Pràctica 2: Neteja i anàlisi de les dades

Adrián Alonso Gonzalo i Alexandre Vidal De Palol

Maig/Juny 2022

Contents

1 Càrrega de llibreries.

En aquesta secció, carregarem les llibreries que s'utilitzaran durant la realització d'aquesta pràctica.

```
library(mice)
library(ggplot2)
library(magrittr)
library(dplyr)
```

2 Descripció i càrrega del dataset.

Els conjunts de dades train.csv i 'test.csv' que es troben a la carpeta 'data/input' d'aquest paquet s'han obtingut del web https://www.kaggle.com/c/titanic.

Aquests conjunts de dades contenen informació sobre la tripulació del Titanic amb 12 (11 el de test) columnes i un total de 891 registres (418 el de test).

Les variables d'aquesta mostra son:

- PassengerId: Número de passatger.
- Survived: Supervivència (0=No, 1=Si).
- Pclass: Classe de tiquet (1=Primera, 2=Segona, 3=Tercera) .
- Name: Nom.
- Sex: Sexe.
- Age: Edat.
- SibSp: Germans / Cónjugues a bord del Titanic.
- Parch: Pares / nens a bord del Titanic.
- Ticket: Número de ticket.
- Fare: Preu del ticket.
- Cabin: Numero de cabina.
- Embarked: Port de embarcament.

A continuació, passem a carregar el fitxer i a mostrar una sèrie de metadades del conjunt que ens donaran una primer idea del joc de dades amb el que estem tractant.

```
# Carreguem els fitxers 'test.csv' i 'train.csv' de la carpeta 'data/input' (indicant que volem els 'st
test_dataset <- read.csv("../data/input/test.csv", stringsAsFactors=TRUE)
train_dataset <- read.csv("../data/input/train.csv", stringsAsFactors=TRUE)

# Mostrem les primeres files dels jocs de dades
head(test_dataset)
```

```
##   PassengerId Pclass                                          Name    Sex  Age
## 1         892      3                              Kelly, Mr. James   male 34.5
## 2         893      3              Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3         894      2                     Myles, Mr. Thomas Francis   male 62.0
## 4         895      3                              Wirz, Mr. Albert   male 27.0
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6         897      3                    Svensson, Mr. Johan Cervin   male 14.0
##   SibSp Parch  Ticket    Fare Cabin Embarked
## 1     0     0  330911  7.8292              Q
## 2     1     0  363272  7.0000              S
## 3     0     0  240276  9.6875              Q
## 4     0     0  315154  8.6625              S
## 5     1     1 3101298 12.2875              S
## 6     0     0    7538  9.2250              S
```

```
head(train_dataset)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
```

```
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
# Creem una nova columna amb el nom del dataset del qual provenen les files
test_dataset['row_type'] <- as.factor('test')
train_dataset['row_type'] <- as.factor('train')

# Creem la columna 'Survived' ja que el joc de dades de test no la conté (introduïrem un valor dummy qu
test_dataset['Survived'] <- 99

# Juntem els dos jocs de dades en un únic dataset per poder netejar els dos a la vegada
dataset <- rbind(train_dataset, test_dataset[,colnames(train_dataset)])

# Mostrem les primeres files del dataset unificat
head(dataset)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked row_type
## 1        A/5 21171  7.2500              S    train
## 2         PC 17599 71.2833   C85        C    train
## 3 STON/O2. 3101282  7.9250              S    train
## 4           113803 53.1000  C123        S    train
## 5           373450  8.0500              S    train
## 6           330877  8.4583              Q    train
```

```r
# Mostrem l'estructura del dataset
str(dataset)
```

```
## 'data.frame':    1309 obs. of  13 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : num  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 929 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
##  $ row_type   : Factor w/ 2 levels "train","test": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Mostrem un resum de les estadistiques de les columnes del dataset
summary(dataset)
```

```
##   PassengerId      Survived         Pclass     
##  Min.   :   1   Min.   : 0.00   Min.   :1.000  
##  1st Qu.: 328   1st Qu.: 0.00   1st Qu.:2.000  
##  Median : 655   Median : 1.00   Median :3.000  
##  Mean   : 655   Mean   :31.87   Mean   :2.295  
##  3rd Qu.: 982   3rd Qu.:99.00   3rd Qu.:3.000  
##  Max.   :1309   Max.   :99.00   Max.   :3.000  
##                                                
##                                 Name          Sex           Age       
##  Connolly, Miss. Kate                :   2   female:466   Min.   : 0.17  
##  Kelly, Mr. James                    :   2   male  :843   1st Qu.:21.00  
##  Abbing, Mr. Anthony                 :   1                Median :28.00  
##  Abbott, Mr. Rossmore Edward         :   1                Mean   :29.88  
##  Abbott, Mrs. Stanton (Rosa Hunt):   1                3rd Qu.:39.00  
##  Abelson, Mr. Samuel                 :   1                Max.   :80.00  
##  (Other)                             :1301                NA's   :263    
##      SibSp            Parch           Ticket          Fare        
##  Min.   :0.0000   Min.   :0.000   CA. 2343:  11   Min.   :  0.000  
##  1st Qu.:0.0000   1st Qu.:0.000   1601    :   8   1st Qu.:  7.896  
##  Median :0.0000   Median :0.000   CA 2144 :   8   Median : 14.454  
##  Mean   :0.4989   Mean   :0.385   3101295 :   7   Mean   : 33.295  
##  3rd Qu.:1.0000   3rd Qu.:0.000   347077  :   7   3rd Qu.: 31.275  
##  Max.   :8.0000   Max.   :9.000   347082  :   7   Max.   :512.329  
##                                   (Other) :1261   NA's   :1        
##             Cabin      Embarked  row_type   
##                :1014    :  2    train:891  
##  C23 C25 C27    :   6   C:270    test :418  
##  B57 B59 B63 B66:   5   Q:123               
##  G6             :   5   S:914               
##  B96 B98        :   4                       
##  C22 C26        :   4                       
##  (Other)        : 271
```

```
# Mostrem el número de files del joc de dades unificat i per separat
dim(dataset)[1]
```

```
## [1] 1309
```

```
dim(dataset[dataset$row_type=='test',])[1]
```

```
## [1] 418
```

```
dim(dataset[dataset$row_type=='train',])[1]
```

```
## [1] 891
```

Observem que que el dataset conté 3 tipus de variables del quals caràcter, numèric i enter.

