Pràctica 2

Aitor Hernández i Anna Mayoral

Juny 2022

---

1. Descripció del dataset.

Per la realització d'aquesta pràctica, treballarem amb el dataset Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).

El dataset conté un registre per cada passatger que viatjava en el Titanic el dia del seu enfonsament. L'estudi és important de cara a determinar quina afectació van tenir els factors socio-econòmics dels passatgers en la seva supervivència en el malaurat accident.

Les dades s'han dividit en dos grups:

- conjunt d'entrenament (train.csv)
- conjunt de prova (test.csv)

Per la realització d'aquesta pràctica només treballarem amb el dataset train.csv ja que és el que conté la variable Survived, que ens indica si el passatger en qüestió va sobreviure o no a l'accident.

1.1 Revisió de les dades, extracció visual d'informació i preparació de les dades

Inicialitzarem les llibreries que utilitzarem durant la realització de la pràctica

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(nortest)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(rpart)
library(rpart.plot)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.0.5
```

```
library(caret)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.5
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5

## Loading required package: carData
```

El primer que realitzarem és la càrrega de les dades:

```
# Lectura de les dades
data_titanic<-read.csv("./train.csv",stringsAsFactors = FALSE,header=T,sep=",")
```

Farem una visualització de les primeres files per comprovar que les dades s'han carregat correctament:

```
# Visualització primeres files
head (data_titanic)
```

```
##   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
```

Efectivament les dades s'han carregat correctament.

Començarem fent una breu anàlisi de les dades ja que ens interessa tenir una idea general de les dades que disposem. i verifiquem l'estructura del joc de dades principal.

```
structure = str(data_titanic)
```

```
## '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       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Veiem que tenim **12** variables i **891** registres

Revisem la descripció de les variables contingudes al fitxer i els tipus de variables carregades. Les organitzem lògicament per donar-los sentit i construïm un petit diccionari de dades utilitzant la documentació auxiliar.

- **PassengerId** identificador únic del passatger.

FETS A ESTUDIAR

- **Survived** és la variable objectiu del nostre anàlisis. Indica la supervivència o no a l'accident (0 = No, 1 = Si).
- **Pclass** indica la clase en la que viatjava el passatger (1 = 1st, 2 = 2nd, 3 = 3rd).
- **Name** nom del passatger.
- **Sex** gènere dels passatger.
- **Age** edat del passatger (observem números decimals, que farà referencia als mesos).
- **SibSp** nombre de germans/cònjuges a bord del Titanic.
- **Parch** nombre de pares/fills a bord del Titanic.
- **Ticket** identificador del ticket.
- **Fare** preu del ticket.
- **Cabin** número del camarot.
- **Embarked** Port d'embarcament del passatger. (C = Cherbourg, Q = Queenstown, S = Southampton)

---

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

La gran majoria d'atributs presents en el conjunt de dades son necessaris per la realització de l'anàlisis ja que ens aporten informació de les característiques que tenien les persones que van sobreviure a l'accident del titanic i les que no van sobreviure. Tot i això podem excloure de l'anàlisis els camps Name, Ticket i Cabin que no ens aportaran informació de valor per l'anàlisis.

```r
# Eliminem les columnes Name, Ticket i Cabin
data_titanic <- data_titanic[, -c(4,9, 11 )]
```

També convertirem els atributs categòrics a factors

```r
# Convertim les variables categòriques a factor
data_titanic$Survived <- as.factor(data_titanic$Survived)
data_titanic$Pclass <- as.factor(data_titanic$Pclass)
data_titanic$Sex <- as.factor(data_titanic$Sex)
data_titanic$Embarked <- as.factor(data_titanic$Embarked)
```

Revisem la trasnformació així com l'eliminació de les variables (Name, Ticket i Cabin)

```r
structure = str(data_titanic)
```

```
## 'data.frame':    891 obs. of  9 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ 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 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Veiem que ara tenim **9** variables i **891** registres

---

3. Neteja de les dades.

En aquesta apartat començarem el processament de les dades per tal d'aconseguir un dataset preparat per inicialitzar l'anàlisis. És de gran interès saber si tenim molts valors nuls (camps buits) i la distribució de valors per variables.

El primer que farem serà estudiar els valors nulls o buits i posteriorment analitzarem els valors extrems d'algunes variables.

3.1 Anàlisis valors buits

Mostrarem per a cada atribut la quantitat de valors perduts de la següent forma:

```r
# Observem alhora si hi ha valors na o buits al dataset
as.list(colSums(is.na(data_titanic) | data_titanic == ""))
```

```
## $PassengerId
## [1] 0
##
## $Survived
## [1] 0
##
## $Pclass
## [1] 0
##
## $Sex
## [1] 0
##
## $Age
## [1] 177
##
## $SibSp
## [1] 0
##
## $Parch
## [1] 0
##
## $Fare
## [1] 0
##
## $Embarked
## [1] 2
```

Observem fàcilment que hi ha valors missing i, per tant, haurem de preparar les dades en aquest sentit. El camps amb valors perduts son Age (n=177) i Embarked (n=2).

Per tal de completar les dades farem servir dos mètodes diferents:

1) Per els valors na de la categoria edat, assignarem la mitjana de les edats:

```
# substituim els valors que falten per la mitjana de "Edat"
data_titanic[is.na(data_titanic$Age)==TRUE, "Age"] <- mean(data_titanic$Age, na.rm = TRUE)
# Eliminem els decimals de l'edat
data_titanic$Age = as.integer(data_titanic$Age)
# Comprovem com ara no hi han valors na
table(is.na(data_titanic$Age))
```

```
##
## FALSE
##   891
```

2) Pel que fa a la variable "Embarked" farem un anàlisis de freqüència per conèixer si podem pre assignar un port, sinó podem, haurem d'eliminar les dues files.

Com no volem eliminar dades, completarem aquests dos registres amb la moda estadística de la variable.

