Pràctica 2: Neteja i anàlisi de les dades

Mireia Olivella i Gabriel Izquierdo

7 de maig de 2020

Índex

1 Descripció del dataset

El conjunt de dades escollit recull informació dels passatgers del titanic, en el que es pot analitzar la supervivència i les característiques d'aquests. Aquest conjunt de dades s'ha obtingut de la web de Kaggle. S'hi pot accedir a partir de l'enllaç que es mostra a continuació:

https://www.kaggle.com/c/titanic

El conjunt de dades utilitzat està format per 1309 registres amb 12 atributs dividit en 2 fitxers CSV, un de *train* i un de *test*, ja que aquest conjunt de dades està preparat per ser utilitzat per tasques de predicció. Els atributs d'aquest conjunt de dades són els següents:

- **passengerId**: identificador dels registres del dataset.
- **survived**: indica si el passatger va sobreviure (0=No, 1=Sí).
- **pclass**: indica la classe en la que viatjava el passatger (1=1a, 2=2a, 3=3a).
- **name**: nom del passatger.
- **sex**: gènere del passatger (*female* o *male*).
- **age**: edat del passatger.
- **sibsp**: número de germans i cònjugues a bord del Titànic.
- **parch**: número de pares i fills a bord del Titànic.
- **ticket**: número del bitllet
- **fare**: preu de compra del bitllet.
- **cabin**: número de cabina on viatjava el passatger.
- **embarked**: port on va embarcar el passatger (C=Cherbourg, Q=Queenstown, S=Southampton).

Aquest conjunt de dades és important perquè representa les dades d'un dels naufragis més infames de la història. A més, ens permet abastir tots els aspectes importants a tenir en compte a l'hora de dur a terme aquesta pràctica.

La pregunta que intenta respondre és la de quins són els factors que van afavorir a un passatger sobreviure al naufragi. Si bé hi havia un element de sort en la supervivència dels passatgers, sembla que alguns grups de persones tenien més probabilitats de sobreviure que d'altres.

2 Neteja de les dades

Abans de començar amb la neteja de les dades, procedim a realitzar les lectures dels fitxers en format CSV en el que es troben. El procediment és el de carregar la informació dels tres fitxers i unir-les posteriorment.

En la secció anterior s'ha parlat de dos fitxers de tipus CSV, i ara se n'ha parlat de tres. Això es deu a que al fitxer `test.csv` li falta un atribut respecte al fitxer `train.csv`, que és el de `Survived`. La informació referent a la supervivència dels passatgers del fitxer `test.csv` es troba en un altre fitxer anomenat `gender_submission.csv` que té només dues columnes: `PassengerId` i `Survived`.

El primer que fem és carregar la informació de tots els fitxers CSV. Després fem un *merge* de les dades de `test.csv` i `gender_submission.csv` utilitzant la funció `merge` amb l'atribut `PassengerId` com a clau comuna entre les dues taules. Per acabar s'uneixen totes les dades de *train* i *test* en un sol *dataframe* utilitzant la funció `rbind`.

```
# Lectura de les dades
titanic_train <- read.csv("../data/train.csv")
titanic_test <- read.csv("../data/test.csv")
titanic_gender_submission <- read.csv("../data/gender_submission.csv")
titanic_test <- merge(titanic_test, titanic_gender_submission, by="PassengerId")
titanic_data <- rbind(titanic_train, titanic_test)
head(titanic_data)
```

```
##   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
```

```
# Tipus de dada assignat a cada camp
sapply(titanic_data, function(x) class(x))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##   "integer"   "integer"   "integer"    "factor"    "factor"   "numeric"
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##   "integer"   "integer"    "factor"   "numeric"    "factor"    "factor"
```

Podem observar que els tipus de dades assignats automàticament per R a les nostres variables no s'acaben de correspondre amb el domini d'aquestes. Aquest és el cas dels atributs `Survived` i `Pclass`. R detecta que es tracta d'un *integer* quan en realitat es tracta d'un *factor*, pel que procedim a assignar-li el tipus que nosaltres volem.

```
# Canvi del tipus del camp 'Survived' i 'Pclass'
titanic_data$Survived <- factor(titanic_data$Survived)
titanic_data$Pclass <- factor(titanic_data$Pclass)
```

2.1 Selecció de les dades d'interès

Totes les variables que tenim en el dataset fan referència a característiques dels passatgers del titanic. Tot i això, podem precindir de les columnes *PassengerId*, *Name*, *Ticket* i *Cabin* ja que no aporten informació rellevant de cara a la pregunta que respon aquest conjunt de dades.

```
# Eliminació de les columes 'PassengerId', 'Name', 'Ticket' i 'Cabin'
titanic_data <- select(titanic_data, -c(PassengerId, Name, Ticket, Cabin))
summary(titanic_data)
```

```
##  Survived Pclass      Sex           Age            SibSp           Parch
##  0:815    1:323   female:466   Min.   : 0.17   Min.   :0.0000   Min.   :0.000
```

```
##  1:494    2:277   male  :843   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
##           3:709                Median :28.00   Median :0.0000   Median :0.000
##                                Mean   :29.88   Mean   :0.4989   Mean   :0.385
##                                3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##                                Max.   :80.00   Max.   :8.0000   Max.   :9.000
##                                NA's   :263
##       Fare          Embarked
##  Min.   :  0.000    :  2
##  1st Qu.:  7.896   C:270
##  Median : 14.454   Q:123
##  Mean   : 33.295   S:914
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
```

2.2 Dades amb elements buits (valors perduts)

Aquest conjunt de dades conté dades amb elements buits representats de dues maneres diferents: amb el valor NA (*Not Available*) i amb un espai en blanc, pel que es procedeix a comprovar quins camps contenen elements buits i en quina quantitat.

```
# Número de valors perduts per camp
colSums(is.na(titanic_data))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0      263        0        0        1        0
```

```
colSums(titanic_data == "")
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0       NA        0        0       NA        2
```

Com es pot observar tenim 2 valors en blanc a la variable *Embarked*, 1 valor NA a *Fare* i 263 valors NA a *Age*.

Primer de tot tractarem els valors en blanc de la variable *Embarked*. Ens basarem en utilitzar una mesura de tendència central i, al tractar-se d'una variable categòrica, utilitzarem la **moda**.

```
# Consulta de la moda de la variable 'Embarked'
mlv(titanic_data$Embarked, method = "mfv")
```

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

Com es pot observar, al ser *S* la moda prenem aquest valor per omplir als valors buits de la variable.

```
# Imputació dels valors buits de la variable 'Embarked'
titanic_data$Embarked[titanic_data$Embarked == ""] = "S"
```

Per tractar el valor perdut a la variable *Fare* s'utilitzarà la **mitjana**.

```
# Imputació dels valors buits de la variable 'Fare'
titanic_data[is.na(titanic_data$Fare),]$Fare <- mean(titanic_data$Fare, na.rm = TRUE)
```

Per acabar, per tractar els valors perduts de la variable *Age* s'utilitzarà la **mitjana**. Per dur a terme la obtenció d'aquesta mitjana, enlloc d'obtenir la mitjana de l'atribut *Age* sencer, tindrem en compte el gènere (*Sex*) i la classe en la que viatjava (*Pclass*). A la gràfica següent es pot observar la relació entre els atributs *Age* i *Pclass* per dones i per homes.

```
# Visualitzem la relació entre les variables 'Age' i 'Pclass'
par(mfrow = c(1,2))
female_people = titanic_data[titanic_data$Sex == "male",]
male_people = titanic_data[titanic_data$Sex == "female",]
boxplot(female_people$Age~female_people$Pclass, main = "Pclass by age (female)",
        xlab = "Pclass", ylab = "Age")
boxplot(male_people$Age~male_people$Pclass, main = "Pclass by age (male)",
        xlab = "Pclass", ylab = "Age")
```

Per tractar els valors perduts tindrem en compte la informació observada a la gràfica anterior. Per realitzar aquesta tasca s'ha creat una funció `AgeMean` per obtenir la mitjana d'edats de les dones i dels homes segons la classe, i després s'ha creat una altra funció assignar als passatgers que tenen l'edat en blanc la mitjana corresponent al seu gènere i a la classe en la que viatjava.

