Pràctica 2: Neteja i anàlisi de les dades

Claudia Puche Garcia i Ernest Panareda Roig

19 de diciembre, 2020

Contents

Introducció

Aquesta activitat consisteix en treballar sobre un conjunt de dades (o de l'anglès, dataset) per tal de respondre un objectiu definit prèviament. Per aquest cas s'ha agafat un conjunt de dades sobre els passatgers del Titànic, el famós transanlàntic que es va enfonsar a l'impactar contra un iceberg. Aquest conjunt de dades conté informació dels passatgers com el nom, l'edat, el sexe o la classe en que viatja.

Amb aquestes dades, l'objectiu que es presenta en aquest estudi és trobar si entre els supervivents o no supervivents de la catàstrofe del Titànic hi ha alguna característica comuna que permeti identificar als passatgers que sobreviuen o no sense considerar aquest fet. És a dir, si és possible saber quins passatgers sobreviurien en cas d'accident abans que passés. Per satisfer aquest objectiu, es definirà un model que permeti realitzar prediccions sobre si un passatger sobreviurà.

Descripció del dataset

Importació de les dades

Les dades que s'utilitzen en aquest estudi s'han obtingut de la plataforma virtual Kaggle (https://www.kaggle.com/c/titanic), i estan organitzades en dos fitxers de format csv: un fitxer anomenat 'train.csv' amb totes les dades d'una part dels passatgers, i un segon fitxer anomenat 'test.csv' amb totes les dades dels passatgers restants sense l'atribut que indica si van sobreviure o no. Aquest fet es deu a que el primer fitxer ('train.csv') està pensat per estudiar la relació entre les diferents característiques dels passatgers i el fet de sobreviure o no, mentre que el segon ('test.csv') està pensat en predir si els passatgers sobreviuran o no en funció de les observacions fetes en el primer.

Dit això, s'importen els datasets dels dos fitxers per començar a descriure les dades que contenen.

```
# Import libraries needed for the whole study.
library(ggplot2)
library(ggpubr)
library(car)
```

```
## Loading required package: carData
```

```
# Import train and test data from the csv files.
d_tr <- read.csv("train.csv", sep = ',', stringsAsFactors = TRUE)
d_tt <- read.csv("test.csv", sep = ',', stringsAsFactors = TRUE)
```

Descripció de les dades

Amb els dos datasets importants, s'estudien les dades que contenen, observant si les aquestes són de tipus text, numèric o categòric.

```
# Explore train data.
str(d_tr)
```

```
## '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 ...
```

```
# Explore test data
str(d_tt)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 5
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

El primer dataset, amb les dades d'entrenament que s'utilitzaran per estudiar els passatgers i crear el model, s'observa que hi ha un total de 891 registres i 12 variables, mentre que per les dades de prova o test hi ha 418 registres i 11 variables. Aquesta variable de menys en el segon conjunt de dades es deu a la informació de si el passatger sobreviu o no, la informació objectiu d'aquest estudi.

Les dades que contenen aquests datasets es defineixen a continuació:

- PassengerId: variable numèrica que conté l'identificador únic per a cada passatger.
- Survival: variable categòrica que indica si el passatger sobreviu (0 = No, 1 = Yes).
- Pclass: variable numèrica que indica la classe en que viatge el passatger (1 = 1st, 2 = 2nd, 3 = 3rd).
- Name: variable categòrica que indica el nom del passatger. Hi ha tantes categories com registres: 891.
- Sex: variable categòrica que indica el gènere del passatger.
- Age: variable numèrica que indica l'edat del passatger, en anys.
- SibSp: variable numèrica que indica el nombre de germans/es i espòs/a que el passatger té a bord del Titànic.
- Parch: variable numèrica que indica el nombre de pares i fills que el passatger té a bord del Titànic.
- Ticket: variable categòrica que indica l'identificador de tiquet del passatger. Conté 363 categories diferents.
- Fare: variable numèrica que indica el preu del tiquet del passatger.
- Cabin: variable categòrica que indica l'identificador de la cambra del passatger.
- Embarked: variable categòrica que indica el port on el passatger va embarcar al Titànic (C = Cherbourg, Q = Queenstown, S = Southampton)

Observant aquestes variables, es pot apreciar la importància d'aquestes per a poder donar resposta a l'objectiu d'aquest estudi i trobar si hi ha algun factor, o factors, que determinin la supervivència dels passatgers en l'accident del Titànic.

Integració i selecció de les dades d'interès

Per a realitzar l'anàlisis sobre quins passatgers van sobreviure i quins no, es considera que no totes les variables que conté el dataset d'entrenament són d'interès.

Per començar, la informació que contenen les variables 'Ticket', 'Fare' i 'Cabin' és molt variada i s'entén que la informació d'interès que poden contenir està continguda en la variable 'Pclass', ja que el número de tíquet, el seu preu i la cambra dependran de si el passatger viatja en primera, segona o tercera classe.

També les variables 'PassengerId' i 'Name', sent úniques per a cada persona, es decideix no seleccionar-les per a l'anàlisis posterior.

Finalment, les variables 'SibSp' i 'Parch' s'han considerat que la única informació important que contenen és si el passatger viatjaba sol o en família. Així doncs, a partir d'aquestes dues variables se'n crea una de nova que s'anomenarà 'Family'. Aquesta variable serà categòrica i contidrà dos valors 'Yes' en cas de que el passatger viatgi en família (que mínim una de les viariable 'SibSp' o 'Parch' sigui diferent a 0), o 'No' en cas que viatgi sol (que les dues variables siguin 0).

```
# Create the new variable 'Family'.
d_tr$Family <- d_tr$Pclass
n <- 1
while (n <= length(d_tr$Family)) {
  if(d_tr$SibSp[n]==0 && d_tr$Parch[n]==0){
    d_tr$Family[n] <- 'No'
  }else{
    d_tr$Family[n] <- 'Yes'
  }
  n <- n+1
}
d_tr$Family <- as.factor(d_tr$Family)
head(d_tr$Family,n=20)
```

```
##  [1] Yes Yes No  Yes No  No  No  Yes Yes Yes Yes No  No  Yes No  No  Yes No  Yes
## [20] No
## Levels: No Yes
```

```
# Convert 'Survived' and 'Pclass' as factors.
d_tr$Survived <- as.factor(d_tr$Survived)
d_tr$Pclass <- as.factor(d_tr$Pclass)
head(d_tr$Survived,n=20)
```

```
##  [1] 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 1 0 1 0 1
## Levels: 0 1
```

```
head(d_tr$Pclass,n=20)
```

```
##  [1] 3 1 3 1 3 3 1 3 3 2 3 1 3 3 3 2 3 2 3 3
## Levels: 1 2 3
```

Amb la variable nova 'Family' creada, es procedeix a seleccionar les variables d'interès per a l'anàlisis, sent aquestes les següents: 'Survived','Pclass','Sex','Age','Family' i 'Embarked'.

```
# Create a new dataset with the selected data.
d_tr_f <- d_tr[c('Survived','Pclass','Sex','Age','Family','Embarked')]
head(d_tr_f, n=5)
```

```
##   Survived Pclass    Sex Age Family Embarked
## 1        0      3   male  22    Yes        S
## 2        1      1 female  38    Yes        C
## 3        1      3 female  26     No        S
## 4        1      1 female  35    Yes        S
## 5        0      3   male  35     No        S
```

Neteja de les dades

Abans de començar amb l'anàlisis, s'ha de realitzar una neteja de les dades a utilitzar, assegurar que aquestes són correctes, d'aquesta manera, els resultats que s'obtinguin amb aquestes dades també ho seran.

Primer de tot, es visualitzen els atributs seleccionats per tal de tenir una primera idea d'aquestes dades.

```r
# Study basic statistics of train data.
summary(d_tr_f)
```

```
##  Survived Pclass      Sex           Age        Family    Embarked
##  0:549    1:216   female:314   Min.   : 0.42   No :537    :  2
##  1:342    2:184   male  :577   1st Qu.:20.12   Yes:354   C:168
##           3:491                 Median :28.00             Q: 77
##                                 Mean   :29.70             S:644
##                                 3rd Qu.:38.00
##                                 Max.   :80.00
##                                 NA's   :177
```

En aquestes dades es pot observar com hi ha la presència de 177 valors buits en la única variable numèrica del dataset 'Age', i en la variable 'Embarked' hi ha dos registres amb un espai en blanc. Tanmateix, observant els valors estadístics a 'Age', sembla que hi pot haver valors extrems en edats grans.

Valors buits

Com bé ja s'ha comentat, s'han detectat valors buits en les variables 'Age' i 'Embarked', tot i que en el segon cas són espais en blanc, no nul · ls. Aquestes dades perdudes no contenen informació, són errors, així que es decidexi que s'hi ha d'imputar valors.

La regla que s'ha decidit més correcte en aquest cas per associar un valor en les dades buides de l'edat, ha estat col · locar la mitjana d'edats de la resta de dades agrupant per a la classe en que viatgen.

En quant als 2 registres amb un espai en blanc a la variable 'Embark', com Southampton ('S') és on embarca més gent (644), es decideix atribuir aquests dos registres a aquest port.

