TITÀNIC

Josep Anton Charles

10 de juny de 2018

DETALLS DE L'ACTIVITAT.

Posar en pràctica la identificació de les dades rellevants per un projecte analític i revisar els processos relacionats amb el tractament de dades (integració, transformació, neteja i validació) per millorar la seva qualitat abans d'aplicar les diferents etapes d'anàlisi.

Els gràfics s'han presentat a mesura que eren necessaris i no s'ha creat un punt especific amb tots ells.

Objectius.

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis multidisciplinars.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació contiguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

1. DESCRIPCIÓ DEL DATASET.

Els registres d'aquest dataset representen els viatgers del Titànic. Les dades s'han aconseguit al següent enllaç de Kaggle.

En total es disposen de 1309 registres repartits en dos fitxers amb la mateixa estructura, un amb dades de 'training' (891 observacions) i un altre de dades de 'test' (amb 418 observacions). Les dades de training disposa de 418 registres amb valor NA per a la variable 'Survived'.

Aquest dataset és important perquè permet aprofundir en quines van ser les raons per les quals una part dels passatgers van sobreviure i una altra part no.

Camps d'aquest conjunt de dades:

```
* PassengerID:  identificatiu
* Name: Nom del passatger
* Survived: Indica si el passatger va sobreviure    (0=NO, 1= Si)
* Pclass:   Clase en la que viatjava    (1= primera, 2= segunda,
3=tercera)
* Sex:  Genere  (male=home, female= dona)
* Age:  Edat    (XX.5 estimada)
* Sibsp:    Numero de germans/ esposas
* Parch:    Número de parients a bord
* Ticket:   Número de ticket
* Fare: Preu del pasatge
* Cabin Número de cabina
* Embarked: Port d'embarcament  C = Cherbourg, Q = Queenstown, S =
Southampton
```

2. INTEGRACIÓ I SELECCIÓ DE LES DADES D'INTERÉS A ANALITZAR.

A partir d'aquest conjunt de dades es planteja la problemàtica de determinar quines variables influeixen més sobre el fet de sobreviure a l'enfonsament del Titànic.

Hi ha una sèrie de dades que semblen importants a l'hora d'intentar analitzar aquesta causa i efecte. A priori tenim: Survived, Pclass, Sex, Age, Fare i Embarked. Per tant descartem:

- PassangerID: és un identificador únic.
- Sibsp i Parch: Encara que participants de Kaggle les utilitzen per a la creació d'una variable "tamany familia", no acabo d'estar convençut amb la seva interpretació i tractament.
- Ticket: Alguns autors de Kaggle la utilitzen per analitzar la coberta en què viatjaven, però hi ha una quantitat molt important de valors nulls i no existeix un criteri fiable per reomplir-los.

Libreries utilitzades.

```
library(dplyr) # data manipulation
library(mice) #imputació de valors perduts a R
library(nortest) #anàlisi de normalitat
library(C50) #arbres de decisió
```

3. NETEJA DE LES DADES.

L'objectiu que perseguim amb aquest treball és determinar quines són les causes més rellevants de cara a determinar la Supervivència al naufragi. No estem imteressats a predir si un individu, en funció de les seves característiques sobreviurà o no.

Per aquesta raó només utilitzarem tots aquells registres sobre els que tenim el valor de la seva Supervivència. En conseqüència unim tots dos arxius (training i test) en un de sol, i ens quedem amb els valors dels que disposem la supervivència.

Podríem considerar que a l'actuar d'aquesta manera estem esbiaixant les dades. Considerem que el fet que una dada estigui a l'arxiu 'train' o 'test' és totalment aleatori, per la qual cosa descartem l'esbiaix.