```
# Taula de freqüències de la variable Embarked
table(data_titanic$Embarked)
```

```
##
##       C   Q   S
##   2 168  77 644
```

Com son únicament dos valors, podem assumir que el port és S (Southampton), que és la variable més freqüent.

```
# Reemplacem els valors buits per S (Southampton)
data_titanic[data_titanic$Embarked=='', "Embarked"] <- 'S'
# Taula de freqüències de la variable Embarked
table(data_titanic$Embarked)
```

```
##
##       C   Q   S
##   0 168  77 646
```

Efectivament, ara ja no tenim valors buits en la variable Embarked.

```
missing <- data_titanic[is.na(data_titanic),]
dim(missing)
```

```
## [1] 0 9
```

Disposem per tant d'un data frame format per 9 variables sense valors nuls.

3.2 Identifica i gestiona els valors extrems.

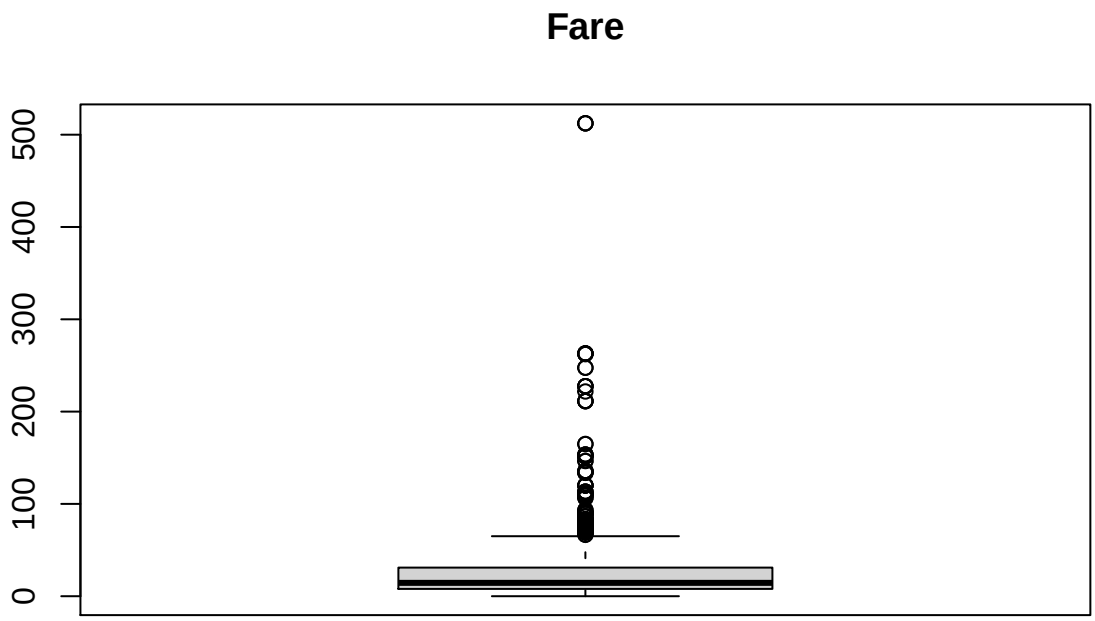
Per identificar valors extrems, primer farem un resum estadístic de les variables amb la funció summary

```
summary(data_titanic)
```

```
##   PassengerId    Survived Pclass      Sex           Age            SibSp
##  Min.   :  1.0   0:549    1:216   female:314   Min.   : 0.00   Min.   :0.000
##  1st Qu.:223.5   1:342    2:184   male  :577   1st Qu.:22.00   1st Qu.:0.000
##  Median :446.0            3:491                Median :29.00   Median :0.000
##  Mean   :446.0                                 Mean   :29.54   Mean   :0.523
##  3rd Qu.:668.5                                 3rd Qu.:35.00   3rd Qu.:1.000
##  Max.   :891.0                                 Max.   :80.00   Max.   :8.000
##      Parch            Fare        Embarked
##  Min.   :0.0000   Min.   :  0.00    :  0
##  1st Qu.:0.0000   1st Qu.:  7.91   C:168
##  Median :0.0000   Median : 14.45   Q: 77
##  Mean   :0.3816   Mean   : 32.20   S:646
##  3rd Qu.:0.0000   3rd Qu.: 31.00
##  Max.   :6.0000   Max.   :512.33
```

La variable Fare, crida l'atenció el valor màxim i per tant farem una visualització gràfica mitjançant un diagrama de caixes.

```r
# Diagrama de caixa Fare
boxplot(data_titanic$Fare, main = "Fare")
```



Tal i com podem observar molt visualment, efectivament existeixen valors molt alts en la variable Fare tot i així podrien ser paquets luxosos del vaixells, per tant el que mirarem és de si es tracta d'un únic valor o quina proporció correspon.

```r
# Agafem els valors majors al tercer quartil
length(data_titanic$Fare[data_titanic$Fare > 31])
```