```r
# Check positions of 'NA' values in 'Age'.
index = apply(d_tr_f, 1, function(x) any(is.na(x)))
# Imputation of NA values from 'Age' with the mean for the corresponding 'Pclass'.
d_tr_f$Embarked_2 <- d_tr_f$Age
n <- 1
while (n <= length(d_tr_f$Age)) {
  if(is.na(d_tr_f$Age[n])){
    if(d_tr_f$Pclass[n] == 1){
      d_tr_f$Age[n] <- summary(d_tr_f$Age[d_tr_f$Pclass == 1])['Mean']
    }else if(d_tr_f$Pclass[n] == 2){
      d_tr_f$Age[n] <- summary(d_tr_f$Age[d_tr_f$Pclass == 2])['Mean']
    }else if(d_tr_f$Pclass[n] == 3){
      d_tr_f$Age[n] <- summary(d_tr_f$Age[d_tr_f$Pclass == 3])['Mean']
    }
  }
  # Imputation of withe values from 'Embarked' with the 'S' value.
  if(d_tr_f$Embarked[n] == ''){
    d_tr_f$Embarked_2[n] <- 'S'
  }else if(d_tr_f$Embarked[n] == 'C'){
    d_tr_f$Embarked_2[n] <- 'C'
  }else if(d_tr_f$Embarked[n] == 'Q'){
    d_tr_f$Embarked_2[n] <- 'Q'
  }else if(d_tr_f$Embarked[n] == 'S'){
    d_tr_f$Embarked_2[n] <- 'S'
  }
  n <- n+1
}
d_tr_f$Embarked <- as.factor(d_tr_f$Embarked_2)
```

```r
d_tr_f <- subset(d_tr_f, select= -c(Embarked_2))
head(d_tr_f$Age[index],n=20)
```

```
##  [1] 25.14062 29.87763 25.14062 25.14062 25.14062 25.14062 38.23344 25.14062
##  [9] 25.14062 25.14062 25.14062 25.14062 25.14062 25.14062 38.23344 38.23344
## [17] 25.14062 25.14062 25.14062 25.14062
```

```r
levels(d_tr_f$Embarked)
```

```
## [1] "C" "Q" "S"
```

Es pot observar que els registres on abans hi havia valors nul · ls en l'edat, ara contenen la mitjana en funció de la classe, i la variable 'Embarked' només conté 3 nivells, ja no té cap espai en blanc.

Valors extrems

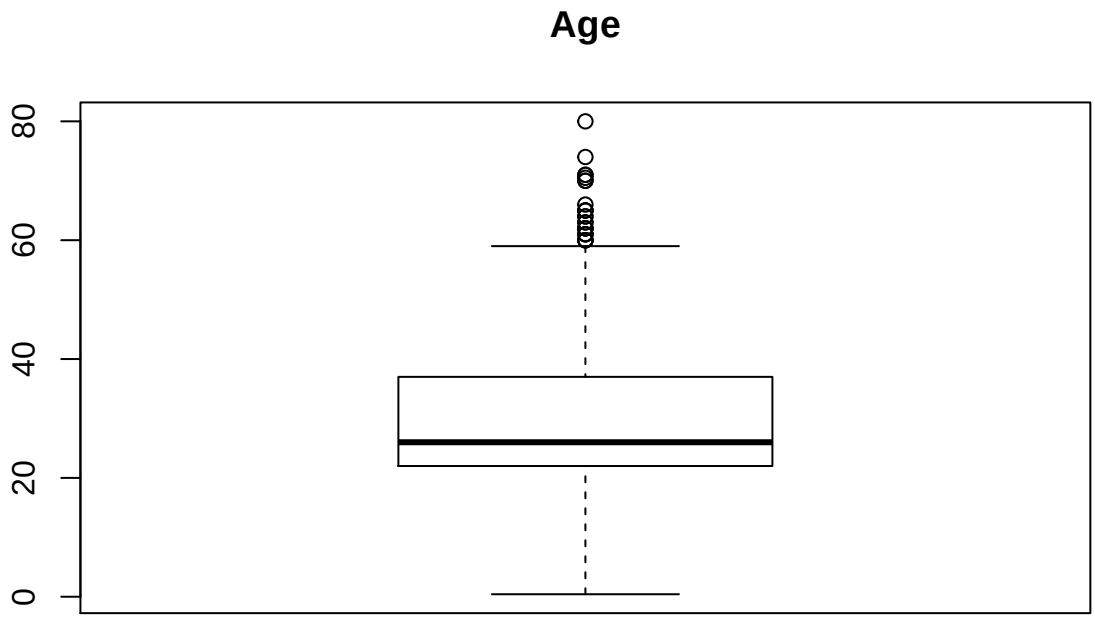
Per estudiar els valors extrems, es calculen aquests i es mostren en una taula.

```r
# Check outliers in the 'Age' variable.
data_attr <- attributes(d_tr_f)$names
outliers <- c()
for(a in data_attr){
  if(is.factor(d_tr_f[a][,1])){
    outlier <- NaN
  }else{
    outlier <- length(boxplot.stats(d_tr_f[a][,1])$out)
  }
  outliers <- append(outliers, outlier)
}

# Creating an output table for the results.
t_outliers <- matrix(outliers, ncol=length(outliers))
colnames(t_outliers) <- data_attr
rownames(t_outliers) <- 'Nº outliers'
t_outliers
```

```
##             Survived Pclass Sex Age Family Embarked
## Nº outliers      NaN    NaN NaN  26    NaN      NaN
```

Sent 'Age' la única variable numèrica, és la única que pot tenir valors atípics, i se'n detecta un total de 26. Per veure si aquests valors extrems s'allunyen molt o no de la resta de dades, es representa gràficament un gràfic de caixa sobre la variable 'Age'.

```r
# Plot a box graphic of the 'Age' variable.
boxplot(d_tr_f$Age, main='Age')
```



En aquesta gràfica es mostren els 26 valors extrems, tots sent valors grans. Aquestes dades en termes generals no es troben massa allunyades de la resta, únicament hi ha un valor vora als 80 anys que està més aïllat. Tot i que 80 anys és un valor possible, es considera que per aquest estudi pot causar problemes en les prediccions, així que s'eliminen tots els registres del detaset que tinguin un valor d'edat superior o igual a 75.

```
# Remove the rows with an age value greater o equal than 75.
d_tr_f <- d_tr_f[-c(d_tr$PassengerId[d_tr_f$Age >= 75]),]
```

Exportació de dades preprocesades.

En aquest punt, les dades seleccionades per a fer l'anàlisis, i definir el model que donarà resposta a l'objectiu d'aquest treball, així que s'exporten per si en futures ocasions es vol tornar a realitzar algun estudi, a partir d'aquestes dades ja es podrà començar, no serà necessari tornar a treballar-les.

```
# Export the preprocesed data in a csv file.
write.csv(d_tr_f, "train_clean.csv")
```

Anàlisis de les dades

Anàlisis bàsic

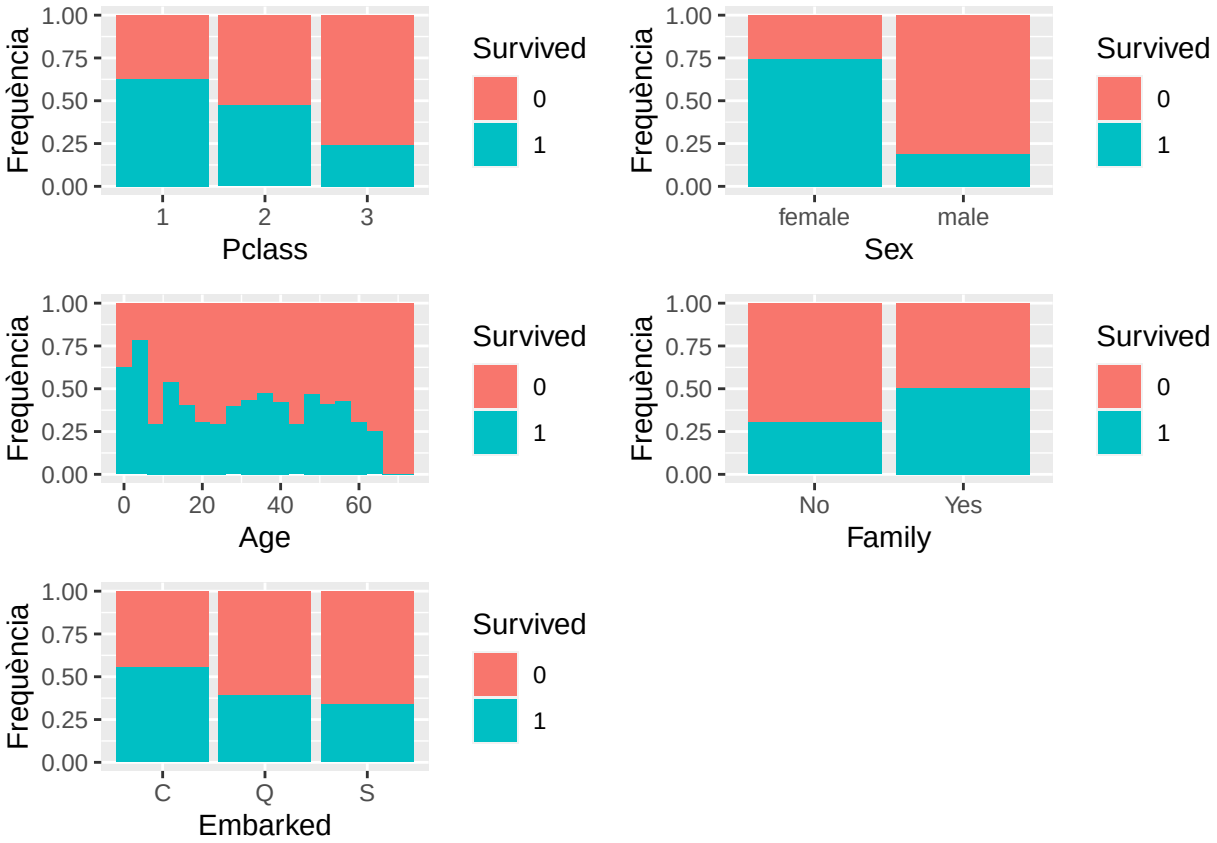
Com a primer anàlisis sobre el conjunt de dades, és interessant observar com estan repartits els passatgers entre supervivents i no supervivients, respecte a la resta de variables seleccionades per a l'estudi.

Per això, es realitzen representacions gràfiques sobre cada variable, separant per a cada grup entre supervivents i no supervivents. Al ser pràcticament tot variables categòriques, les representacions es realitzen en gràfics de barres. En el cas de la variable 'Age', es representen en un histograma, agrupant les columnes en intervals de

4.

Com el cas d'estudi és saber quins passatgers són candidats a sobreviure o no, les representacions es realitzen sobre la freqüència, és a dir, en percentatge, per poder observar així la relació entre supervivència en cada valor de la resta de variables. Per tenir una informació més acurada, s'acompanyen les representacions amb les taules que contenen aquesta infromació.