Un cop units i seleccionats els registres, convertim les variebles character en factors i observem les dades i analitzem els valor NA.

```
train <- read.csv('train.csv', stringsAsFactors = F)
test  <- read.csv('test.csv', stringsAsFactors = F)

full  <- bind_rows(test,train) # bind training & test data

#---comprovem que les variables que són de tipus character, les convertim
a factors
full[sapply(full, is.character)] <- lapply(full[sapply(full,
is.character)], as.factor)

#---visualitzem les dades-------------------------------------
head(full)
```

```
##   PassengerId Pclass                                          Name
Sex
## 1         892      3                              Kelly, Mr. James
male
## 2         893      3              Wilkes, Mrs. James (Ellen Needs)
female
## 3         894      2                     Myles, Mr. Thomas Francis
male
## 4         895      3                              Wirz, Mr. Albert
male
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist)
female
## 6         897      3                    Svensson, Mr. Johan Cervin
male
##    Age SibSp Parch  Ticket    Fare Cabin Embarked Survived
## 1 34.5     0     0  330911  7.8292              Q       NA
## 2 47.0     1     0  363272  7.0000              S       NA
## 3 62.0     0     0  240276  9.6875              Q       NA
## 4 27.0     0     0  315154  8.6625              S       NA
```

```
## 5 22.0     1     1 3101298 12.2875             S       NA
## 6 14.0     0     0    7538  9.2250             S       NA

#---analitzem el tipus de dada de cada camp---------------------
sapply(full,function(x) class(x))

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

Anàlisis de valors perduts.

En total ens falta l'edat de 263 passatgers, 1 preu del billet i no sabem de 418 passatgers si van sobreviure o no.

```
#---comprovem el numero de valors desconeguts-------------------
sapply(full,function(x) sum(is.na(x)))

## PassengerId      Pclass        Name         Sex         Age
SibSp
##           0           0           0           0         263
0
##       Parch      Ticket        Fare       Cabin    Embarked
Survived
##           0           0           1           0           0
418
```

Sobreviure.

Tal com hem comentat al començament d'aquest apartat, l'objectiu d'aquest treball és analitzar les dades disponibles i en particular analitzar aquelles variables que més influeixen en la supervivència dels passatgers.

El que no volem és omplir les dades de la supervivència amb cap mètode, donat que segons el mètode utilitzat influirà de manera important en la nostra anàlisi. Així traiem del nostre dataset totes aquelles dades de les quals no disposem de si el passatger va o no sobreviure.

```
survivedInfo <- full %>%  filter(!is.na(Survived))
#---comprovem el numero de valors desconeguts-------------------
sapply(survivedInfo,function(x) sum(is.na(x)))

## PassengerId      Pclass        Name         Sex         Age
SibSp
##           0           0           0           0         177
0
```

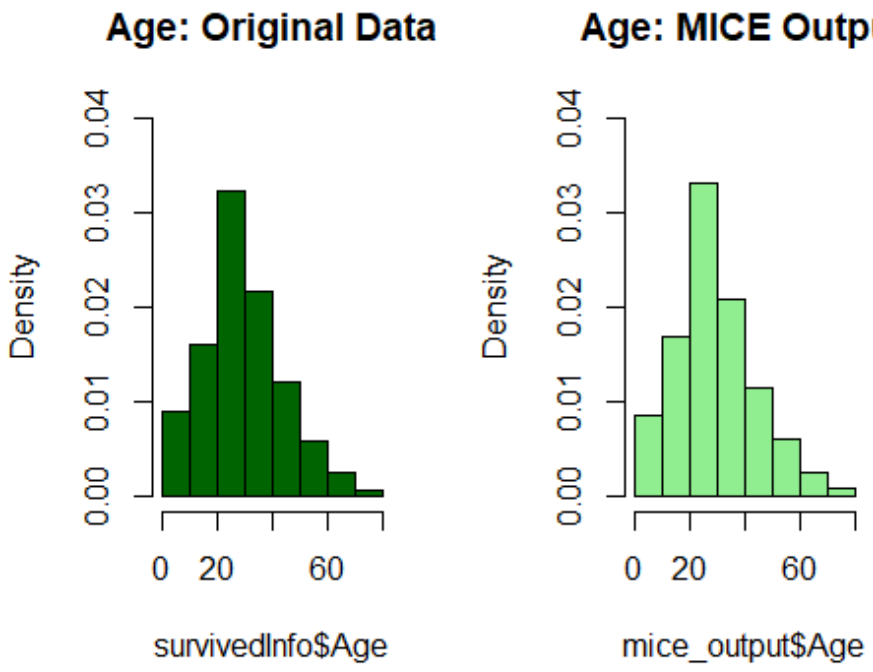
```
##       Parch      Ticket        Fare       Cabin    Embarked
Survived
##           0           0           0           0           0
0
```

Edat.

Ara només ens falta l'edat de 177 passatgers. Per determinar l'edat d'aquest passatger utilitzarem la llibreria 'mice'. Em baso en diferents aportacions generades a la competició de Kaggle