```
## [1] 222
```

Com son forces valors, entenem que aquests preus corresponent a paquets luxosos del vaixell i per tant la dada es considera correcte.

---

4. Anàlisi de les dades.

Per tenir un major coneixement de les dades, utilitzarem les eines de visualització per començar a veure relacions entre les variables del dataset.

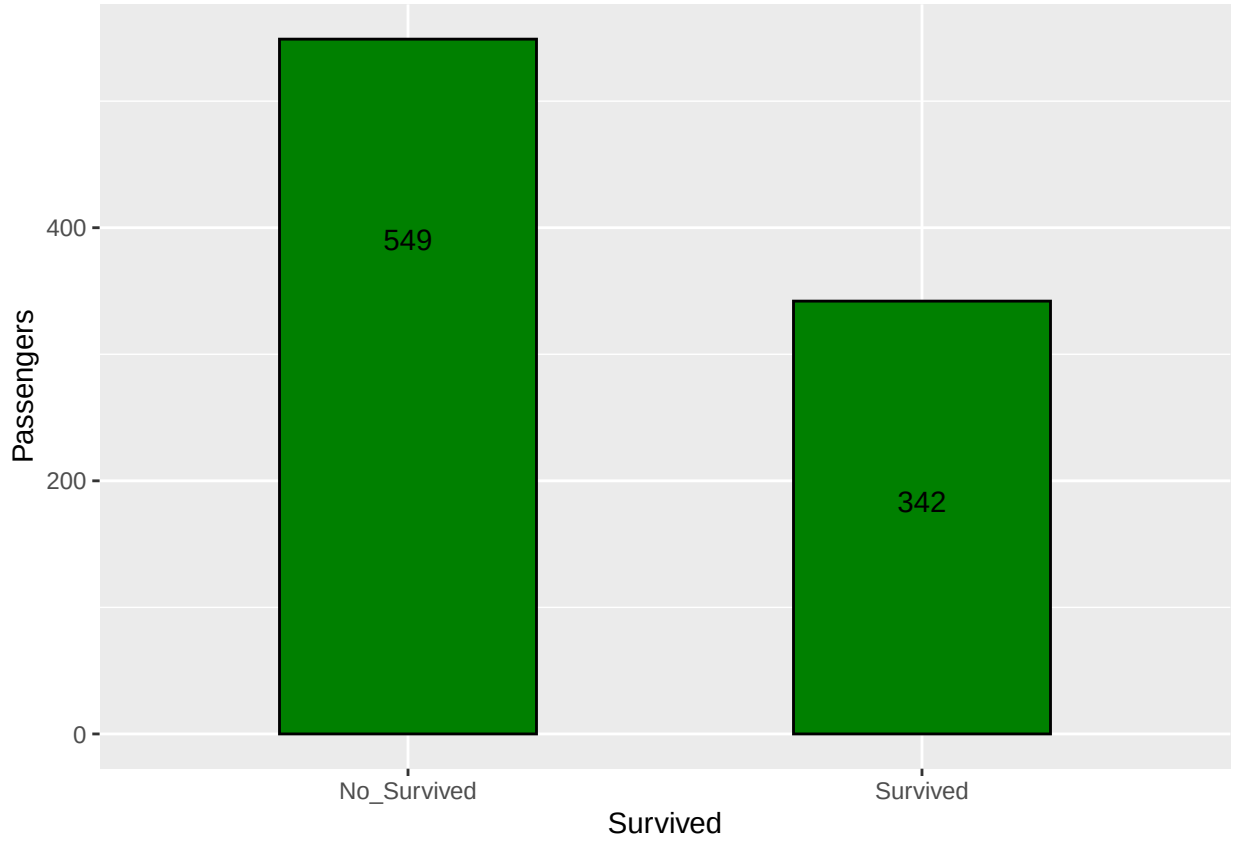
4.1 Selecció dels grups de dades que es volen analitzar/comparar

Com que el que ens interessa és descriure la relació entre la supervivència en l'accident i les variables realitzarem una sèrie de diagrames de barres i de taules de contingència que ens proporcionaran molta informació.

Per a això, d'una banda graficarem mitjançant diagrames de barres la quantitat de morts i supervivents segons la classe en la qual viatjaven, l'edat o el sexe. D'altra banda, per a obtenir les dades que estem graficant utilitzarem la comanda table per a les dues variables que ens proporciona una taula de contingència.

El primer que analitzarem és quanta gent va sobreviure

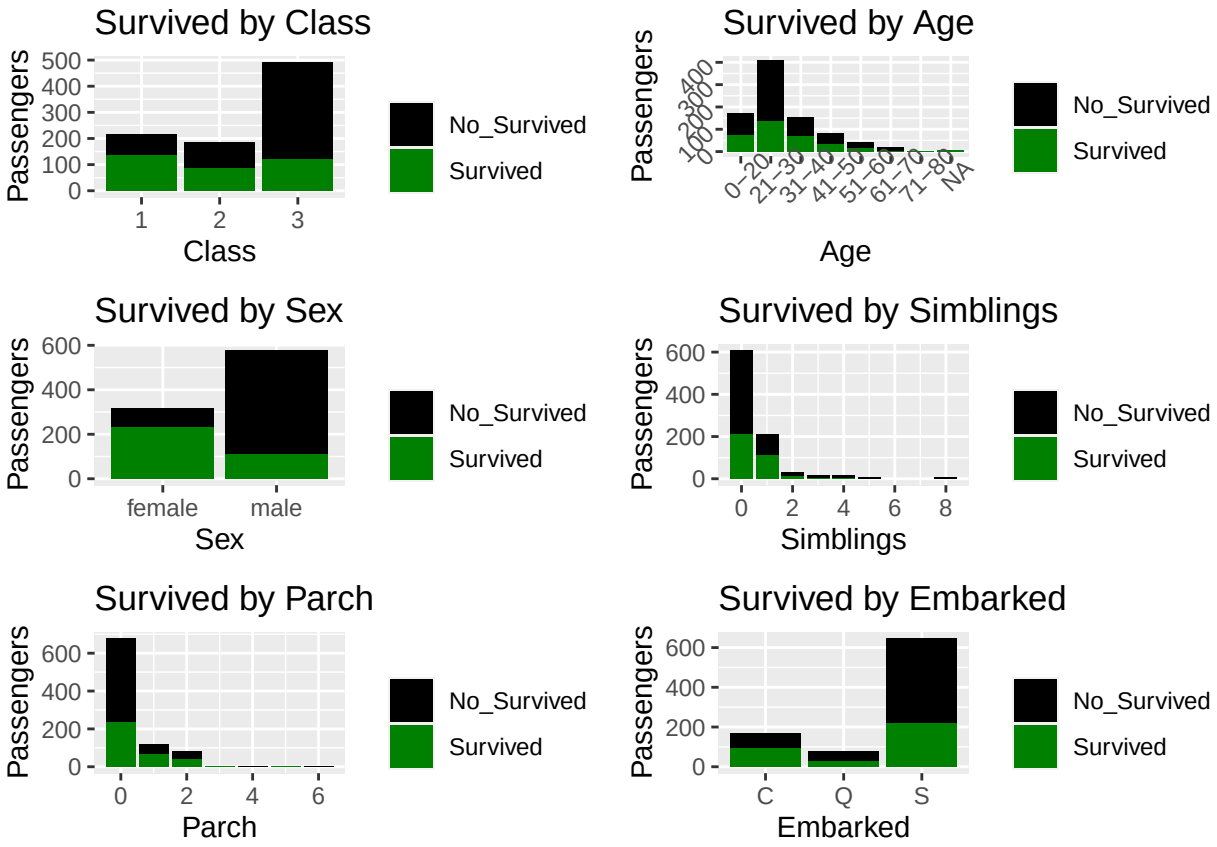
```
# Discretitzem variable Survived i crearem una nova Supervivència
data_titanic["Supervivencia"] <- factor(ifelse(data_titanic$Survived == 0 , "No_Survived", "Survived"))
# Analitzem graficament el número de persones que van o no sobreviure
ggplot(data_titanic, aes(x = Supervivencia)) +
    geom_bar(width = 0.5, fill = "#008000", color = "#000000") +
    geom_text(stat = "count", aes(label = stat(count)), vjust = 10) +
    labs(x="Survived", y="Passengers")
```



En aquest gràfic podem observar com 549 persones NO van sobreviure mentre que 342 van sobreviure a l'accident del Titanic.

