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.

El conjunt de dades 'train.csv' que es troba a la carpeta 'data/input' d'aquest paquet s'ha obtingut del web https://www.kaggle.com/c/titanic.

Aquest conjunt de dades contene informació sobre la tripulació del Titanic amb 12 columnes i un total de 891 registres.

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 el fitxer 'train.csv' de la carpeta 'data/input' (indicant que volem els 'strings' com 'fac
dataset <- read.csv("../data/input/train.csv", stringsAsFactors=TRUE)

# Mostrem les primeres files del joc 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
## 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
```

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

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  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/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581
##  $ 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/ 681 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/ 148 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 ...
```

Observem que el dataset conté 3 tipus de variables, caràcter (factor), numèric i enter.

Aquest dataset és interessant per la nostre recerca d'informació ja que estem intentat veure si la variabñe **Age** va tenir un impacte molt gran o no entre la gent que va sobreviure a la tragèdia del Titanic o no. Apart de l'edat, estem interessats en veure quines són les variables amb més pes en la variancia del valor de la columna **Survived**.

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 del dataset.

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 l'anàlisi de les dades.

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.

```r
# 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
## 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
```

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.

```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          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

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

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           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'.

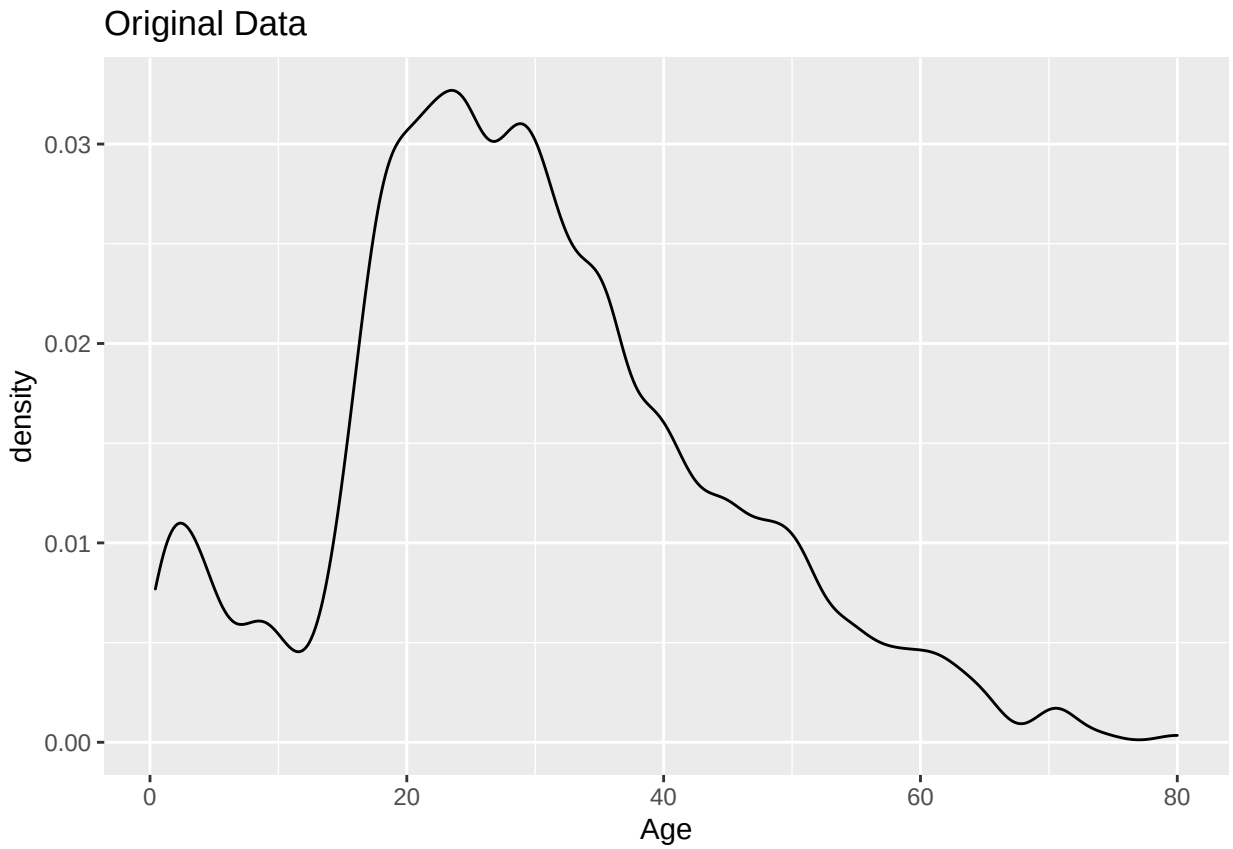
En aquesta secció tractem la columna 'Age' amb l'objectiu de inputar nous valors que en aquells registres on el seu valor és buit o no informat.

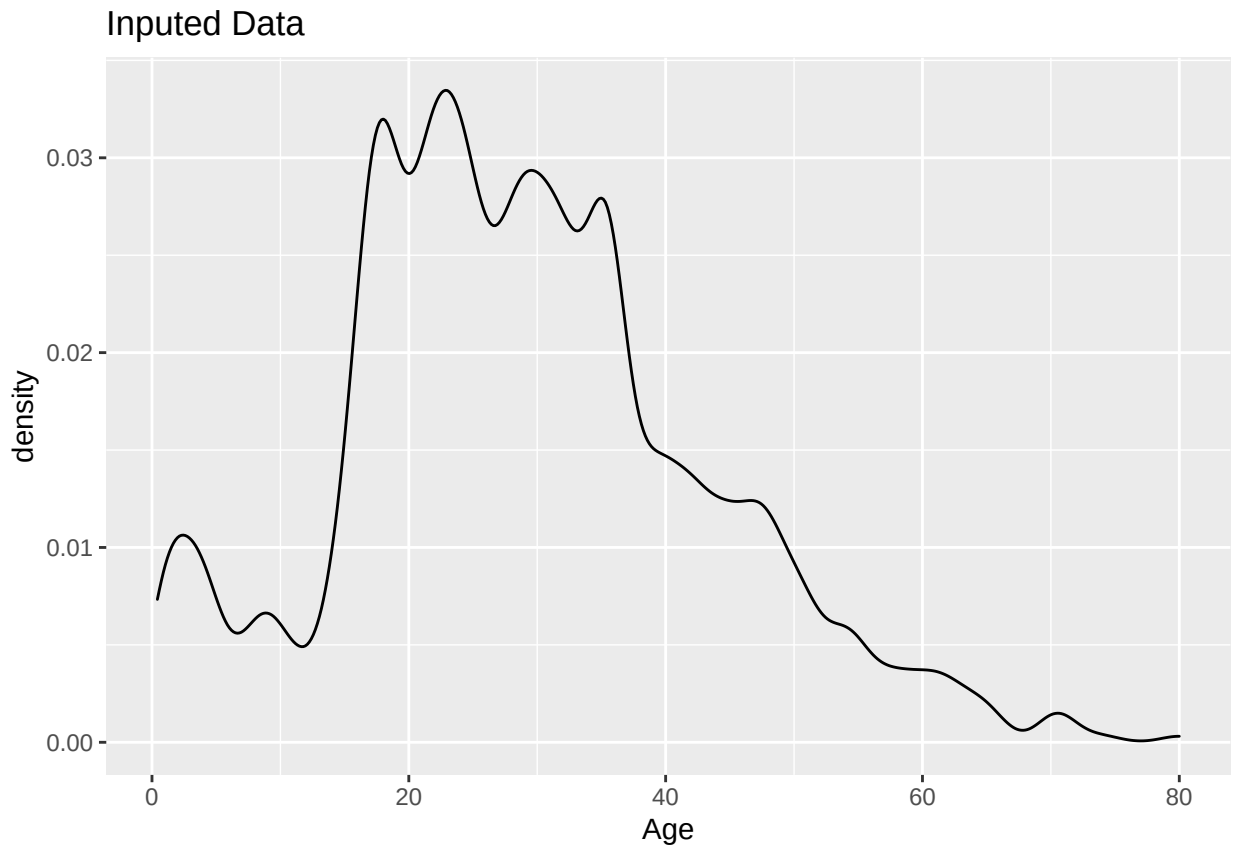
```
# Transformem la variable 'dataset' a data.frame.
dataset <- as.data.frame(dataset)