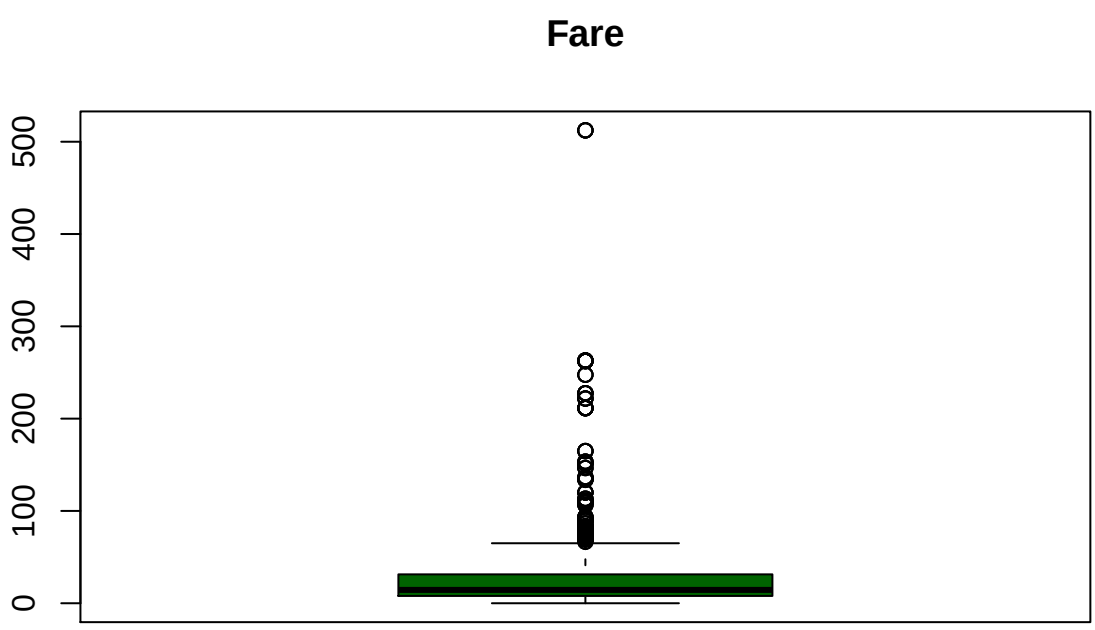
```
# Funció per obtenir el camp 'Mean' del resultat de la funció 'summary'
AgeMean <- function(age) {
  round(summary(age)['Mean'])
```

```
}

female_mean_ages = tapply(female_people$Age, female_people$Pclass, AgeMean)
male_mean_ages = tapply(male_people$Age, male_people$Pclass, AgeMean)

# Funció per obtenir un valor de mitjana d'edat segons els camps 'Sex' i 'Pclass'
AgeImpute <- function(row) {
  sex <- row['Sex']
  age <- row['Age']
  pclass <- row['Pclass']
  value <- age
  if (is.na(age)) {
    if (sex == "female") {
      value <- female_mean_ages[pclass]
    } else {
      value <- male_mean_ages[pclass]
    }
  }
  return(as.numeric(value))
}

titanic_data$Age <- apply(titanic_data[, c("Sex", "Age", "Pclass")], 1, AgeImpute)
```

2.3 Identificació de valors extrems

Els valors extrems o outliers són registres que destacant per ser molt distants al valor central del conjunt. Generalment es considera un outlier quan el seu valor es troba allunyat 3 desviacions estàndars respecte la mitjana, un instrument gràfic que ens permet visualitzar ràpidament aquests valors són els diagrames de caixes. Una altre forma de detectar-los a R, es mitjançant la funció `boxplot.stats()`.

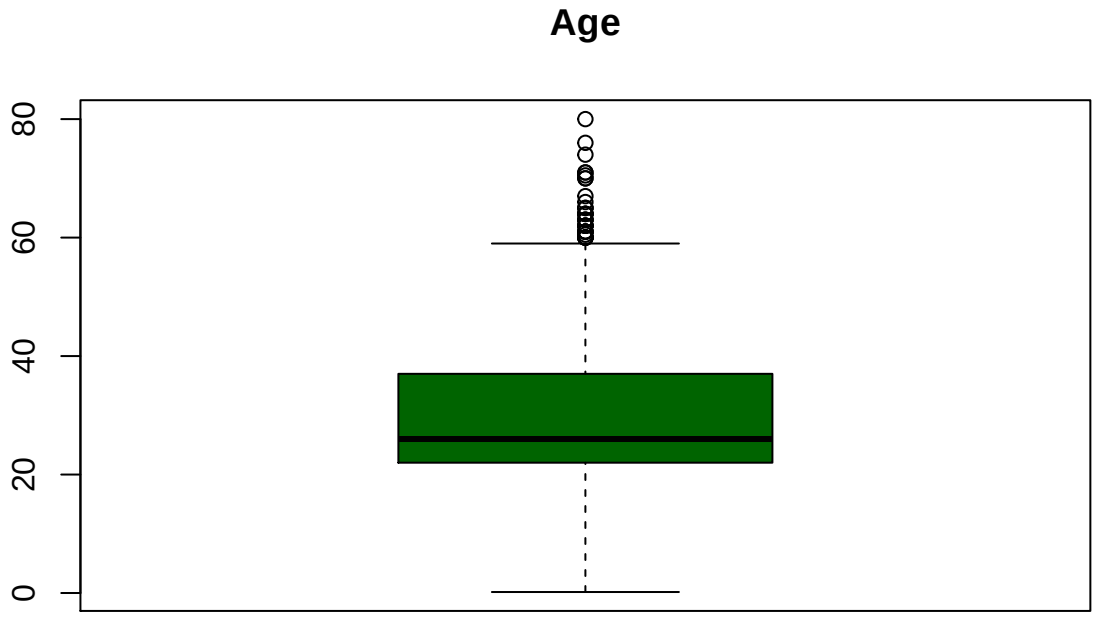
```
fare.bp<-boxplot(titanic_data$Fare, main="Fare", col="darkgreen")
```




```
boxplot.stats(titanic_data$Fare)$out
```

```
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
##   [9] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500
##  [17]  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000
##  [25]  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500
##  [33]  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750
##  [41] 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000
##  [49] 263.0000  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
##  [57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792
##  [65]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
##  [73]  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000
##  [81]  78.2667 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500
##  [89]  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250
##  [97] 211.3375 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
## [105] 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
## [113]  89.1042 164.8667  69.5500  83.1583  82.2667 262.3750  76.2917 263.0000
## [121] 262.3750 262.3750 263.0000 211.5000 211.5000 221.7792  78.8500 221.7792
## [129]  75.2417 151.5500 262.3750  83.1583 221.7792  83.1583  83.1583 247.5208
## [137]  69.5500 134.5000 227.5250  73.5000 164.8667 211.5000  71.2833  75.2500
## [145] 106.4250 134.5000 136.7792  75.2417 136.7792  82.2667  81.8583 151.5500
## [153]  93.5000 135.6333 146.5208 211.3375  79.2000  69.5500 512.3292  73.5000
## [161]  69.5500  69.5500 134.5000  81.8583 262.3750  93.5000  79.2000 164.8667
## [169] 211.5000  90.0000 108.9000
```

```
Age.bp<-boxplot(titanic_data$Age, main="Age", col="darkgreen")
```




```
boxplot.stats(titanic_data$Age)$out
```

```
##  [1] 66.0 65.0 71.0 70.5 61.0 62.0 63.0 65.0 61.0 60.0 64.0 65.0 63.0 71.0 64.0
## [16] 62.0 62.0 60.0 61.0 80.0 70.0 60.0 60.0 70.0 62.0 74.0 62.0 63.0 60.0 60.0
## [31] 67.0 76.0 63.0 61.0 60.5 64.0 61.0 60.0 64.0 64.0
```

Si ens fixem en els valors extrems resultants, en el cas d'Age, són valors que poden donar-se perfectament, ja que podem tenir persones de 80 anys com a passatgers.

En el cas de Fare, són valors que també es poden haver donat, ja que el preu que hagi pogut pagar cada passatger pot tenir una gran oscil · lació, i es poden donar valors de 0 a 500 perfectament.

Es per això, que tot i haver-los detectat, hem decidit no tractar-los de manera diferent a com han estat recollits.

2.4 Exportació de les dades preprocessades

Una vegada hem realitzat sobre el conjunt de dades original els procediments d'integració, validació i neteja de les dades, procedim a guardar aquestes en un nou fitxer anomenat `titanic_data_clean.csv`.

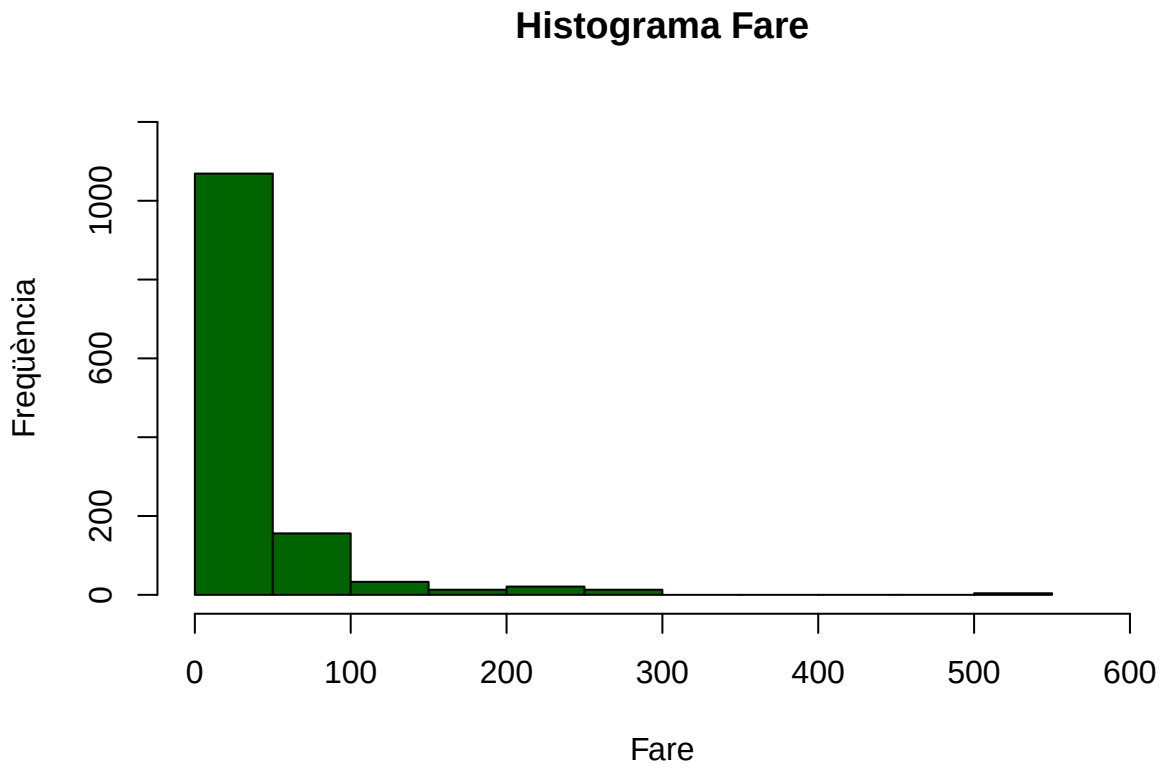
```
write.csv(titanic_data, "../data/titanic_data_clean.csv")
```