```
# Plot different bar and histogram graphic with the variables vs Survived.
layout(matrix(c(1,2,
                3,3,
                4,5), 3,3, byrow = TRUE))
Pclass_plot <- ggplot(data = d_tr_f,aes(x=Pclass,fill=Survived)) +
               geom_bar(position="fill") + ylab("Frequència")
Sex_plot <- ggplot(data = d_tr_f,aes(x=Sex,fill=Survived)) +
            geom_bar(position="fill") + ylab("Frequència")
Age_plot <- ggplot(data = d_tr_f,aes(x=Age,fill=Survived)) +
            geom_histogram(binwidth = 4,position="fill") + ylab("Frequència")
Family_plot <- ggplot(data = d_tr_f,aes(x=Family,fill=Survived)) +
               geom_bar(position="fill") + ylab("Frequència")
Embarked_plot <- ggplot(data = d_tr_f,aes(x=Embarked,fill=Survived)) +
                 geom_bar(position="fill") + ylab("Frequència")

figure <- ggarrange(Pclass_plot,
                    Sex_plot,
                    Age_plot,
                    Family_plot,
                    Embarked_plot,
                    ncol = 2, nrow = 3)
figure
```




```
# Create a tables with the percentage of survivers for each variable, except 'Age'.
data_attr <- c('Pclass', 'Sex', 'Family', 'Embarked')

for(a in data_attr){
  t<-table(d_tr_f[a][,1],d_tr_f$Survived)
  for (i in 1:dim(t)[1]){
      t[i,]<-t[i,]/sum(t[i,])*100
  }
  print(t)
}
```

```
##
##            0        1
##   1 37.20930 62.79070
##   2 52.71739 47.28261
##   3 75.76375 24.23625
##
##                  0        1
##   female 25.79618 74.20382
##   male   81.25000 18.75000
##
##              0        1
##   No  69.77612 30.22388
##   Yes 49.43503 50.56497
##
##            0        1
##   C 44.64286 55.35714
```

```
##   Q 61.03896 38.96104
##   S 66.20155 33.79845
```

Les observacions sobre els resultats per a cada variable s'enumeren a continuació:

- Pclass: per als passatgers de primera classe, el percentatge de supervivents és molt superior al 50%, mentre que per als de segona es situa lleugerament per sota al 50%, significant que hi ha més no supervivents de segona classe. En quant als passatgers de tercera classe, el número de supervivents no arriba a 1/4 del total. Doncs, a millor classe, sembla que les probabilitats de sobreviure augmenten.
- Sex: la diferència de relació entre supervivents respecte homes i dones mostra una diferència evident. El percentatge de supervivents per dones és de casi un 75%, mentre que per homes no arriba al 20%. En aquest cas, no hi ha dubte que el fet de ser dona és un factor molt important per sobreviure.
- Age: per a l'edat dels passatgers, s'observa una clara majoria de supervivents per menors de 10 anys, mentre que per a majors de 60 predominen els no supervivents. Per a la resta d'edats, la relació de supervivents-no va variant entre 30/45-70/55, sent en tots els casos majoria de no supervivents. Aquesta variable sembla ser rellevant en quant a la supervivència per les edats extremes.
- Family: per als passatgers en família, s'observa que la relació entre supervivents és més favorable (pràcticament del 50-50) que per als que no, on en aquest passatgers que viatgen sols hi ha una clara majoria de no supervivents. Així doncs, sembla que els passatgers amb família han de tenir més èxit a sobreviure.
- Embarked: per als passatgers que han embarcat a Cherbourg mostren una relació de supervivència de 45-55, guanyant els supervivents, però en els casos de Queenstown i Southampton s'observa una relació similar, on els supervivents representen menys del 40%. Aquest factor sembla no ser massa important a l'hora de determinarla supervivència dels passatgers, tot i que en un punt d'embarc s'aprecia una millor relació en quant a la supervivència.

Selecció dels grups de dades que es volen analitzar/comparar

Un cop realitzat l'anàlisis bàsic sobre les dades, es pot definir el plà a seguir per determinar quins passatgers són més probables de sobreviure.

Tenint en compte que entre les dades seleccionades únicament hi ha 6 variables, sent una d'elles la variable que es vol explicar en aquest estudi, s'analitzaran totes i cadascuna de les variables.

Els passos a seguir en aquest anàlisis a relaitzar seran: començant analitzant la variable numèrica 'Age', després la resta de variables categòriques, i finalment es crearà el model per a realitzar prediccions sobre la supervivència dels passatgers. A continuació, es detallen aquests passos a seguir:

- Contrast d'hipòtesis sobre la mitjana poblacional de l'edat entre els passatgers supervivents, i els no supervivents.
- Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers homes i passatgeres dones.
- Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers amb família i sense.
- Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers de primera classe i de segona.
- Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers de segona classe i de tercera.
- Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers embarcats a Cherbourg i Queenstown.
- Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers embarcats a Queenstown i Southampton.
- Càlcul d'un model de regressió logística prenent la supervivència com a variable a explicar i la resta com a variables explicatives.

Per a aquests casos enumerats, i per tots els testos que es realitzin d'ara en endavant en aquest estudi, es pren un nivell de confiança del 95%.

Comprovació de la normalitat i homogeneïtat de la variància

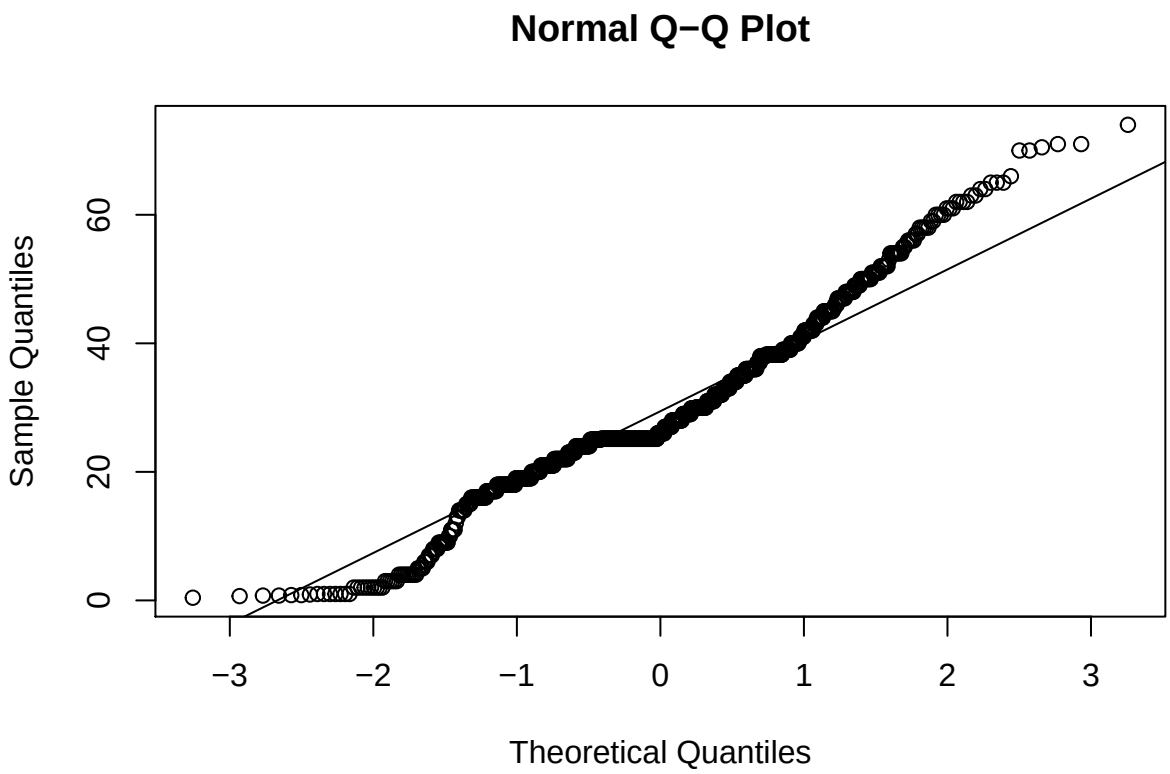
Per a poder realitzar un contrast d'hipòtesis sobre la mitjana poblacional en la variable 'Age', primer és necessàri estudiar si la distribució d'aquestes dades és normal i si la variància és homogènia (homocedasticitat), és a dir, si la variància entre els passatgers supervivents i no es pot considerar igual en ambdós grups.

Començant per comprovar si la variable 'Age' segueix una distribució normal, s'aplica el test de Shapiro-Wilk sobre aquesta variable, així com es representa el gràfic Q-Q.