```
#---determinem una llavor----------------------
set.seed(129)

#---Utilitzem la imputació amb mice excloent variables no rellevants:
mice_mod <- mice(survivedInfo[, !names(survivedInfo) %in%
c('PassengerId','Name','Ticket','Cabin')], method='rf')

##
##  iter imp variable
##   1   1  Age
##   1   2  Age
##   1   3  Age
##   1   4  Age
##   1   5  Age
##   2   1  Age
##   2   2  Age
##   2   3  Age
##   2   4  Age
##   2   5  Age
##   3   1  Age
##   3   2  Age
##   3   3  Age
##   3   4  Age
##   3   5  Age
##   4   1  Age
##   4   2  Age
##   4   3  Age
##   4   4  Age
##   4   5  Age
##   5   1  Age
##   5   2  Age
##   5   3  Age
##   5   4  Age
##   5   5  Age

#---guardem l' output complert-------------------------
mice_output <- complete(mice_mod)

#---Comparem la distribució de l'edat original i la calculada-----------
```

```
--
#---Fem un plot de les distribucions de l'edat--------------
par(mfrow=c(1,2))
hist(survivedInfo$Age, freq=F, main='Age: Original Data',
     col='darkgreen', ylim=c(0,0.04))
hist(mice_output$Age, freq=F, main='Age: MICE Output',
     col='lightgreen', ylim=c(0,0.04))
```




```
par(mfrow=c(1,2))

#---Creem una variable nova per poder-les comparar.
survivedInfo$AgeModel <- mice_output$Age

#---analitzam la correlació entre totes dues variables------------------
----------
cor(survivedInfo$Age,survivedInfo$AgeModel, use='complete.obs')

## [1] 1

cor.test(survivedInfo$Age,survivedInfo$AgeModel, use='complete.obs')

##
##  Pearson's product-moment correlation
##
## data:  survivedInfo$Age and survivedInfo$AgeModel
## t = Inf, df = 712, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  1 1
## sample estimates:
## cor
##   1
```

Podem comprovar que totes dues gràfiques tenen una distribució similar. Hem posat els valors que ens ha donat el model 'mice' en una nova variable i analitzem la correlació entre edat original i la calculada. El coeficient de correlació és 1 i el seu p_valor=0 (em quedo amb h1: la correlació és diferent de 0).

Anàlisi dels valors en blanc a Embarked.

La variable Embarked té dos valors en blanc, els registres 62 i 830, són dones que van pagar 80 (fare).

Fem un plot dels imports del passatge per cada port d'embarcament. Amb aquestes dades podem determinar que el port d'embarcament més probable sigui C 'Cherbourg'.

```
#---escollim les variables rellevants en un nou dataframe
df<-survivedInfo[,c(2,4,9,11,12,13)]