3 Anàlisi de les dades

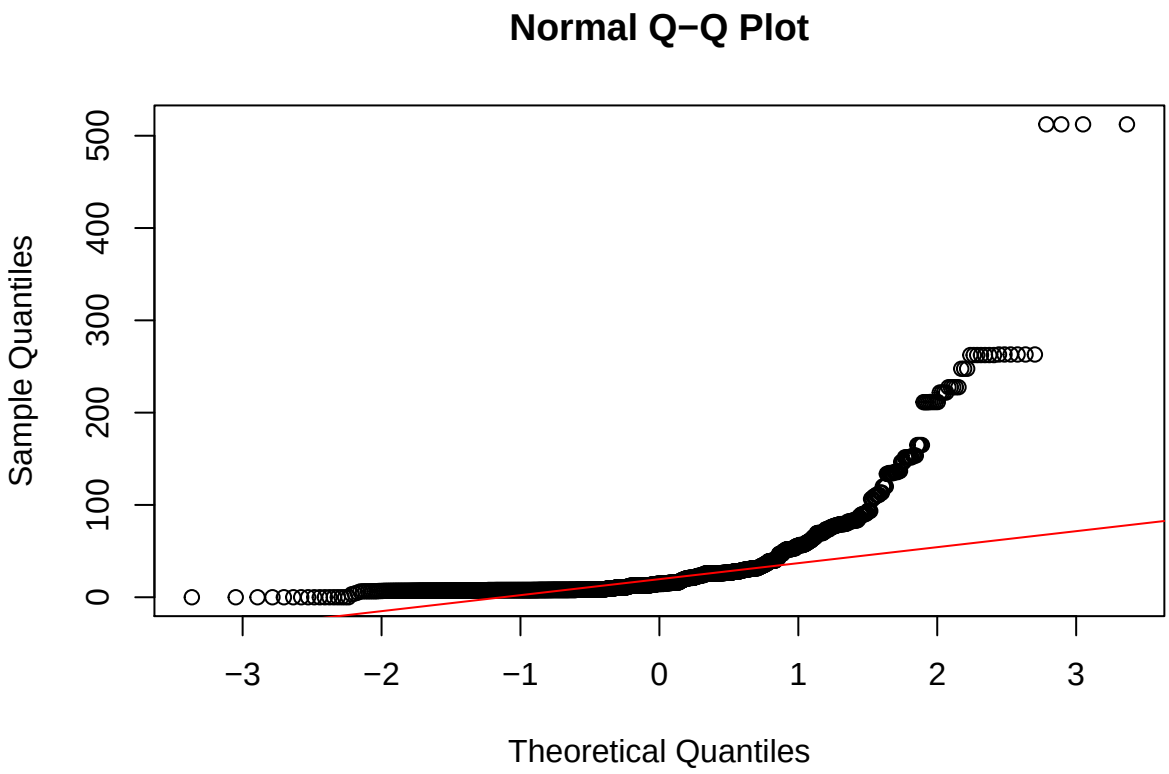
3.1 Comprovació de la normalitat i homogeneïtat de la variància

Per comprovar si les variables *Age* i *Fare* segueixen una distribució normal utilitzem la funció `qqnorm` per tenir una aproximació. Aquesta funció comparara els quartils de la distribució observada amb els quartils teòrics d'una distribució normal. Com més s'aproximen a les dades d'una normal, més alineats es trobaran els punts al voltant de la recta.

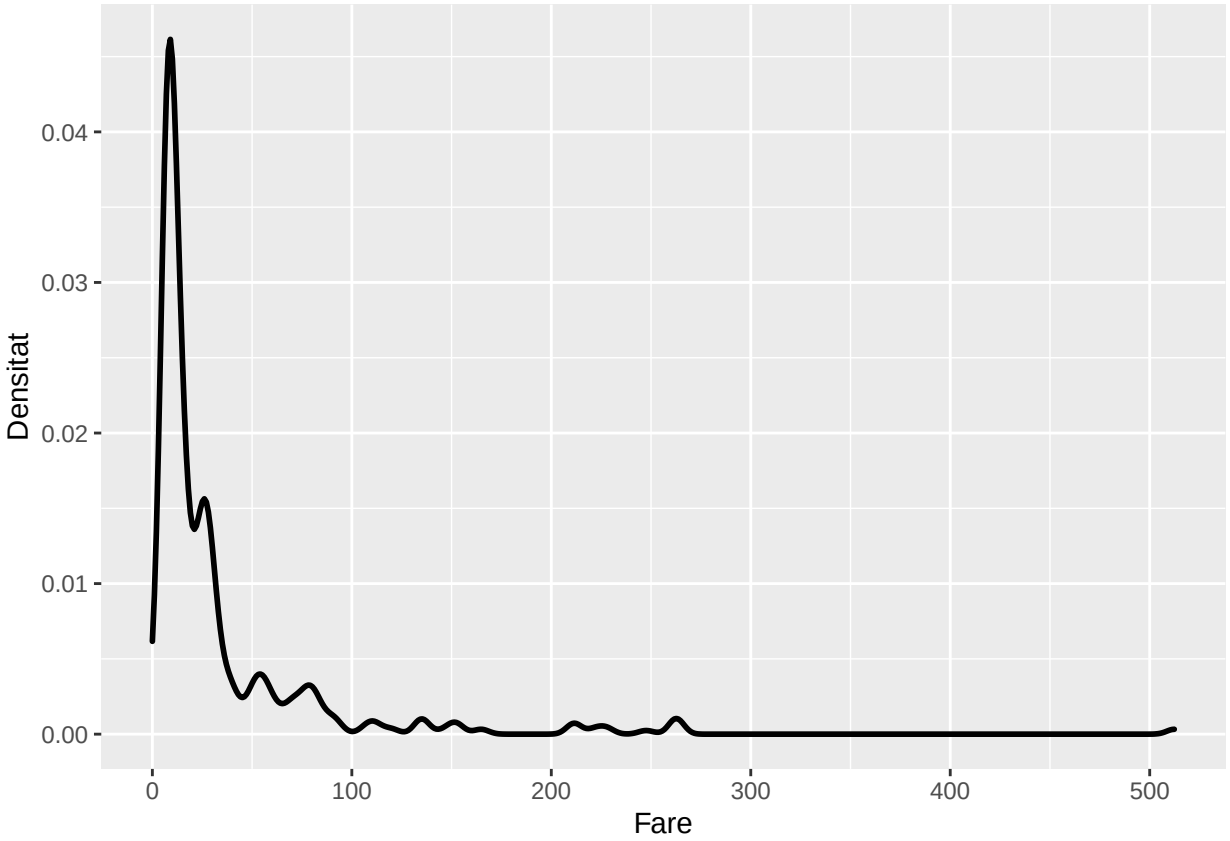
```
# Representació de la distribució de la variable 'Fare' mitjançant un histograma
hist(x = titanic_data$Fare, main = "Histograma Fare",
     xlab = "Fare", ylab = "Freqüència", col = "darkgreen",
     ylim = c(0,1200), xlim = c(0,600))
```




```
qqnorm(titanic_data$Fare)
qqline(titanic_data$Fare, col = "red")
```




```
ggplot(titanic_data, aes(Fare)) + geom_density(size = 1, alpha = 0.6) + ylab("Densitat")
```



Mitjançant els gràfics anteriors, podem veure que hi ha força desviació en alguns trams, i per tant, possibles evidències de que no segueix una distribució normal. Ho contrasterem mitjançant el Test *Lilliefors* (asumeix mediana i variança poblacionals desconegudes).

- Hipòtesis nul.la: les dades procedeixen d'una distribució normal.
- Hipòtesis alterantiva: no procedeixen d'una distribució normal.