```
# Apply the Shapiro-Wilk test and the Q-Q graphic to 'Age'.
shapiro.test(d_tr_f$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d_tr_f$Age
## W = 0.96481, p-value = 8.04e-14
```

```
qqnorm(d_tr_f$Age)
qqline(d_tr_f$Age)
```



El test Shapiro-Wilk dona un p-value de $8{,}04 \cdot 10^{-14}$ (molt inferior a 0,05), així que es rebutja l'hipòtesis de que la variable 'Age' segueix una distribució normal. A més, en la gràfica Q-Q, els punts s'allunyen significativament de la recta.

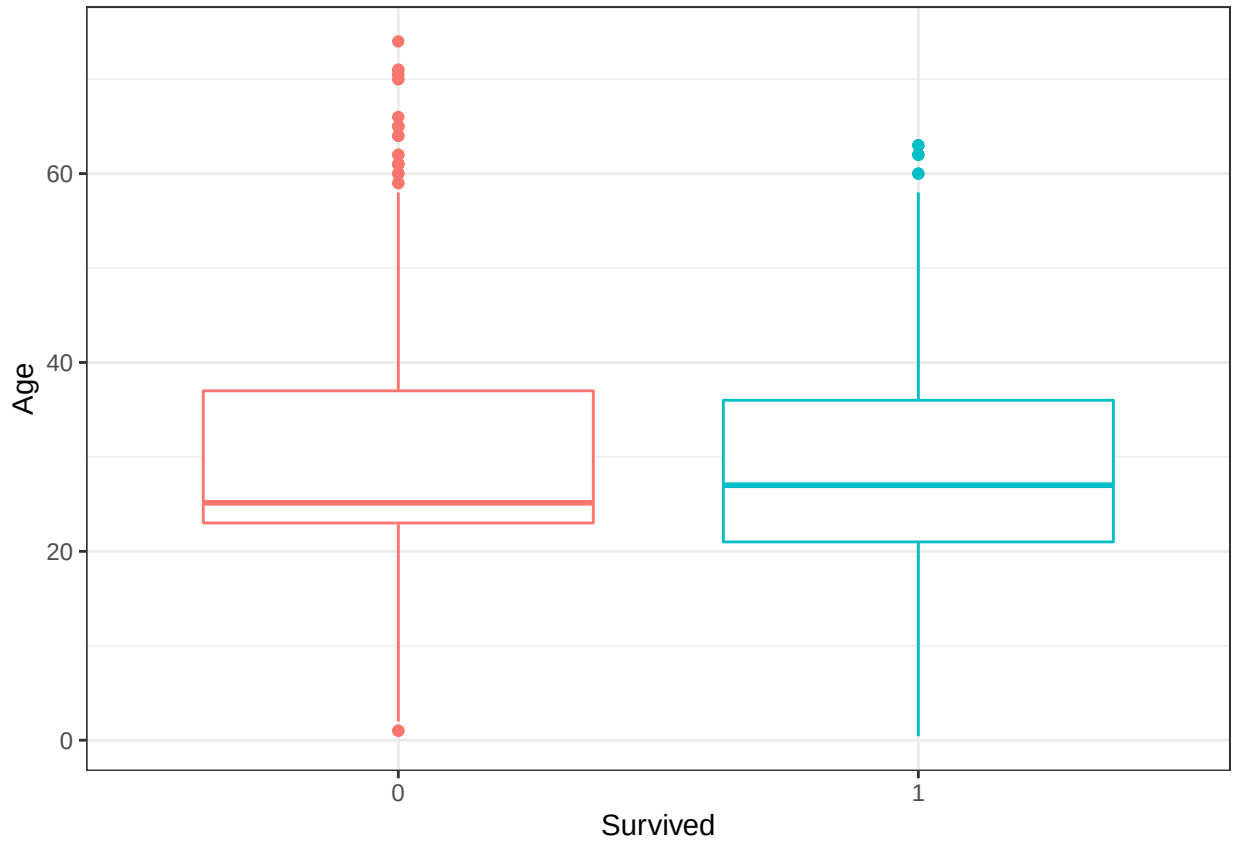
Doncs sabent que les dades sobre l'edat dels passatgers no es distribueixen de forma normal, es pot procedir

a comprovar la homogeneïtat de la variància en aquesta mateixa variable 'Age'.

Per estudiar l'homocedasticitat de les dades sobre l'edat, primer es representen els diagrames de caixes per ambdós grups, per tenir una idea de si les mostres entre aquests dos grups es distribueixen de la mateixa manera. I per tenir una resposta definitiva a saber si les variàcnies són iguals, tenint en compte que la variable 'Age' no segueix una distribució normal, s'aplica el test de Levene sobre la mediana.

Estudiem homocedasticitat (homogeneitat de la variància). https://www.cienciadedatos.net/documentos/9_homogeneidad_de_varianza_homocedasticidad.html

```
# Plot the boxe graphics for 'Age' in the each two Survive groups.
ggplot(data = d_tr_f, aes(x = Survived, y = Age, colour = Survived)) +
  geom_boxplot() +
  theme_bw() +
  theme(legend.position = "none")
```




```
# Apply the Levene test on 'Age' grouping with 'Survived'
leveneTest(y = d_tr_f$Age, group = d_tr_f$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   1  4.3147 0.03807 *
##       888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observant la posició de la mediana en els diagrames de caixes respecte als quartils, s'aprecia una diferència significativa entre els dos grups. I finalment, en el test de Levene, tot i que per poc, el p-value obtingut de 0,038 és inferior al nivell de significació 0,05. Això implica que es rebutja l'hipòtesis nul · la de que les dues

variàncies són iguals, i per tant, les variàncies entre les edats dels dos grups (supervivents i no supervivents) són diferent.

Contrast d'hipòtesis sobre la mitjana poblacional de l'edat entre els passatgers supervivents, i els no supervivents.

Les hipòtesis nul · la i alternativa d'aquest contrast d'hipòtesi són:

$$H_0: \quad \mu_{survived} = \mu_{no\_survived}$$

$$H_1: \quad \mu_{survived} < \mu_{no\_survived}$$

Això vol dir que es vol demostrar que la mitjana poblacional de l'edat entre els passatger supervivents és igual a mitjana d'edat dels no supervivents.

```
# Apply an hipotesis test that checks if the mean age for survivers
# is inferior than the mean for no survivers.
t.test(d_tr_f$Age[d_tr_f$Survived==1],
       d_tr_f$Age[d_tr_f$Survived==0],
       alternative="less",
       var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  d_tr_f$Age[d_tr_f$Survived == 1] and d_tr_f$Age[d_tr_f$Survived == 0]
## t = -1.6575, df = 680.99, p-value = 0.04894
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.009515644
## sample estimates:
## mean of x mean of y
##  28.29686  29.81917
```

Es pot observar que el p-valor representa un valor de 0.048 que és inferior a 0.05, per tant, es rebutja la hipòtesi nul · la de que la mitjana poblacional és igual entre els supervivents i els no supervivents. Tot i que, s'ha establert que el nivell de confiança és del 95% s'hauria de dir que si es reestableix a un 97% s'acceptaria la hipòtesi, ja que el p-valor és elevat i es troba en el límit. Per tant, es confirma que la mitjana de l'edad que sobreviuen és inferior a la mitjana d'edat d'aquells no supervivents.

Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers homes i passatgeres dones.

