PR2: Neteja i anàlisi de les dades

Tipologia i cicle de vida de les dades

Meritxell Bosch i Marta Martínez

6 de juny, 2020

Contents

1 Descripció del dataset **3**

1.1 Reducció del dataset 5

2 Neteja de les dades **5**

2.1 Valors perduts 5

2.2 Valors Extrems 8

3 Anàlisi de les dades **9**

3.1 Homogeneïtat de la variància: homocedasticidad 9

3.1.1 Variància 9

3.1.2 Homocedasticidad 10

3.2 Proves estadístiques 10

3.2.1 Preparació prèvia de les variables 10

3.2.2 Relacions de dependència 11

3.2.2.1 Sex 11

3.2.2.2 Pclass 12

3.2.2.3 Age 12

3.2.3 Contrast d'hipòtesi 13

3.2.4 Models de regressió 13

3.2.4.1 Survived i PClass 13

3.2.4.2 Survived i Sex 15

3.2.4.3 Survived i Age 15

3.2.4.4 Regressió logística multivariable 17

3.2.5 Predicció 18

3.2.5.1 Predicció d'un passatger 18

3.2.5.2 Predicció del dataset test 18

4 Gràfics i taules **21**

4.1 Percentages de supervivents i no supervivents 22

4.1.1 Al conjunt d'entrenament 22

4.1.2 A les prediccions 22

4.2 Gràfiques del conjunt d'entrenament 22

4.3 Gràfiques del conjunt de test amb la predicció 29

5 Fitxer final **35**

6 Fonts consultades **36**

```r
# Carreguem els paquets R que utilitzarem
library(data.table)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
# install.packages("arules")
library(arules)
```

```
## Warning: package 'arules' was built under R version 3.6.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
#install.packages("DescTools")
#install.packages("oddsratio")
library(oddsratio)
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.6.2
```

```
##
## Attaching package: 'DescTools'
```

```
## The following object is masked from 'package:car':
##
##     Recode
```

```
## The following object is masked from 'package:data.table':
##
##     %like%
```

```
dir<-getwd()
setwd(dir)
```

---

1 Descripció del dataset

```
# Obrim i guardem el joc de dades test i train
test <- read.csv('test.csv',stringsAsFactors = FALSE)
train <- read.csv('train.csv', stringsAsFactors = FALSE)
```

Podríem haver unit els dos jocs de dades en un només amb: myData <- bind_rows(train,test).

Pero finalment treballarem només amb les dades de train.

```
myData <- train
filas=dim(train)[1]

# Verifiquem l'estructura del joc de dades
str(myData)
```

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

```
summary(myData)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
# Verifiquem que no hi ha IDs repetits
myData[duplicated(myData$PassengerId), ]
```

```
##  [1] PassengerId Survived    Pclass      Name        Sex         Age
##  [7] SibSp       Parch       Ticket      Fare        Cabin       Embarked
## <0 rows> (or 0-length row.names)
```

1.1 Reducció del dataset

Creem un nou joc de dades només amb les columnes que ens interessen:

```
data <- select(myData, -PassengerId, -Name, -SibSp, -Parch,
               -Ticket, -Fare, -Cabin, -Embarked)

# Verifiquem l'estructura del joc de dades
str(data)
```

```
## 'data.frame':    891 obs. of  4 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : chr  "male" "female" "female" "female" ...
##  $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
```

---

2 Neteja de les dades

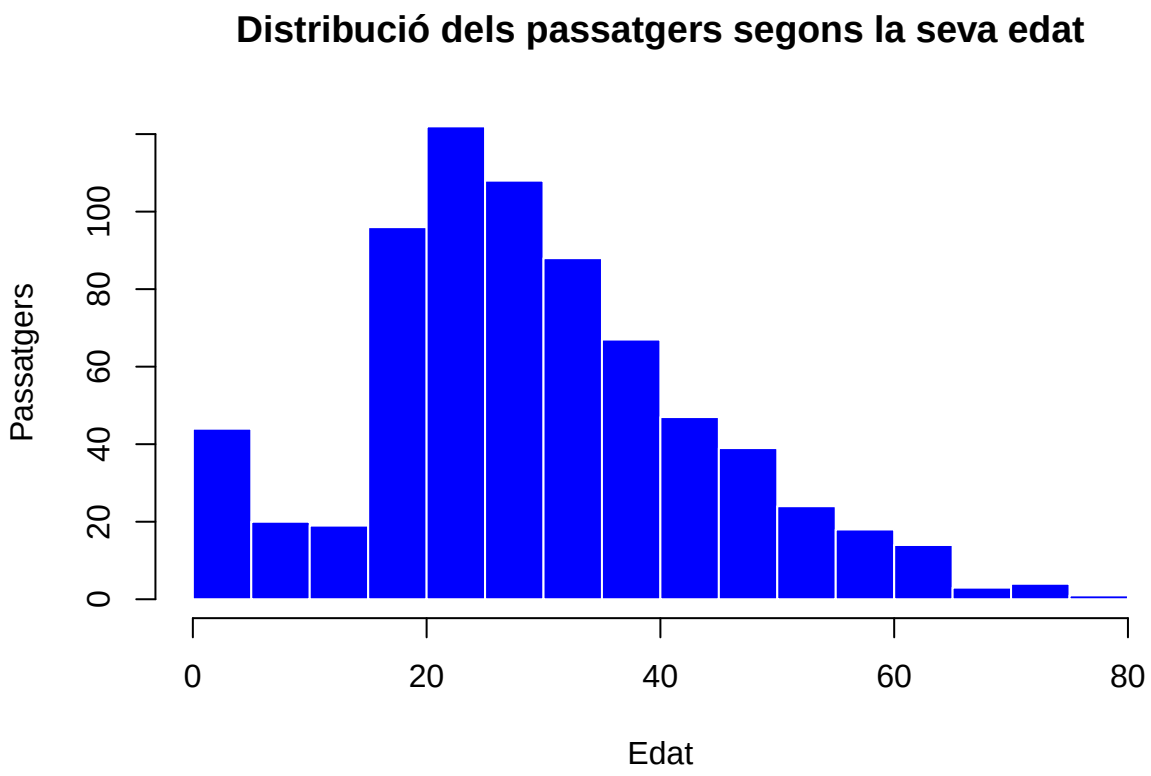
2.1 Valors perduts

```
# busquem valors perduts
colSums(is.na(data))
```

```
## Survived   Pclass      Sex      Age
##        0        0        0      177
```

Els valors perduts de l'edat es poden substituir per dades estadísticament coherents, però per decidir com es fa la substitució, és necessari veure la distribució de l'edat.

```
hist(data$Age,main="Distribució dels passatgers segons la seva edat",
     col="blue",border="white",breaks=15,xlab="Edat",ylab="Passatgers")
```



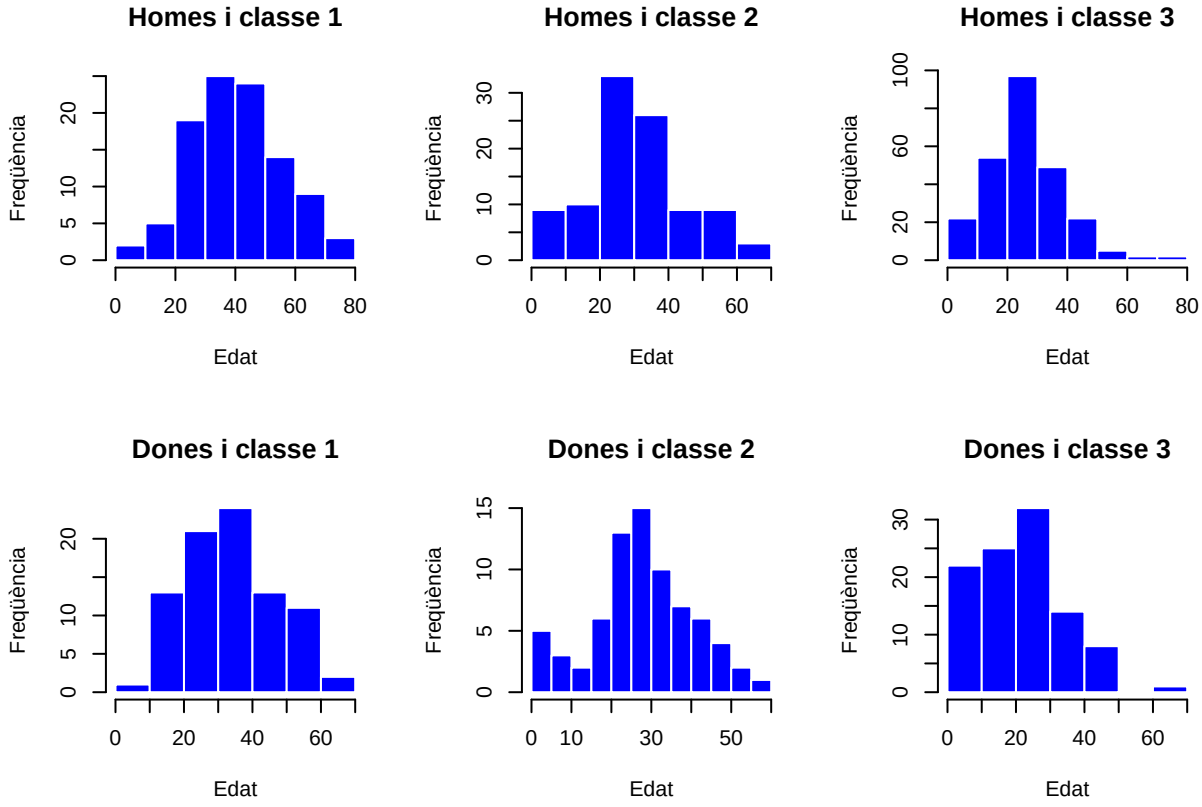

```
# Comprovem la distribució dels individus
table(data$Sex, data$Pclass)
```

```
##
##             1   2   3
##   female  94  76 144
##   male   122 108 347
```