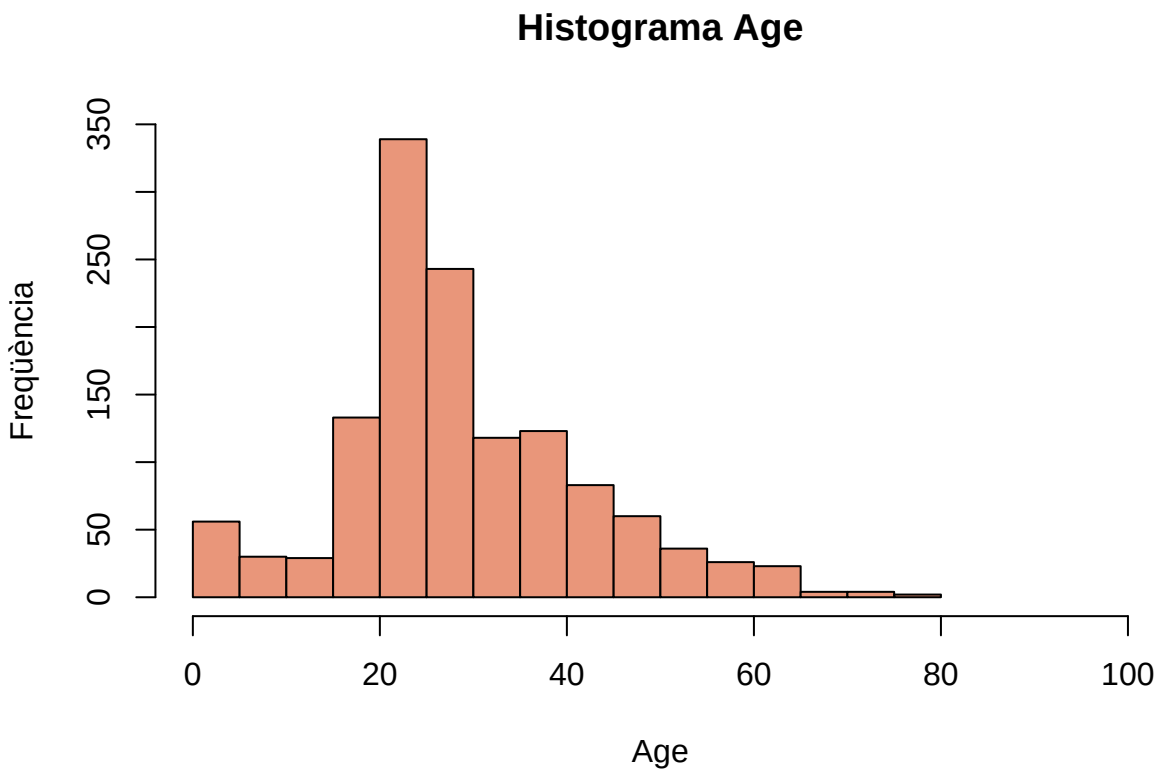
```
# Test Lilliefors
lillie.test(x = titanic_data$Fare)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  titanic_data$Fare
## D = 0.2858, p-value < 2.2e-16
```

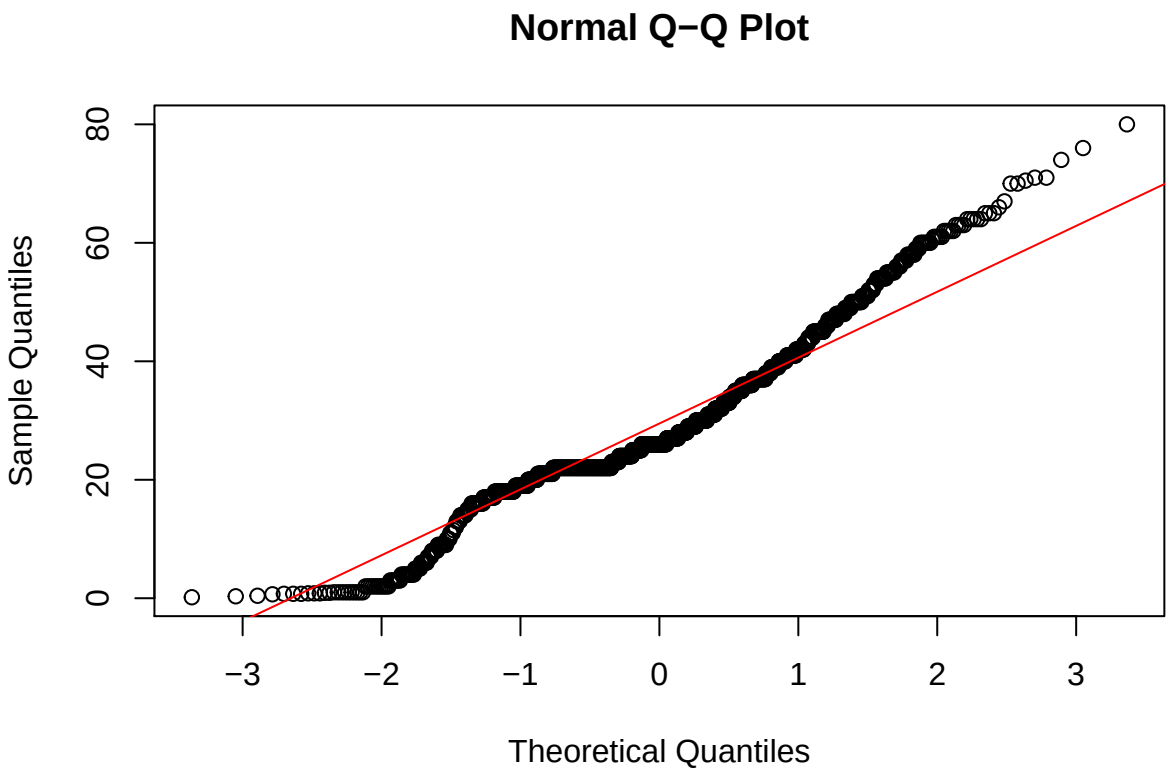
Rebutjem la hipòtesis nul · la perquè la diferència és estadísticament significativa, és a dir, amb un 95% de confiança podem dir que la variable *Fare* no segueix una distribució normal.

Repetim el mateix procediment per la variable *Age*:

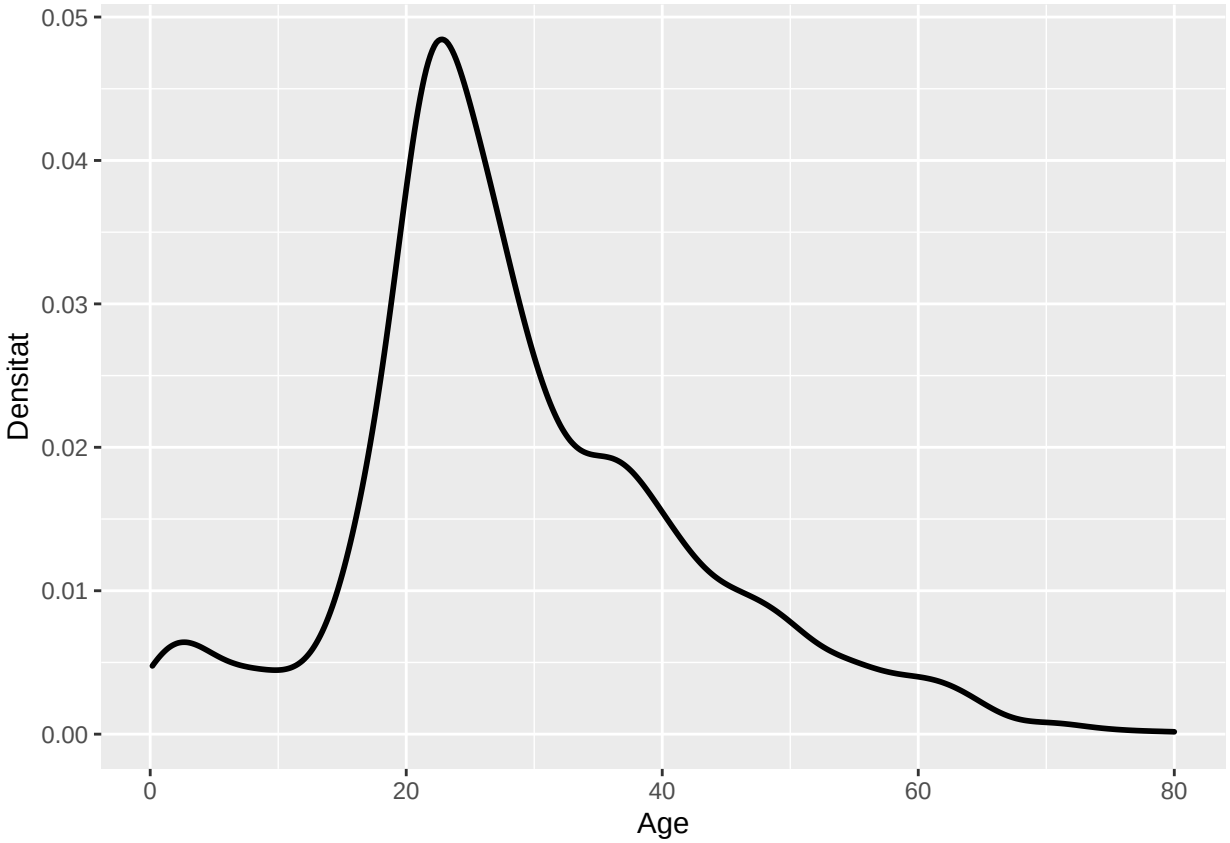
```
# Representació de la distribució de la variable 'Age' mitjançant un histograma
hist(x = titanic_data$Age, main = "Histograma Age",
     xlab = "Age", ylab = "Freqüència", col = "darksalmon",
     ylim = c(0,350), xlim = c(0,100))
```




```
qqnorm(titanic_data$Age)
qqline(titanic_data$Age, col = "red")
```




```
ggplot(titanic_data, aes(Age)) + geom_density(size = 1, alpha = 0.6) + ylab("Densitat")
```

