(title="Estat col·laboració en funció del canal de captació",
       x= "Origen",
       y="Frequència") +scale_fill_manual(values=c("green3","red3"))
```



Aquesta gràfica és molt interessant ja que si que ens aporta informació rellevant. Observem com el canal D2D (Door-to-door = porta freda), el 75% de les altes ja estan donades de baixa. El segon canal de captació és el F2F (Face-to-face = captació al carrer), on observem que gairebé 1 de cada 2 socis que entren per aquest canal, ja està de baixa. En canvi, els socis captats per digital o tlmk tenen un percentatge de baixes menor que els altes dos canals de captació. Finalment, com és lògic aquelles persones que han vingut físicament a la nostra oficina (SEDE) son els més fidels.

Ara crearem una base de dades auxiliar amb els registres donats de baixa per facilitar-nos la anàlisis d'aquests segment en concret:

```
# Creem un nou dataset amb tots els registres de baixa
bajasData <- sociosData[sociosData$Estado=="Baja",]
```

Fem un ràpid anàlisis d'aquesta base de dades:

```
# Resum
str(bajasData)
```

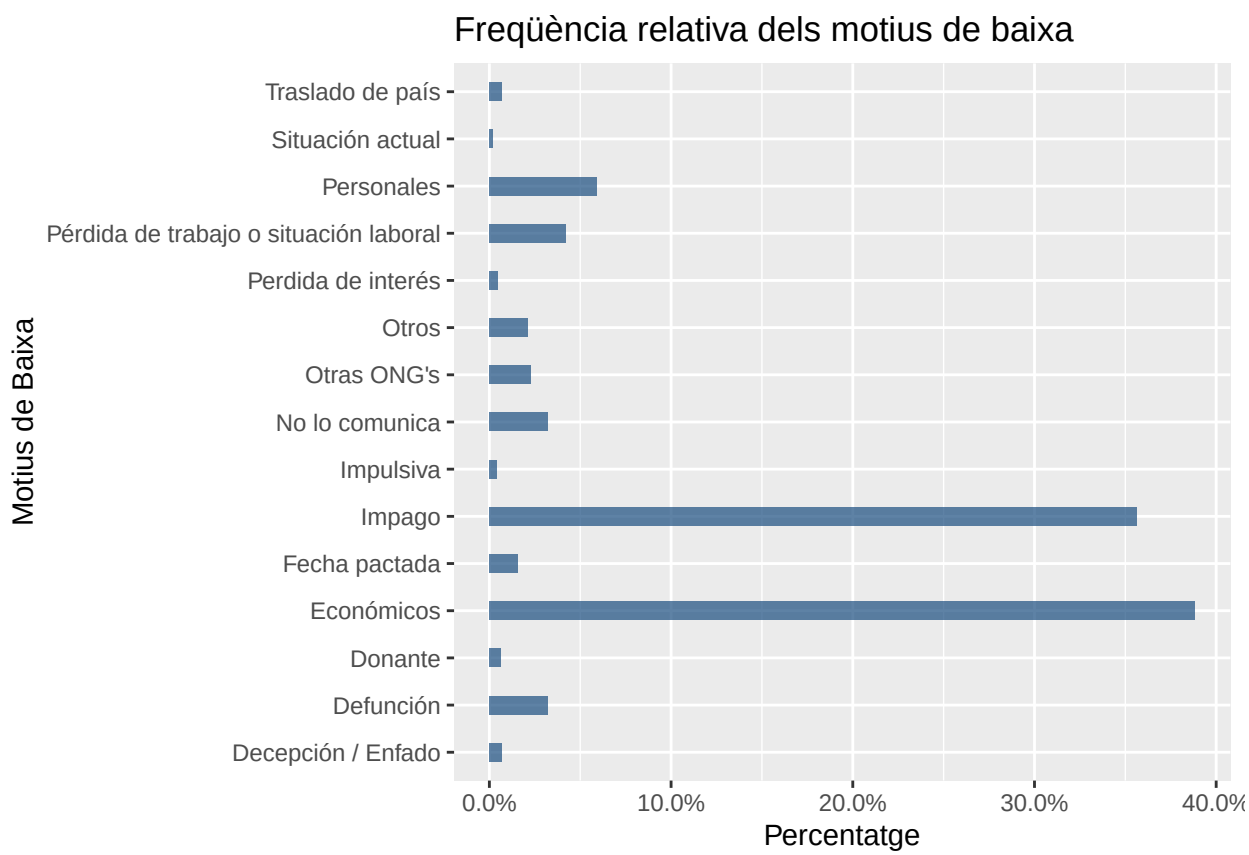
```
## tibble [26,653 x 17] (S3: tbl_df/tbl/data.frame)
##  $ ID                    : num [1:26653] 3 73 519 588 589 596 601 602 603 604 ...
##  $ Estado                : chr [1:26653] "Baja" "Baja" "Baja" "Baja" ...
##  $ Origen                : chr [1:26653] "TLMK" "TLMK" "TLMK" "TLMK" ...
##  $ Genero                : chr [1:26653] "H" "H" "M" "H" ...
##  $ Fecha_Alta            : POSIXct[1:26653], format: "2017-07-12" "2016-12-10" ...
##  $ Fecha_Baja            : POSIXct[1:26653], format: "2019-04-24" "2021-09-07" ...
##  $ Motivo_Baja           : chr [1:26653] "Defunción" "Defunción" "Otras ONG's" "Decepción / Enfado"
```