#---quants regsitres tenim en blanc------
aggregate(Fare~Embarked, data=df, length)
```

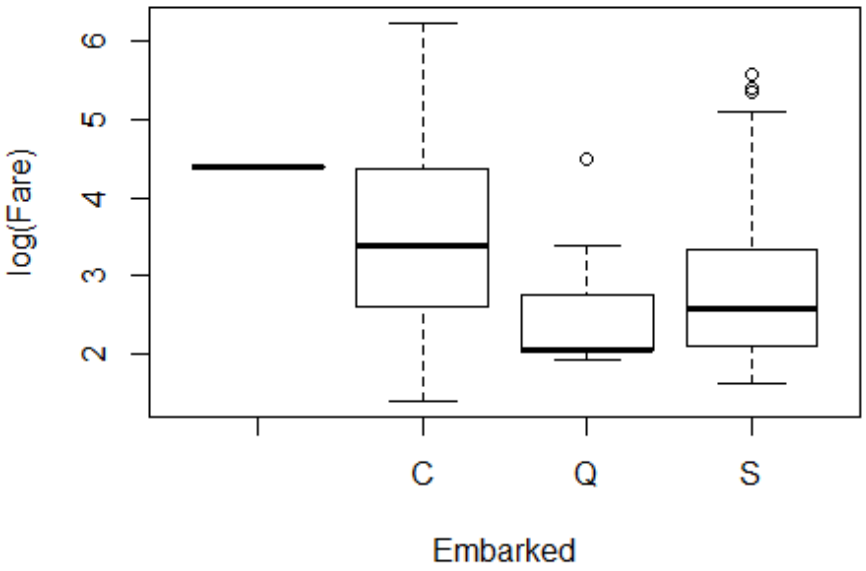
```
##   Embarked Fare
## 1             2
## 2        C  168
## 3        Q   77
## 4        S  644
```

```
#---quins números de registre són
df[df$Embarked=='',]
```

```
##     Pclass    Sex Fare Embarked Survived AgeModel
## 62       1 female   80                1       38
## 830      1 female   80                1       62
```

```
#---a quin port hauríem d'assignar-los
plot(log(Fare)~Embarked, data = df)
```

```
## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
z$out[z
## $group == : Outlier (-Inf) in boxplot 4 is not drawn
```




```
#---assignem els valors---------
df$Embarked[62]<-'C'
df$Embarked[830]<-'C'
df$Embarked<-factor(df$Embarked)
```

Anàlisi dels 0 a Fare.

La variable edat no disposa de 0, en canvi la variable Fare sí. Considerem que els 0 representa la falta d'una dada.

Creem un vector amb els índexs dels registres que tenen Fare =0. Veiem que tots són homes i embarcats a S 'Southampton'. Determinem la mitjana de l'import del passatge per a aquest tipus de registres i els assignen als que tenen valor 0.

```
#---elemnents del dataframe que tenen  valor 0 a Fare.-----
df[df$Fare==0,]
```

```
##     Pclass  Sex Fare Embarked Survived AgeModel
## 180      3 male    0        S        0       36
## 264      1 male    0        S        0       40
## 272      3 male    0        S        1       25
## 278      2 male    0        S        0       31
## 303      3 male    0        S        0       19
## 414      2 male    0        S        0       22
## 467      2 male    0        S        0       21
## 482      2 male    0        S        0       31
## 598      3 male    0        S        0       49
```

```
## 634      1 male    0        S        0       47
## 675      2 male    0        S        0       16
## 733      2 male    0        S        0       28
## 807      1 male    0        S        0       39
## 816      1 male    0        S        0       45
## 823      1 male    0        S        0       38

#---vector amb els index que tenen Fare = 0---------------
fare0<-which(df$Fare %in% 0)

#---calcul de les diferents mitjanes en funció de la classe i el port
d'embarcament----
mitjanas<-
aggregate(Fare~Pclass,df[df$Sex=='male'&&df$Embarked=='S',],mean)

#substitucio dels 0 per les mitjanes en funció de la classe i el port
d'embarcament----
for (i in 1:length(fare0)){
  #i=1
  n=fare0[i]
  fareClass<-df$Pclass[n]
  df$Fare[n]<-mitjanas$Fare[mitjanas$Pclass==fareClass]
}
```

Anàlisi d' outsiders.

Disposem de dues variables numèriques l'edat i l'import del passatge.

A l'edat ens apareixen edats per sobre del 70 anys, però considerem que encara que siguin persones grans podrien estar interessades a fer el viatge. Considerem que són correctes.

Respecte a l'import del passatge, la dispersió és molt gran, però podem pensar que dintre de la primera classe podrien haver-hi preus molts diferents en funció de nivells de luxe (a la pel·lícula de Di Caprio s'observa que determinats viatgers disposaven d'un nivell de vida molt elevat), per la qual cosa també acceptem els outsiders.