S'observa que els individus no estan repartits de forma equitativa entre classes socials i sexe. Per tant, per trobar un valor de substitució el millor és tenir en compte individus similars (mateix sexe i classe social).

```
par(mfrow=c(2,3))
hist(data$Age[(data$Sex=="male")&(data$Pclass==1)],
     main = "Homes i classe 1",
     xlab = "Edat",
     ylab = "Freqüència",col="blue",border="white")
hist(data$Age[(data$Sex=="male")&(data$Pclass==2)],
     main = "Homes i classe 2",
     xlab = "Edat",
     ylab = "Freqüència",col="blue",border="white")
hist(data$Age[(data$Sex=="male")&(data$Pclass==3)],
     main = "Homes i classe 3",
     xlab = "Edat",
     ylab = "Freqüència",col="blue",border="white")
hist(data$Age[(data$Sex=="female")&(data$Pclass==1)],
```

```
     main = "Dones i classe 1",
     xlab = "Edat",
     ylab = "Freqüència",col="blue",border="white")
hist(data$Age[(data$Sex=="female")&(data$Pclass==2)],
     main = "Dones i classe 2",
     xlab = "Edat",
     ylab = "Freqüència",col="blue",border="white")
hist(data$Age[(data$Sex=="female")&(data$Pclass==3)],
     main = "Dones i classe 3",
     xlab = "Edat",
     ylab = "Freqüència",col="blue",border="white")
```



Excepte pel cas de les dones i classe 2, cap de les distribucions d'edat té distribució que s'assembli a una normal, per tant, es calcula la mediana de cada grup per substituïr el valor perdut.

```
mediana <- median(data$Age[!is.na(data$Age)],na.rm=T)
mediana_m_1 <- median(data$Age[!is.na(data$Age)&(data$Sex=="male")&
                               (data$Pclass==1)],na.rm=T)
mediana_m_2 <- median(data$Age[!is.na(data$Age)&(data$Sex=="male")&
                               (data$Pclass==2)],na.rm=T)
mediana_m_3 <- median(data$Age[!is.na(data$Age)&(data$Sex=="male")&
                               (data$Pclass==3)],na.rm=T)

mediana_f_1 <- median(data$Age[!is.na(data$Age)&(data$Sex=="female")&
                               (data$Pclass==1)],na.rm=T)
mediana_f_2 <- median(data$Age[!is.na(data$Age)&(data$Sex=="female")&
```

```
                                    (data$Pclass==2)],na.rm=T)
mediana_f_3 <- median(data$Age[!is.na(data$Age)&(data$Sex=="female")&
                                    (data$Pclass==3)],na.rm=T)
```

S'observa que la mediana general és propera a 3 de les categories però molt allunyada de les altres 3, o sigui que reforça que hem d'aplicar la mediana a cada grup per separat.

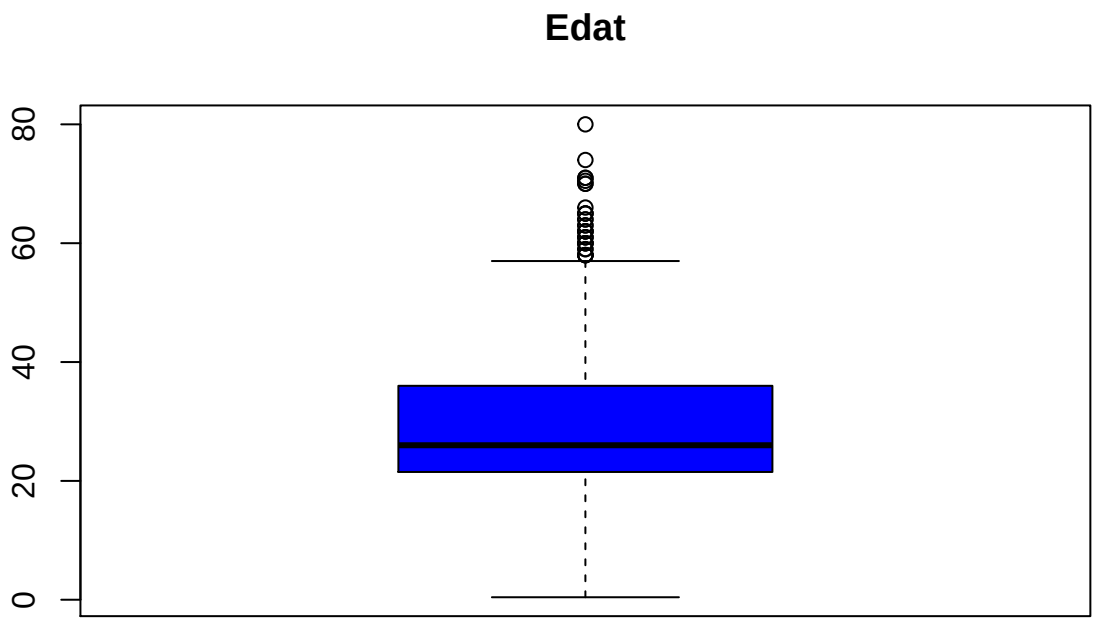
```
data$Age[is.na(data$Age)&(data$Sex=="male")&(data$Pclass==1)] <- mediana_m_1
data$Age[is.na(data$Age)&(data$Sex=="male")&(data$Pclass==2)] <- mediana_m_2
data$Age[is.na(data$Age)&(data$Sex=="male")&(data$Pclass==3)] <- mediana_m_3
data$Age[is.na(data$Age)&(data$Sex=="female")&(data$Pclass==1)] <- mediana_f_1
data$Age[is.na(data$Age)&(data$Sex=="female")&(data$Pclass==2)] <- mediana_f_2
data$Age[is.na(data$Age)&(data$Sex=="female")&(data$Pclass==3)] <- mediana_f_3

# Comprovem que ja no tenim valors perduts
colSums(is.na(data))
```

```
## Survived   Pclass      Sex      Age
##        0        0        0        0
```

2.2 Valors Extrems

Detecció de la presència de valors extrems en la variable 'edat'.

```
boxplot(data$Age,col="blue",border="black", main = "Edat")
```



```r
# Identificació els valors numèrics de les mostres que tenen valors extrems

ages <- boxplot.stats(data$Age)$out
sprintf('Edat mínima:  %i' , min(ages))
```

```
## [1] "Edat mínima:  58"
```

```r
sprintf('Edat màxima:  %i' , max(ages))
```

```
## [1] "Edat màxima:  80"
```

---

3 Anàlisi de les dades

S'executa del test de Shapiro per a comprobar si la variable numèrica Age té distribució normal.

```r
# Ja s'havia fet l'histograma anteriorment
age.test <- shapiro.test(data$Age)
print(age.test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Age
## W = 0.96548, p-value = 1.118e-13
```

3.1 Homogeneïtat de la variància: homocedasticidad

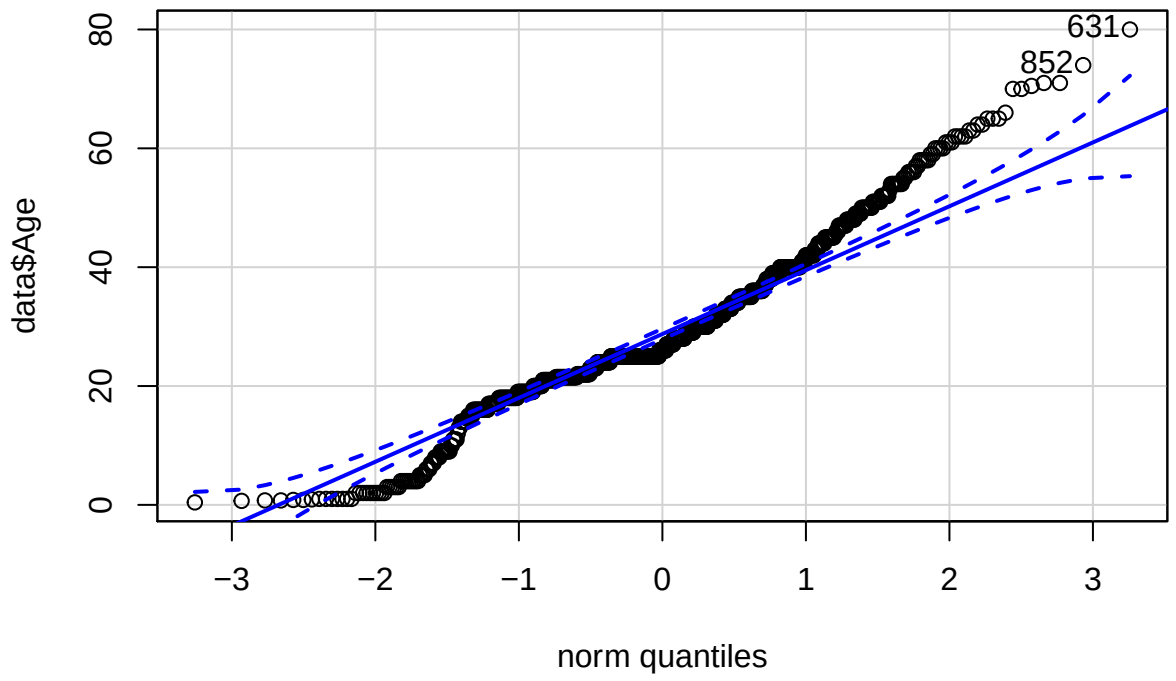
3.1.1 Variància

```r
# Comencem per fer el test de de la variança sobre
var.test(x = data[data$Survived == "0", "Age"],
         y = data[data$Survived == "1", "Age"] )
```

```
##
##  F test to compare two variances
##
## data:  data[data$Survived == "0", "Age"] and data[data$Survived == "1", "Age"]
## F = 0.83704, num df = 548, denom df = 341, p-value = 0.06553
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6897026 1.0114736
## sample estimates:
## ratio of variances
##          0.8370418
```