```
# Discretitzem la variable age i creem una nova variable discreta edat
data_titanic["edat"] <- cut(data_titanic$Age, breaks = c(0, 20,30,40,50,60,70,80,100), labels = c("0-20
```

```
# Supervivència per classe
q1 <- ggplot(data_titanic,aes(Pclass,fill=Supervivencia))+geom_bar() +labs(x="Class", y="Passengers")+ g

# Supervivència per edat
q2 <- ggplot(data_titanic,aes(edat,fill=Supervivencia))+geom_bar() +labs(x="Age", y="Passengers")+ guid

# Supervivència per gènere
q3 <- ggplot(data_titanic,aes(Sex,fill=Supervivencia))+geom_bar() +labs(x="Sex", y="Passengers")+ guides

# Supervivència per unitat familiar
q4 <- ggplot(data_titanic,aes(SibSp,fill=Supervivencia))+geom_bar() +labs(x="Simblings", y="Passengers"

# Supervivència per unitat familiar
q5 <- ggplot(data_titanic,aes(Parch,fill=Supervivencia))+geom_bar() +labs(x="Parch", y="Passengers")+ gu

# Supervivència per Embarked
q6 <- ggplot(data_titanic,aes(Embarked,fill=Supervivencia))+geom_bar() +labs(x="Embarked", y="Passenger

grid.arrange(q1, q2, q3, q4, q5, q6, ncol = 2, nrow = 3)
```



A continuació treballarem amb taules contingència per enriquir l'anàlisis

```
tabla_SST <- table(data_titanic$Sex, data_titanic$Supervivencia)
tabla_SST
```

```
##
```

```
##          No_Survived Survived
##   female          81      233
##   male           468      109
```

```
prop.table(tabla_SST, margin = 1)
```

```
##
##          No_Survived  Survived
##   female   0.2579618 0.7420382
##   male     0.8110919 0.1889081
```

D'aquests gràfics obtenim informació molt valuosa que complementem amb les taules de contingència. Per exemple si tenim en compte el percentatge de supervivència respecte al seu sexe observem com la taxa de mort en homes és molt major (el 81,1% dels homes van morir mentre que en dones aquest percentatge baixa a 25,8%).

```
tabla_SCT <- table(data_titanic$Pclass,data_titanic$Supervivencia)
tabla_SCT
```

```
##
##     No_Survived Survived
##   1          80      136
##   2          97       87
##   3         372      119
```

```
prop.table(tabla_SCT, margin = 1)
```

```
##
##     No_Survived  Survived
##   1   0.3703704 0.6296296
##   2   0.5271739 0.4728261
##   3   0.7576375 0.2423625
```

Referent a la classe en la qual viatjaven, els passatgers que viatjaven en primera classe van ser els únics que el percentatge de supervivència era major que el de mortalitat. El 62,96% dels viatgers de primera classe va sobreviure, el 47,2% dels quals viatjaven en segona classe mentre que dels viatgers de tercera només van sobreviure un 24,23%.

```
tabla_SAT <- table(data_titanic$edat,data_titanic$Supervivencia)
tabla_SAT
```

```
##
##         No_Survived Survived
##   0-20           98       75
##   21-30         272      136
##   31-40          86       69
##   41-50          51       33
##   51-60          25       17
##   61-70          14        4
##   71-80           3        1
##   +80             0        0
```