3 Integració i selecció de les dades d'interès a analitzar.

El procés d'integració i selecció de les dades es realitzarà al llarg del procés de neteja i anàlisi de les diverses variables del conjunt de dades d'entrenament del dataset (on la columna 'row_type' és igual a 'train').

En aquest procés es pretén anar analitzant les diferents variables en el procés de neteja i anàlisi, i en funció de les característiques que es vagin observant de les diverses variables es prendrà la decisió d'utilitzar un conjunt seleccionat el qual pugui ser útil per a la predicció del model i la comprovació amb el conjunt de test o validació.

El resultat del projecte pot respondre a possibles causes de mort dels tripulants que no van sobreviure a la tragèdia del Titanic, permetent establir models d'inferència sobre les causes relatives a la mortalitat entre diversos tipus de passatgers. Per altra banda la implementació de un model d'interès sobre quines han sigut variables que han influït més o menys en la supervivència del naufragi.

```
# Mostrem un subconjunt de dades
head(dataset)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                                   Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked row_type
## 1        A/5 21171  7.2500              S    train
## 2         PC 17599 71.2833   C85        C    train
## 3 STON/O2. 3101282  7.9250              S    train
## 4           113803 53.1000  C123        S    train
## 5           373450  8.0500              S    train
## 6           330877  8.4583              Q    train
```

4 Neteja de dades.

4.1 Valors buits (missing values).

En aquesta secció, farem un petit anàlisi sobre l'existència de valors buits o valors no informats en el nostre joc de dades. A partir de l'identificació de columnes amb valors buits, aplicarem diverses tècniques per inputar valors en els registres que contenen columnes amb aquestes característiques.

4.1.1 Identificació de les columnes amb valors buits.

En aquesta secció identificarem les columnes que contenen aquest tipus de valors.

```
# Mostrem el número de registres buits per cada columna
apply(dataset=="",2, sum)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
##    row_type
##           0
```

```
apply(is.na(dataset),2, sum)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
##    row_type
##           0
```

En els resultats anteriors podem observar que les variables amb valors buits són **Age**, **Cabin** i **Embarked**.

4.1.2 Ajust de la variable 'Age'.

En aquesta secció tractem la columna 'Age' amb l'objectiu de inputar nous valors que en aquells registres on el seu valor és buit o no informat.

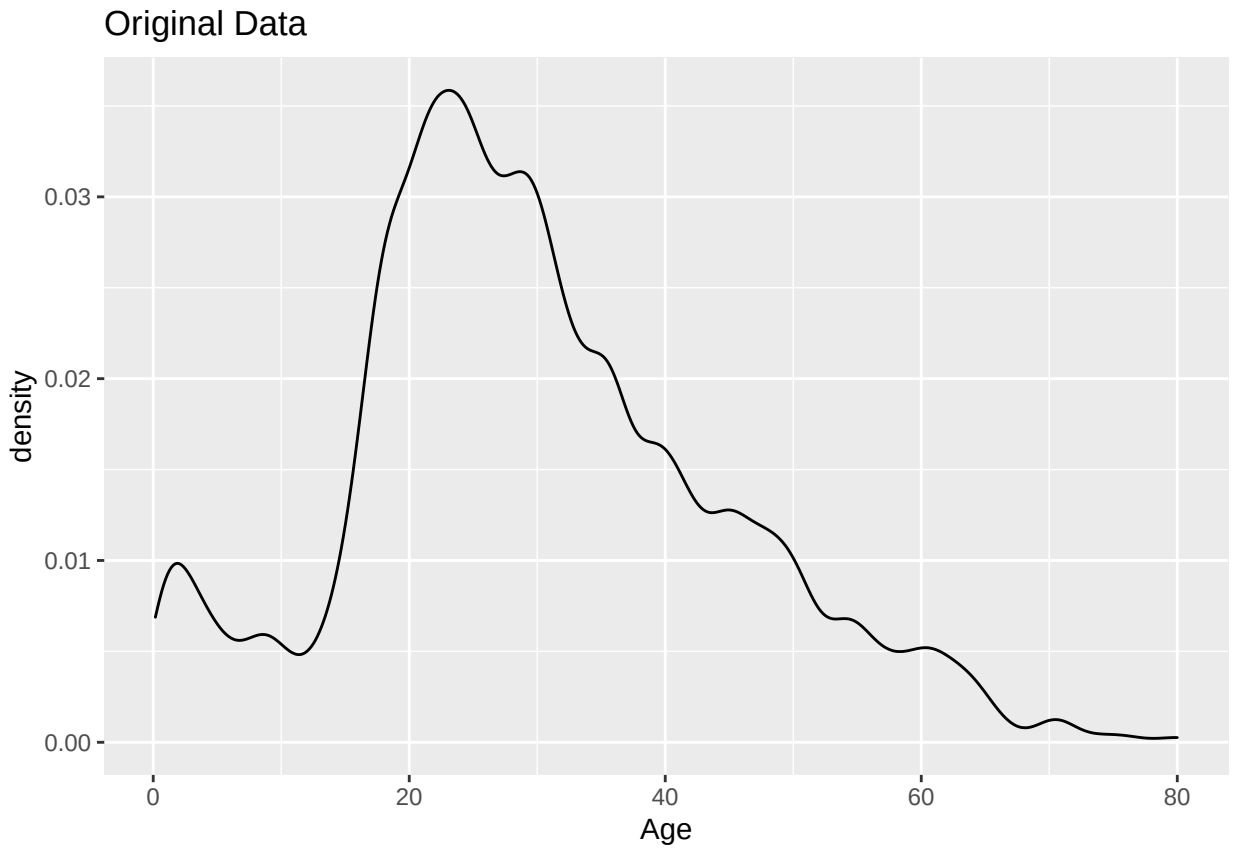
```
# Transformem la variable 'dataset' a data.frame.
dataset <- as.data.frame(dataset)

# Inputem els valors d'edat que falten amb l'ajuda del paquet MICE
input <- mice(
  dataset[, !names(dataset) %in%
            c('PassengerId','Name','Ticket','Cabin','Survived'
              ,'Assigned','FL','FT','ticketlength','one','two'
              ,'three','four','five','six','seven')], rfPackage = "randomForest")
```