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

```
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 'Cabin'.

Com gran part de les observacions conté la columna **Cabin** sense informar, 687 de les 891 observacions, hem decidit eliminar aquesta columna del nostre joc de dades.

```
# Eliminem la columna 'Cabin'
dataset$Cabin <- NULL
```

4.1.4 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)
```

```
## factor(0)
## 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: 0 x 2
## # ... with 2 variables: Embarked <fct>, Fare <dbl>
```

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

```
## # A tibble: 0 x 2
## # ... with 2 variables: Embarked <fct>, Fare <dbl>
```

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'
```

4.1.5 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.

```
# 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    Embarked
##           0           0           0           0           0
```

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

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

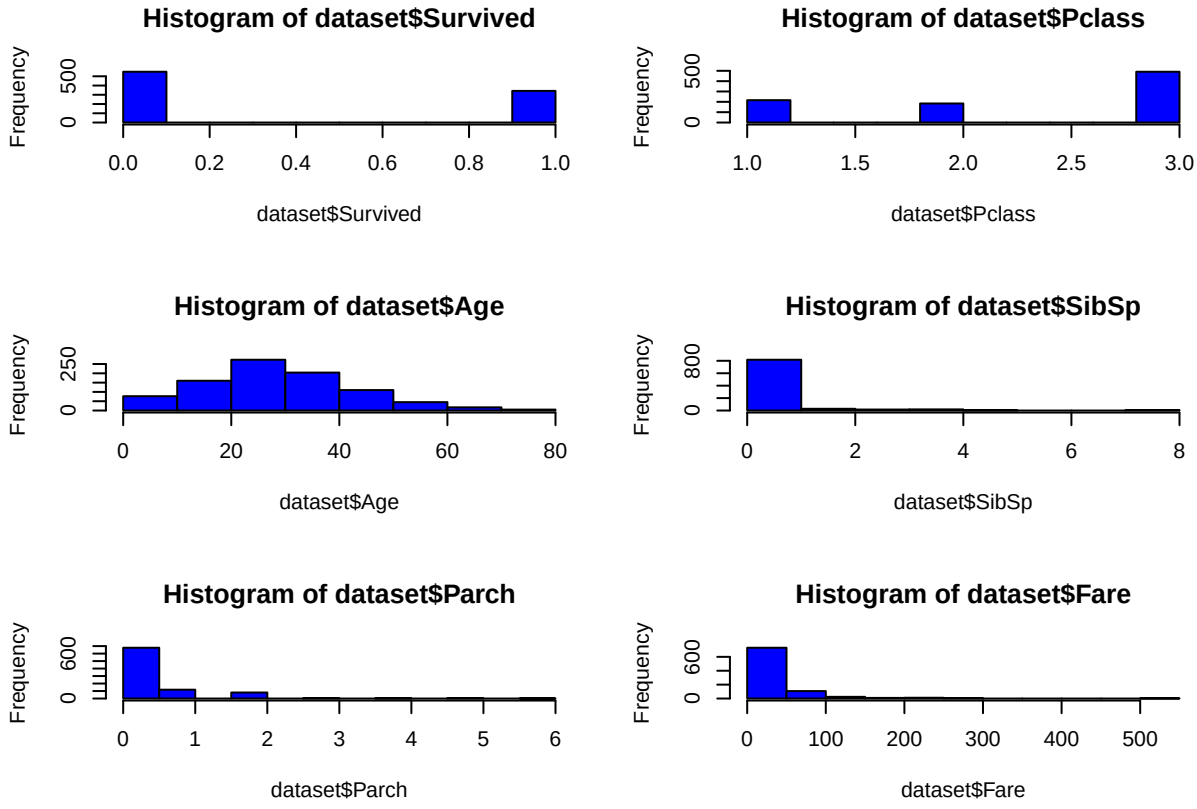
4.2 Valors extrems (outliers).

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.