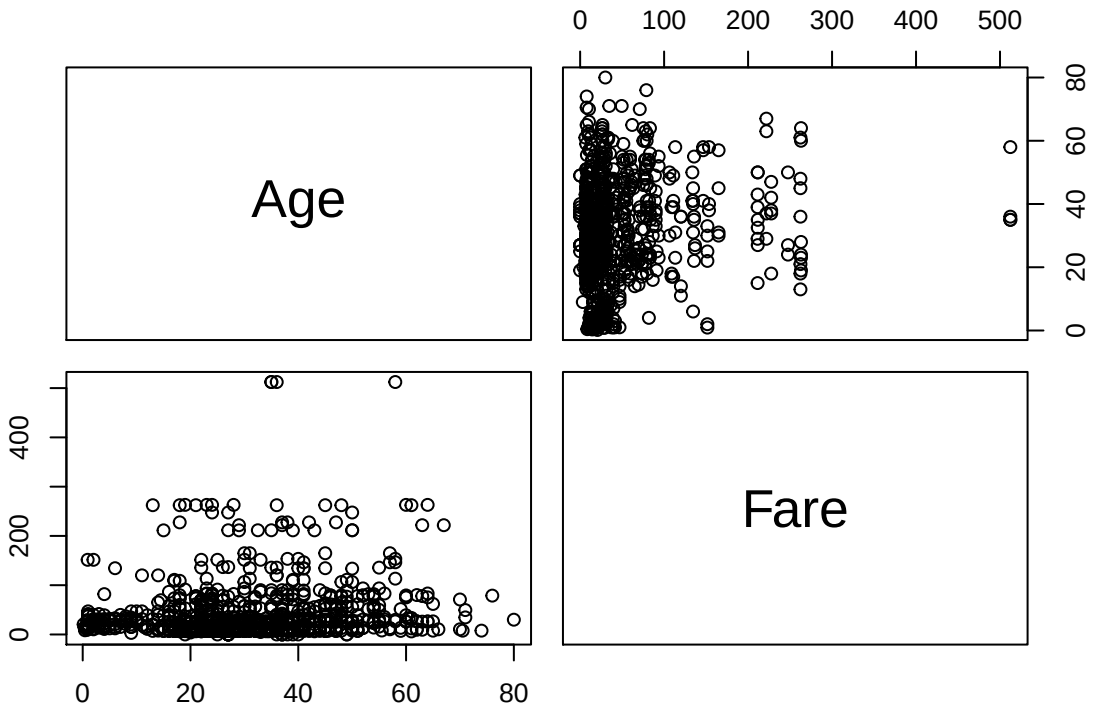



```
lillie.test(x = titanic_data$Age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  titanic_data$Age
## D = 0.11525, p-value < 2.2e-16
```

Podem veure de nou amb la variable *Age* com la diferència és estadísticament significativa. Per tant, rebutgem la hipòtesis nul · la i afirmem, amb un 95% de confiança, que la variable *Age* no segueix una distribució normal.

```
pairs(titanic_data[, c(4,7)])
```



Finalment comprovarem l'homoscedasticitat de les dades, és a dir, la igualtat de variàncies per *Fare* i *Age*.

Al no tenir seguretat que provinguin d'una població normal hem utilitzat el test de *Levene* amb la mediana com a mesura de centralitat, juntament amb el test no paramètric *Fligner-Killeen* que també es basa en la mediana.

- Hipòtesis nul · la: la variança és constant.
- Hipòtesis alterantiva: la variança no és constant.

```
aggregate(Fare~Survived, data = titanic_data, FUN = var)
```

```
##   Survived     Fare
## 1        0 1217.107
## 2        1 4705.192
```

```
aggregate(Age~Survived, data = titanic_data, FUN = var)
```

```
##   Survived      Age
## 1        0 161.2441
## 2        1 198.1852
```

```
# Test Levene
levene <- filter(.data = titanic_data, Survived %in% c("0", "1"))
leveneTest(y = levene$Fare, group = levene$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    1  56.493 1.043e-13 ***
##       1307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y = levene$Age, group = levene$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value  Pr(>F)
## group    1  5.1627 0.02324 *
##       1307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Test Fligner-Killeen
fligner.test(Fare ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 128.39, df = 1, p-value < 2.2e-16
```

```
fligner.test(Age ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 4.8298, df = 1, p-value = 0.02797
```

En els dos tests realitzats es pot observar que tant per la variable *Fare* com per la variable *Age* es rebutja la hipòtesis nul · la. Amb un nivell de confiança del 95% podem concloure que en ambdós grups la variança no és constant.

3.2 Aplicació de proves estadístiques

3.2.1 Contrast d'hipòtesis

Ens interessa descriure la relació entre la supervivència i les variables *Age*, *Pclass* i *Sex*.

En primer lloc hem dut a terme un gràfic mitjançant diagrames de barres amb la quantitat de morts i supervivents segons la classe en la que viatjaven, l'edat o el gènere.

```
plotByClass <- ggplot(titanic_data, aes(Pclass, fill = Survived)) + geom_bar() +
  labs(x = "Class", y = "Passengers") + guides(fill = guide_legend(title = "")) +
  scale_fill_manual(values = c("darksalmon", "darkseagreen4")) + ggtitle("Survived by Class")

plotByAge <- ggplot(titanic_data, aes(Age, fill = Survived)) + geom_bar() +
```

```
  labs(x = "Age", y = "Passengers") + guides(fill = guide_legend(title = "")) +
  scale_fill_manual(values = c("darksalmon", "darkseagreen4")) + ggtitle("Survived by Age")

plotBySex <- ggplot(titanic_data, aes(Sex, fill = Survived)) + geom_bar() +
  labs(x = "Sex", y = "Passengers") + guides(fill = guide_legend(title = "")) +
  scale_fill_manual(values = c("darksalmon", "darkseagreen4")) + ggtitle("Survived by Sex")

grid.arrange(plotByClass, plotByAge, plotBySex, ncol = 2)
```

Criteris d'èxit A continuació per tal de conèixer les característiques del passatgers i la possibilitat de supervivència hem realitzat una sèrie de test.

1. Van sobreviure més del 50% dels passatgers? Existeix diferència significativa per un nivell de significació del 5%?

Per poder contrastar la hipòtesis, utilitzarem el test binominal exacte.

- H0: la proporció és major del 50%.
- H1: la proporció no és major.

```
tableSurvived <- table(titanic_data$Survived)
prop.table(table(titanic_data$Survived))
```

```
##
##         0         1
## 0.6226127 0.3773873
```

```r
binom.test(x = c(494, 815), alternative = "less", conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  c(494, 815)
## number of successes = 494, number of trials = 1309, p-value < 2.2e-16
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.3999954
## sample estimates:
## probability of success
##              0.3773873
```

Amb un nivell de confiança del 95% podem concloure que no va sobreviure més del 50% dels passatgers.

2. Hi ha diferència significativa entre la proporció d'homes i dones que van sobreviure?

- H0: les dues variables són independents.
- H1: les dues variables no són independents.

Per esbrinar si hi ha diferència hem executat el test de *Fisher*, el qual ens permet estudiar si existeix asociació entre dues variables qualitatives.

```r
table_Sex <- table(titanic_data$Sex, titanic_data$Survived)
prop.table(table(titanic_data$Sex, titanic_data$Survived), margin = 1)
```

```
##
##                  0         1
##   female 0.1738197 0.8261803
##   male   0.8706999 0.1293001
```

```r
fisher.test(table_Sex, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_Sex
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02255471 0.04318701
## sample estimates:
## odds ratio
## 0.03139796
```

```r
# Si fem el test X² també és significatiu
chisq.test(x = table_Sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_Sex
## X-squared = 617.31, df = 1, p-value < 2.2e-16
```

```
chisq.test(x = table_Sex)$residuals
```

```
##
##                   0          1
##   female -12.278067  15.770495
##   male     9.128705 -11.725314
```

Amb un 95% de confiança podem rebutjar el test i, per tant, afirmar que les dues variables estan relacionades. Concretament s'esperava un 11.7% més d'homes que sobrevisques i un 15.8% menys de dones.

3. Hi ha diferències en la supervivencia segons la classe en la que viatjaven?

- H0: les dues variables són independents.
- H1: les dues variables no són independents.

```
table_Class <- table(titanic_data$Pclass, titanic_data$Survived)
prop.table(table(titanic_data$Pclass, titanic_data$Survived), margin = 1)
```

```
##
##             0         1
##   1 0.4241486 0.5758514
##   2 0.5776173 0.4223827
##   3 0.7306065 0.2693935
```

```
fisher.test(table_Class, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_Class
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
# Si fem el test X² també és significatiu
chisq.test(x = table_Class)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_Class
## X-squared = 91.724, df = 2, p-value < 2.2e-16
```

```r
chisq.test(x = table_Class)$residuals
```

```
##
##              0          1
##   1 -4.5203721  5.8061669
##   2 -0.9490707  1.2190286
##   3  3.6442905 -4.6808887
```

Podem afirmar de nou amb un 95% de confiança que hi ha relació entre ambdues variables, on s'esperava un 4.7% més de supervivents de la classe 3, en canvi de la classe 1 s'esperava un 5.8% menys.