```
##  $ Cuota_Anualizada      : num [1:26653] 1008 600 60 40 120 ...
##  $ Num_Devoluciones_Total: num [1:26653] 1 3 0 10 11 1 0 0 1 0 ...
##  $ Edad                  : num [1:26653] 83 85 61.8 63 45 ...
##  $ Country               : chr [1:26653] "Spain" "Spain" "Spain" "Spain" ...
##  $ Provincia             : chr [1:26653] "Barcelona" "Barcelona" "Vizcaya" "Tarragona" ...
##  $ CCAA                  : chr [1:26653] "Catalunya" "Catalunya" "País Vasco" "Catalunya" ...
##  $ CP                    : chr [1:26653] "08035" "08034" "48340" "43580" ...
##  $ Aceptar_recibir_info  : num [1:26653] 0 0 1 1 1 1 1 1 1 1 ...
##  $ Idioma                : chr [1:26653] "CAT" "CAT" "ESP" "CAT" ...
##  $ Profesion             : chr [1:26653] "Desconegut" "PENS" "Desconegut" "NS/NC" ...
```

Veiem que tenim les mateixes **17** variables però el número de registres s'ha reduït fins els **26.653** registres

Ara analitzarem les categories dels motius de baixa per veure si ens pot donar alguna pista sobre quina variable més podem analitzar:

```
# Gràfic il·lustratiu dels motius de baixa
ggplot(data=bajasData, aes(x = Motivo_Baja)) +
  geom_bar(width = 0.4, fill=rgb(0.1,0.3,0.5,0.7), aes(y = (..count..)/sum(..count..))) +
  scale_x_discrete("Motius de Baixa") +
  scale_y_continuous("Percentatge",labels=scales::percent) +
  labs(title = "Freqüència relativa dels motius de baixa")+
  coord_flip()
```



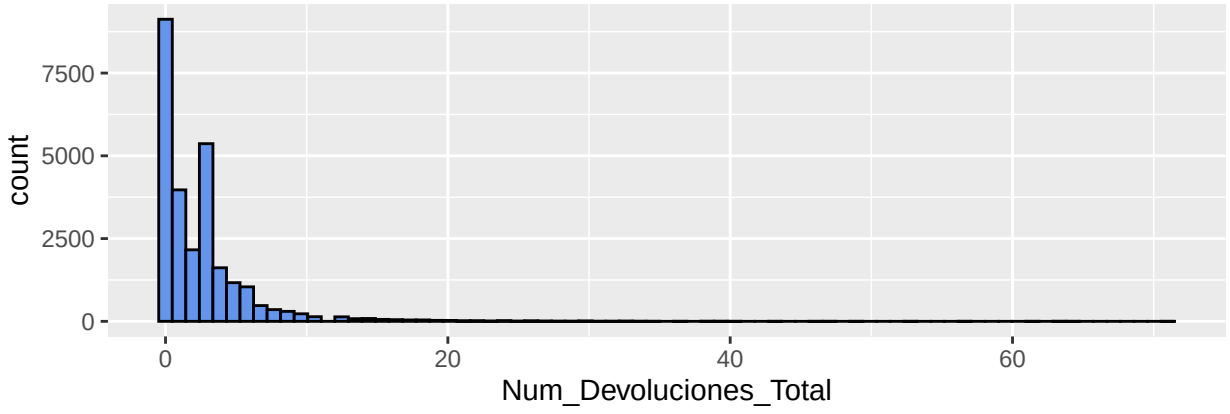
Observem com els motius de baixa més destacats son **Impago** i **econòmics** per tant ara procedirem a analitzar quotes i devolucions per veure si aquestes variables son determinants a l'hora de detectar una baixa.

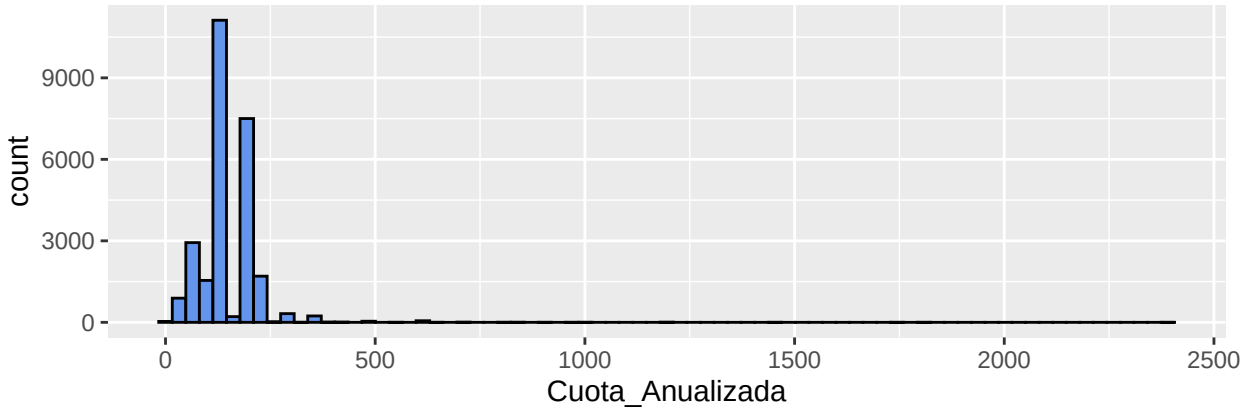
Primer de tot, farem un anàlisis de les devolucions, és a dir, al número de quotes que el soci ens ha retornat per veure si les baixes tendeixen a retornar més quotes i per tant ser una variable important i determinant:

```
summary(bajasData[c("Num_Devoluciones_Total","Cuota_Anualizada")])
```

```
##  Num_Devoluciones_Total Cuota_Anualizada
##  Min.   : 0.000         Min.   :  10
##  1st Qu.: 0.000         1st Qu.: 120
##  Median : 2.000         Median : 144
##  Mean   : 2.557         Mean   : 147
##  3rd Qu.: 3.000         3rd Qu.: 180
##  Max.   :71.000         Max.   :2400
```

```
#Crearem una llista per mostrar els atributs que interessen.
histList<- list()
n = c("Num_Devoluciones_Total", "Cuota_Anualizada")
bajasDataAux= bajasData %>% select(all_of(n))
for(i in 1:ncol(bajasDataAux)){
  col <- names(bajasDataAux)[i]
  ggp <- ggplot(bajasDataAux, aes_string(x = col)) +
    geom_histogram(bins = 75, fill = "cornflowerblue", color = "black")
      histList[[i]] <- ggp  # afegim cada plot a la llista buida
}
 multiplot(plotlist = histList, cols = 1)
```





Destaquem que hi han col · laboradors que no han realitzat mai una devolució, però la mitjana es que realitzin 2 devolucions. El valor màxim son 71 devolucions.

Pel que fa a la quota anual, observem com el valor mínim son 10 euros mentre que la mitja està als 144 euros anuals. Observem com el valor màxim son 2.400 euros anuals.

Una altre variable interessant a analitzar pot ser el temps de permanència dels col · laboradors. Son baixes ràpides? Depèn del canal de captació?

Per la realització d'aquest punt, creem una nova variable, calculant la diferencia en dies de la data de baixa i la data d'alta

```
# Nova variable en el dataset de les baixes
bajasData$Tiempo <- as.numeric(difftime(bajasData$Fecha_Baja, bajasData$Fecha_Alta, units = "days"))
str(bajasData$Tiempo)
```