```
# Mostrem els histogrames de cada variable numèrica
par(mfrow=c(3,2))
hist(dataset$Survived, col = "blue")
hist(dataset$Pclass, col = "blue")
hist(dataset$Age, col = "blue")
hist(dataset$SibSp, col = "blue")
hist(dataset$Parch, col = "blue")
hist(dataset$Fare, col = "blue")
```




```
# Mostrem el summary de les variables numèriques
summary(dataset$Survived)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3838  1.0000  1.0000
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.309   3.000   3.000
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   20.00   28.00   29.16   36.75   80.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.523   1.000   8.000
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3816  0.0000  6.0000
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.91   14.45   32.20   31.00  512.33
```

```r
# Mostrem les freqüències d'algunes variables numèriques (les que sabem que són més agrupades)
table(dataset$Survived)
```

```
##
##   0   1
## 549 342
```

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

```
##
##   1   2   3
## 216 184 491
```

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

```
##
##   0   1   2   3   4   5   8
## 608 209  28  16  18   5   7
```

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

```
##
##   0   1   2   3   4   5   6
## 678 118  80   5   4   5   1
```

```
head(table(dataset$Fare))
```

```
##
##      0 4.0125      5 6.2375 6.4375   6.45
##     15      1      1      1      1      1
```

```
tail(table(dataset$Fare))
```

```
##
## 221.7792  227.525 247.5208  262.375      263 512.3292
##        1        4        2        2        4        3
```

D'aquesta informació, observem el següent:

- **Survived**: Tots els valors són o bé 0 o bé 1. 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 0.42 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 7 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 1 observacions amb un valor allunyat de la resta com és el valor 6. 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 3 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 29.57, 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 aquestes observacions.

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**.

```
# 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] 891  11
```

```r
# 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] 888  11
```

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

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

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

5 Joc de dades final.

En aquesta secció, guardem el joc de dades un cop feta la neteja.