```
#---analisi d'outsiders-----------------------------------------------
-
col.names=colnames(df)
for (i in 1:ncol(df)) {
  if(i==1) cat('outsiders per variables\n')
  if (is.integer(df[,i]) | is.numeric(df[,i])) {
    cat(col.names[i])
    cat('\n')
    cat(boxplot.stats(df[,i])$out)
    cat('\n')
  }
}
```

```
## outsiders per variables
## Pclass
##
## Fare
## 71.2833 263 146.5208 82.1708 76.7292 80 83.475 73.5 263 77.2875
247.5208 73.5 77.2875 79.2 66.6 69.55 69.55 146.5208 69.55 113.275
76.2917 90 83.475 90 79.2 86.5 512.3292 79.65 84.15469 153.4625 135.6333
77.9583 78.85 91.0792 151.55 247.5208 151.55 110.8833 108.9 83.1583
262.375 164.8667 134.5 69.55 135.6333 153.4625 133.65 66.6 134.5 263
75.25 69.3 135.6333 82.1708 211.5 227.525 73.5 120 113.275 90 120 263
81.8583 89.1042 91.0792 90 78.2667 151.55 86.5 108.9 93.5 221.7792
106.425 71 106.425 110.8833 227.525 79.65 110.8833 79.65 79.2 78.2667
153.4625 77.9583 84.15469 69.3 76.7292 73.5 113.275 133.65 73.5 512.3292
76.7292 211.3375 110.8833 227.525 151.55 227.525 211.3375 512.3292 78.85
262.375 71 86.5 120 77.9583 211.3375 79.2 69.55 120 84.15469 84.15469
93.5 84.15469 80 83.1583 69.55 89.1042 164.8667 69.55 83.1583
## Survived
##
## AgeModel
## 66 71 70.5 70.5 80 71 80 70 70 74
```

Últims passos amb el dataframe.

Ara podem veure que ja tenim tot el dataset net. Fem les últimes modificacions.

```
#---comprovem de nou el número de valors desconeguts-------------------
sapply(survivedInfo,function(x) sum(is.na(x)))
```

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

```
#---convertim en factors les variables necessàries i altres modificacions
df$Pclass<-factor(df$Pclass)
df$Survived<-factor(df$Survived)
levels(df$Survived)<-c('died', 'survived')
df$Age<-df$AgeModel
df<-df[,-6]
```

Exportem les dades definitives.

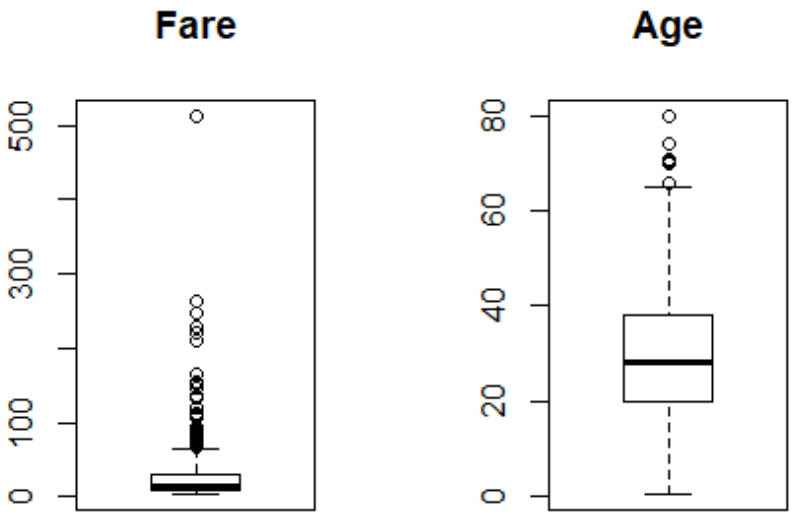
Una vegada que sobre el dataset original hem realitzat tots els procediments d'integració, validació i neteja, procedim a guardar les dades en un nou fitxer denominat "titanic_dataframe_clean.csx".