El segon contrast d'hipòtesi serà sobre la proporció de supervivents dones i homes. Per tant, les hipòtesis nul · la i alternativa són:

$$H_0: \quad p_{female} = p_{male}$$

$$H_1: \quad p_{female} > p_{male}$$

La hipòtesi nul · la descriu que la proporció de dones supervivents és igual a la proporció d'homes supervivents mentre que la hipòtesi alternativa descriu que la proporció de dones és superior a la d'homes supervivents.

```r
# Print a table of the number of survivers vs the sex.
table(d_tr_f$Sex, d_tr_f$Survived)
```

```
##
##            0   1
##   female  81 233
##   male   468 108
```

```r
# Apply an hipotesis test that checks if the proportion of survivers
# is greater for women than is for men.
x1 <- d_tr_f[d_tr_f$Sex=="female",]
x2 <- d_tr_f[d_tr_f$Sex=="male",]
n1 <- length(x1$Survived)
n2 <- length(x2$Survived)
p1 <- sum(x1$Survived==1)/length(x1$Survived)
p2 <- sum(x2$Survived==1)/length(x2$Survived)
success<-c(p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 264.41, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.5059079 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7420382 0.1875000
```

A la taula ja es pot observar que les dones que sobreviuen són 233 de 314 mentre que la proporció d'homes que sobreviu és de 108 de 576. Addicionalment, el contrast d'hipòtesi afirma que amb un nivell de confiança del 95% es rebutja la hipòtesi nul · la de que les proporcions són iguals, ja que el p-valor corresponent és de 2.2e-16, inferior a 0.05. Per tant, es confirma que la proporció de dones supervivents és més elevada que la proporció d'homes.

Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers amb família i sense.

El tercer contrast d'hipòtesi també serà sobre la proporció, tot i que en aquest cas es voldrà comparar la proporció de passatgers que van sobreviure que tenien familia a bord amb aquells que van sobreviure i que no tenien familia.

$$H_0: \quad p_{family} = p_{non-family}$$

$$H_1: \quad p_{family} > p_{non-family}$$

En aquest cas la hipòtesi nul · la afirma que la proporció de supervivents que tenen família és igual a la proporció que no tenen família. De lo contrari, la hipòtesi alternativa descriu que la proporció de supervivents amb família és superior a la proporció de supervivents sense família dintre del Titanic.

```r
# Print a table of the number of survivers vs if the passenger
# has family on board
table(d_tr_f$Family, d_tr_f$Survived)
```

```
##
##         0   1
##   No  374 162
##   Yes 175 179
```

```r
# Apply an hipotesis test that checks if the proportion of survivers
# is greater for people with family on board than is for the ones that don't.
x1 <- d_tr_f[d_tr_f$Family=="Yes",]
x2 <- d_tr_f[d_tr_f$Family=="No",]
n1 <- length(x1$Survived)
n2 <- length(x2$Survived)
p1 <- sum(x1$Survived==1)/length(x1$Survived)
p2 <- sum(x2$Survived==1)/length(x2$Survived)
success<-c(p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 37.323, df = 1, p-value = 5.004e-10
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1488678 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.5056497 0.3022388
```

Observant la taula es pot comprobar que els que supervivents que no tenen família són 162 de 532 mentre que els supervivents que sí tenen família respresenten 179 sobre 354. Tot i que si s'observa el contrast d'hipòtesi es pot comprobar que aquest obté un p-valor igual a 5.004e-10 inferior a 0.05, per tant, es pot afirmar que ambdues proporcions són diferents. Es confirma llavors que amb un nivell de confiança del 95% la proporció de supervivents amb família és superior a la proporció de supervivents sense família.

Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers de primera classe i de segona.

A continuació es procedirà a realitzar un contrast d'hipòtesi que compari la proporció de supervivents de primera classe amb els de segona classe. Per tant, les hipòtesis nul · la i alternativa a demostrar són:

$$H_0: \quad p_{class1} = p_{class2}$$

$$H_1: \quad p_{class1} > p_{class2}$$

La hipòtesi nul · la descriu la igualtat entre la proporció de supervivents de primera classe amb els supervivents de segona classe mentre que la hipòtesi alternativa afirma que la proporció de supervivents que viatjen en primera classe és superior als supervivents de segona classe.

```r
# Print a table of the number of survivers vs the class
table(d_tr_f$Pclass, d_tr_f$Survived)
```

```
##
##       0   1
##   1  80 135
##   2  97  87
##   3 372 119
```

```r
# Apply an hipotesis test that checks if the proportion of survivers
# is greater for people in first clas than is for the people in second.
x1 <- d_tr_f[d_tr_f$Pclass==1,]
x2 <- d_tr_f[d_tr_f$Pclass==2,]
n1 <- length(x1$Survived)
n2 <- length(x2$Survived)
p1 <- sum(x1$Survived==1)/length(x1$Survived)
p2 <- sum(x2$Survived==1)/length(x2$Survived)
success<-c(p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 9.6609, df = 1, p-value = 0.0009411
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.07380815 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.6279070 0.4728261
```

Primerament es visualitzen les dades mitjançant la taula en la que es pot observar que sobreviuen 135 passatgers de primera classe de 215 front a 87 supervivents de segona classe d'un total de 184. A continuació es visualitza el constrast d'hipòtesi que compara ambdues proporcions. El p-valor del model és igual a 0.00094, per tant, es rebutja la hipòtesi nul · la de que la proporció de supervivents de primera i segona classe són iguals, ja que és inferior a 0.05. Com era d'esperar, la proporció de supervivents de primera classe és superior a la proporció dels de segona classe.

Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers de segona classe i de tercera.

Per extreure conclusió sobre la proporció de supervivents entre classes s'ha de comparar també la tercera classe. Per tant, el següent contrast d'hipòtesi compararà les proporcions entre els supervivents de segona i tercera classe. Les hipòtesi nul · la i alternativa d'aquest model seràn:

$$H_0: \quad p_{class2} = p_{class3}$$

$$H_1: \quad p_{class2} > p_{class3}$$

La hipòtesi nul · la afirma que la proporció de supervivents entre els passatgers de segona classe i tercera són iguals mentre que la alternativa descriu que la proporció dels supervivents de segona classe és superior als de

tercera.

```
# Apply an hipotesis test that checks if the proportion of survivers
# is greater for people in the second class than is for the third.
x1 <- d_tr_f[d_tr_f$Pclass==2,]
x2 <- d_tr_f[d_tr_f$Pclass==3,]
n1 <- length(x1$Survived)
n2 <- length(x2$Survived)
p1 <- sum(x1$Survived==1)/length(x1$Survived)
p2 <- sum(x2$Survived==1)/length(x2$Survived)
success<-c(p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 33.525, df = 1, p-value = 3.518e-09
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1620752 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.4728261 0.2423625
```

A la taula mostrada al contrast d'hipòtesi anterior es pot observar que els supervivents de tercera classe són 119 d'un total de 491. El constrast d'hipòtesi confirma que el rebutja la hipòtesi nul · la, ja que el p-valor és igual a 3.518e-09 inferior a 0.05. D'aquesta manera es confirma amb un 95% de confiança, que la proporció de supervivents de segona classe és més gran que la proporció de supervivents de tercera. Mitjançant aquest contrast d'hipòtesi i l'anterior, es pot concloure que els passatgers de primera classe són els que tenen més probabilitats de sobreviure la catàstrofe seguit dels de segona classe i per últim els de tercera.

Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers embarcats a Cherbourg i Queenstown.

A continuació es procedirà a analitzar la proporció de supervivents en funció de la ciutat d'embarcament. Inicialment, es començarà a fer un contrast d'hipòtesi que compari la proporció de supervivents que van embarcar a Cherbourg amb els que van embarcar a Queenstown. Per tant, les hipòtesis nul · la i alternativa seràn:

$$H_0: \quad p_{Cherbourg} = p_{Queenstown}$$

$$H_1: \quad p_{Cherbourg} > p_{Queenstown}$$

La hipòtesi nul · la defensa que la proporció de supervivents que embarca a Cherbourg i Queenstown és la mateixa, mentre que la alternativa afirma que la proporció de supervivents de Cherbourg és superior que la Queenstown.

```
# Print a table of the number of survivers vs the embarked place.
table(d_tr_f$Embarked, d_tr_f$Survived)
```

```
##
##       0   1
```

```
##   C  75  93
##   Q  47  30
##   S 427 218
```

```
# Apply an hipotesis test that checks if the proportion of survivers
# is greater for people that embarked in Cherbourg than is for the
# ones that did in Queenstown.
x1 <- d_tr_f[d_tr_f$Embarked=='C',]
x2 <- d_tr_f[d_tr_f$Embarked=='Q',]
n1 <- length(x1$Survived)
n2 <- length(x2$Survived)
p1 <- sum(x1$Survived==1)/length(x1$Survived)
p2 <- sum(x2$Survived==1)/length(x2$Survived)
success<-c(p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 5.6778, df = 1, p-value = 0.00859
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.05289373 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5535714 0.3896104
```

Mitjançant la taula es pot comprobar que els passatgers que sobreviuen que embarquen a Cherbourg són 93 de 168 mentre que els que sobreviuen i embarquen a Queenstown són 47 de 77. El contrast d'hipòtesi deixa veure que el p-valor és igual a 0.008 inferior a 0.05. D'aquesta manera, es rebutja la hipòtesi nul · la i amb un nivell de confiança del 95% es confirma que la proporció de supervivents que embarquen a Cherbourg és superior a la proporció de Queenstown.

Contrast d'hipòtesis sobre la proporció de supervivents entre passatgers embarcats a Queenstown i Southampton.