```r
prop.table(tabla_SAT, margin = 1)
```

```
##
##         No_Survived  Survived
##   0-20    0.5664740 0.4335260
##   21-30   0.6666667 0.3333333
##   31-40   0.5548387 0.4451613
##   41-50   0.6071429 0.3928571
##   51-60   0.5952381 0.4047619
##   61-70   0.7777778 0.2222222
##   71-80   0.7500000 0.2500000
##   +80
```

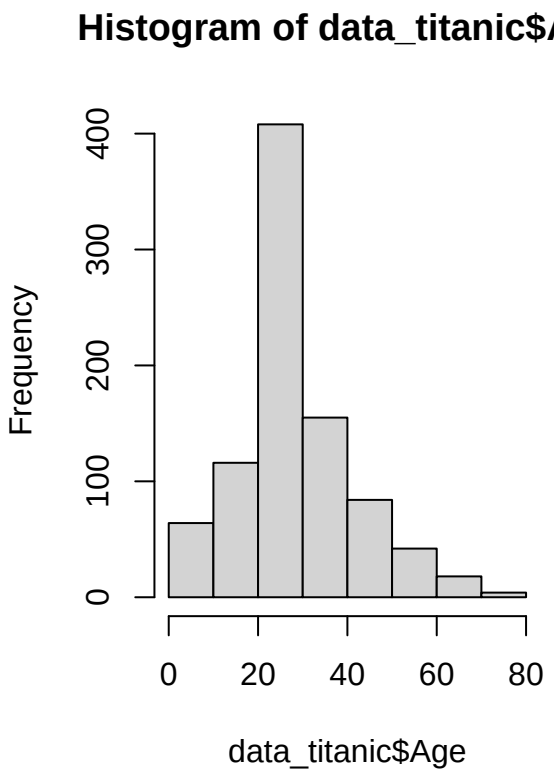
Per a finalitzar, destaquem que la presència de passatgers majors de 20 anys era molt major que la gent més jove i que la taxa de supervivència en els menors de 20 anys és la major (45,81%)

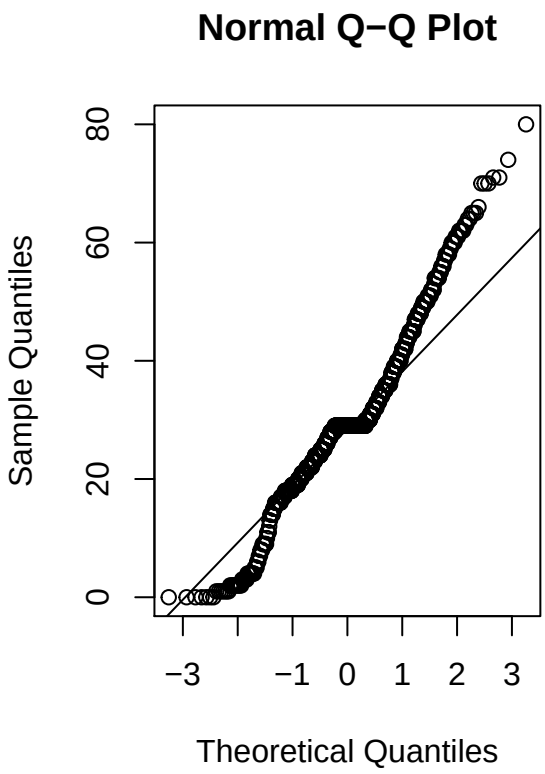
4.2 Comprovació de la normalitat i homogeneïtat de la variància.

Normalitat

El primer que realitzarem serà comprovar la normalitat de les variables numèriques

```r
# Gràfics
par(mfrow=c(1,2))
hist(data_titanic$Age) # histograma de l'edat
qqnorm(data_titanic$Age) # gràfic quantile
qqline(data_titanic$Age)
```

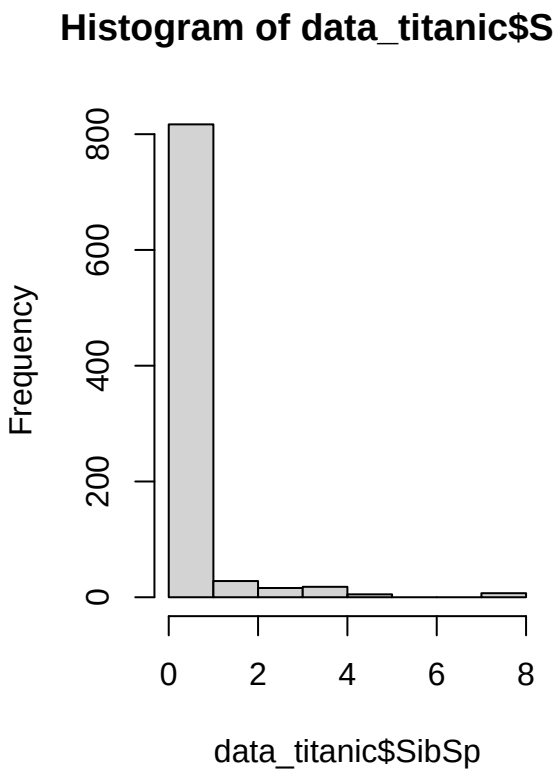


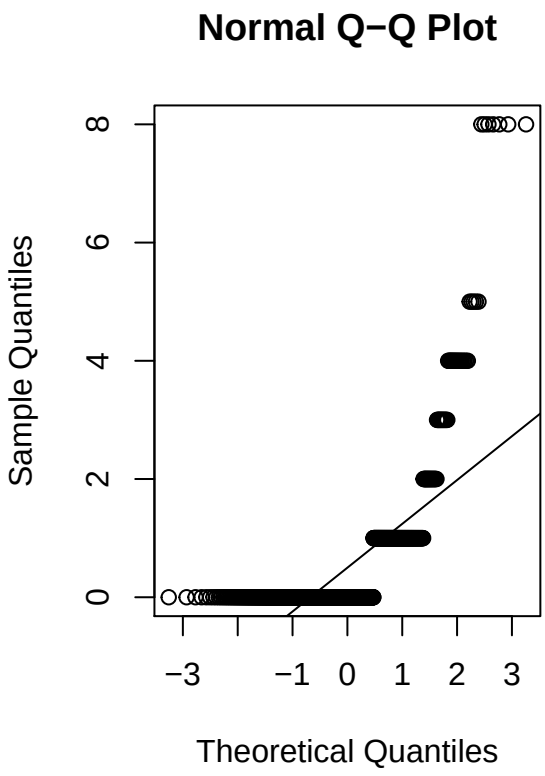



```
# Contrast de normalitat
lillie.test(data_titanic$Age) #contrast
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_titanic$Age
## D = 0.14631, p-value < 2.2e-16
```