```
#---exportem les dades definitives-------------------------------------
-------
write.csv(df, file = 'titanic_dataFrame_clean.csv', row.names = F)
```

4. ANÀLISI DE LES DADES.

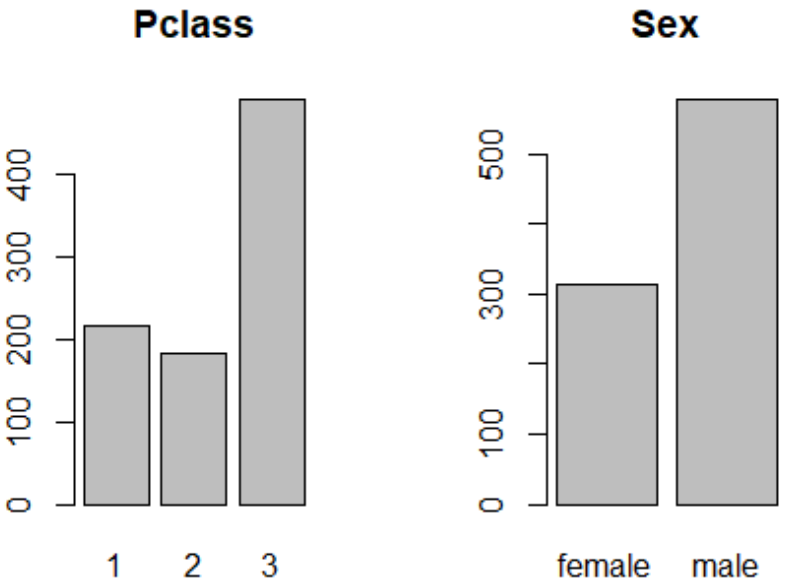
Comencem fent boxplots per les variables numèriques i barplots per les categòriques.

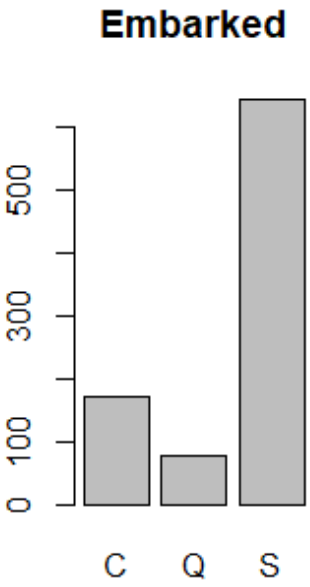
```
#--- grafic boxplot per les variable numeriques-------------------------
-----
col.names<-colnames(df)
par(mfrow=c(1,2))
for (i in 1:ncol(df)) {
  #i=1
  if(i==1) cat('boxplot per variables numeriques\n')
  if (!is.factor(df[,i])) {
    boxplot(df[,i], main=col.names[i])
  }
}

## boxplot per variables numeriques
```

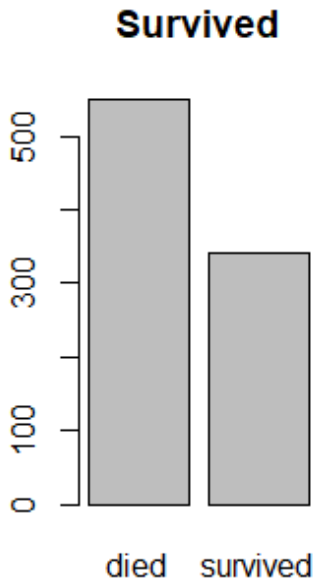



```
par(mfrow=c(1,1))

#---gràfics per variables categòriques---------------------------------
--------
```

```r
par(mfrow=c(1,2))
i=1
for (i in 1:ncol(df)) {
  if(i==1) cat('barplot per variables categoriques\n')
  if (is.factor(df[,i])) {
    barplot(table(df[,i]), main=col.names[i])
  }
}

## barplot per variables categoriques
```








```
par(mfrow=c(1,1))
```

Anàlisi de Normalitat per les variables numèriques.

Ara analitzem si les variables numèriques es comporten com una normal. El gràfic de la variable Fare ja ens indica que difícilment es comportarà como una Normal, Age té més possibilitats.

Apliquem l' **ad.test**. El resultat del test en confirma que cap variable és comporta com una normal.