3.1.2 Homocedasticidad

```
# I el Gràfic QQ
qqPlot(data$Age)
```




```
## [1] 631 852
```

3.2 Proves estadístiques

3.2.1 Preparació prèvia de les variables

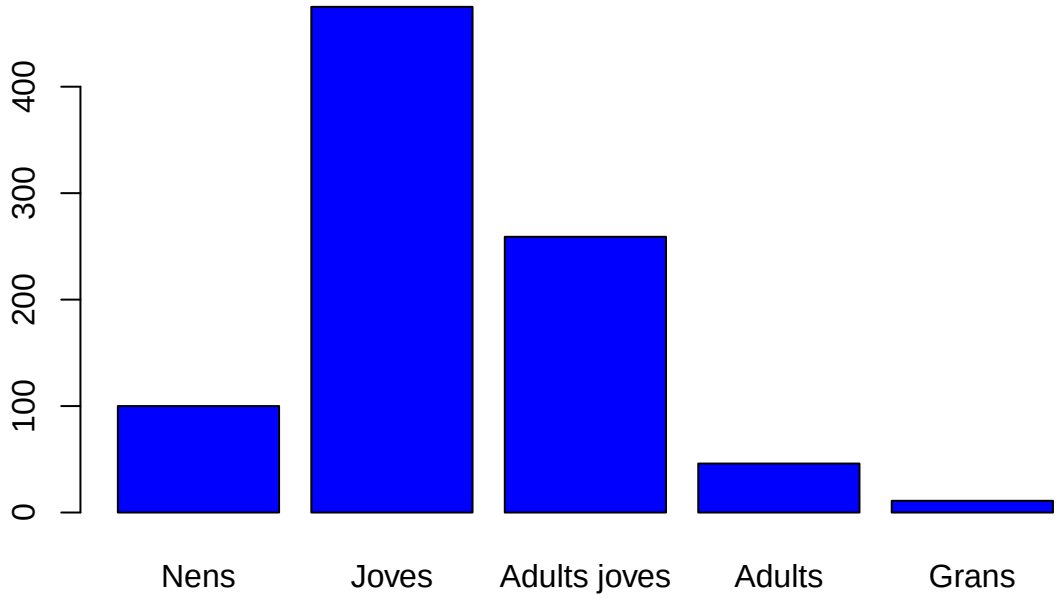
Factorització de les variables 'Survived', 'Pclass' i 'Sex'.

```
cols<-c("Survived","Pclass","Sex")
for (i in cols){
          data[,i] <- as.factor(data[,i])
}
```

```
summary(data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   21.50   26.00   29.11   36.00   80.00
```

```r
# Discretització de l'edat

data$Age.factor <- cut(data$Age, breaks=c(0, 16, 31, 51, 64, 100),
                       labels=c('Nens', 'Joves', 'Adults joves', 'Adults', 'Grans'))

plot(data$Age.factor, col = "blue")
```



Cal saber l'incidència sobre els que es salven, així que ens cal assegurar l'ordre de la variable factoritzada 'Survived'.

```r
# Reordenem
data$Survived <- relevel(data$Survived, ref = "0")
```

3.2.2 Relacions de dependència

Aplicació del el chi-squared-test a la taula de contingència. La hipòtesi nul · la d'aquest test és que no hi ha relació entre les variables (variables són independents), i l'alternativa és que sí que hi ha una relació establerta.

3.2.2.1 Sex

```r
# test chi-quadrat Sex
(tbl1 = table(data$Sex, data$Survived))
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

S'aplica el chi-squared test a la taula de contingència. Com anteriorment, la hipòtesi nul · la d'aquest test és que no hi ha relació entre les varibles, i l'alternativa és que sí que hi ha una relació establerta.

```
chisq.test(table(data$Sex, data$Survived))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(data$Sex, data$Survived)
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Com que el p-valor (2.2e-16) és inferior al valor significatiu 0.05, es rebutja la hipòtesi nul · la d'independència.

Per tant, es conclou que les dues variables s'associen estadísticament de forma significativa.

3.2.2.2 Pclass

```
# test chi-quadrat Pclass

chisq.test((table(data$Pclass, data$Survived)))
```

```
##
##  Pearson's Chi-squared test
##
## data:  (table(data$Pclass, data$Survived))
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

Com que el p-valor (2.2e-16) és inferior al significat 0.05, es rebutja la hipòtesi nul · la d'independència.

Per tant, es conclou que les dues variables s'associen estadísticament de forma significativa.

3.2.2.3 Age

```
# test chi-quadrat Age

chisq.test((table(data$Age.factor, data$Survived)), simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  (table(data$Age.factor, data$Survived))
## X-squared = 23.368, df = NA, p-value = 0.0009995
```

Com que el p-valor (0.003498) és inferior al significat .05, ees rebutja la hipòtesi nul · la d'independència.

Per tant, es conclou que les dues variables s'associen estadísticament de forma significativa.

3.2.3 Contrast d'hipòtesi

Volem saber si és cert que es salva més gent menor de 30 anys.

H~0: = ~0

H~1: < ~0

on = 30

```
# Subdividim el nostre conjunt de dades
data_survived = data[data$Survived == "1",]
data_nosurvived = data[data$Survived == "0",]
```

Per poder aplicar proves per contrast d'hipòtesis paramètriques, com la prova t de Student: 1) Les variables de les dades analitzades han d'estar normalment distribuïdes. 2) Les variàncies d'aquestes variables han de romandre constants al llarg del rang observat d'alguna altra variable.

Quan no sigui així, es pot optar per utilitzar una alternativa no paramètrica, com les proves de Wilcoxon o Mann-Whitney.

La diferència principal entre el test de suma de rangs o test U de Mann-Whitney i el test de rangs i signes de Wilcoxon és que en el primer les mostres són independents i en el segon que les dades es troben emparellades.

En aquest cas, les dades són independents, per tant, escollim el test U de Mann-Whitney (Mann–Whitney–Wilcoxon, Wilcoxon rank-sum test o Wilcoxon–Mann–Whitney).

```
# Test wilcox.test
res <- wilcox.test(x = data_survived$Age, y = data_nosurvived$Age, alternative = "less",
                   mu = 30, paired = FALSE, conf.int = 0.95)
res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_survived$Age and data_nosurvived$Age
## W = 8469.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 30
## 95 percent confidence interval:
##          -Inf 3.052763e-05
## sample estimates:
## difference in location
##             -0.9999991
```

Obtenim un p-value < 2.2e-16 i per tant podem assumir que és més probable sobreviure per al menors de 30 anys.

3.2.4 Models de regressió

3.2.4.1 Survived i PClass

```
# Survived  i PClass
model.glm <- glm(Survived ~ Pclass, data = data,family = binomial )
summary(model.glm)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4094  -0.7450  -0.7450   0.9619   1.6836
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5306     0.1409   3.766 0.000166 ***
## Pclass2      -0.6394     0.2041  -3.133 0.001731 **
## Pclass3      -1.6704     0.1759  -9.496  < 2e-16 ***
## ---
## 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: 1083.1  on 888  degrees of freedom
## AIC: 1089.1
##
## Number of Fisher Scoring iterations: 4
```

```
odds <- or_glm(data = data, model = model.glm, CI = 0.95 )
odds
```

```
## # A tibble: 2 x 5
##   predictor oddsratio `CI_low (2.5)` `CI_high (97.5)` increment
##   <chr>         <dbl>          <dbl>            <dbl> <chr>
## 1 Pclass2       0.528          0.353            0.786 Indicator variable
## 2 Pclass3       0.188          0.133            0.265 Indicator variable
```

En quin percentatge es veu modificada la probabilitat de sobreviure segons la classe?

La variació en la probabilitat de sobreviure es pot representar en una regressió logística com la que tenim com a 1-odds %, en aquest cas, on l'intercept recull els factors més favorables. D'aquesta manera, el odds, en % representa el decrement de probabilitat de sobreviure:

```
myodds <-odds[1:2,1:2]
myodds["%"] <- (1 - myodds[2]) * 100
myodds
```

```
## # A tibble: 2 x 3
##   predictor oddsratio   `%`
##   <chr>         <dbl> <dbl>
## 1 Pclass2       0.528  47.2
## 2 Pclass3       0.188  81.2
```

La probabilitat de la primera classe es troba a l'intercepte.

A segona classe la probabilitat millora decau un 47,2% i tercera classe un 81,2%. És a dir, com millor classe, més probabilitat de sobreviure,

3.2.4.2 Survived i Sex

```
# Regressió
# Survived  i Sex
model.glm <- glm(Survived ~ Sex, data = data,family = binomial )
summary(model.glm)
```

```
##
## Call:
## glm(formula = Survived ~ Sex, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6462  -0.6471  -0.6471   0.7725   1.8256
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
## Sexmale      -2.5137     0.1672 -15.036  < 2e-16 ***
## ---
## 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:  917.8  on 889  degrees of freedom
## AIC: 921.8
##
## Number of Fisher Scoring iterations: 4
```

```
odds <- or_glm(data = data, model = model.glm, CI = 0.95 )
odds
```

```
## # A tibble: 1 x 5
##   predictor oddsratio `CI_low (2.5)` `CI_high (97.5)` increment
##   <chr>         <dbl>          <dbl>            <dbl> <chr>
## 1 Sexmale       0.081          0.058            0.112 Indicator variable
```

```
myodds <-odds[1,1:2]
myodds["%"] <- (1 - myodds[2]) * 100
myodds
```

```
## # A tibble: 1 x 3
##   predictor oddsratio   `%`
##   <chr>         <dbl> <dbl>
## 1 Sexmale       0.081  91.9
```

En el cas del sexe masculí, la probabilitat de sobreviure decau un 91,9%.

3.2.4.3 Survived i Age

```r
# Regressió
# Survived  i Age
model.glm <- glm(Survived ~ Age.factor, data = data,family = binomial )
summary(model.glm)
```

```
##
## Call:
## glm(formula = Survived ~ Age.factor, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2637  -0.8958  -0.8958   1.3087   2.1899
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.2007     0.2010   0.998   0.3181
## Age.factorJoves          -0.9065     0.2234  -4.057 4.97e-05 ***
## Age.factorAdults joves   -0.5041     0.2371  -2.126   0.0335 *
## Age.factorAdults         -0.5521     0.3607  -1.531   0.1258
## Age.factorGrans          -2.5033     1.0677  -2.345   0.0190 *
## ---
## 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: 1162.7  on 886  degrees of freedom
## AIC: 1172.7
##
## Number of Fisher Scoring iterations: 4
```