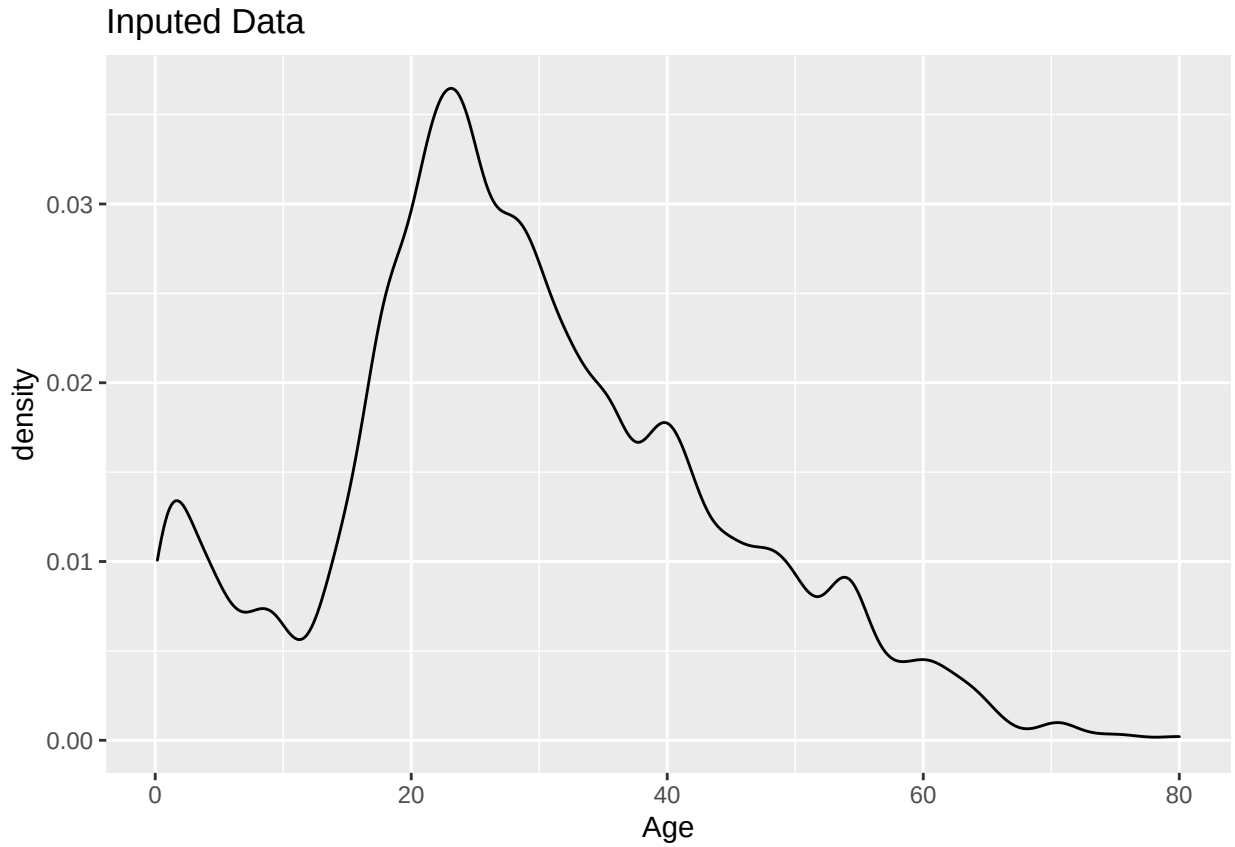
```
##
##  iter imp variable
```

```
##   1   1  Age  Fare
##   1   2  Age  Fare
##   1   3  Age  Fare
##   1   4  Age  Fare
##   1   5  Age  Fare
##   2   1  Age  Fare
##   2   2  Age  Fare
##   2   3  Age  Fare
##   2   4  Age  Fare
##   2   5  Age  Fare
##   3   1  Age  Fare
##   3   2  Age  Fare
##   3   3  Age  Fare
##   3   4  Age  Fare
##   3   5  Age  Fare
##   4   1  Age  Fare
##   4   2  Age  Fare
##   4   3  Age  Fare
##   4   4  Age  Fare
##   4   5  Age  Fare
##   5   1  Age  Fare
##   5   2  Age  Fare
##   5   3  Age  Fare
##   5   4  Age  Fare
##   5   5  Age  Fare
```

```
trained_mouse <- complete(input)
```

A continuació, crearem dos histogrames amb la finalitat de comprovar que els valors generats per el paquet 'mice' no degraden la qualitat del nostre joc de dades.