Per extreure conclusions definitives sobre el lloc d'embarcament caldria incloure els van embarcar a Southampton. Dit això, aquest últim contrast d'hipòtesi compararà els supervivents que van embarcar a Queenstown amb els de Southampton. La hipòtesi nul · la i la hipòtesi alternativa s'escriuen a continuació:

$$H_0: \quad p_{Queenstown} = p_{Southampton}$$

$$H_1: \quad p_{Queenstown} > p_{Southampton}$$

La hipòtesi nul · la descriu que la proporció de supervivients que embarquen a Queenstown és igual a la proporció de supervivents que embarquen a Southampton mentre que la hipòtesi alternativa confirma que la proporció de Queenstown és superior a la de Southampton.

```
# Apply an hipotesis test that checks if the proportion of survivers
# is greater for people that embarked in Queenstown than is for the
# ones that did in Southampton.
```

```r
x1 <- d_tr_f[d_tr_f$Embarked=='Q',]
x2 <- d_tr_f[d_tr_f$Embarked=='S',]
n1 <- length(x1$Survived)
n2 <- length(x2$Survived)
p1 <- sum(x1$Survived==1)/length(x1$Survived)
p2 <- sum(x2$Survived==1)/length(x2$Survived)
success<-c(p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 0.813, df = 1, p-value = 0.1836
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.04478268  1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.3896104 0.3379845
```

Mitjançant el contrast d'hipòtesi sobre la proporció es pot determinar que s'accepta la hipòtesi nul · la de que la proporció de supervivents a Queenstown i Southampton és igual, ja que el p-valor és igual a 0.18 que és superior a 0.05. Per tant, amb un nivell de confiança del 95% es por confirmar que ambdues proporcions són iguals.

Càlcul d'un model de regressió logística prenent la supervivència com a variable a explicar i la resta com a variables explicatives

En aquest apartat s'aplica un model de regressió logística en la que la variable depenent és Survived i les explicatives totes les altres.

```r
# Create a logistic regression method to explain 'Survived' with all the others.
model1 <- glm(formula = Survived~Pclass+Sex+Age+Family+Embarked,
              data = d_tr_f, family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Family + Embarked,
##     family = binomial, data = d_tr_f)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6381  -0.6455  -0.4034   0.6262   2.5383
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.173831   0.464191   8.992  < 2e-16 ***
## Pclass2     -1.053998   0.272973  -3.861 0.000113 ***
## Pclass3     -2.447573   0.268608  -9.112  < 2e-16 ***
## Sexmale     -2.614162   0.197522 -13.235  < 2e-16 ***
## Age         -0.038835   0.007993  -4.859 1.18e-06 ***
```

```
## FamilyYes   -0.102489   0.192465  -0.533 0.594373
## EmbarkedQ   -0.073101   0.373767  -0.196 0.844940
## EmbarkedS   -0.545272   0.235297  -2.317 0.020483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1184.74  on 889  degrees of freedom
## Residual deviance:  791.28  on 882  degrees of freedom
## AIC: 807.28
##
## Number of Fisher Scoring iterations: 5
```

Es pot observar que totes les variables explicatives són estadísticament significatives a un nivell de confiança del 95% menys FamilyYes i EmbarkedQ. Aquestes dues últimes variables no tenen cap asterisc al final de la fila, per tant, això indica que no representen valor estadístic dintre del model. Tot i així, per poder interpretar més exhaustivament la regressió logística caldria calcular i interpretar els odds ratio. Addicionalment, també es pot comprobar que tots els coeficients de la regressió són negatius, això vol dir que tots ells afecten negativament al fet de sobreviure a la catàtrofe del Titanic.

Afegir que aquest model calculat té un valor pel criteri d'Akaike (AIC) de 807,28. Aquest valor no té cap sentit si no es pot comparar amb altres models. Tot i que no s'ha mostrat el codi per simplificar, aquest model de regressió logística obtingut utilitzant totes les variables seleccionades, és el que presenta un valor de AIC menor, és a dir, s'han observat els valors d'aquest criteri en models prenent només algunes de les variables, i el valor del criteri d'Akaike resultaba major. Així doncs, aquest model 1 és el que presenta un millor nivell d'ajust.

Càlcul i interpretació dels odds ratio

```
# Calculate the odds ratio  in the model1.
exp(coefficients(model1))
```

```
## (Intercept)     Pclass2     Pclass3     Sexmale         Age   FamilyYes
## 64.96388472  0.34854152  0.08650328  0.07322914  0.96190897  0.90258781
##   EmbarkedQ   EmbarkedS
##  0.92950693  0.57968422
```

A continuació es farà una breu interpretació dels odds ratio de cada una de les variables significatives dintre del model:

- Pclass2: per cada passatger que viatja en segona classe, l'odds de sobreviure disminueix en un 0.348
- Pclass3: per cada passatger que viatja en tercera classe, l'odds de sobreviure disminueix en un 0.086
- Sexmale: si el passatger és un home, l'odds de sobreviure disminueix en un 0.073
- Age: per cada any addicional que té un passatger, l'odds de sobreviure disminueix en un 0.961
- FamilyYes: si el passatger té família, l'odds de sobreviure disminueix en un 0.902
- EmbarkedQ: si el passatger va embarcar al port de Queenstown, l'odds de sobreviure disminueix en 0.929
- EmbarkedS si el passatger va embarcar al port de Southamptons, l'odds de sobreviure disminueix en 0.57

Una vegada interpretats els coeficients, s'ha de mencionar que la variable que té més pes en aquest model és el sexe, ja que per l'odds ratio és el més petit. Per tant, la probabilitat de sobreviure canvia més que qualsevol altre variable. A continuació s'avaluarà la precisió del model.

Matriu de confusió

Mitjançant la matriu de confusió del model de regressió logística es pot observar quantes observacions s'han predit correctament i quines no.

```
# Define the confusion matrix for model1.
pred <- ifelse(test = model1$fitted.values > 0.5, yes = 1, no = 0)
conf_mat <- table(model1$model$Survived, pred,
                          dnn = c("observations", "predictions"))
conf_mat
```

```
##             predictions
## observations   0   1
##            0 467  82
##            1  99 242
```

```
# Calculate the precision of model1.
(467+242)/(467+242+99+82)*100
```

```
## [1] 79.66292
```

Es pot concloure que el model de regressió logística té una precisió del 79.66%.

Visualització del model

A continuació es crearà un gràfic per cada variable analitzada anteriorment en funció de l'edat de cada passatger. Les variables a analitzar seran el sexe, la classe en la que viatjava el passatger, el port d'embarcament i si els passatger té família o no. Aquestes gràfiques tenen l'objectiu de mostrar l'incidència de cada variable dins del model representant-les en funció de l'edat a l'eix de les X i en funció de la probabilitat de supervivència a l'eix de les Y.

```
# Calculate the data to plot the male curve.
ages <- seq(from = min(d_tr_f$Age),
               to = max(d_tr_f$Age), by = 0.5)
sex_m <- as.factor(rep(x = "male", length(ages)))
sex_f <- as.factor(rep(x = "female", length(ages)))
fam <- as.factor(rep(x = "Yes", length(ages)))
pclass <- as.factor(rep(x = '1', length(ages)))
emb <- as.factor(rep(x = "C", length(ages)))
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_m,
                                          Age = ages,
                                          Family = fam,
                                          Embarked = emb),
                        type = "response")

data_male <- data.frame(Age = ages,
                        Sex = sex_m,
                        Survived = predictions)


# Calculate the data to plot the female curve.
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_f,
                                          Age = ages,
```

```
                                        Family = fam,
                                        Embarked = emb),
                        type = "response")
data_female <- data.frame(Age = ages,
                        Sex = sex_f,
                        Survived = predictions)

# Bind the data for the curves created.
data <- rbind(data_male, data_female)

# Create the graphic of the predictions between sexs.
sex_plot <-  ggplot(data = d_tr_f,
                    aes(x = Age, y = as.numeric(as.character(Survived)),
                    color = Sex)) +
                    geom_point() +
                    geom_line(data = data, aes(y = Survived)) +
                    geom_line(data = data, aes(y = Survived)) +
                    theme_bw() +
                    labs(title = "P. Survived vs Age vs Sex",
                         y = "P(Survived)") +
                    theme(plot.title = element_text(size = 10))
```

```
# Calculate the data to plot the 1st class curve.
pclass_2 <- as.factor(rep(x = '2', length(ages)))
pclass_3 <- as.factor(rep(x = '3', length(ages)))
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                        Sex = sex_m,
                                        Age = ages,
                                        Family = fam,
                                        Embarked = emb),
                        type = "response")

data_class1 <- data.frame(Age = ages,
                          Pclass = pclass,
                          Survived = predictions)


# Calculate the data to plot the 2nd class curve.
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass_2,
                                        Sex = sex_m,
                                        Age = ages,
                                        Family = fam,
                                        Embarked = emb),
                        type = "response")

data_class2 <- data.frame(Age = ages,
                          Pclass = pclass_2,
                          Survived = predictions)

# Calculate the data to plot the 3rd class curve.
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass_3,
```

```
                                        Sex = sex_m,
                                        Age = ages,
                                        Family = fam,
                                        Embarked = emb),
                        type = "response")

data_class3 <- data.frame(Age = ages,
                          Pclass = pclass_3,
                          Survived = predictions)

# Bind the data for the curves created.
data <- rbind(data_class1, data_class2, data_class3)

# Create the graphic of the predictions between classes.
pclass_plot <- ggplot(data = d_tr_f,
               aes(x = Age, y = as.numeric(as.character(Survived)),
               color = Pclass)) +
               geom_point() +
               geom_line(data = data, aes(y = Survived)) +
               geom_line(data = data, aes(y = Survived)) +
               geom_line(data = data, aes(y = Survived)) +
               theme_bw() +
               labs(title = "P. Survived vs Age vs Pclass",
                    y = "P(Survived)") +
               theme(plot.title = element_text(size = 10))
```