```
##  num [1:26653] 651 1732 443 200 797 ...
```

```
summary(bajasData$Tiempo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -13.0   126.0   329.0   523.4   755.0  3347.0
```

Realitzant aquest anàlisis de temporalitat, observem que el mínim és un valor negatiu i per tant que hi han dades incorrectes ja que la data de baixa no pot ser anterior a la baixa d'alta.

Per solucionar aquest error, procedirem a modificar aquests valors per 0, entent que la data de baixa va ser el mateix dia que la data d'alta

```
# visualitxem les dades amb els error:
bajasData %>%
  filter(bajasData$Tiempo<0)
```

```
## # A tibble: 1 x 18
##       ID Estado Origen Genero Fecha_Alta          Fecha_Baja
##    <dbl> <chr>  <chr>  <chr>  <dttm>              <dttm>
## 1 965062 Baja   F2F    M      2022-05-31 00:00:00 2022-05-18 00:00:00
## # ... with 12 more variables: Motivo_Baja <chr>, Cuota_Anualizada <dbl>,
## #   Num_Devoluciones_Total <dbl>, Edad <dbl>, Country <chr>, Provincia <chr>,
## #   CCAA <chr>, CP <chr>, Aceptar_recibir_info <dbl>, Idioma <chr>,
## #   Profesion <chr>, Tiempo <dbl>
```

Observem com és unicament un registre on efectivament la "Fecha_Baja" es abans a la "Fecha_Alta". En aquests casos modificarem la variable Tiempo i li assignarem el valor 0

```
# Si Temps < 0 <- 0
bajasData$Tiempo[ bajasData$Tiempo < 0 ] <- 0
# comprovació que ho he solucionat
bajasData %>%
  filter(bajasData$Tiempo<0)
```

```
## # A tibble: 0 x 18
## # ... with 18 variables: ID <dbl>, Estado <chr>, Origen <chr>, Genero <chr>,
```

```
## #   Fecha_Alta <dttm>, Fecha_Baja <dttm>, Motivo_Baja <chr>,
## #   Cuota_Anualizada <dbl>, Num_Devoluciones_Total <dbl>, Edad <dbl>,
## #   Country <chr>, Provincia <chr>, CCAA <chr>, CP <chr>,
## #   Aceptar_recibir_info <dbl>, Idioma <chr>, Profesion <chr>, Tiempo <dbl>
```

```
# Anàlisis estadístic
summary(bajasData$Tiempo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   126.0   329.0   523.4   755.0  3347.0
```