Original Data
density
0.03
0.02
0.01
0.00
0
20
40
60
80
Age



Observem que els dos gràfics són raonablement semblants, per tant procedim a reemplaçar les dades dels valors inputats als originals.

```
# Insertem a la columna 'Age' del dataset original la nova columna calculada amb la lliberia 'mice'
dataset$Age <- trained_mouse$Age
```

4.1.3 Ajust de la variable 'Embarked'.

Com hem vist anteriorment, hi ha dos valors de la variable **Embarked** que falten. Per trobar el valor d'aquestes dues observacions, procedim a verificar-ho amb l'ajuda dels valors de la variable **Cabin**.

A partir de les dades, podem comprobar que totes les cabines que comencen amb **B** es van embarcar des de les ciutats de Southampton o Charbourg.

```
# Mostrem els valors únics de la variable 'Embark' quan el nom de la cabina comença per 'B'
unique(dataset[grep("*^B", dataset$Cabin),]$Embarked)
```

```
## [1] C   S
## Levels:  C Q S
```

A més a més, podem veure que els bitllets de viatge de tipus **B** costen al voltant de 80 USD, que és molt similar a la tarifa mitja o mitjana dels passatgers **S**.

```
# Calcul de la mitja i la mitjana de les cabines que comencen per 'B' (agrupat pels valors de la variabl
dataset[grep("*^B", dataset$Cabin),] %>% group_by(Embarked) %>% summarize_each(funs(mean),Fare)
```

```
## # A tibble: 3 x 2
##   Embarked  Fare
##   <fct>    <dbl>
## 1 ""        80
## 2 "C"      167.
## 3 "S"       78.6
```

```
dataset[grep("*^B", dataset$Cabin),] %>% group_by(Embarked) %>% summarize_each(funs(median),Fare)
```

```
## # A tibble: 3 x 2
##   Embarked  Fare
##   <fct>    <dbl>
## 1 ""        80
## 2 "C"       91.1
## 3 "S"       82.3
```

Per tant, procedim a inputar els dos valors perduts de **Embarked** com a tipus **S**.

```
# Inputem el valor 'S' en les dues observacions amb valors buits
dataset$Embarked[c(62, 830)] <- 'S'
```

4.1.4 Ajust de la variable 'Fare'.

En aquest cas, com només es tracta d'un registre que no conté la dada, el que farem serà introduïr la mitjana de la resta de valors d'aquesta columna. Aquesta tècnica d'inputació de dades és molt freqüent quan la quantitat de valors no informats és petita i quan l'atribut és del tipus numèric.

Primer de tot, busquem on està localitzat el valor perdut de la variable **Fare**.

```
# Trobem quina és la posició de l'observació que conté el valor buit/NULL
which(is.na(dataset$Fare))
```

```
## [1] 1044
```

A continuació, trobem la mitjana del total dels tiquets excloent el valor perdut.

```
# Calculem la mitjana dels valors de la columna 'Fare' eliminant el registre amb valor buit
mean_fare <- mean(dataset$Fare, na.rm = TRUE)
```

Per últim, inputem el el valor obtingut al valor perdut de la columna **Fare**.

```
# Definim el valor de l'observació número '1044' amb el valor de la mitjana abans calculat
dataset$Fare[c(1044)] <- mean_fare
```

4.1.5 Comprovació de valors buits.

Com podem observar en el càlcul que computarem a continuació, la quantitat de valors no informats després del tractament és de zero observacions.

```r
# Mostrem el número de registres buits per cada columna
apply(dataset=="",2, sum)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0        1014           0
##    row_type
##           0
```

```r
apply(is.na(dataset),2, sum)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
##    row_type
##           0
```

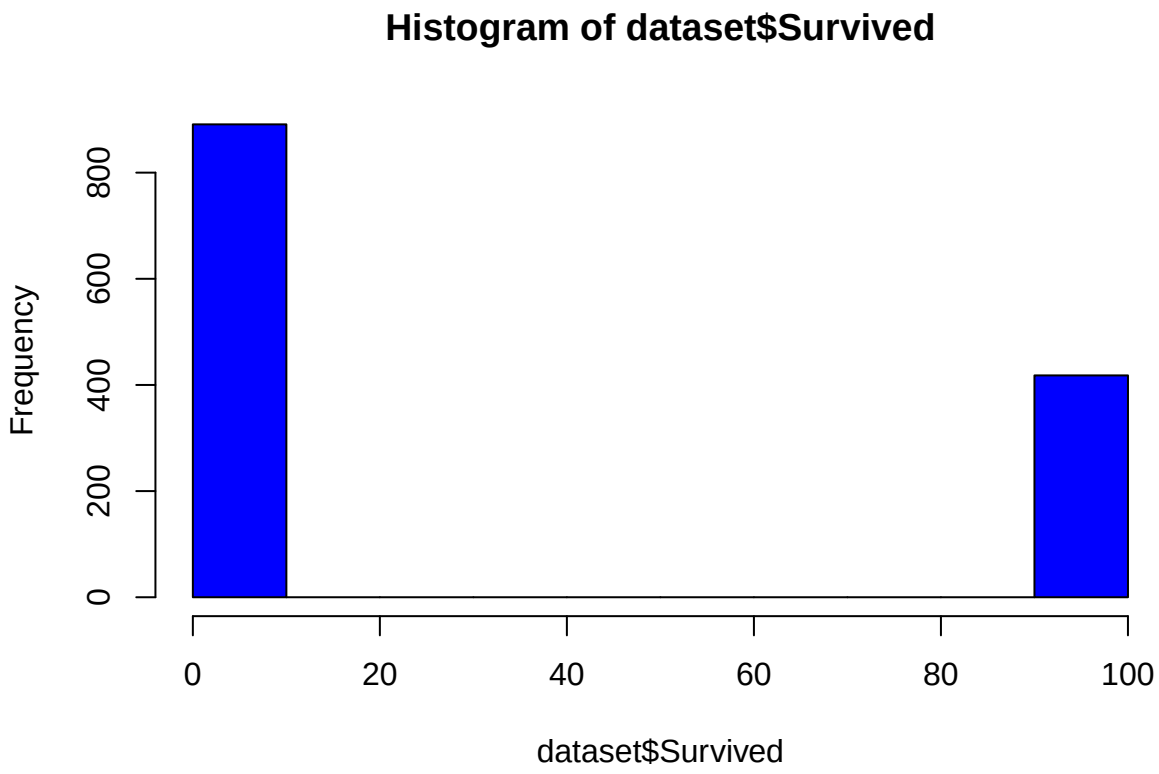
4.2 Valors extrems (outliers).