```r
# Guardem el joc de dades a la carpeta 'data/output'
write.csv(dataset,'../data/output/train.csv', row.names = TRUE)
```

6 Anàlisi de les dades.

6.1 Selecció de grups a analitzar/comparar.

Per a aquest estudi, utilitzarem les files del joc de dades i farem una comparació entre la població de persones que va sobreviure i les que no (**Survived**) per veure com va poder influenciar la variable edat en aquests fets (**Age**).

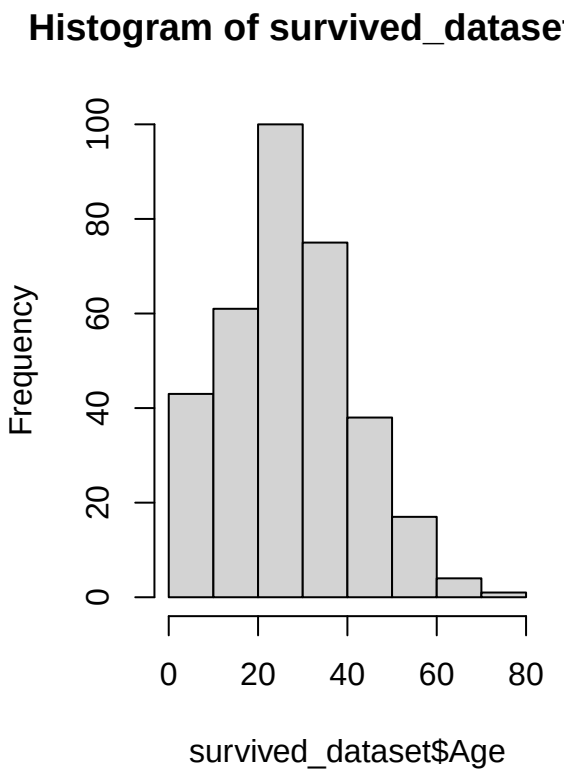
A continuació, ens disposem a dividir el joc de dades en 2 (un per cada valor de la columna **Survived**).

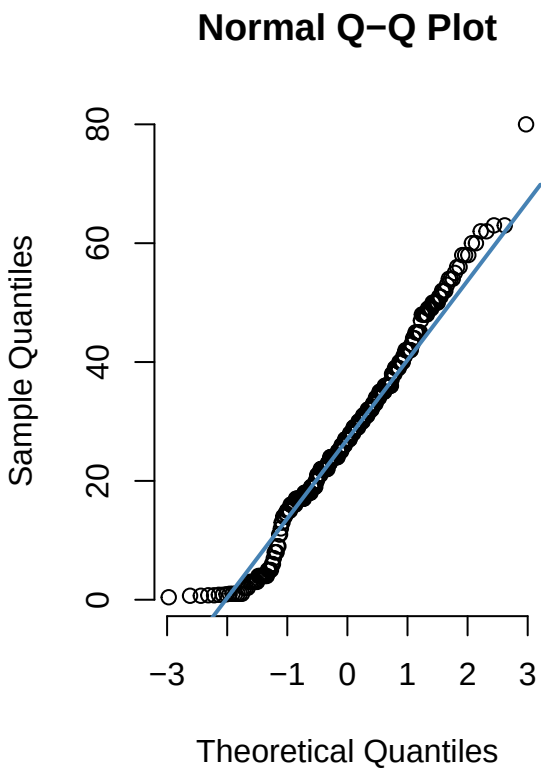
```r
# Selecció dels grups a analitzar
survived_dataset <- dataset[dataset$Survived == 1, ]