```
str(bajasData$Tiempo)
```

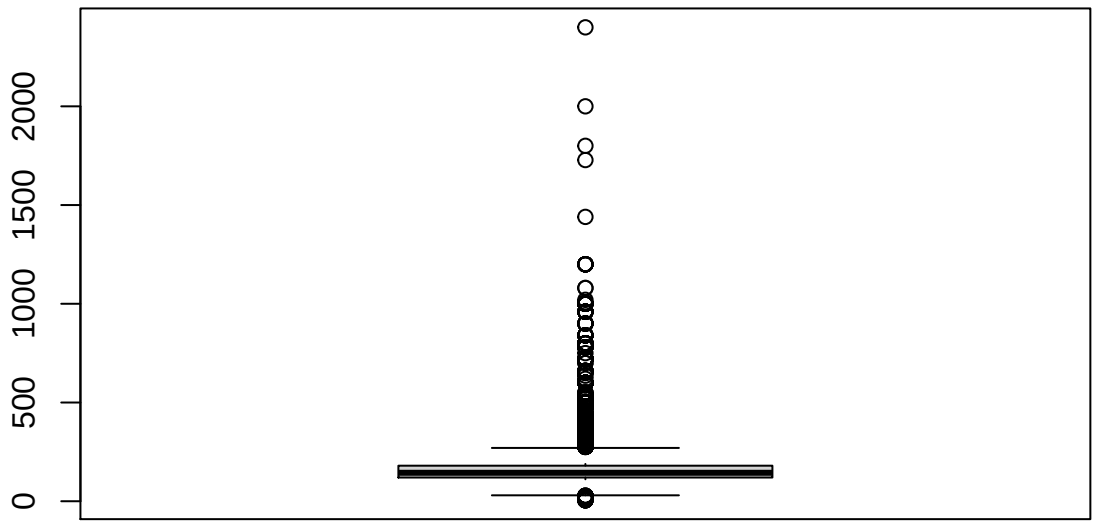
```
##  num [1:26653] 651 1732 443 200 797 ...
```

Ara si tornem a analitzar la variable "Tiempo", veiem com no hi han valors negatius, el mínim és 0 és a dir son altes i baixes en el mateix moment, la mitjana son 329 dies (és a dir abans de l'any) i el màxim és de 9 anys (3.347 dies).

Anàlisis d'alguna variable més

A continuació realitzarem algun anàlisis addicional.

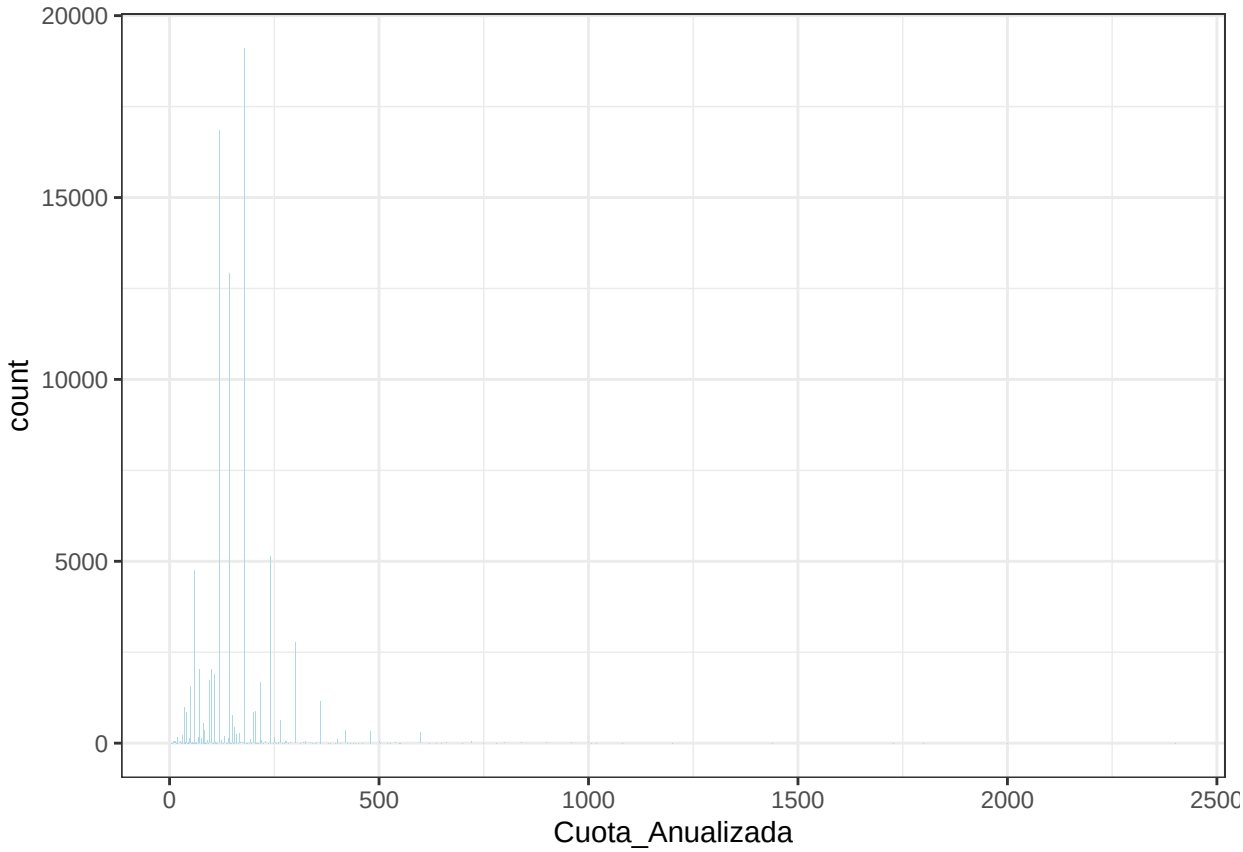
```
# boxplot quota anual
boxplot (sociosData$Cuota_Anualizada)
```



En el boxplot de la quota anual, no es pot apreciar res a causa de les observacions amb valors elevats

Ara analitzarem si segueix un patró de normalitat:

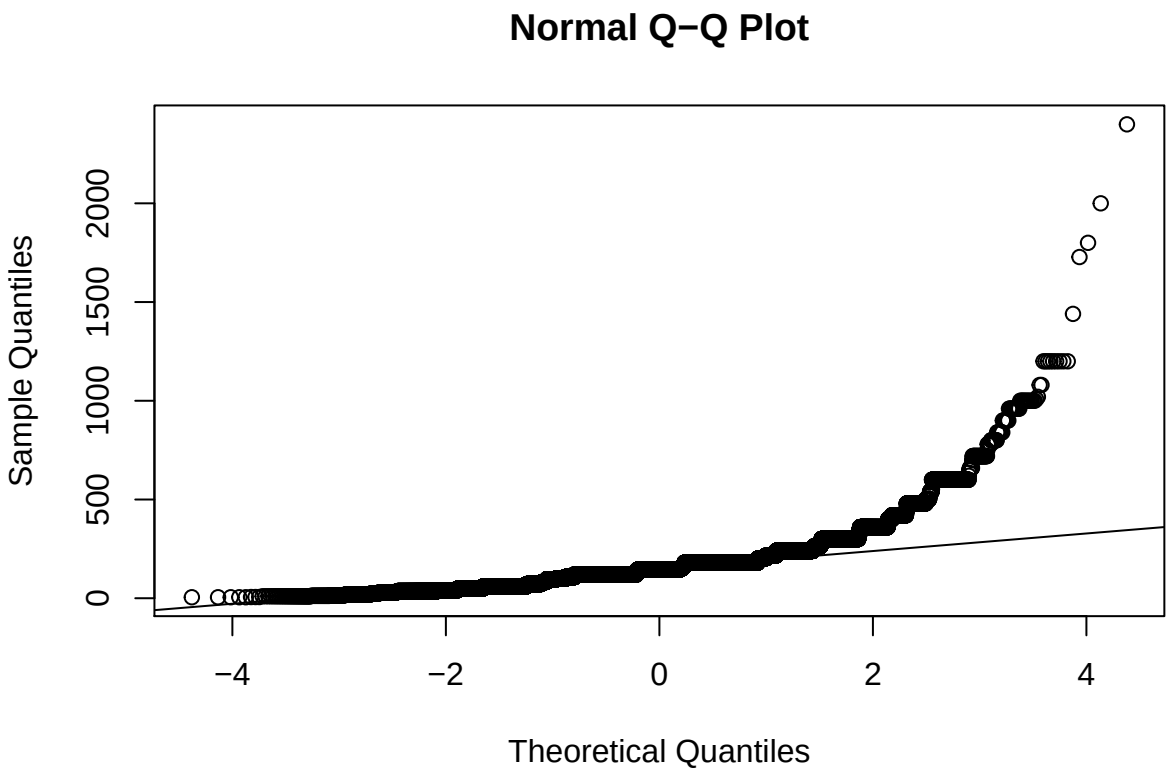
```
# Carreguem la llibreria necessària
library (ggplot2)