L'estudi de valors extrem el farem només en les variables del tipus quantitatiu. Això és així ja que, per les variables del tipus qualitatiu, és molt difícil saber que vol dir que un valor esta fora del que es considera 'normal' (o similar a la resta).

4.2.1 Identificació de les columnes amb valors extrems

A continuació, passem a mostrar una sèrie de gràfiques i taules d'estadístiques que ens ajudaran a identificar aquells atributs amb valors extrems.

```r
# Mostrem l'histograma i un resum de la variable 'Survived'
hist(dataset$Survived, col = "blue")
```



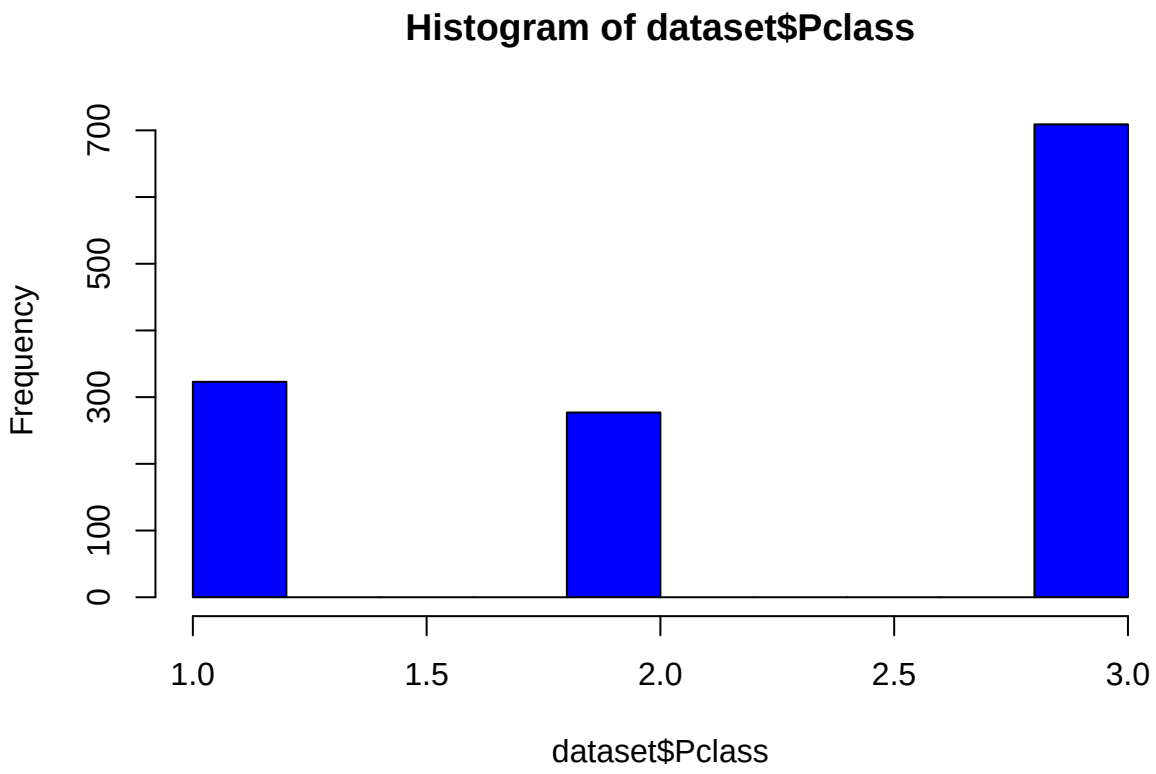

```
summary(dataset$Survived)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    1.00   31.87   99.00   99.00
```

```
table(dataset$Survived)
```

```
##
##   0   1  99
## 549 342 418
```

```
# Mostrem l'histograma i un resum de la variable 'Survived'
hist(dataset$Pclass, col = "blue")
```