```r
odds <- or_glm(data = data, model = model.glm, CI = 0.95 )
odds
```

```
## # A tibble: 4 x 5
##   predictor            oddsratio `CI_low (2.5)` `CI_high (97.5)` increment
##   <chr>                    <dbl>          <dbl>            <dbl> <chr>
## 1 Age.factorJoves          0.404          0.26             0.625 Indicator vari~
## 2 Age.factorAdults jo~     0.604          0.378            0.96  Indicator vari~
## 3 Age.factorAdults         0.576          0.281            1.16  Indicator vari~
## 4 Age.factorGrans          0.082          0.004            0.451 Indicator vari~
```

```r
myodds <-odds[1:4,1:2]
myodds["%"] <- (1 - myodds[2]) * 100
myodds
```

```
## # A tibble: 4 x 3
##   predictor              oddsratio   `%`
##   <chr>                      <dbl> <dbl>
## 1 Age.factorJoves            0.404  59.6
## 2 Age.factorAdults joves     0.604  39.6
## 3 Age.factorAdults           0.576  42.4
## 4 Age.factorGrans            0.082  91.8
```

En relació a l'edat, segons el grup tenim aquesta minoració de la possibilitat de sobreviure respecte al grup dels nens: - joves: 59,6% - Adults joves : 39,6% - Adults:42,4% - Grans: 91,8%

3.2.4.4 Regressió logística multivariable

```
# Regressió logística múltiple
modelm.glm <- glm(Survived ~ Age.factor + Pclass + Sex, data = data, family = binomial)
summary(modelm.glm)
```

```
##
## Call:
## glm(formula = Survived ~ Age.factor + Pclass + Sex, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6604  -0.6690  -0.4267   0.6471   2.3121
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               3.5093     0.3700   9.485  < 2e-16 ***
## Age.factorJoves          -0.9660     0.2798  -3.453 0.000555 ***
## Age.factorAdults joves   -1.2166     0.3207  -3.794 0.000148 ***
## Age.factorAdults         -1.7957     0.4782  -3.755 0.000173 ***
## Age.factorGrans          -2.7258     1.1143  -2.446 0.014435 *
## Pclass2                  -1.0860     0.2613  -4.155 3.25e-05 ***
## Pclass3                  -2.3042     0.2553  -9.027  < 2e-16 ***
## Sexmale                  -2.5897     0.1871 -13.844  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  804.42  on 883  degrees of freedom
## AIC: 820.42
##
## Number of Fisher Scoring iterations: 5
```

```
odds_glm <- or_glm(data = data, model = modelm.glm, CI = 0.95 )
odds_glm
```

```
## # A tibble: 7 x 5
##   predictor            oddsratio `CI_low (2.5)` `CI_high (97.5)` increment
##   <chr>                    <dbl>          <dbl>            <dbl> <chr>
## 1 Age.factorJoves          0.381          0.219            0.657 Indicator vari~
## 2 Age.factorAdults jo~     0.296          0.157            0.553 Indicator vari~
## 3 Age.factorAdults         0.166          0.064            0.418 Indicator vari~
## 4 Age.factorGrans          0.065          0.003            0.411 Indicator vari~
## 5 Pclass2                  0.338          0.201            0.561 Indicator vari~
## 6 Pclass3                  0.1            0.06             0.163 Indicator vari~
## 7 Sexmale                  0.075          0.052            0.108 Indicator vari~
```

```
myodds_glm <-odds_glm[1:7,1:2]
myodds_glm["%"] <- (1 - myodds_glm[2]) * 100
myodds_glm
```

```
## # A tibble: 7 x 3
##   predictor              oddsratio   `%`
##   <chr>                      <dbl> <dbl>
## 1 Age.factorJoves            0.381  61.9
## 2 Age.factorAdults joves     0.296  70.4
## 3 Age.factorAdults           0.166  83.4
## 4 Age.factorGrans            0.065  93.5
## 5 Pclass2                    0.338  66.2
## 6 Pclass3                    0.1    90
## 7 Sexmale                    0.075  92.5
```

Els predictors Nens, Classe 1 i sexe femení es troben a l'intercepte i són els casos més favorables per la supervivència (tots els odds, són menors que 1)

Per tant, els predictors empitjoren tots la possibilitat de supervivència en els percentatges següents: - Joves: 61,9% - Adults joves: 70,4% - Adults: 83,4% - Grans: 93,5% - 2a Classe: 66,2% - 3a Classe: 90% - Sexe masculí: 92,5%

3.2.5 Predicció

3.2.5.1 Predicció d'un passatger

```
test[3,]
```

```
##   PassengerId Pclass                     Name  Sex Age SibSp Parch Ticket
## 3         894      2 Myles, Mr. Thomas Francis male  62     0     0 240276
##     Fare Cabin Embarked
## 3 9.6875              Q
```

```
# Predicció d'un passatger
passatger_1 = data.frame(Sex = "male", Pclass = "2", Age.factor = "Adults")
pred_p1 = predict(modelm.glm, passatger_1, type= "response")
pred_p1
```

```
##         1
## 0.1232326
```

Aquest passatger té un 12,32% de probabilitats de sobreviure.

3.2.5.2 Predicció del dataset test

```
data_test <- select(test, -PassengerId, -Name, -SibSp, -Parch, -Ticket, -Fare,
                    -Cabin, -Embarked)

colSums(is.na(data_test))
```

```
## Pclass    Sex    Age
##      0      0     86
```

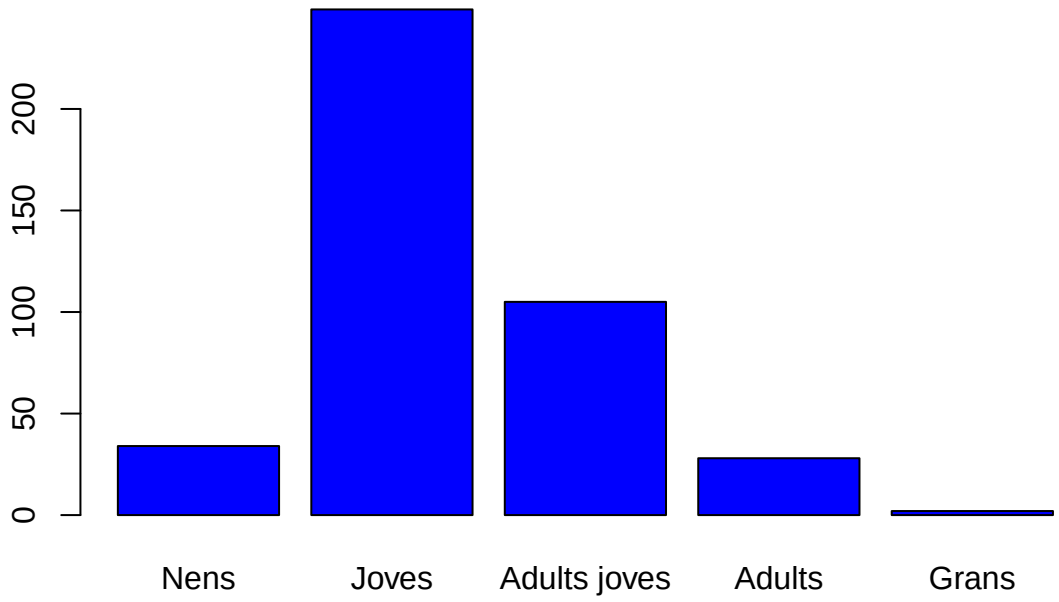
No calculem medianes noves, imputem les calculades en el grup anterior.

```
# Imputem
data_test$Age[is.na(data_test$Age)&(data_test$Sex=="male")&
                (data_test$Pclass==1)] <- mediana_m_1
data_test$Age[is.na(data_test$Age)&(data_test$Sex=="male")&
                (data_test$Pclass==2)] <- mediana_m_2
data_test$Age[is.na(data_test$Age)&(data_test$Sex=="male")&
                (data_test$Pclass==3)] <- mediana_m_3
data_test$Age[is.na(data_test$Age)&(data_test$Sex=="female")&
                (data_test$Pclass==1)] <- mediana_f_1
data_test$Age[is.na(data_test$Age)&(data_test$Sex=="female")&
                (data_test$Pclass==2)] <- mediana_f_2
data_test$Age[is.na(data_test$Age)&(data_test$Sex=="female")&
                (data_test$Pclass==3)] <- mediana_f_3
# Comprovem que ja no tenim valors perduts
colSums(is.na(data_test))
```

```
## Pclass    Sex    Age
##      0      0      0
```

Cal aplicar els mateixos procesos a les variables, en aquest cas, discretitzar i factoritzar:

```
# Factoritzem les variables 'Pclass' i 'Sex'
cols<-c("Pclass","Sex")
for (i in cols){
          data_test[,i] <- as.factor(data_test[,i])
}
# Discretitzem l'edat
data_test$Age.factor <- cut(data_test$Age, breaks=c(0, 16, 31, 51, 64, 100),
                       labels=c('Nens', 'Joves', 'Adults joves', 'Adults', 'Grans'))
plot(data_test$Age.factor, col = "blue")
```