# Visualització gràfica de la normalitat
ggplot(sociosData,aes(x=Cuota_Anualizada)) +
geom_bar(fill="lightblue") +
theme_bw()
```




```
#contrast normalitat
library(nortest)
lillie.test(sociosData$Cuota_Anualizada)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  sociosData$Cuota_Anualizada
## D = 0.20816, p-value < 2.2e-16
```

```
# gràfic quantile
qqnorm(sociosData$Cuota_Anualizada)
qqline(sociosData$Cuota_Anualizada)
```



Tal i com era d'esperar la quota anual no es distribueix segons una distribució normal.

Conclusió

Per finalitzar farem unes breus conclusions de les dades estudiades.

Les dades que s'han treballant durant aquesta pràctica provenen d'una font de dades sobre els col · laboradors d'una entitat del tercer sector (Fundació / ONG). Tots els registres tenen un identificador únic i poden estar actius o de baixa. A més tenen un sèrie de de variables que enriqueixen les dades (origen de captació, quotes, devolucions...)

Les dades han estat revisades i sembla que estan ben informades, a més d'estar força netes i ben documentades. Els camps amb valors buits no m'han generat gaires problemes.

Destaquem la bona salut de la base de dades dels col · laboradors on el 68% estan actius mentre que només un 32% està de baixa. També hem observat com 6 de cada 10 nous sòcies son dones i per tant el perfil del captació hauria de ser una dona d'uns 62 anys.

També hem observat com hi han canals de captació que tenen un percentatge de baixa més elevat i això s'ha de tenir en compte per l'estratègia de creixement de l'entitat.

Amb aquestes conclusions, fem la visualització!