```
#comprovació de la normalitat de les variables numèriques----------
----------

alpha=0.05 #en general utilitzarem aquest valor durant tot el treball
#i=1
for (i in 1:ncol(df)) {
  if(i==1) cat('variables que no segueixen una distribucio normal:\n')
  if (is.integer(df[,i]) | is.numeric(df[,i])) {
    p_valor=ad.test(df[,i])$p.value
    if (p_valor<alpha)
      cat(col.names[i])
    cat('\n')
  }
}

## variables que no segueixen una distribucio normal:
## Fare
## Age

#Podem comprovar que cap variable numèrica es comporta com a una Normal--
----------
```

Anàlisi de la diferència d'edat per sexe.

Volem comprovar si els homes i les dones eren de la mateixa edat.

La mitjana d'edat dels homes és de 30.6 i el de les dones 27.8. Veiem que hi ha certa diferència, la pregunta és si és significativa aquesta diferència.

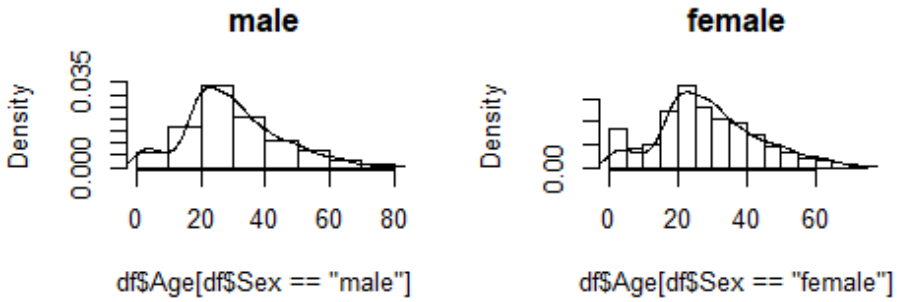
Encara que sabem que la variable edat no es comporta com una normal, suposarem que sí ho fa i aplicarem el **t.test**. El resultat del p-valor del t.test és 0.006475, al ser més petit que alpha (0.05) ens quedem amb la h1: Hi ha diferència entre les edats.

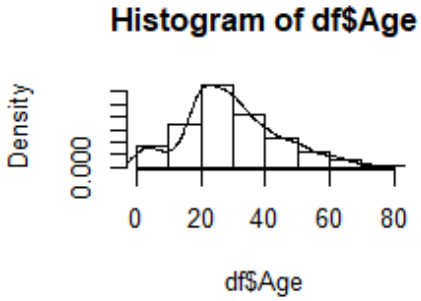
També podem comprovar la diferència de variança amb el **var.test**, que ens dóna un p-valor de 0.8818. Ens quedem amb la h0: el rati de variances és igual a 1 (són iguals).

Però com que sabem que no es comporta com una normal hem d'aplicar una prova no paramètrica, el **wilcox.test**. El resultat del seu p-valor és 0.01669, també inferior a alpha.

Podem concloure que **"ELS HOMES SON MÉS GRANS QUE LES DONES"**.

```r
#-----TWO SAMPLE TEST-------------------------------
#---Anem a comparar l'edat entre homes i dones

#analitzam els histogrames-----------------------
par(mfrow=c(2,2))
hist(df$Age[df$Sex=='male'], probability = T, main = 'male')
lines(density(df$Age))
hist(df$Age[df$Sex=='female'], probability = T, main = 'female')
lines(density(df$Age))
hist(df$Age, probability = T)
lines(density(df$Age))
par(mfrow=c(1,1))
```






```r
#---comparem les mitjanes--------------------------
mean(df$Age[df$Sex=='male'])
```

```
## [1] 30.59562
```

```r
mean(df$Age[df$Sex=='female'])
```

```
## [1] 27.82166
```

```r
#---Suposem Normalitat de la variable AGE-------------
t.test(Age~Sex, data = df)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  Age by Sex
## t = -2.7317, df = 639.12, p-value = 0.006475
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.7680443 -0.7798741
## sample estimates:
## mean in group female   mean in group male
##             27.82166             30.59562
```