```
summary(data_test$Age.factor)
```

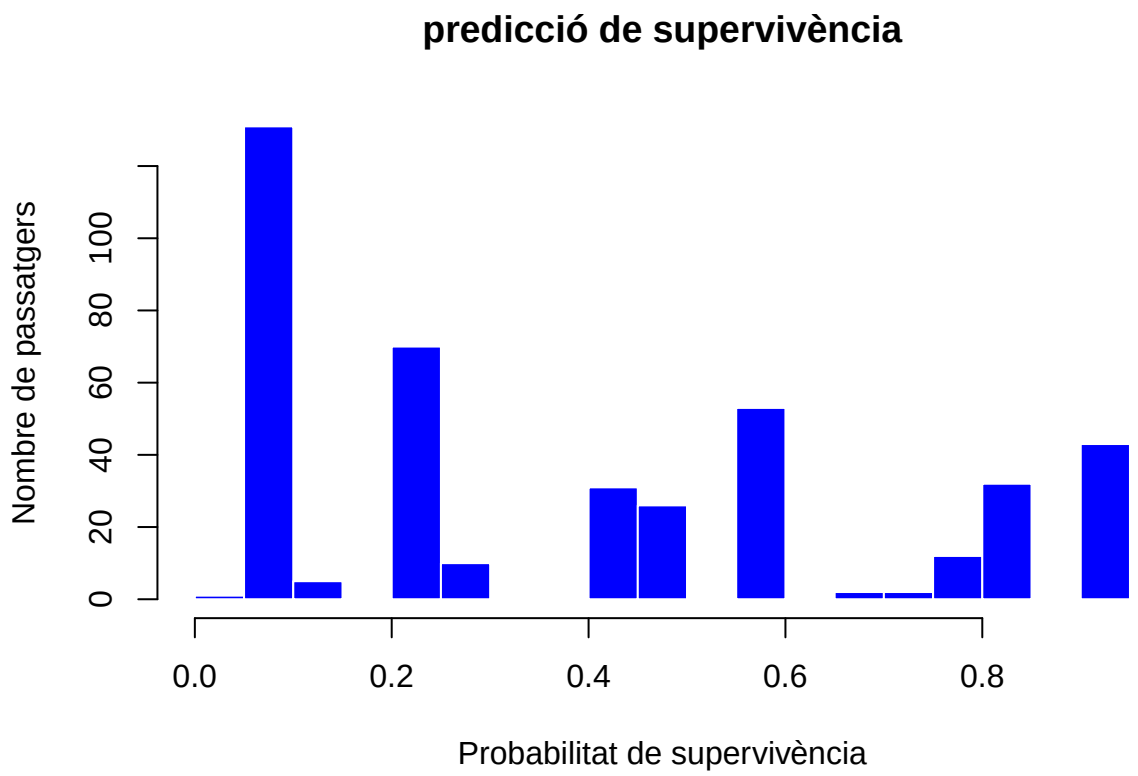
```
##         Nens        Joves Adults joves       Adults        Grans
##           34          249          105           28            2
```

```
# Revisem que no hi hagi valors nuls
colSums(is.na(data_test))
```

```
##     Pclass        Sex        Age Age.factor
##          0          0          0          0
```

Predicció del contingut del fitxer test:

```
# Generem un dataframe de valors
tots_passatgers <- data.frame(Sex = data_test$Sex, Pclass = data_test$Pclass,
                              Age.factor = data_test$Age.factor)

# Executem la predicció. Com que és binaria fem servir type = response
pred_pt = predict(modelm.glm, tots_passatgers, type = "response")

# Visualitzem resultats

hist(pred_pt,main="predicció de supervivència", col="blue", border="white",
     breaks=15, xlab="Probabilitat de supervivència",ylab="Nombre de passatgers")
```



Assignar valors a la columna del dataset: Si volguéssim participar al concurs, es mantindria com a objecte independent.

```
# Carreguem els valors de predicció al data set
data_test$Survived <- pred_pt

# Hem de decidir a partir de quin llindar de probabilitat considerem que sobreviu o no.

llindar = 0.5

data_test$Survived[which(data_test$Survived < llindar)]<- 0
data_test$Survived[which(data_test$Survived >= llindar)]<- 1

# Recompte dels supervivents en el set de test
table(data_test$Survived)
```

```
##
##   0   1
## 274 144
```

---

4 Gràfics i taules

Preparació de la variable Survived per a que sigui més entenedora:

```r
#Factoritzem la variable "Survived" del conjunt de test
data_test$Survived <- as.factor(data_test$Survived)

# Reordenem per a les gràfiques
data$Survived <- relevel(data$Survived, ref = "0")
data_test$Survived <- relevel(data_test$Survived, ref = "0")


# Relevel de les varibles Survives i Age factoritzat
levels(data$Survived ) <- c('No', 'Sí')
levels(data_test$Survived) <- c('No', 'Sí')
```

4.1 Percentages de supervivents i no supervivents

4.1.1 Al conjunt d'entrenament

```r
tabla1 <- setDT(data)[, .(Percentatge = 100 *.N / nrow(data)), by = Survived]
tabla1
```

```
##    Survived Percentatge
## 1:       No    61.61616
## 2:       Sí    38.38384
```

4.1.2 A les prediccions

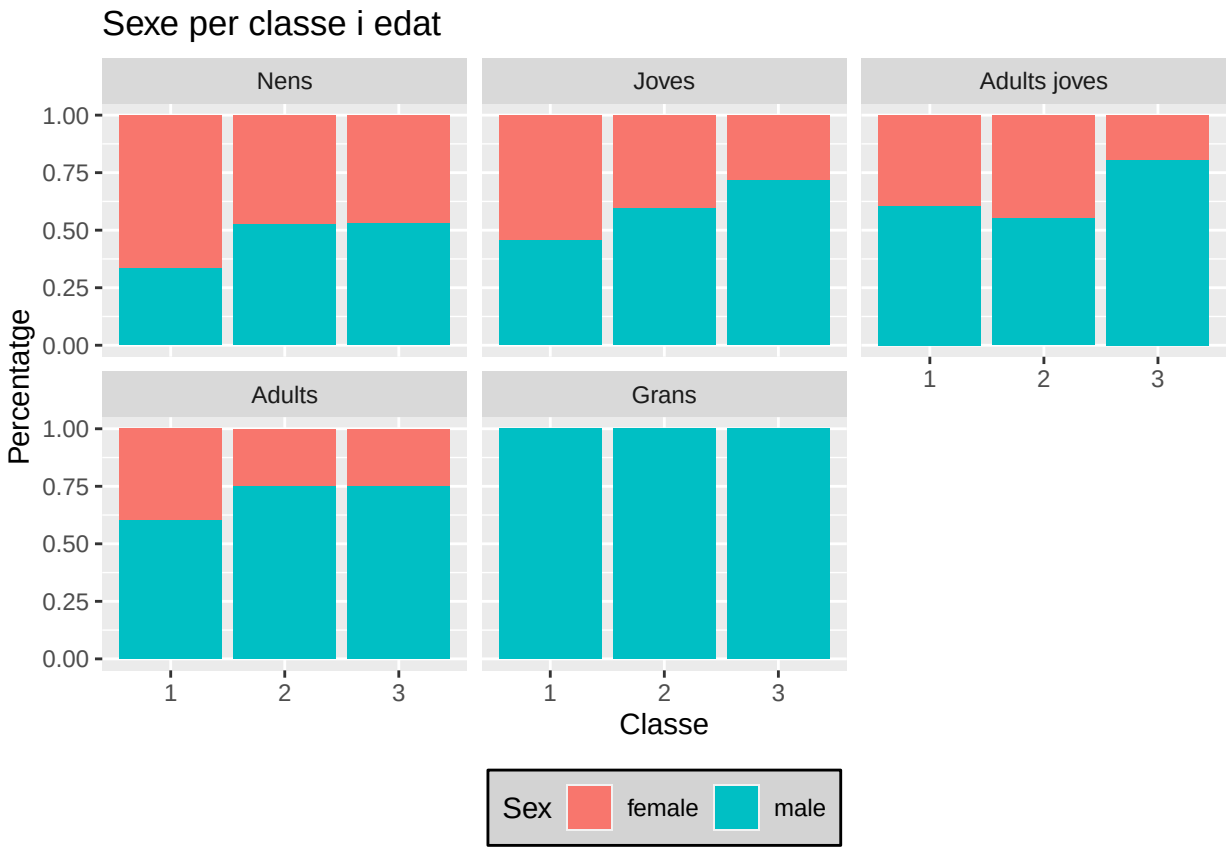
```r
tabla2 <- setDT(data_test)[, .(Percentatge = 100 *.N / nrow(data_test)), by = Survived]
tabla2
```