did_not_survive_dataset <- dataset[dataset$Survived == 0, ]
```

6.2 Normalitat i homogeneïtat de la variància.

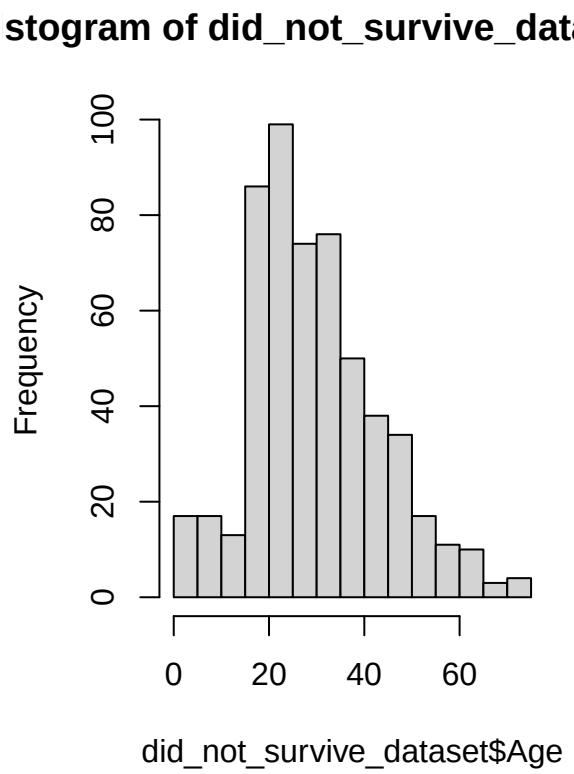
En aquest apartat, comprovem que les poblacions provenen de mostres normalment distribuïdes en quant a la columna **Age**. Per fer-ho, construïm els histogrames d'ambdues poblacions conjuntament amb el QQ-plot.

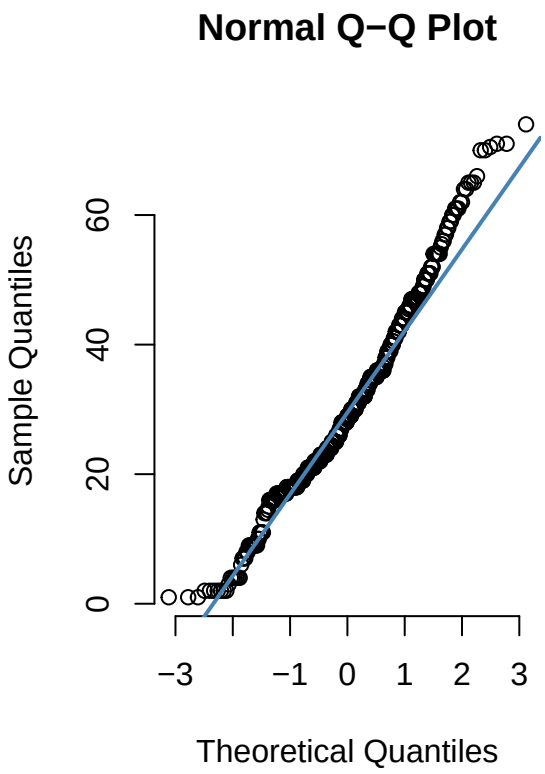
```r
# Generem l'histograma i el QQ-plot pel grup 'survived_dataset'
par(mfrow=c(1,2))
hist(x = survived_dataset$Age)
qqnorm(survived_dataset$Age, pch = 1, frame = FALSE)
qqline(survived_dataset$Age, col = "steelblue", lwd = 2)
```