```
# Gràfics
par(mfrow=c(1,2))
hist(data_titanic$SibSp) # histograma de Sib
qqnorm(data_titanic$SibSp) # gràfic quantile
qqline(data_titanic$SibSp)
```

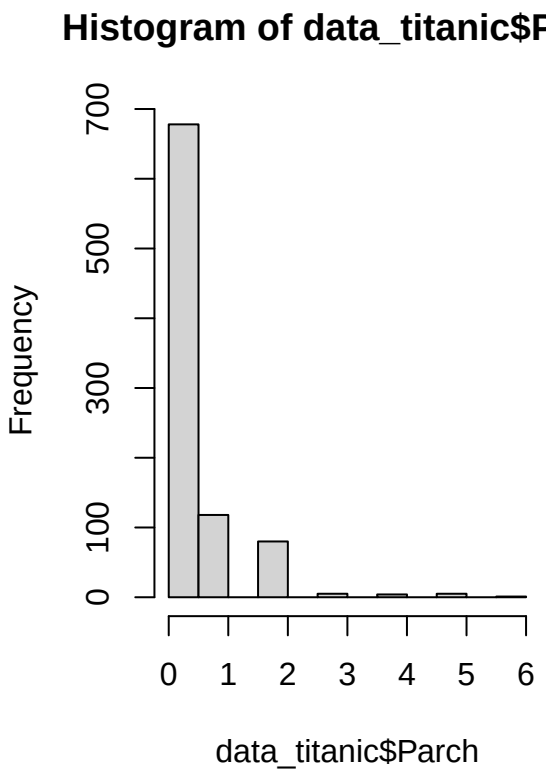


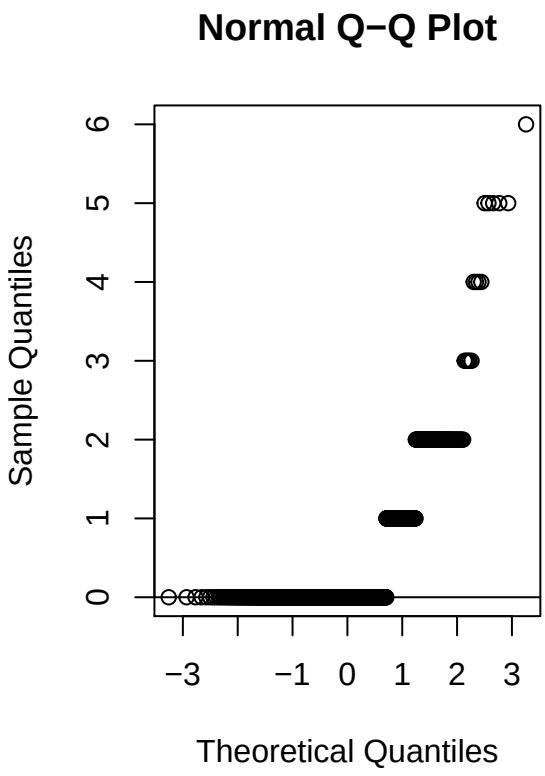



```
# Contrast normalitat
lillie.test(data_titanic$SibSp) #contrast
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_titanic$SibSp
## D = 0.36473, p-value < 2.2e-16
```

```
# Gràfics
par(mfrow=c(1,2))
hist(data_titanic$Parch) # histograma de Parch
qqnorm(data_titanic$Parch) # gràfic quantile
qqline(data_titanic$Parch)
```

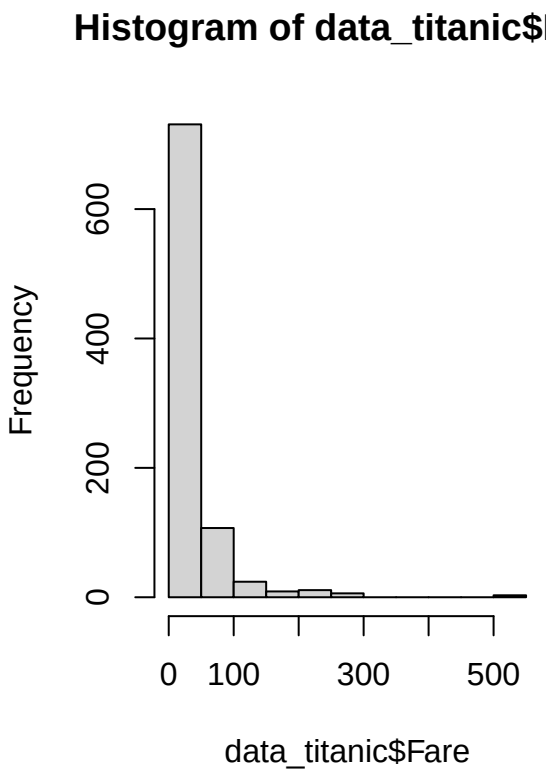


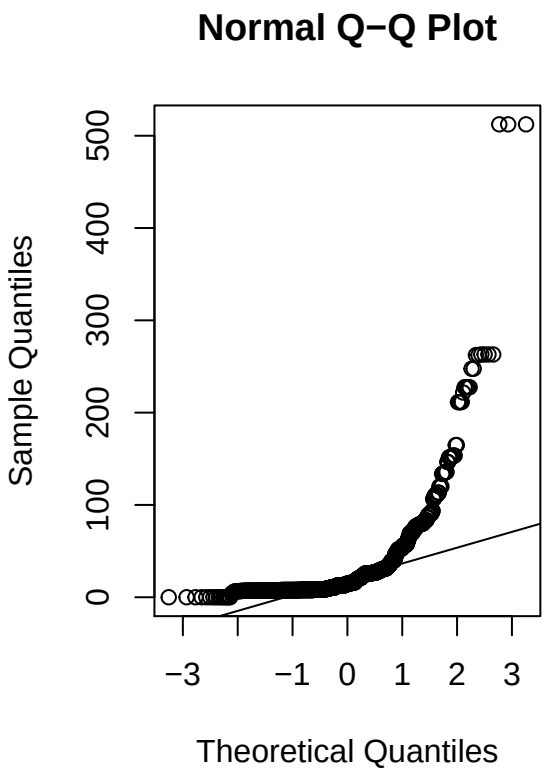