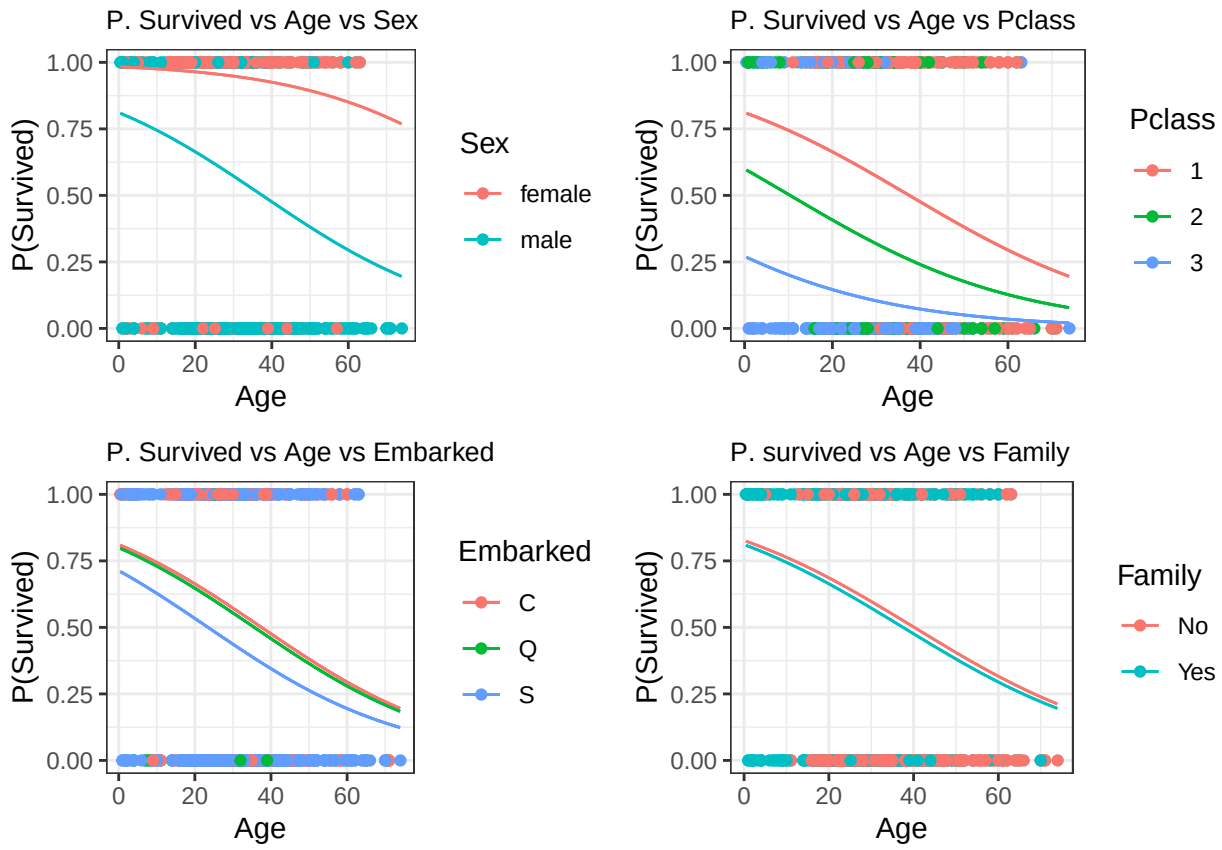
```
# Calculate the data to plot the 'C' curve.
emb_q <- as.factor(rep(x = "Q", length(ages)))
emb_s <- as.factor(rep(x = "S", length(ages)))
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_m,
                                          Age = ages,
                                          Family = fam,
                                          Embarked = emb),
                        type = "response")

data_embC <- data.frame(Age = ages,
                        Embarked = emb,
                        Survived = predictions)


# Calculate the data to plot the 'Q' curve.
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_m,
                                          Age = ages,
                                          Family = fam,
                                          Embarked = emb_q),
                        type = "response")

data_embQ <- data.frame(Age = ages,
                        Embarked = emb_q,
                        Survived = predictions)
```

```r
# Calculate the data to plot the 'S' curve.
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_m,
                                          Age = ages,
                                          Family = fam,
                                          Embarked = emb_s),
                        type = "response")

data_embS <- data.frame(Age = ages,
                        Embarked = emb_s,
                        Survived = predictions)

# Bind the data for the curves created.
data <- rbind(data_embC, data_embQ, data_embS)

# Create the graphic of the predictions between embarked places.
emb_plot <-  ggplot(data = d_tr_f,
             aes(x = Age, y = as.numeric(as.character(Survived)),
             color = Embarked)) +
             geom_point() +
             geom_line(data = data, aes(y = Survived)) +
             geom_line(data = data, aes(y = Survived)) +
             geom_line(data = data, aes(y = Survived)) +
             theme_bw() +
             labs(title = "P. Survived vs Age vs Embarked",
                  y = "P(Survived)") +
             theme(plot.title = element_text(size = 10))
```

```r
# Calculate the data to plot the passengers with family curve.
fam_no <- as.factor(rep(x = "No", length(ages)))
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_m,
                                          Age = ages,
                                          Family = fam,
                                          Embarked = emb),
                        type = "response")

data_fam <- data.frame(Age = ages,
                       Family = fam,
                       Survived = predictions)


# Calculate the data to plot the passengers without family curve.
predictions <- predict(object = model1,
                       newdata=data.frame(Pclass = pclass,
                                          Sex = sex_m,
                                          Age = ages,
                                          Family = fam_no,
                                          Embarked = emb),
                        type = "response")
```

```r
data_fam_no <- data.frame(Age = ages,
                          Family = fam_no,
                          Survived = predictions)


# Bind the data for the curves created.
data <- rbind(data_fam, data_fam_no)

# Create the graphic of the predictions between passengers with family and not.
family_plot <- ggplot(data = d_tr_f,
               aes(x = Age, y = as.numeric(as.character(Survived)),
               color = Family)) +
               geom_point() +
               geom_line(data = data, aes(y = Survived)) +
               geom_line(data = data, aes(y = Survived)) +
               theme_bw() +
               labs(title = "P. survived vs Age vs Family",
                    y = "P(Survived)") +
               theme(plot.title = element_text(size = 10))
```

```r
# Plot all 4 graphics of the model1 created together
figure <- ggarrange(sex_plot, pclass_plot, emb_plot, family_plot, ncol = 2, nrow = 2)
figure
```



Afegint informació als tests d'hipòtesi anteriors, es grafiquen aquestes variables segons l'edat i es conclou que:

- Totes les variables resulten sensibles a l'edat, de tal manera que en tots els casos, quan l'edat augmenta

la probabilitat de sobreviure disminueix.

- El fet de que el passatger sigui dona o home resulta significatiu a l'hora de predir si el passatger sobreviurà o no. D'aquesta manera, es confirma que si el passatger és una dona té més probabilitats de sobreviure. A més, l'edat de la dona no resulta tan sensible com l'edat de l'home.
- La classe en la que els passatgers viatjen també resulta sensible a l'edat, tot i que, els que viatgen en primera classe tenen més probabilitats de sobreviure que els de segona i tercera. De la mateixa manera que els que viatjen en segona classe tenen més probabilitat de sobreviure que els de tercera.
- En referència al port d'embarcament es pot confirmar que no existeix cap diferència en la probabilitat de sobreviure entre embarcar al port de Cherbourg i Queenstown. Tot i que els passatgers que embarquen al port de Southampton sí que experimenten una probabilitat lleugerament inferior.
- Finalment, tal i com s'havia observat a la regressió logística no es troba cap diferència significatvia entre tenir família o no, ja que ambdues línies segueixen la mateixa tendència.