A continuació es discretitzarà la variable *Age*. El nombre d'intervals escollits és 4 i s'utilitzarà el mètode d'igual freqüència per tal de mantenir sempre la mateixa freqüència.

```r
dis1 <- table(discretize(x = titanic_data$Age, method = "frequency", breaks = 4, include.lowest = TRUE))
dis1
```

```
##
## [0.17,22)   [22,26)   [26,37)   [37,80]
##       290       297       394       328
```

```r
titanic_data$AgeD[titanic_data$Age < 22] <- "Menors de 22 anys"
titanic_data$AgeD[titanic_data$Age >= 22 & titanic_data$Age < 26] <- "Entre 22 i 25 anys"
titanic_data$AgeD[titanic_data$Age >= 26 & titanic_data$Age < 37] <- "Entre 26 i 37 anys"
titanic_data$AgeD[titanic_data$Age >= 37] <- "Majors de 37"
```

Tot seguit fem de la nova variable un factor.

```r
titanic_data$AgeD <- factor(
  titanic_data$AgeD,
  ordered = FALSE,
  levels = c(
    "Menors de 22 anys",
    "Entre 22 i 25 anys",
    "Entre 26 i 37 anys",
    "Majors de 37"
  )
)
summary(titanic_data$AgeD)
```

```
##  Menors de 22 anys Entre 22 i 25 anys Entre 26 i 37 anys       Majors de 37
##                290                297                394                328
```

4. Hi ha diferències en la supervivència segons l'edat?

- H0: les dues variables són independents.
- H1: les dues variables no són independents.

```r
table_AgeD <- table(titanic_data$AgeD, titanic_data$Survived)
prop.table(table(titanic_data$AgeD, titanic_data$Survived), margin = 1)
```

```
##
##                              0         1
##   Menors de 22 anys  0.5793103 0.4206897
##   Entre 22 i 25 anys 0.7744108 0.2255892
##   Entre 26 i 37 anys 0.5406091 0.4593909
##   Majors de 37       0.6219512 0.3780488
```

```
fisher.test(table_AgeD, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_AgeD
## p-value = 1.139e-09
## alternative hypothesis: two.sided
```

```
# Si fem el test X² també és significatiu
chisq.test(x = table_AgeD)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_AgeD
## X-squared = 42.717, df = 3, p-value = 2.826e-09
```

```
chisq.test(x = table_AgeD)$residuals
```

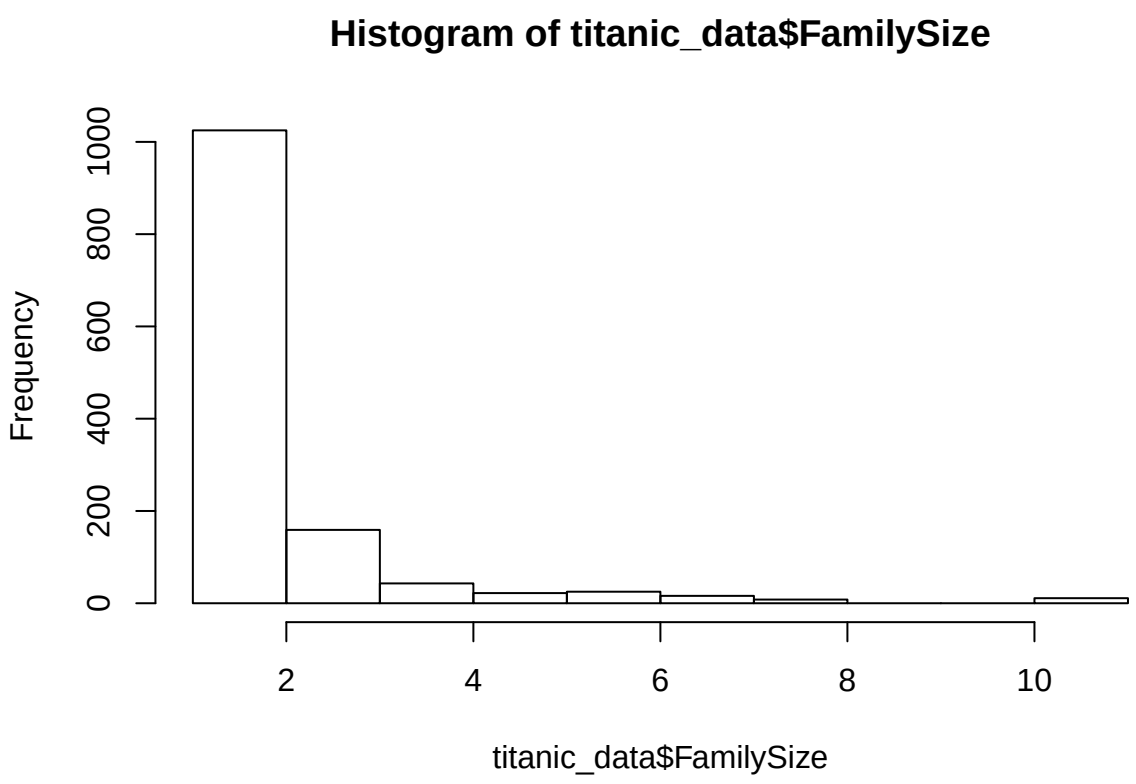
```
##
##                               0           1
##   Menors de 22 anys  -0.93454743  1.20037427
##   Entre 22 i 25 anys  3.31539715 -4.25844350
##   Entre 26 i 37 anys -2.06286950  2.64964130
##   Majors de 37       -0.01518213  0.01950061
```

Podem afirmar de nou amb un 95% de confiança que hi ha relació entre ambdues variables. S'esperava un 4,3% més de supervivents en la franja d'edat 22-25 anys, en canvi s'esperava un 2.6% menys en la de 26-37 anys.

Finalment, ens interessava conèixer si la probabilitat de sobreviure tenia alguna relació amb el tamany de la família?

En primer lloc, el que hem fet es crear una nova variable anomenada *FamilySize*, que és la suma de *SibSp* i *Parch*. Hem analitzat si tenia valors extrems, si seguia una distribució normal i si la variança era constant entre els diferents grups.

```
# Creació nova variable
titanic_data$FamilySize <- titanic_data$SibSp + titanic_data$Parch + 1;
hist(titanic_data$FamilySize)
```




```
boxplot.stats(titanic_data$FamilySize)$out
```

```
##   [1]  5  7  6  5  7  6  4  6  4  8  6  7  8  4  5  6  4  7  5 11  6  6  6  5 11
##  [26]  7  4 11  5  7  7  6  6  4  4  5 11  6  6  5  8  4  5  4  5  6  6  4  4  4
##  [51]  4  8  5  4  4  7  7  5  4  4  7  4  4  6  6  6  4  8  8  4  6  4  5  5  4
##  [76]  4  5  4  6  4 11  4  7  6  6 11  7  4 11  6  4  5  4  4  4  6  6  5  6  4
## [101]  5  8  8  5  4  7  5  7  4 11  7  4  4  4  4  4 11  4  4 11 11  7  4  5  5
```

```
fligner.test(Age ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 4.8298, df = 1, p-value = 0.02797
```

```
lillie.test(x = titanic_data$FamilySize)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  titanic_data$FamilySize
## D = 0.31514, p-value < 2.2e-16
```

Podem veure que les famílies més grans, comptant pares, fills, germans i parelles, és de 11, número que s'ha donat per vàlid.

Mitjançant els diferents test es pot veure que no segueix una distribució normal i que no presenta una variança constant.

El que es farà a continuació serà discretitzar la variable en 3 grups.

```
summary(titanic_data$FamilySize)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.884   2.000  11.000
```

```
# Discretitzem la variable 'FamíliaSize'
titanic_data$FamilySizeD[titanic_data$FamilySize < 2] <- "Adult sol"
titanic_data$FamilySizeD[titanic_data$FamilySize >= 2 & titanic_data$FamilySize < 5] <- "Famílies de dos
titanic_data$FamilySizeD[titanic_data$FamilySize >= 5] <- "Famílies amb més de 4 membres"
```

Tot seguit fem de la nova variable un factor.

```
titanic_data$FamilySizeD <- factor(
  titanic_data$FamilySizeD,
  ordered = FALSE,
  levels = c(
    "Adult sol",
    "Famílies de dos a 4 membres",
    "Famílies amb més de 4 membres"
  )
)

summary(titanic_data$FamilySizeD)
```

```
##                     Adult sol   Famílies de dos a 4 membres
##                           790                           437
## Famílies amb més de 4 membres
##                            82
```

- H0:Les variables són independents, és a dir, el fet de sobreviure no varia segons el tamany de la unitat familiar.
- H1: Les variables són dependents.

```
table_Family <- table(titanic_data$FamilySizeD, titanic_data$Survived)
prop.table(table(titanic_data$FamilySizeD, titanic_data$Survived), margin = 1)
```

```
##
##                                       0         1
##   Adult sol                     0.7075949 0.2924051
##   Famílies de dos a 4 membres   0.4393593 0.5606407
##   Famílies amb més de 4 membres 0.7804878 0.2195122
```

```r
fisher.test(table_Family, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_Family
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```r
# Si fem el test X² també és significatiu
chisq.test(x = table_Family)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_Family
## X-squared = 95.437, df = 2, p-value < 2.2e-16
```

```r
chisq.test(x = table_Family)$residuals
```

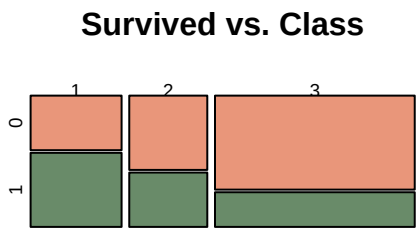
```
##
##                                       0         1
##   Adult sol                    3.027142 -3.888196
##   Famílies de dos a 4 membres -4.854939  6.235900
##   Famílies amb més de 4 membres  1.811806 -2.327164
```

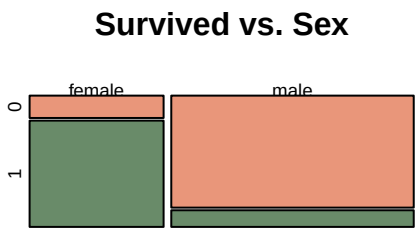
Es pot veure amb un nivell de significació del 5% com el tamany de la unitat familiar també va influir. Van sobreviure un 6.2% més de famílies de 2-4 membres del que s'esperava. En canvi, s'esperava que un 3.9% més d'adults sols sobrevisqués.