```
##    Survived Percentatge
## 1:       No    65.55024
## 2:       Sí    34.44976
```

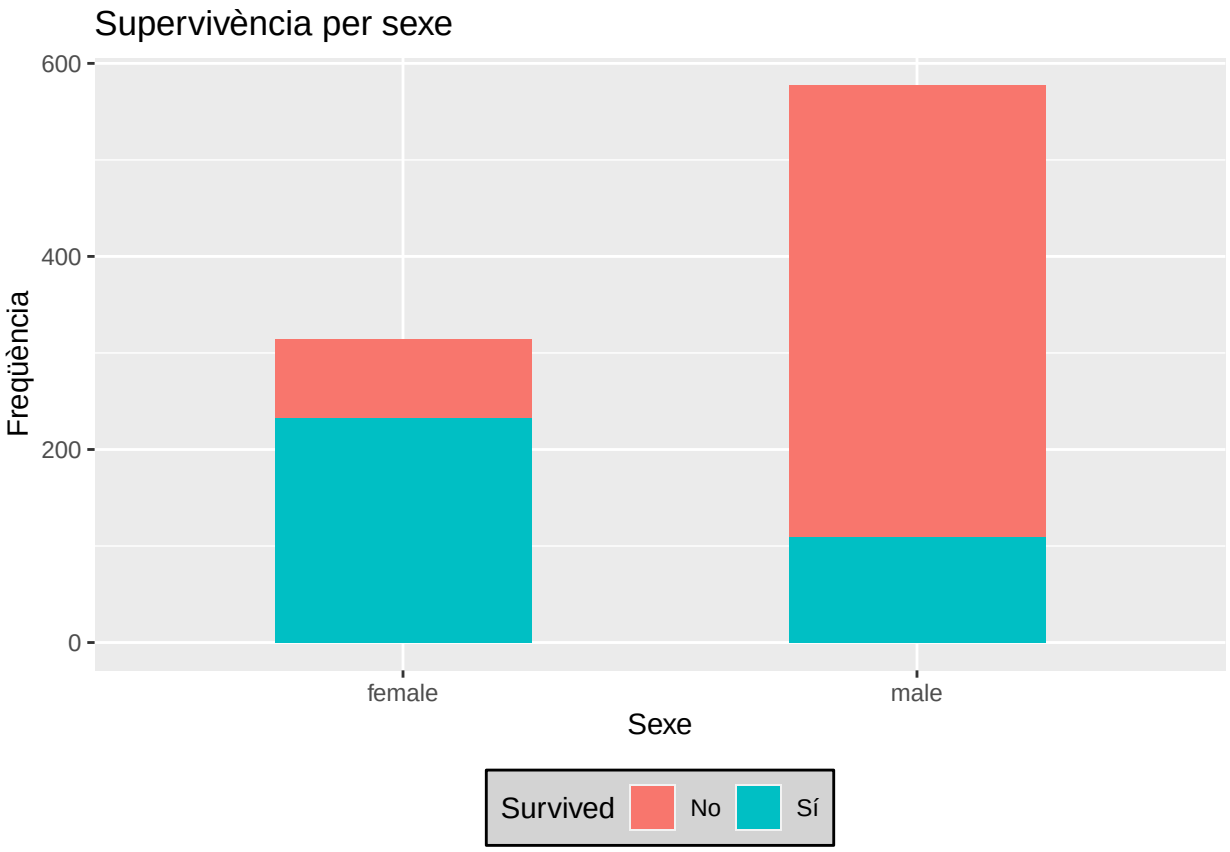
4.2 Gràfiques del conjunt d'entrenament

Distribució de la supervivència en relació al sexe, la classe i l'edat.

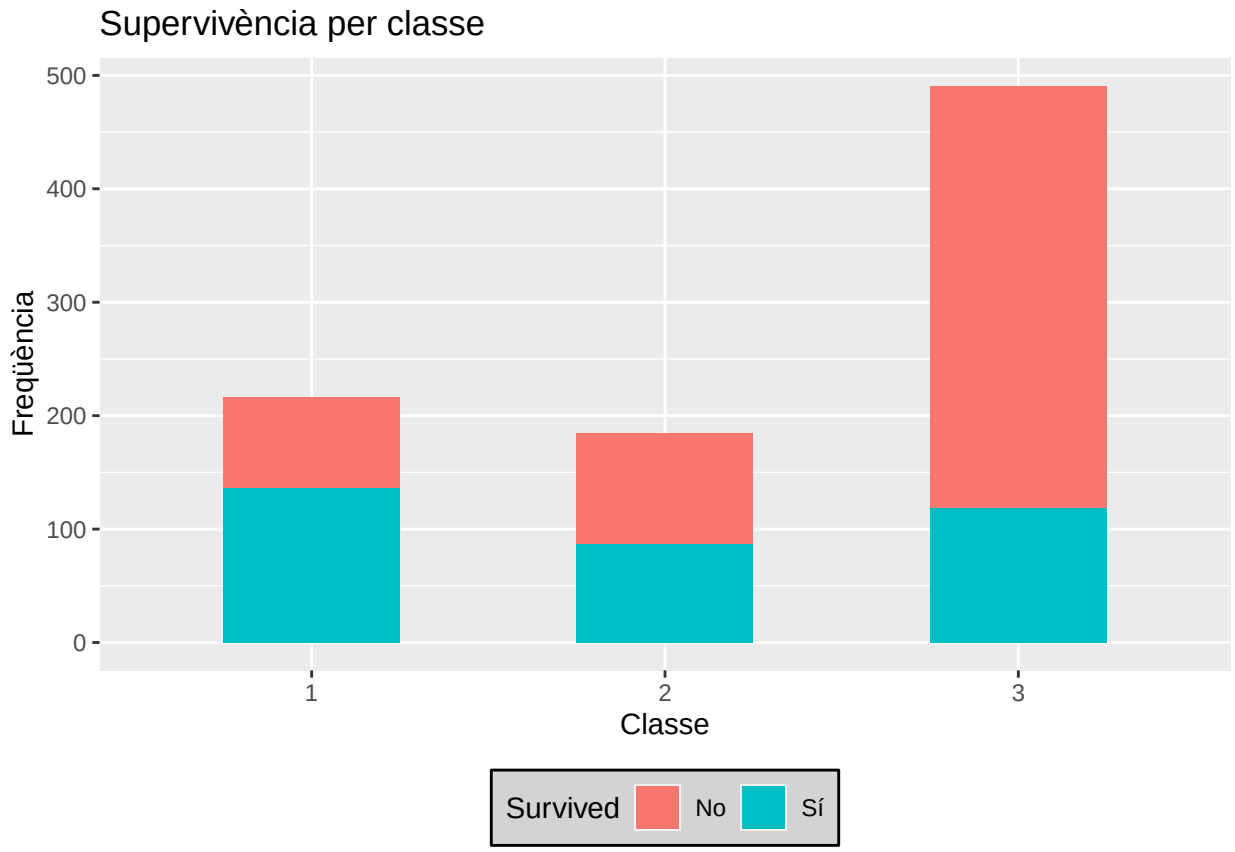
```r
ggplot(data = data,aes(x= Pclass,fill=Sex))+geom_bar(position="fill") +
  facet_wrap(~Age.factor) + labs(y= "Percentatge", x = "Classe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black"))+
  ggtitle("Sexe per classe i edat")
```



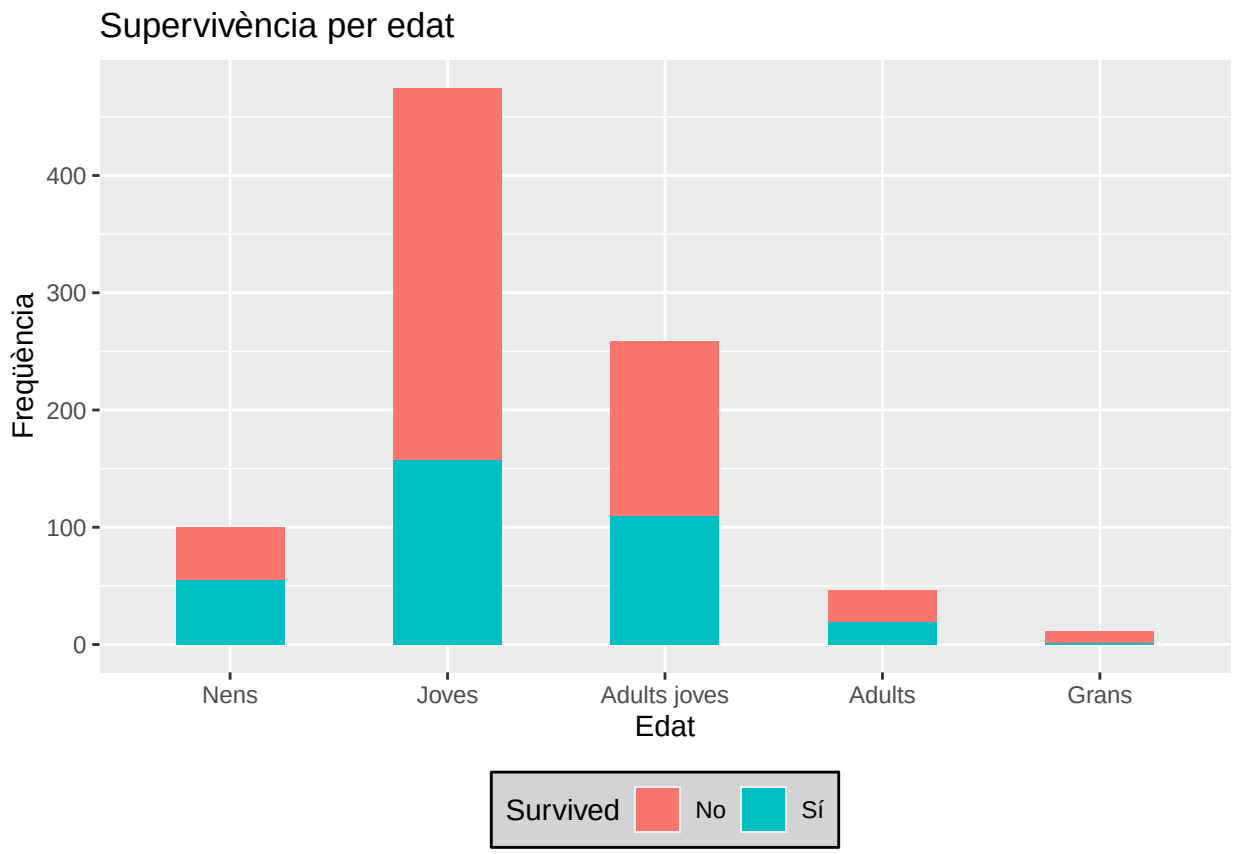

```
ggplot(data=data,aes(x=Sex,fill=Survived))+geom_bar(width=0.5)+
  labs(y= "Freqüència", x = "Sexe")+
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey", size=0.5,
                                         linetype="solid",colour ="black"))+
  ggtitle("Supervivència per sexe")
```




```
ggplot(data=data,aes(x=Pclass,fill=Survived))+geom_bar(width=0.5) +
  labs(y= "Freqüència", x = "Classe") +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey", size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència per classe")
```




```
ggplot(data=data,aes(x=Age.factor,fill=Survived))+geom_bar(width=0.5) +
  labs(y= "Freqüència", x = "Edat")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència per edat")
```