Prediccions sobre les dades de test

Integració i selecció de les dades d'interès

A continuació es procedirà realitzar tots els canvis necessaris per treballar amb l'arxiu test de les dades amb l'objectiu de predir si els passatgers sobreviuran o no.

```
# Create the 'Family' variable for the test data.
d_tt$Family <- d_tt$Pclass
n <- 1
while (n <= length(d_tt$Family)) {
  if(d_tt$SibSp[n]==0 && d_tt$Parch[n]==0){
    d_tt$Family[n] <- 'No'
  }else{
    d_tt$Family[n] <- 'Yes'
  }
  n <- n+1
}
d_tt$Family <- as.factor(d_tt$Family)
head(d_tt$Family,n=20)
```

```
##  [1] No  Yes No  No  Yes No  No  Yes No  Yes No  No  Yes Yes Yes Yes No  No  Yes
## [20] No
## Levels: No Yes
```

```
# Convert 'Pclass'  as a factor.
d_tt$Pclass <- as.factor(d_tt$Pclass)
head(d_tt$Pclass,n=20)
```

```
##  [1] 3 3 2 3 3 3 3 2 3 3 3 1 1 2 1 2 2 3 3 3
## Levels: 1 2 3
```

Es seleccionen les variables d'interés per realitzar la predicció.

```
# Select the data needed from test data to use the model1.
d_tt_f <- d_tt[c('Pclass','Sex','Age','Family','Embarked')]
head(d_tt_f, n=5)
```

```
##   Pclass    Sex  Age Family Embarked
## 1      3   male 34.5     No        Q
## 2      3 female 47.0    Yes        S
## 3      2   male 62.0     No        Q
```

```
## 4      3   male 27.0     No        S
## 5      3 female 22.0    Yes        S
```

Neteja de les dades

```
# Study basic statistics of test data.
summary(d_tt_f)
```

```
##  Pclass      Sex           Age        Family    Embarked
##  1:107   female:152   Min.   : 0.17   No :253   C:102
##  2: 93   male  :266   1st Qu.:21.00   Yes:165   Q: 46
##  3:218                Median :27.00             S:270
##                       Mean   :30.27
##                       3rd Qu.:39.00
##                       Max.   :76.00
##                       NA's   :86
```

S'imputen els valors nuls del conjunt de dades per la mitjana en funció de la classe com s'ha fet anteriorment amb les dades d'entrenament

```
# Imputation of NA values from 'Age' with the mean for the corresponding 'Pclass'.
n <- 1
while (n <= length(d_tt_f$Age)) {
  if(is.na(d_tt_f$Age[n])){
    if(d_tt_f$Pclass[n] == 1){
      d_tt_f$Age[n] <- summary(d_tt_f$Age[d_tt_f$Pclass == 1])['Mean']
    }else if(d_tt_f$Pclass[n] == 2){
      d_tt_f$Age[n] <- summary(d_tt_f$Age[d_tt_f$Pclass == 2])['Mean']
    }else if(d_tt_f$Pclass[n] == 3){
      d_tt_f$Age[n] <- summary(d_tt_f$Age[d_tt_f$Pclass == 3])['Mean']
    }
  }
  n <- n+1
}
summary(d_tt_f$Age)['NA\'s']
```

```
## <NA>
##
```

Una vegada realitzats els canvis es procedeix a computar els resultats.

Predicció de supervivència

A continuació es procedirà a realitzar les prediccions del conjunt de dades de test, utilitzant el model obtingut. D'aquesta manera, si el resultat és superior o igual a una probabilitat de 0.5 s'interpretarà com que el passatger sobreviu, en cas contrari, s'interpretarà que el passatger no sobreviu.

```
# Predict the survived status for each passenger in the test data.
predictions <- predict(object = model1,
                       newdata=d_tt_f,
                        type = "response")

# Convert the predictions in categories by the rule of separeting in 0.5.
n <- 1
while (n <= length(predictions)) {
  if(predictions[n] >= 0.5){
```

```
    predictions[n] = 1
  }else{
    predictions[n] = 0
  }
  n <- n+1
}
head(predictions, n=20)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  0  0  0  0  1  0  1  0  1  0  0  0  1  0  1  1  0  0  1  0
```

Es comproba que per cada línia s'extreu un valor 0 o 1 corresponent a si el passatger determinat no sobreviu (0) o sobreviu (1).

```
# Create a dataset with the id and the survive status for each passenger in the test data.
d_result <- data.frame(PassengerId = d_tt$PassengerId, Survived = predictions)
head(d_result, n=10)
```

```
##    PassengerId Survived
## 1          892        0
## 2          893        0
## 3          894        0
## 4          895        0
## 5          896        1
## 6          897        0
## 7          898        1
## 8          899        0
## 9          900        1
## 10         901        0
```

Exportació dels resultats

Un cop obtinguts els resultat d'efectuar prediccions sobre les dades de test aplicant el model 1, es procedeix a importar aquestes dades a un arxiu csv.

```
# Export the result of the prediction in a csv file.
write.csv(d_result, "test_result.csv")
```

Conclusions

Arribats a aquest punt, on l'estudi sobre el conjunt de dades amb informació dels passatgers del Titànic ja està finalitzat, es poden enumerar un seguit de conclusions. Aquestes es mostren a continuació:

- Aquest treball es treballa amb dades sobre els passatgers del Titànic per osbervar si hi ha alguna relació entre els passatgers supervivents i no supervivents.
- El conjunt de dades importat inicialment amb dades sobre els passatgers del Titànic contenia un total de 12 variables (PassengerId, Survival, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked), i només se n'han seleccionat 6 (Survival, Pclass, Sex, Age, Family, Embarked) d'aquestes, les que s'han considerat que contenien informació rellevant per observar si hi ha relació entre passatgers supervivents. La variable 'Family' és resultat d'unir la informació que contenen les variables 'SibSp' i 'Parch', sent una variable categòrica amb dos únics valors possibles 'Yes' o 'No', per si el passatger viatge sol o en família (pares, fills, germans o parella).
- En el procés de neteja de dades s'ha detectat un gran nombre de valors perduts en la variable 'Age', els quals s'han imputat per la mitjana de la resta de passatgers segons a la classe de viatge en cada

cas, intentant així mantenir el màxim de dades possibles per l'estudi sense alterar els resultats que se'n puguin extreure d'aquestes. També s'han detectat cert valors extrems en edats altes i s'ha eliminat tots els registres amb passatgers d'edat superior a 75 anys, procurant evitar un esbiax en els resultats per edats altes. Les dades seleccionades, s'han exportat un cop netejades.

- Sobre les dades seleccionades s'hi han aplicat diferents mètodes d'anàlisis per observar l'efecte de cada variable sobre el fet de si un passatger sobreviu o no. Aquests mètodes han estat: contrast d'hipòtesis sobre la mitjana poblacional en dues mostres independents (i de variàncies diferents), contrast d'hipòtesis sobre la proporció en dues mostres independents, i un model de regressió logística.
- Els resultats obtinguts en els diferents contrastos d'hipòtesis aplicats han mostrat que totes les variables seleccionades (Pclass, Sex, Age, Family, Embarked) són explicatives a l'hora de determinar la supervivència d'un passatger. Remarcar que per l'edat s'ha obtingut un resultat molt ajustat a l'hora de determinar si l'edat mitjana dels passatgers supervivents és diferent dels que no, però tot i això aquest resultat mostra que sí que hi ha diferència entre les mitjanes. I per al lloc d'embarcament, no s'ha observat efecte per determinar la supervivència entre passatgers que ho han fet a Queenstown i Southampton.
- En quant al model de regressió logística obtingut, s'han utilitzat totes les variables explicatives seleccionades, sent aquest el model que mostra un millor nivell d'ajust.
- Estudiant aquest model s'ha observat que el fet de que un passatger viatgi en familia o no, o bé si el lloc d'embarcament d'aquest és Queenstown o Cherbourg, no es consideren explicatives per determinar la probabilitat de supervivència. Per a la resta de casos, incloent que el passatger embarqui a Cherbourg o a Southampton, es consideren explicatives.
- Observant els *odds ratio* del model, destaca la variable referent al sexe del passatger, sent la que més efecte té sobre la probabilitat de supervivència. El fet de que un passatger sigui home respecte a que sigui dona, la probabilitat de supervivència disminueix en un 0,073.
- La precisió d'aquest model de regressió logística té una precissió d'un 79,66%.
- S'han realitzat diferents representacions gràfiques del model, prenent sempre l'edat en l'eix X i la probabilitat de supervivència en l'eix Y, per tal d'observar l'evolució de cada variable prenent els diferents valors possibles. D'aquestes representacions es pot extreure el següent:
  - Totes les variables resulten sensibles a l'edat, de tal manera que en tots els casos, quan l'edat augmenta la probabilitat de sobreviure disminueix.
  - El fet de que el passatger sigui dona o home resulta significatiu a l'hora de predir si el passatger sobreviurà o no. D'aquesta manera, es confirma que si el passatger és una dona té més probabilitats de sobreviure. A més, l'edat de la dona no resulta tan sensible com l'edat de l'home.
  - La classe en la que els passatgers viatjen també resulta sensible a l'edat, tot i que, els que viatgen en primera classe tenen més probabilitats de sobreviure que els de segona i tercera. De la mateixa manera que els que viatjen en segona classe tenen més probabilitat de sobreviure que els de tercera.
  - En referència al port d'embarcament es pot confirmar que no existeix cap diferència en la probabilitat de sobreviure entre embarcar al port de Cherbourg i Queenstown. Tot i que els passatgers que embarquen al port de Southampton sí que experimenten una probabilitat lleugerament inferior.
  - Finalment, tal i com s'havia observat a la regressió logística no es troba cap diferència significatvia entre tenir família o no, ja que ambdues línies segueixen la mateixa tendència.
- Per acabar, s'utilitza el model obtingut per predir quin passatger sobreviurà o no en les dades de test sobre passatgers del Titànic sense la variable de supervivència. Per això s'ha determinat que un passatger sobreviu si la seva probabilitat és igual o major a 0,5, i que no sobreviu en cas contrari. Aleshores, amb els resultats d'aquestes prediccions, s'han extret les dades amb l'identificador del passatger i la variable que indica si sobreviu o no.

Participació dels compoenents del grup

Contribuciones	Firma
Investigació prèvia	CPG i EPR
Redacció de les respostes	CPG i EPR
Desenvolupament codi	CPG i EPR