```
# Contrast normalitat
lillie.test(data_titanic$Parch) #contrast
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_titanic$Parch
## D = 0.44298, p-value < 2.2e-16
```

```
# Gràfics
par(mfrow=c(1,2))
hist(data_titanic$Fare) # histograma de Fare
qqnorm(data_titanic$Fare) # gràfic quantile
qqline(data_titanic$Fare)
```






```
# Contrast normalitat
lillie.test(data_titanic$Fare) #contrast
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_titanic$Fare
## D = 0.28185, p-value < 2.2e-16
```

Si usem un nivell de significança de alpha = 0.05, podem veure que en tots els tests anteriors es rebutja la hipòtesi nul · la amb un nivell del confiança del 95%, ja que p_valor < 0.05 en tots els casos. Per tant, podem dir que les variables quantitatives d'aquest conjunt de dades (Age, SibSp, Parch i Fare) no segueixen una distribució normal. Tot i això veiem que la variable Age es troba a prop de la normalitat, per tant, com tenim una quantitat d'observacions prou gran podem assumir que aquesta variable segueix una distribució normal basant-nos en el teorema central del límit.

Homogeneïtat de variances

A continuació comprovarem la homogeneïtat per la variables Age, com hem suposat normalitat, utilitzarem el test de Levene.

```
# Variable Age
leveneTest(Age ~ Pclass, data = data_titanic)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  17.341 4.096e-08 ***
##       888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Age ~ Sex, data = data_titanic)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.8795 0.3486
##       889
```

```
leveneTest(Age ~ Embarked, data = data_titanic)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  9.3296 9.777e-05 ***
##       888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Age ~ Survived, data = data_titanic)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  5.5002 0.01923 *
##       889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

S'observa com la variable Age presenta heteroscedasticitat amb Pclass, Embarked i Survived, i homoscedasticitat amb la variable Sex.

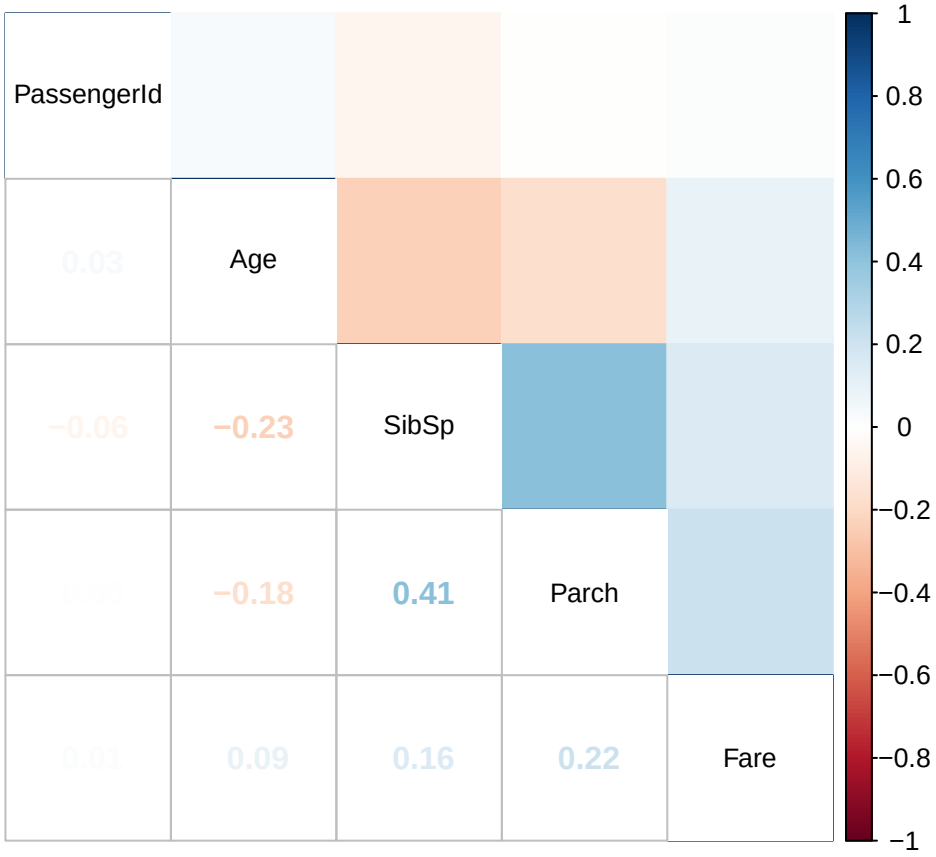
4.3 Aplicació de proves estadístiques per comparar els grups de dades

4.3.1 Matriu de correlacions

Aplicarem en primer lloc la matriu de correlacions entre les variables numèriques.

```
# Creem la correlació de variables numèriques
data_numeric = data_titanic[, sapply(data_titanic, is.numeric)]
M = cor(data_numeric, use="pairwise.complete.obs")
```

```
# Printem la matriu de correlacions
corrplot.mixed(M, lower = "number", upper="color", tl.pos = "d", tl.col = "black", tl.cex=0.85)
```



4.3.2 Model de regresió logístic

La regresió logística ens permet predit el resultat d'una variable categòrica en base a les variables independients o predictores. En el nostre cas, generarem un model de regresió logística, i analitzarem els paràmetres del mateix, així com les variables que tenen un pes més gran en la seva predicció.