```
ggplot(data = data,aes(x=Sex,fill=Survived))+geom_bar(position="fill") +
  facet_wrap(~Pclass) + labs(y= "Percentatge", x = "Sexe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència per sexe i classe")
```

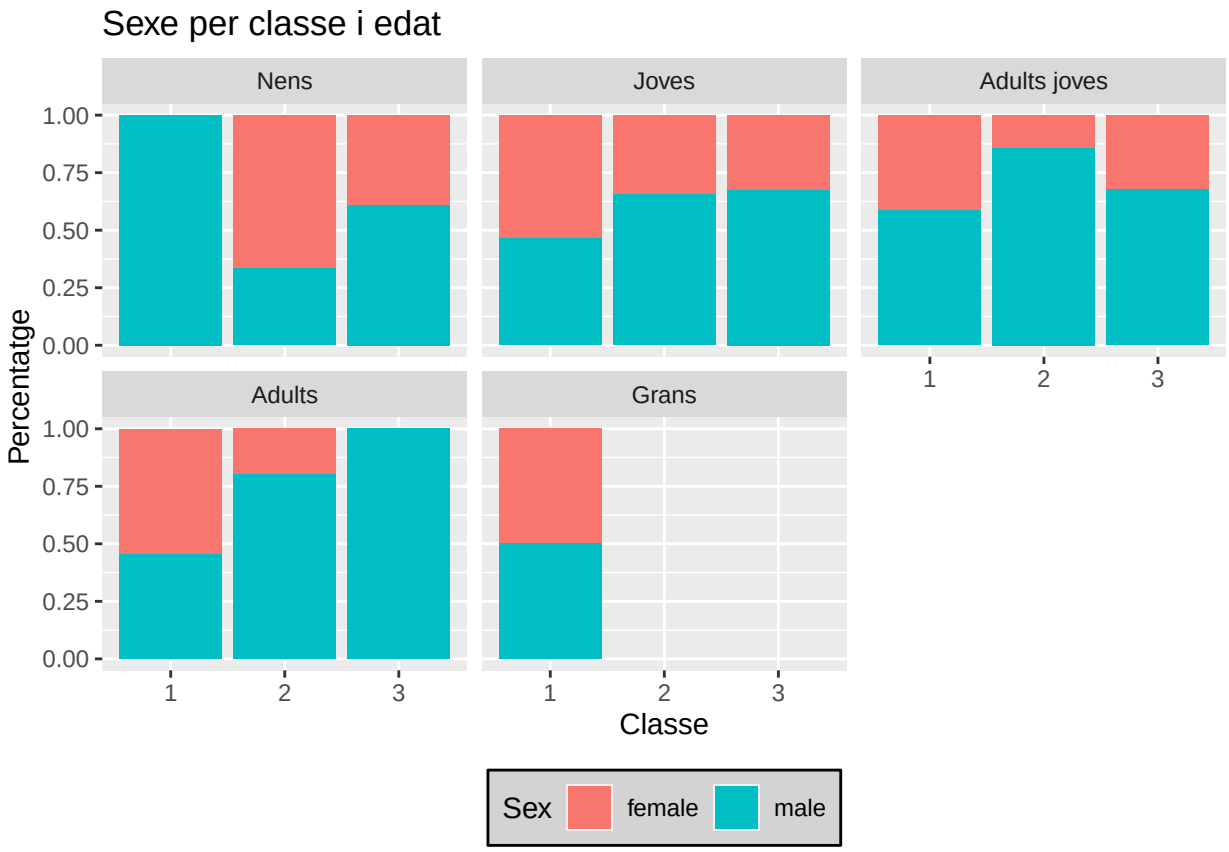

```
ggplot(data = data,aes(x=Pclass,fill=Survived))+geom_bar(position="fill") +
  facet_wrap(~Age.factor) + labs(y= "Percentatge", x = "Classe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència per classe i edat")
```


```
ggplot(data = data,aes(x=Sex,fill=Survived))+geom_bar(position="fill") +
  facet_wrap(~Age.factor) + labs(y= "Percentatge", x = "Sexe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència per sexe i edat")
```

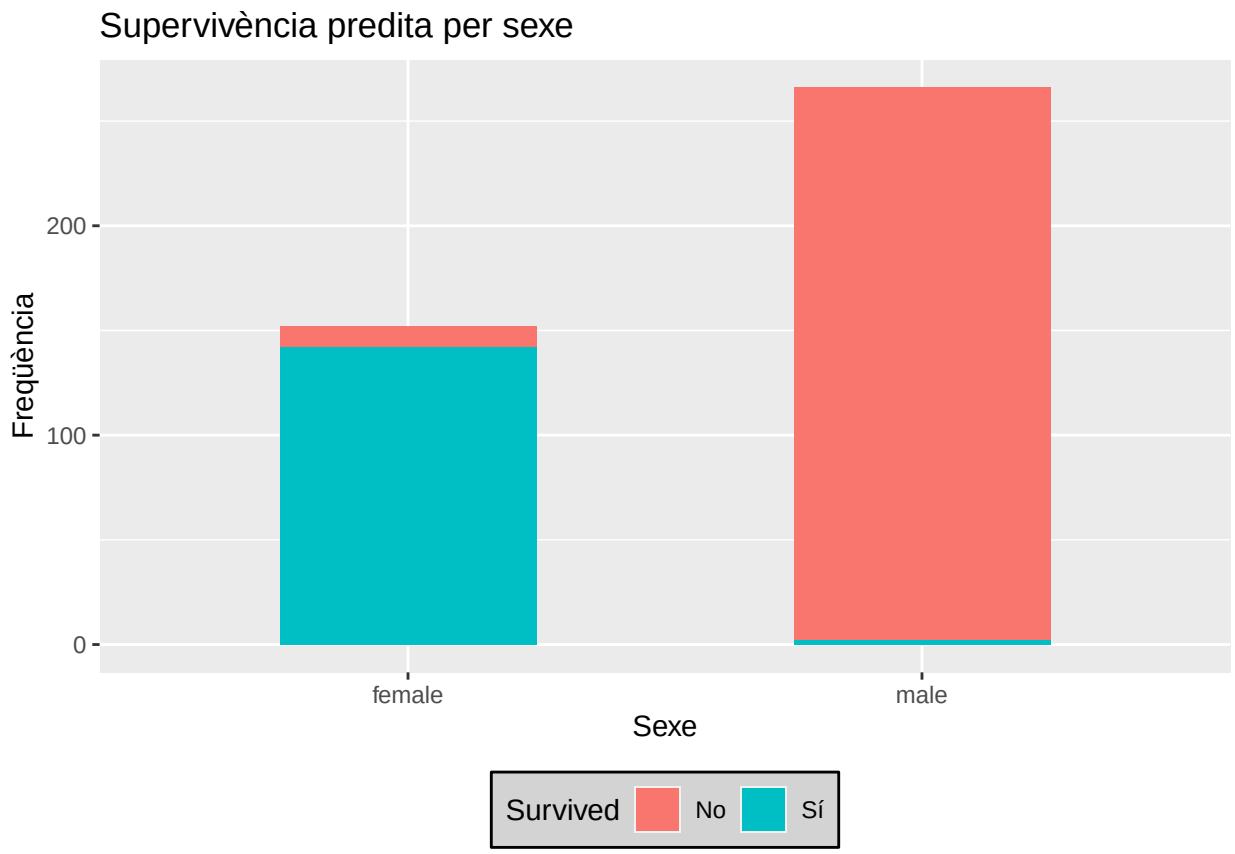
4.3 Gràfiques del conjunt de test amb la predicció

Distribució de la supervivència predita en relació al sexe, la classe i l'edat.

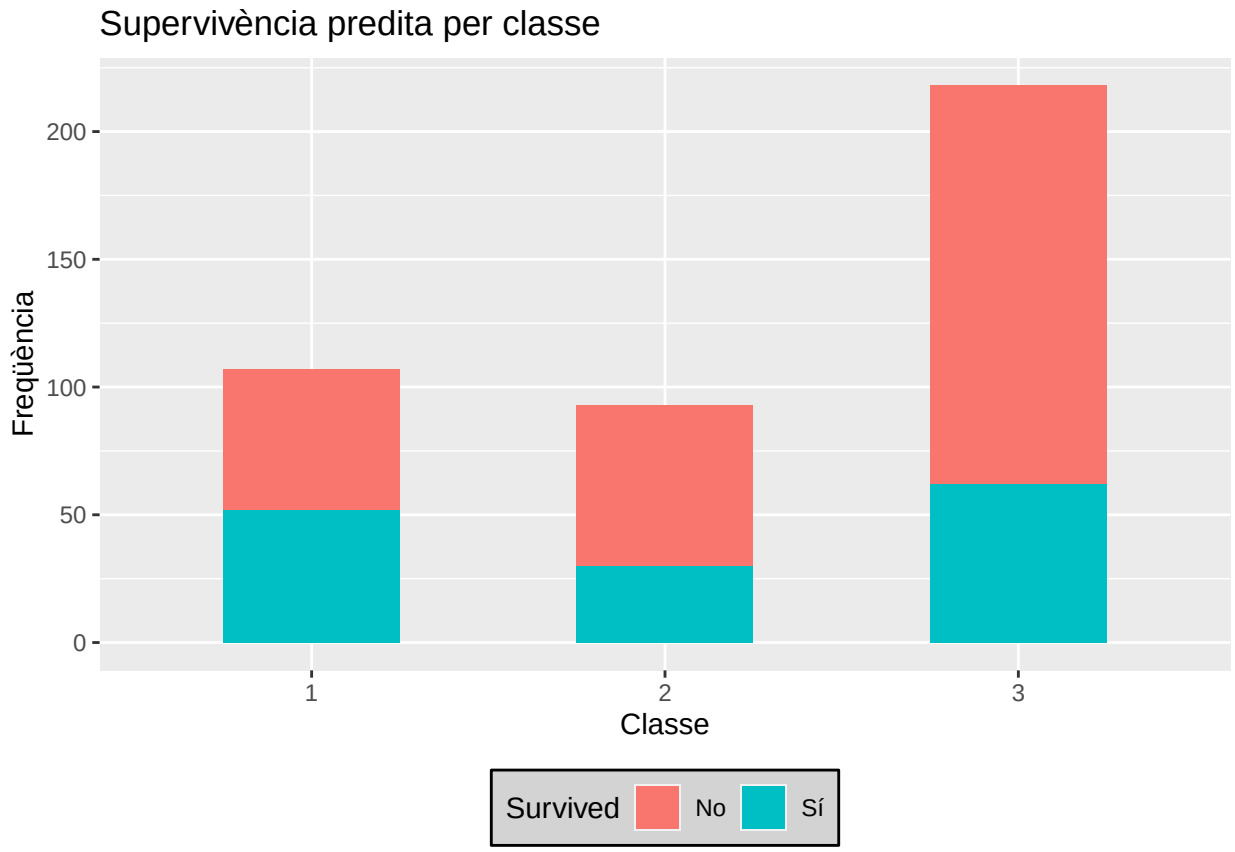
```
ggplot(data = data_test,aes(x= Pclass,fill=Sex))+geom_bar(position="fill") +
  facet_wrap(~Age.factor) + labs(y= "Percentatge", x = "Classe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Sexe per classe i edat")
```



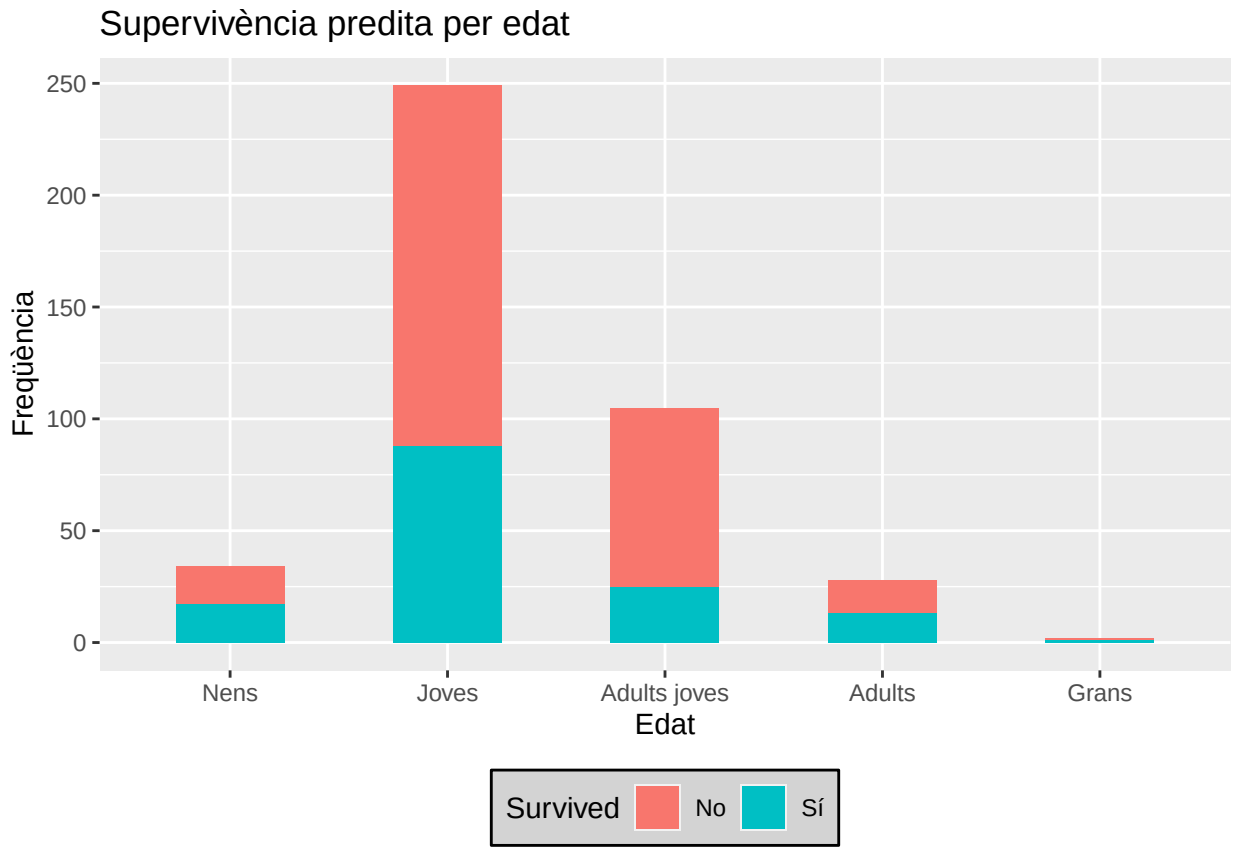