```r
# Generem l'histograma i el QQ-plot pel grup 'did_not_survive_dataset'
par(mfrow=c(1,2))
hist(x = did_not_survive_dataset$Age)
qqnorm(did_not_survive_dataset$Age, pch = 1, frame = FALSE)
qqline(did_not_survive_dataset$Age, col = "steelblue", lwd = 2)
```





Com podem veure als histogrames i als QQ-plots, podem assumir que ambdos grups provenen d'una mostra normalment distribuïda. L'histograma s'assimila molt a la funció de densitat d'una variable amb distribució normal i el QQ-plot ens mostre com els valors de la mostra és troben sobre la QQ-line. Per tant, podem concloure amb l'ajuda d'aquest gràfics que la variable **Age** de les dues poblacions escollides segueix una **distribució normal**.

A continuació, passem a comprovar la homogeneïtat de la variància o també coneguda com a **homocedasticitat**. Com sabem que ambdues mostres segueixen una distribució normal, el que farem serà aplicar el F-Test. Aquesta prova d'R ens permetrà comparar les variances dels dos grups.

```
# Realització del test i mostra dels resultats
set.seed(1)
result <- var.test(Age ~ Survived, data = dataset)
print(result)
```

```
##
##  F test to compare two variances
##
## data:  Age by Survived
## F = 0.88829, num df = 548, denom df = 338, p-value = 0.2212
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7315102 1.0739126
## sample estimates:
## ratio of variances
##          0.8882918
```

Com podem comprovar, el **p-valor** és superior a 0,05. És a dir, podem, amb un 95% de nivell de confiança descartar la hipòtesi alternativa i quedar-nos amb la hipòtesi nul · la. La hipòtesi nul · la d'aquest test menciona que les **variancies dels dos grups són similars**.

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

A continuació, mostrem diverses proves fetes amb la finalitat d'extreure informació sobre l'impacte de la variable **Age** en el fet de sobreviure o no.

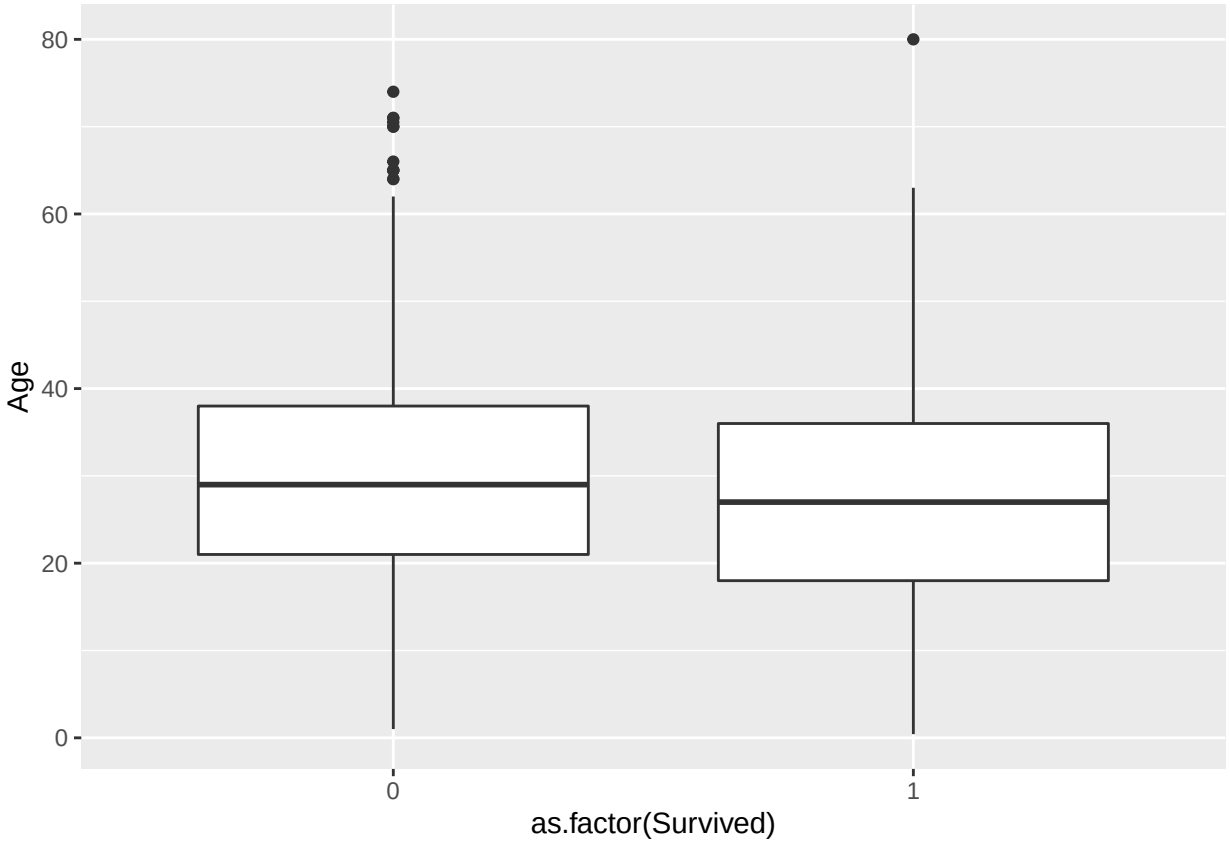
6.3.1 T-test.

Començem realitzant la prova **t-test**, que té com a objectiu determinar si les mitjanes de dos grups són iguals o no ho són. El test assumeix que els grups provenen de distribucions normals amb les mateixes variances. Com ja hem comprovat aquests dos fets, procedim a executar el test.

```
# Calculem el mínim de files entre els dos jocs de dades per utilitzar la mateixa quantitat al T-test
min_rows <- min(dim(survived_dataset)[1], dim(did_not_survive_dataset)[1])
t_test_dataset <- rbind(survived_dataset[1:min_rows, ], did_not_survive_dataset[1:min_rows, ])

# Realitzem el t-test
T_test <- t.test(Age ~ Survived, data = t_test_dataset, paired = TRUE)
T_test
```

```
##
##  Paired t-test
##
## data:  Age by Survived
## t = 2.4303, df = 338, p-value = 0.01561
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4842267 4.5963632
## sample estimates:
## mean of the differences
##                2.540295
```

```
# Il·lustrem el t-test
ggplot(dataset, aes(as.factor(Survived), Age)) + geom_boxplot()
```



Com podem comprovar, el **p-valor** esta per sobre de 0.05. És a dir, amb un 95% de nivell de confiança podem confirmar/concloure que no hi ha diferencia entre les mitjanes de les dues poblacions. Això vol dir, que, en el fet de sobreviure o no, no va tenir un impacte gran l'edat de les persones que anaven a bord.

6.3.2 ANOVA.

La segona prova que executarem s'anomena **ANOVA**, one-way analysis of variance. Aquesta prova també compara les mitjanes entre grups (com ja hem fet amb la prova anterior).

A continuació, executem el test i mostrem els resultats.