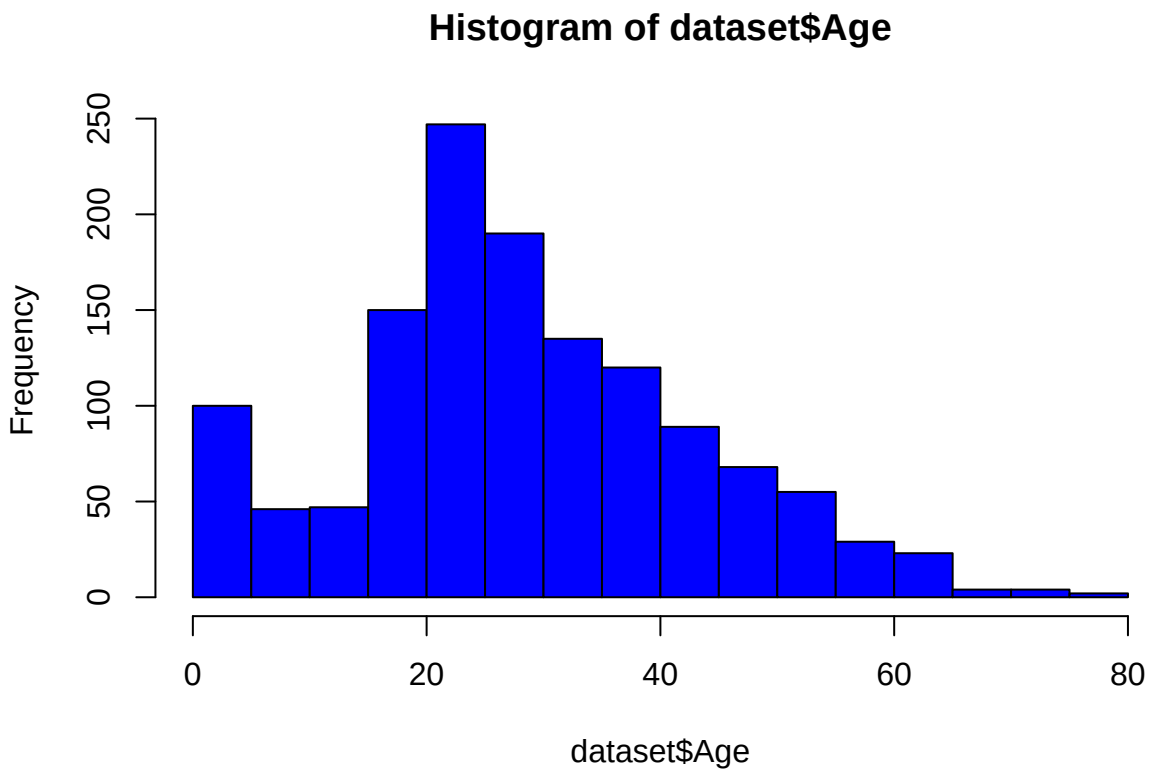

```
summary(dataset$Pclass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.295   3.000   3.000
```

```
table(dataset$Pclass)
```

```
##
##   1   2   3
## 323 277 709
```

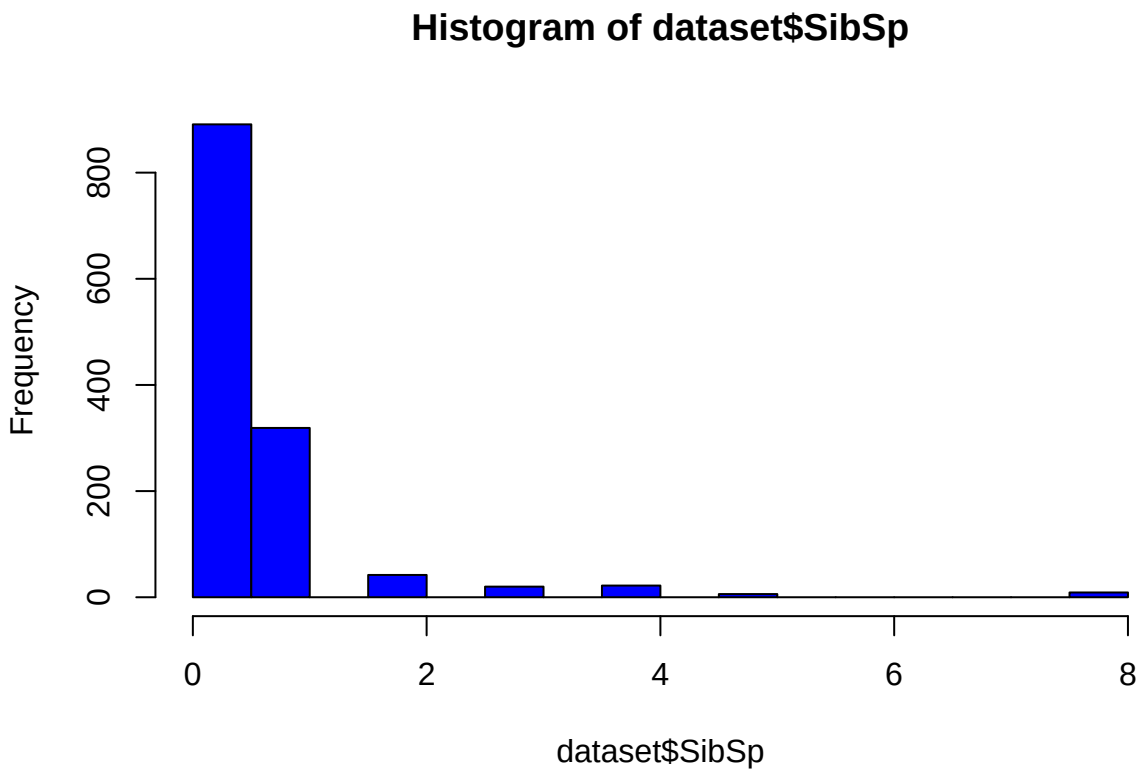
```
# Mostrem l'histograma i un resum de la variable 'Age'
hist(dataset$Age, col = "blue")
```




```
summary(dataset$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.17   20.00   27.00   28.90   39.00   80.00
```

```
# Mostrem l'histograma i un resum de la variable 'SibSp'
hist(dataset$SibSp, col = "blue")
```