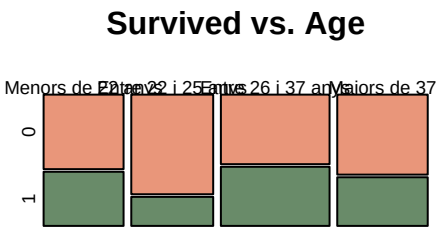
Mitjançant els gràfics de barres, les taules de contingencia i els tests realitzats podem concloure:

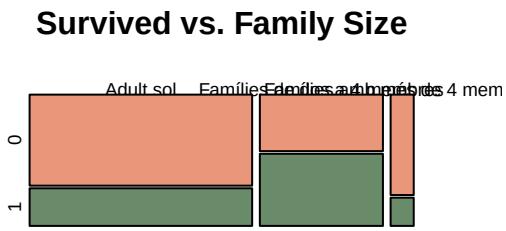
- La proporció d'homes i dones que van sobreviure és força diferent(homes: 109, dones: 385). Si ens fixem en el % respecte el seu gènere, per les dones és del 83% mentre que pels homes és del 13%.
- Referent a la classe en la que viatjaven, si ens fixem en el gràfic, el nombre de persones que més van sobreviure són els que viatjaven en 3a classe. Cal dir que el nombre de passatgers d'aquesta classe és molt major. Si ens fixem en el % dins de cada classe, són els de 1a classe els que tenen una ràtio més alta de supervivència.
- Cal destacar també que la proporció d'adults sols és més del 50%, i que la franja on hi trobem més viatjants és la franja d'edat entre 26 i 37 anys.
- Després de realitzar els 4 tests es pot observar que les diferències són significatives, i que tant l'edat, la classe en la que viatjaven, el gènere com el tamany de la unitat familiar van ser significatius per la supervièncía.

```r
par(mfrow=c(2,2))
plot(table_Class, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Class")
plot(table_Sex, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Sex")
plot(table_AgeD, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Age")
plot(table_Family, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Family Size")
```





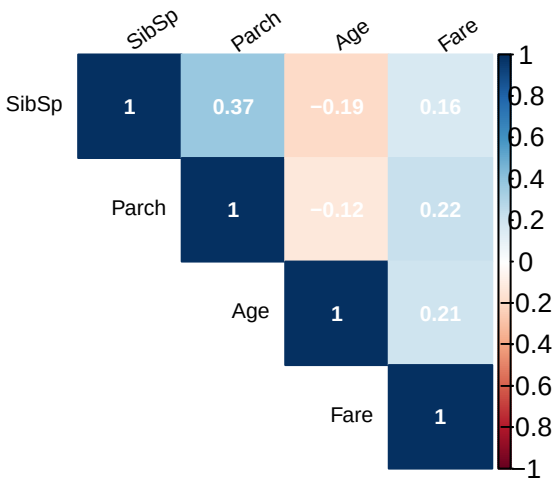


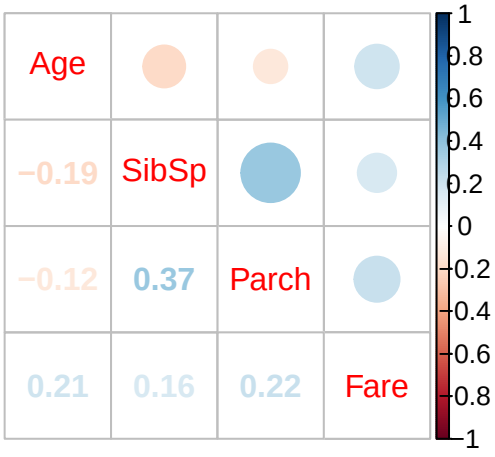


3.2.2 Matriu de correlació

A continuació hem analitzat la relació entre les diferents característiques dels passatgers, tot calculant prèviament la matriu de correlació i guardant-la en un objecte.

```
# Creació de la matriu de correlació
corr_data <- titanic_data[, c("Age", "SibSp", "Parch", "Fare")]
M <- cor(corr_data)
par(mfrow = c(1,2))
corrplot(M, method = "color", type = "upper",
         addCoef.col = "white", number.cex = 0.7,
         tl.col="black", tl.srt = 35, tl.cex = 0.7,
         order = "hclust")
corrplot.mixed(M)
```





Però no podem dir si són significativament diferent de 0, és a dir, no tenim evidencies estadístiques. Per saber-ho cal dur a terme una prova de significació. Amb la següent instrucció podem veure la matriu anterior i els p-value, on en la majoria dels casos hi ha correlació, els p-value són especialment petits.

```
rcorr(as.matrix(corr_data))
```

```
##          Age SibSp Parch Fare
## Age    1.00 -0.19 -0.12 0.21
## SibSp -0.19  1.00  0.37 0.16
## Parch -0.12  0.37  1.00 0.22
## Fare   0.21  0.16  0.22 1.00
##
## n= 1309
##
##
## P
##        Age SibSp Parch Fare
## Age         0     0     0
## SibSp  0          0     0
## Parch  0    0           0
## Fare   0    0     0
```

En tots els casos el p-value es (=0) molt petit, és a dir, que és estadísticament significatiu. Destaquen les relacions entre el preu del bitllet i l'edat, i entre el preu del bitllet i la mida de la família. En aquestes relacions la correlació és positiva, de manera que a major edat major ha sigut el preu del bitllet, i el mateix passa amb la mida de la família i el preu del bitllet.

3.2.3 Regressió logística

Volem predir el fet de sobreviure o no, de manera que ens trobem amb una variable discreta, concretament binaria (0,1). Si utilitzéssim un model lineal per predir un grup binàri estariem obtenint un model erroni.

```
#Selecció de dades per la regressió:
titanic_data <- select(titanic_data, -c(FamilySize, AgeD, SibSp, Parch, FamilySizeD))
# Divisió del conjunt de dades en dos subconjunts, un de train i l'altre de test
train <- titanic_data[1:667,]
test <- titanic_data[668:889,]

# Creació del model de predicció
model <- glm(Survived ~., family = binomial(link = 'logit'), data = train)
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3695  -0.7028  -0.4174   0.6429   2.4537
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.831559   0.560527   6.836 8.16e-12 ***
## Pclass2     -0.983816   0.354678  -2.774  0.00554 **
## Pclass3     -2.412016   0.362961  -6.645 3.02e-11 ***
## Sexmale     -2.604617   0.217012 -12.002  < 2e-16 ***
## Age         -0.028102   0.008781  -3.200  0.00137 **
## Fare        -0.003626   0.002796  -1.297  0.19467
## EmbarkedQ    0.127313   0.413628   0.308  0.75824
## EmbarkedS   -0.480851   0.269026  -1.787  0.07388 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 891.99  on 666  degrees of freedom
## Residual deviance: 611.22  on 659  degrees of freedom
## AIC: 627.22
##
## Number of Fisher Scoring iterations: 4
```

```
# Predicció de les dades
result <- predict(model, newdata = test, type = 'response')
result <- ifelse(result > 0.5, 1, 0)
fitted.probabilities <- predict(model, newdata = test, type = 'response')
fitted.results <- ifelse(fitted.probabilities > 0.5, 1, 0)

# Matriu de confusió
confusionMatrix(table(fitted.results, test$Survived))
```

```
## Confusion Matrix and Statistics
##
##
## fitted.results   0   1
##              0 124  24
##              1  17  57
##
##                Accuracy : 0.8153
##                  95% CI : (0.7579, 0.8641)
##     No Information Rate : 0.6351
##     P-Value [Acc > NIR] : 3.481e-09
##
##                   Kappa : 0.5941
##
##  Mcnemar's Test P-Value : 0.3487
##
##             Sensitivity : 0.8794
##             Specificity : 0.7037
##          Pos Pred Value : 0.8378
##          Neg Pred Value : 0.7703
##              Prevalence : 0.6351
##          Detection Rate : 0.5586
##    Detection Prevalence : 0.6667
##       Balanced Accuracy : 0.7916
##
##        'Positive' Class : 0
##
```

Mitjançant els resultats del model podem veure com el fet de pertanyer a la classe 2 o 3 està relacionat amb el fet de sobreviure, com també el genère, on el fet de ser home té un efecte negatiu igual que viatjar en 2a i 3a classes.