```
ggplot(data=data_test,aes(x=Sex,fill=Survived))+geom_bar(width=0.5) +
  labs(y= "Freqüència", x = "Sexe") +
  theme(legend.position="bottom"
        ,legend.background = element_rect(fill="lightgrey", size=0.5,
                                          linetype="solid",colour ="black")) +
  ggtitle("Supervivència predita per sexe")
```




```
ggplot(data=data_test,aes(x=Pclass,fill=Survived))+geom_bar(width=0.5) +
  labs(y= "Freqüència", x = "Classe") +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey", size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència predita per classe")
```

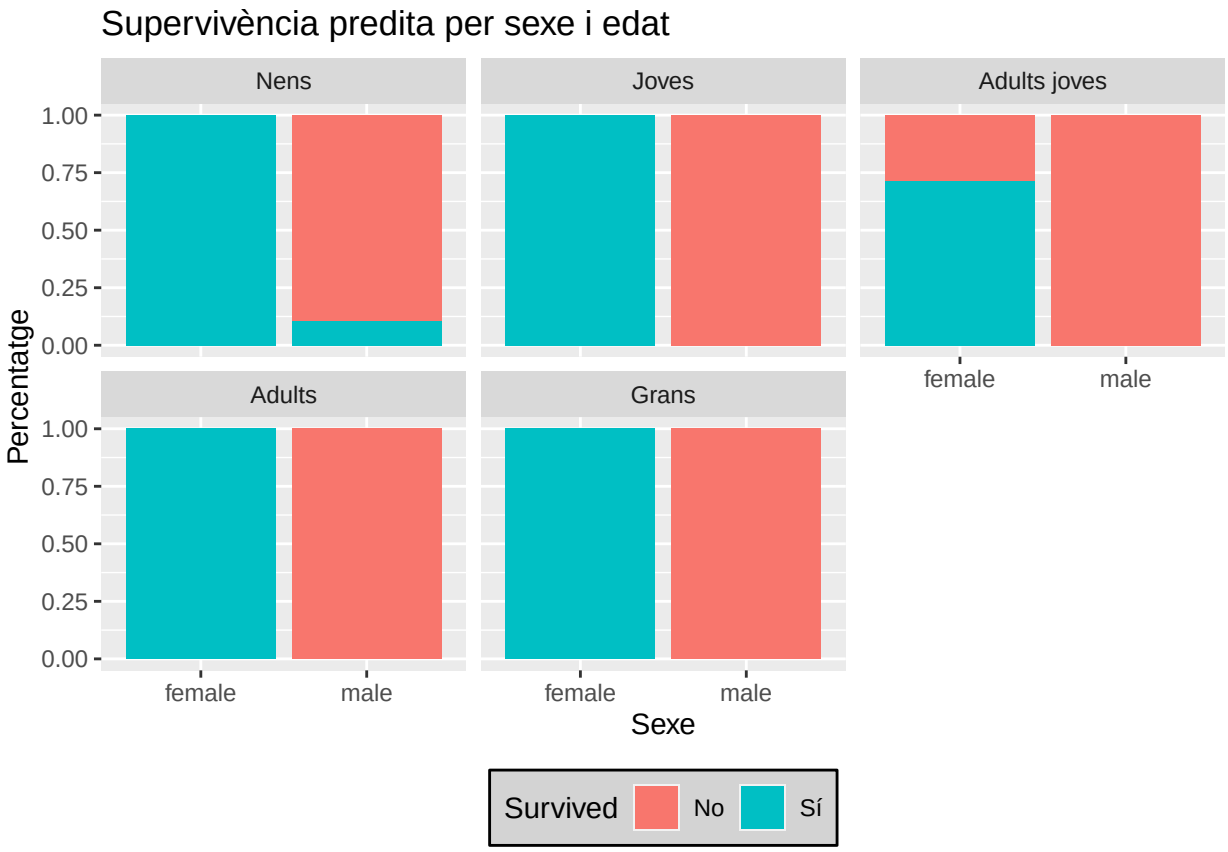



```
ggplot(data=data_test,aes(x=Age.factor,fill=Survived))+geom_bar(width=0.5) +
  labs(y= "Freqüència", x = "Edat")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència predita per edat")
```




```
ggplot(data = data_test,aes(x=Sex,fill=Survived))+geom_bar(position="fill") +
  facet_wrap(~Pclass) + labs(y= "Percentatge", x = "Sexe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència predita per sexe i classe")
```


```
ggplot(data = data_test,aes(x=Sex,fill=Survived))+geom_bar(position="fill") +
  facet_wrap(~Age.factor) + labs(y= "Percentatge", x = "Sexe")  +
  theme(legend.position="bottom",
        legend.background = element_rect(fill="lightgrey",size=0.5,
                                         linetype="solid",colour ="black")) +
  ggtitle("Supervivència predita per sexe i edat")
```



5 Fitxer final

Fitxer amb el conjunt d'entrenament i el de test amb les prediccions

```
# Unió de les dades originals amb les dades de test amb els resultats de la predicció
finalData <- bind_rows(data,data_test)

# Modificació de les etiqutes del factor edat per a que no es perdin els talls fets
levels(finalData$Age.factor) <- c("0-16", "17-31", "32-51", "52-64", "65+")

# Revisió de la integritat del fitxer
str(finalData)
```

```
## Classes 'data.table' and 'data.frame':   1309 obs. of  5 variables:
##  $ Survived  : Factor w/ 2 levels "No","Sí": 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 25 54 2 27 14 ...
##  $ Age.factor: Factor w/ 5 levels "0-16","17-31",..: 2 3 2 3 3 2 4 1 2 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# Generació de l'arxiu de sortida
write.csv(finalData, file = "titanic_final.csv", row.names = TRUE)
```

6 Fonts consultades

Chi-squared Test of Independence Recurs en línea: http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence

How do I interpret the AIC. 2018. [Web] R-blogger. Recurs en línea: https://www.r-bloggers.com/how-do-i-interpret-the-aic/

López Cano, Emilio. Ejemplo de Regresión Logística (modelo logit) con R.2017. [Web] RPubs. Recurs en línea: https://rpubs.com/emilopezcano/logit

Schratz, Patrick. Calculate Odds Ratios Of Generalized Linear (Mixed) Models. [Web] rdocumentation. Recurs en línea: https://www.rdocumentation.org/packages/oddsratio/versions/2.0.0/topics/or_glm