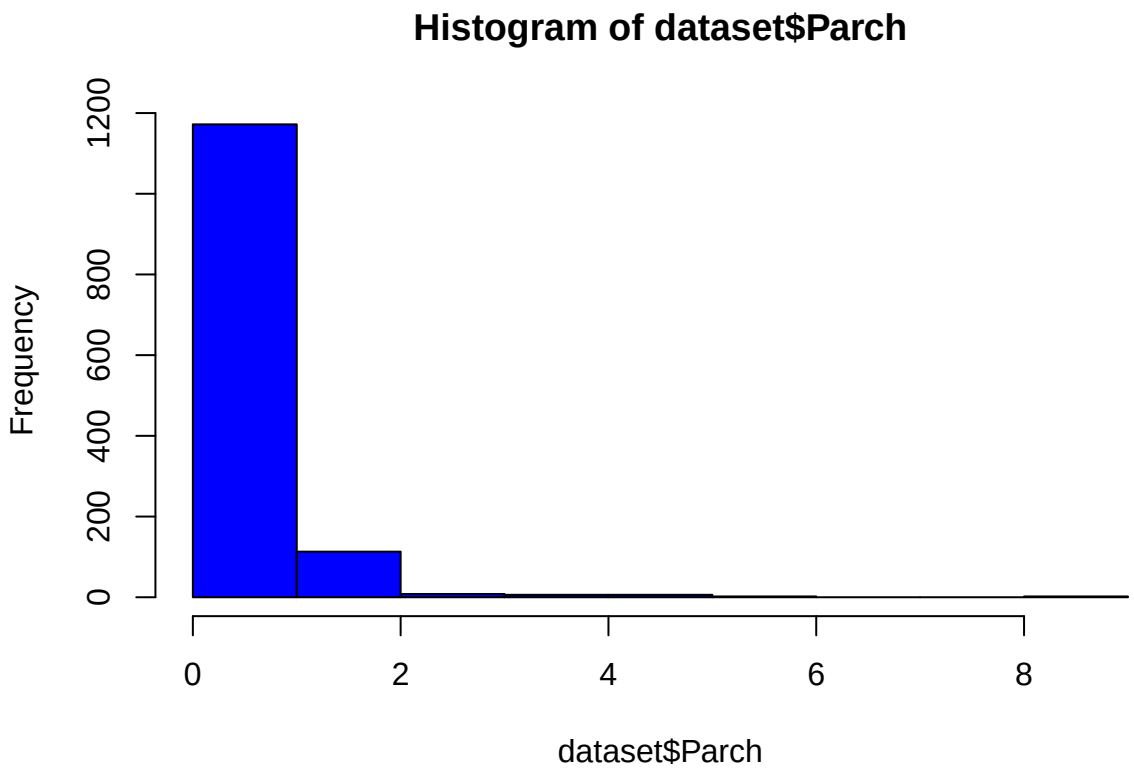

```
summary(dataset$SibSp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4989  1.0000  8.0000
```

```
table(dataset$SibSp)
```

```
##
##   0   1   2   3   4   5   8
## 891 319  42  20  22   6   9
```

```
# Mostrem l'histograma i un resum de la variable 'Parch'
hist(dataset$Parch, col = "blue")
```



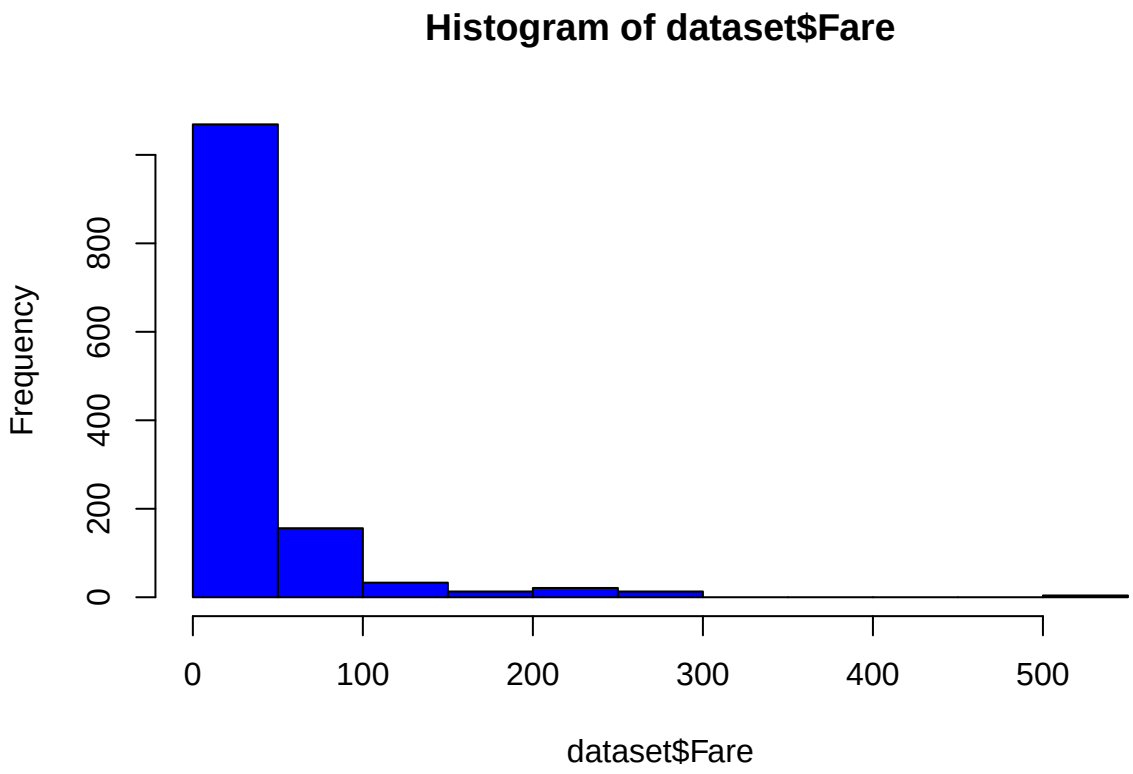

```
summary(dataset$Parch)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.385   0.000   9.000
```

```
table(dataset$Parch)
```

```
##
##    0    1    2    3    4    5    6    9
## 1002  170  113    8    6    6    2    2
```

```
# Mostrem l'histograma i un resum de la variable 'Fare'
hist(dataset$Fare, col = "blue")
```




```
summary(dataset$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.896  14.454  33.295  31.275 512.329
```