```
# Realitzem el test ANOVA i mostrem els resultats
ANOVA_test <- aov(Age ~ Survived, data = dataset)
summary(ANOVA_test)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Survived      1   1439  1439.4   7.287 0.00708 **
## Residuals   886 175012   197.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De manera similar al test anterior, en aquest, el **p-valor** és superior a 0.05. És a dir, amb un 95% de nivell de confiança podem confirmar/concloure que no hi ha diferencia entre les mitjanes de les dues poblacions. Això vol dir, que, en el fet de sobreviure o no, no va tenir un impacte gran l'edat de les persones que anaven a bord, com ja haviem esmentat anteriorment amb l'ajut del **t-test**.

6.3.3 Regressió logistica.

Per últim, tot i que no és, de manera estrica, un test estadístic de comparació, realitzarem una regressió logistica per comprovar fins quin punt, la variació de la variable **Survived** pot ser explicada per la variació dels valors en la variable **Age**.

A continuació, construïm el model i mostrem un resum del mateix.

```
# Creem el model de regressió logistica i mostrem un resum del mateix
model <- glm(Survived ~ Age + Sex + Pclass + Embarked,family=binomial(link = 'logit'), data=dataset)
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Embarked, family = binomial(link = "logit"),
##     data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6603  -0.6275  -0.3970   0.6531   2.5251
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.298318   0.487424  10.870  < 2e-16 ***
## Age         -0.037940   0.007157  -5.301 1.15e-07 ***
## Sexmale     -2.548125   0.188460 -13.521  < 2e-16 ***
## Pclass      -1.237677   0.130480  -9.486  < 2e-16 ***
## EmbarkedQ   -0.077704   0.373244  -0.208   0.8351
## EmbarkedS   -0.475788   0.233338  -2.039   0.0414 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1180.89  on 887  degrees of freedom
## Residual deviance:  787.51  on 882  degrees of freedom
## AIC: 799.51
##
## Number of Fisher Scoring iterations: 5
```

Com s'observa amb la interpretació del model aplicat podem verificar que les variables class, Sex són clarament significatives ja que ens ho indiquen amb el simbol ***, concretament Sex(male), Pclass(3) i Embarked(S) ja que el valor del Z-value és més gran que 2 ja sigui en positiu o en negatiu.

7 Conclusions.

Les conclusions que podem extreure amb els anàlisis realitzats són les següents:

- Els anàlisis desenvolupats ens han permès trobar la resposta a les preguntes que ens plantejavem al principi de la pràctica: quin paper va jugar la variable **Age** en el fet de que les persones del Titanic sobrevisquessin? quines van ser les variables amb més impacte en aquest fet?

- Després d'aplicar i executar la prova t-test en les mostres de població podem concloure que, amb un 95% de nivell de confiança, no existeix diferència entre les mitjanes de les poblacions, d'aquesta manera provem que l'edat no va ser una variable rellevent en quant a la supervivència del Titanic.
- Aquesta mateixa conclusió també la corroborem amb la prova ANOVA (one-way analysis of variance) ja que el p-valor obté un resultat clarament superior a 0.05.
- El model de regressió logístic lineal creat per predir la probabilitat que les diverses variables siguin significatives en quan a la supervivencia del naufragi, ens expliquen que el fet de ser home o dona i de pertanyer a una classe o a una altra va ser clarament significatiu en relació a la supervivència del Titanic ja que el p-valor de les variables **Sex** i **Pclass** són superior a 2 ja sigui en positiu o negatiu.

8 Contribucions.

Contribucions	**Firma**
Investigació prèvia	A.A.G/A.V.P
Redacció de les respostes	A.A.G/A.V.P
Desenvolupament codi	A.A.G/A.V.P