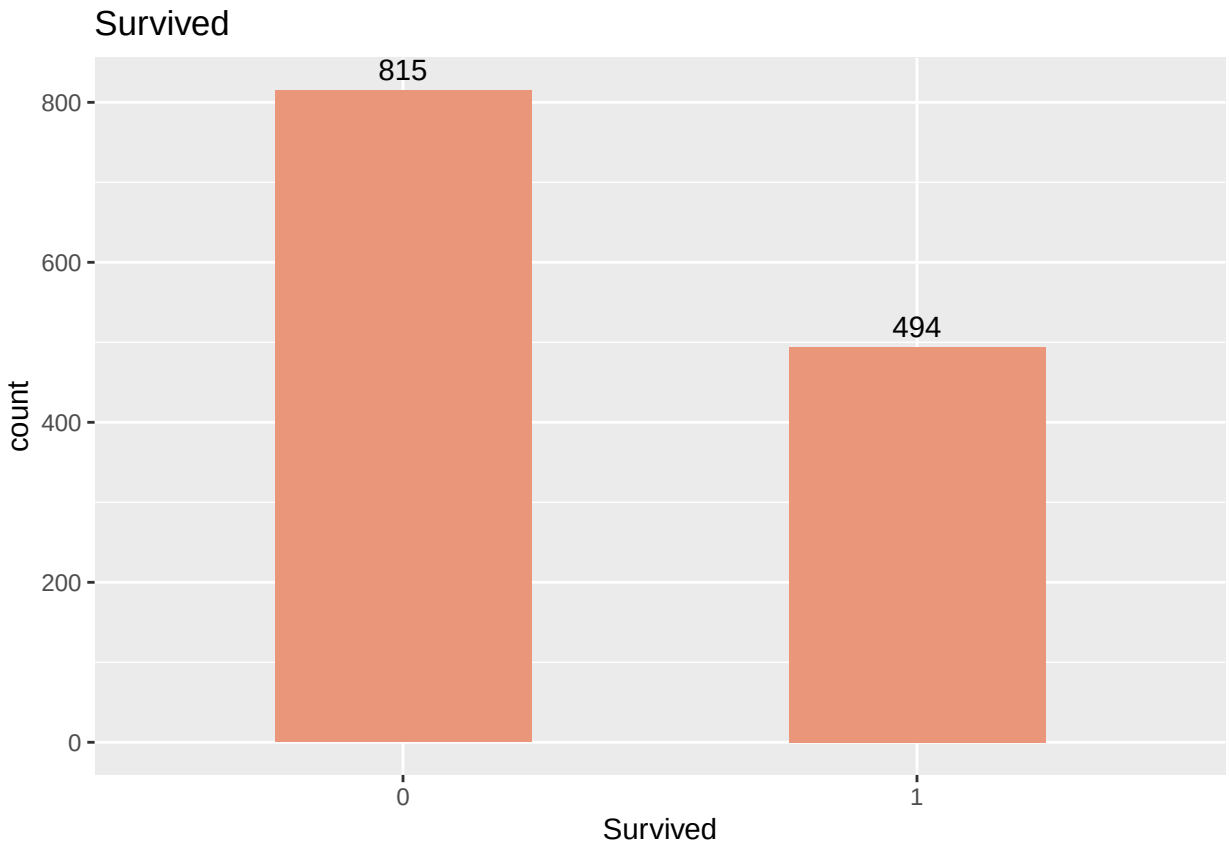
L'edat també té un efecte negatiu en la supervivencia: a major edat menor probabilitat de sobreviure.

Mitjançant l'intercept, podem veure el que hem anat confirmant amb els test d'independència i els gràfics, i és que el fet de ser dona i viatjar a la classe 1 té una major probabilitat de supervivència.

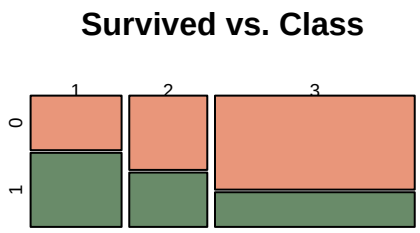
A través de la matriu de confusió es pot veure com el model té un 82% de precisió en la predicció.

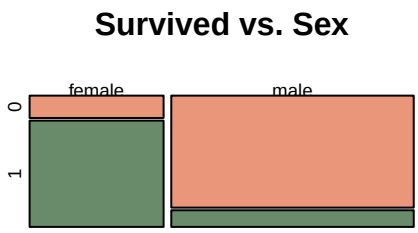
4 Representació dels resultats

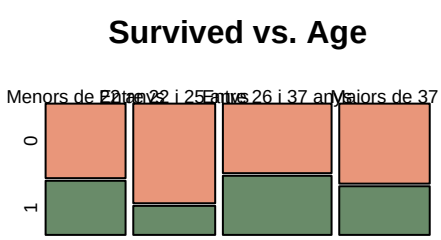
```r
#Supervivents
ggplot(titanic_data, aes(x = Survived)) +
  geom_bar(width = 0.5, fill = "darksalmon") +
  geom_text(stat = 'count', aes(label = stat(count)), vjust = -0.5) + ggtitle("Survived")
```

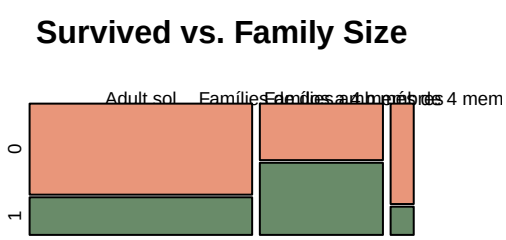



```
#Relació variables i supervivència
par(mfrow = c(2,2))
plot(table_Class, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Class")
plot(table_Sex, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Sex")
plot(table_AgeD, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Age")
plot(table_Family, col = c("darksalmon","darkseagreen4"), main = "Survived vs. Family Size")
```



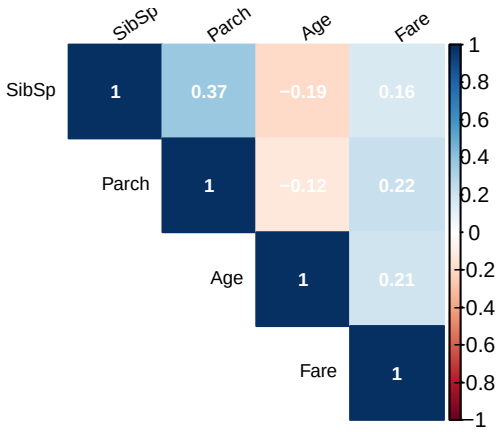







```
#Correlació
corrplot(M, method = "color", type = "upper",
         addCoef.col = "white", number.cex = 0.7,
         tl.col="black", tl.srt = 35, tl.cex = 0.7,
         order = "hclust")

#Matriu de confusió
table(test$Survived, fitted.results)
```

```
##    fitted.results
##       0   1
##   0 124  17
##   1  24  57
```



5 Conlusió

Mitjançant els diferents anàlisi estadístics podem concloure:

- Que els passatgers que van sobreviure són menys del 50%, concretament només representen un 37% del total. *La proporció d'homes i dones que van sobreviure és força diferent, on predominen les dones. Si ens fixem en el % respecte el seu gènere, per les dones és del 83% mentre que pels homes és tan sols del 13%.
- Referent a la classe en la que viatjaven, el nombre de persones que més van sobreviure són els que viatjaven en 3a classe. Cal dir que el nombre de passatgers d'aquesta classe és molt major. Si ens fixem en el % dins de cada classe, són els de 1a classe els que tenen una ràtio més alta de supervivència.
- Cal destacar també que la proporció d'adults sols és més del 50%, i que la franja on hi trobem més viatjants és la franja d'edat entre 26 i 37 anys.
- Em pogut comprovar com tan l'edat, la classe en la que viatjaven, el gènere, com el tamany de la unitat familiar van ser significatius per la superviència, és a dir, hi ha relació entre totes elles i el fet de sobreviure.
- Concretament, van sobreviure més dones que homes, la classe amb majors supervivents és la Classe 1, pel que fa l'edat s'esperava més supervivents entre la franja dels 22-25 d'anys de la que es va produir, i finalment, pel que fa el tamany familiar també va tenir un paper important, on destaquen les famílies de 2-4 membres destacan com a supervivents, m'entre que pel que fa els adults que viatjeven sols s'esperava una ràtio de supervients major.
- Mitjaçant la regressió logística, hem pogut comprovar novament, el que hem anat veient al llarg de tots els gràfics, taules i test, i és doncs, que la supervivència no va ser igual per tots els passatgers, que

a major edat, els homes, com el fet de viatjar en les classes 2 i 3 van tenir un paper negatiu, fent que la probabilitat de sobreviure fos menor.
- També podem extreure que el preu pagat del tiquet i la porta per la que van embarcar no va ser significativa per la supervivència.

6 Recursos

1. Subirats Maté, L., Pérez Trenanz, D. O., & Calvo González, M. (2019). Introducció a la neteja i anàlisi de dades. UOC.
2. Amat Rodrigo, J. (2016). RPubs - Análisis de Normalidad: gráficos y contrastes de hipótesis. https://rpubs.com/Joaquin_AR/218465
3. Homogeneity of variance. http://www.cookbook-r.com/Statistical_analysis/Homogeneity_of_variance/
4. aggregate function | R Documentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate
5. Shirokanova, A., & Volchenko, O. (2019). RPubs - Chi squared. https://rpubs.com/ovolchenko/chisq2
6. fisher.test function | R Documentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fisher.test
7. Amat Rodrigo, J. (2016). RPubs - Test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran. https://rpubs.com/Joaquin_AR/220579
8. Ramos Lorenzo, C. (2019). RPubs - Logistic Regression. https://rpubs.com/MrCristianrl/500969