D'aquesta informació, observem el següent:

- **Survived**: Tots els valors són o bé 0 o bé 1. Els valors 99 són els que hem inputat nosaltres quan hem fet el *merge* del joc de dades de 'test' amb el de 'train'. No hi ha cap fora dels valors esperats i, per tant, no eliminarem cap registre en base a aquest atribut.
- **Pclass**: Tots els valors són o bé 1 o bé 2 o bé 3. Les diferents classes de tiquet. No hi ha cap fora dels valors esperats i, per tant, no eliminarem cap registre en base a aquest atribut.
- **Age**: El mínim és 0.17 i el màxim és 80. No hi ha cap edat que cridi l'atenció com per considerar-la fora de l'esperat i eliminar-la del joc de dades.
- **SibSp**: Gran part dels valors són enters entre 0 i 1. Hi ha un 9 observacions amb un valor allunyat de la resta com és el valor 8. Tot i això, és un valor que seria possible ja que existeixen families numerosas amb aquesta quantitat de fills. En el nostre cas, no eliminarem aquestes observacions ja que no les considerem extremes (tot i que sí poc probables).
- **Parch**: Gran part dels valors són enters entre 0 i 1. Hi ha 2 observacions amb un valor allunyat de la resta com és el valor 9. Tot i això, és un valor que seria possible ja que existeixen families numerosas amb aquesta quantitat de fills. En el nostre cas, no eliminarem aquestes observacions ja que no les considerem extremes (tot i que sí poc probables).

- **Fare**: Existeixen 4 observacions amb un valor extremadament allunyat de la resta. Aquest valor és el valor 512.329 que és el màxim de la variable. Com hem pogut observar al resum d'estadístiques, la mitjana dels valors d'aquesta columna és 33.295, és a dir, es troba molt lluny de la tendència de valors (també de la mediana i dels quartils). És per aquest motiu que eliminarem aquesta observació i recalcularem la mitjana per introduïrla en els valors que originalment eren buits (ja que abans ho haviem fet amb una mitjana esbiaixada).

Cal apuntar que tot i que hi hagi d'altres valors de la variable **Fare** que semblin extrems, creiem que es poden arribar a donar i és per aquest motiu que els mantindrem.

4.2.2 Ajust de la variable 'Fare'

A continuació eliminarem el registre que conté el valor extrem en la variable **Fare** i recalcularem la mitjana per introduïrla en els valors que originalment eren buits (ja que abans ho haviem fet amb una mitjana esbiaixada).

```
# Calculem el màxim de la variable 'Fare'
max_fare <- max(dataset$Fare)

# Mostrem les dimensions del 'dataset' abans d'eliminar les observacions
dim(dataset)
```

```
## [1] 1309   13
```

```
# Eliminem els registres on el valor és igual al màxim
dataset <- dataset[dataset$Fare != max_fare,]

# Mostrem les dimensions del 'dataset' després d'eliminar les observacions
dim(dataset)
```

```
## [1] 1305   13
```

```
# Calculem el màxim de la variable 'Fare'
max(dataset$Fare)
```

```
## [1] 263
```

```
# Calculem la mitjana de la variable 'Fare'
mean_fare <- mean(dataset$Fare)

# Introduïm el nou valor en l'observació que originalment era buida (l'observació 1044)
dataset$Fare[c(1044)] <- mean_fare
```

4.3 Altres accions per a la neteja del joc de dades.

Primerament, observem que la primera variable "PassengerId" no és res més que un identificador, per tant procedim a eliminar-la del conjunt de dades ja que no ens interessa per l'estudi.

```r
# eliminació de la primera columna PassengerId
dim(unique(dataset$PassengerId))
```

```
## NULL
```

```r
dataset <- dataset[,-1]
```

Observem que les variables 'Survived' i 'Pclass' són de tipus enter, però la seva funció es indicar una categoria. Per tant, procedim a convertir-les en tipu factor.

```r
# transformació de les variables

dataset$Survived <- as.factor(dataset$Survived)
class(dataset$Survived)
```

```
## [1] "factor"
```

```r
dataset$Pclass <- as.factor(dataset$Pclass)
class(dataset$Pclass)
```

```
## [1] "factor"
```

5 Anàlisi de les dades.

5.1 Selecció de grups a analitzar/comparar.

```r
# Placeholder
```

5.2 Normalitat i homogeneïtat de la variància.

```r
# Placeholder
```

5.3 Aplicació de proves estadístiques per a la comparació de grups.

```r
# Placeholder
```

6 Representació de resultats (taules i gràfiques).

```r
# Placeholder
```

7 Conclusions.