Muntem el model

```
# Model de regresió logística
classifier = glm(Survived ~ Pclass + Age + SibSp + Parch + Fare, data = data_titanic, family = "binomial
summary(classifier)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age + SibSp + Parch + Fare,
##     family = "binomial", data = data_titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2717  -0.8287  -0.6771   1.0206   2.2789
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.548602   0.359716   4.305 1.67e-05 ***
## Pclass2     -0.647267   0.261671  -2.474  0.01338 *
```

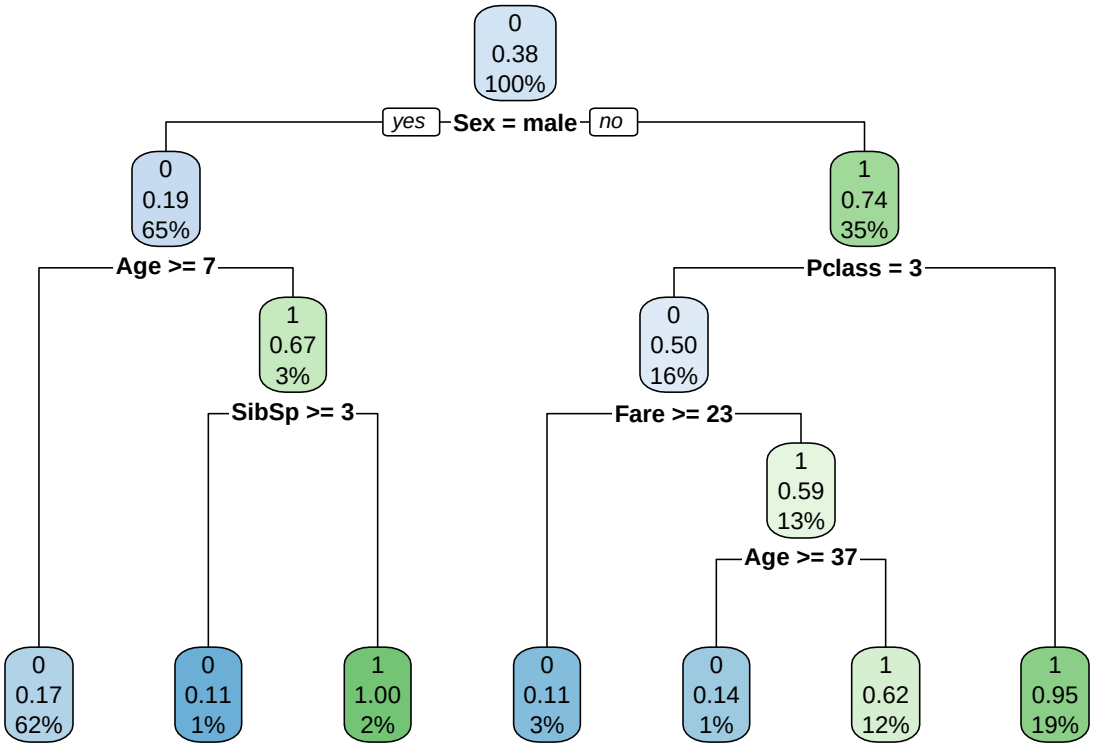
```
## Pclass3     -1.766255   0.262123  -6.738 1.60e-11 ***
## Age         -0.037384   0.006678  -5.598 2.17e-08 ***
## SibSp       -0.237867   0.088944  -2.674  0.00749 **
## Parch        0.209010   0.103699   2.016  0.04385 *
## Fare         0.005593   0.002673   2.093  0.03637 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1031.8  on 884  degrees of freedom
## AIC: 1045.8
##
## Number of Fisher Scoring iterations: 4
```

Veiem que amb el model de regresió logarítmic, la variable PClass és la que té un valor PValue més petit, seguit de la Edat i el número de fills. Per tant, segons el model son les variables que tenen un pes més significatiu en la variable objectiu 'survived'

4.3.3 Arbre de decisió

Per a crear un arbre de decisió, utilitzarem la funció rpart()

```
# Separem en train i test
partition <- createDataPartition(y = data_titanic$Survived,
                                 p = 0.7, list = F)
train_data <- data_titanic[partition, ]
test_data <- data_titanic[-partition, ]
# creem el model d'arbre de decisió
fit <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, data=data_titanic, method=
rpart.plot(fit)
```



Com podem veure, les variables que s'han utilitzat per a construir el model son Sex, Age, PClass, SibSipb i Fare

Per últim, farem una predicció amb l'arbre que hem generat.

```
set.seed(1)
prediction <- predict(fit, newdata = test_data, type = "class")
confusionMatrix(prediction, as.factor(test_data$Survived), positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 152  30
##          1  12  72
##
##                Accuracy : 0.8421
##                  95% CI : (0.7926, 0.8838)
##     No Information Rate : 0.6165
##     P-Value [Acc > NIR] : 6.804e-16
##
##                   Kappa : 0.6545
##
##  Mcnemar's Test P-Value : 0.008712
##
##             Sensitivity : 0.7059
##             Specificity : 0.9268
```

```
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.8352
##              Prevalence : 0.3835
##          Detection Rate : 0.2707
##    Detection Prevalence : 0.3158
##       Balanced Accuracy : 0.8164
##
##        'Positive' Class : 1
##
```

Veiem que la precisió del model d'arbre de decisió és del 84.27%.

---

5. Conclusions

En base a l'estudi realitzat, s'han trobat certs resultats que poden ser d'interes:

- En primer lloc, mitjançant el model de regresió logístic hem vist que les variables categóriques que més influeixen a l'hora de determinar si el passatger va sobreviure o no son la classe que ocupava el passatger al vaixell i l'edat.
- En segon lloc, mitjançant el model d'arbres de decisió hem pogut veure que som capaços de determinar amb una precisió d'un 84.27% si el passatger sobreviu o no. El model utilitza el Sexe, la Classe i l'Edat com els factors més determinants. Per exemple, veiem que els varons de menys de 6.5 anys es van salvar en la seva gran majoria.

Encara que es podria fer un estudi molt més exhaustiu sobre les característiques dels supervivents, creiem que si que hem pogut contestar les preguntes que ens plantejavem a l'hora d'iniciar l'estudi.

---

6. Fitxer de sortida

```r
# export del fitxer d'analisis en format csv
write.csv(data_titanic, file = "titanic.csv", row.names = FALSE)
```

Contribucions	Firma
Investigació prèvia	AH, AM
Redacció de respostes	AH, AM
Desenvolupament codi	AH, AM