```
#---anem a comparar les seves variances--------------
#h1: El ratio de la varianza es diferente de 1
var.test(Age~Sex, data = df)
```

```
##
##  F test to compare two variances
##
## data:  Age by Sex
## F = 1.0139, num df = 313, denom df = 576, p-value = 0.8818
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8366934 1.2356083
## sample estimates:
## ratio of variances
##           1.013881
```

```
#---En comptes de fer servir el t.test he de fer servir el wilcox.test
#h1:les dues medianas son diferents
wilcox.test(Age~Sex, data = df)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Age by Sex
## W = 81808, p-value = 0.01669
## alternative hypothesis: true location shift is not equal to 0
```

Anàlisi de la diferència d'edat per classe en què viatjaven

Analitzarem si hi ha diferència d'edat entre els pasatgers en funció de la classe en què viatgen. Utilitzarem l'anàlisi **ANOVA** sota l'assumpció que hi ha igualtat de variances.

Calculem la mitjana d'edat per classe. Ens dóna 38.8 anys en primera, 29,7 en segona i 25.5 en tercera. Sembla que els més grans viatgen en primera i el joves en 3a, la pregunta torna a ser si aquesta diferència és prou significativa.

Generem un regressió entre les dues variables i apliquem l'**ANOVA**. El p-valor ens dóna 0. Ens quedem amb h1: hi ha diferència d'edat entre els viatgers en funció de la classe.

La pregunta ara és entre quines classes la diferència és significativa. La regressió ens diu que la mitjana de la primera classe és el Intercept (38.8) i els coeficients dels altres

valors de la variable, la diferència respecte a aquella (-9.2 vs classe2 i -13.3 vs. classe3).

Per saber si las diferents diferències són significatives apliquem **pairwise.t.test** amb el mètode de **Bonferroni** i tots els p-valors són iguals a 0. La conclusió és que els més gran viatgen en primera i els més joves en 3a.

Tot això seria cert si les variances fossin iguals. Així que comprovem si existeix tal diferència de variances. Utilitzem el **bartlett.test** i el p_valor ens dóna 0.02519 ens quedem amb h1: hi ha diferències entre les variances.

Com que no podem assumir la igualtat de variances, hem de fer servir un test no paramètric **kruskal.test**. El p-valor d'aquest test és 0 i ens quedem amb h0: existeixen diferències entre les mitjanes.

Podem concloure que **"ELS MES GRANS VIATGEN EN PRIMERA I ELS MES JOVES EN TERCERA".**

```
#-----ANALISI DE LA VARIANÇA I KRUSHAL-WALIS TEST-----------------------
-----
#---ONE WAY ANALISIS-------------------------
#---calculem la mitjana d'edat per classe-------------------------------
------
aggregate(Age~Pclass, data=df,mean)

##   Pclass      Age
## 1      1 38.83528
## 2      2 29.68386
## 3      3 25.53853

#---creem un regressió entre totes dues variables i calculem l' ANOVA----
-----
anova(lm(Age~Pclass, data = df))

## Analysis of Variance Table
##
## Response: Age
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Pclass      2  26523   13262  73.283 < 2.2e-16 ***
## Residuals 888 160696     181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#---ara necessitem saber entre quins grups hi ha diferència-------------
---------
lm(Age~Pclass, data = df)

##
## Call:
## lm(formula = Age ~ Pclass, data = df)
##
```

```
## Coefficients:
## (Intercept)      Pclass2      Pclass3
##      38.835       -9.151      -13.297
```

```
pairwise.t.test(df$Age,df$Pclass,p.adjust.method = 'bonferroni')#tots el
p-valors=0
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$Age and df$Pclass
##
##   1       2
## 2 6.5e-11 -
## 3 < 2e-16 0.0011
##
## P value adjustment method: bonferroni
```

```
#conclusió: Existen diferències i els més grans viatjaven en 1a i els més
joves en 3a

#---comprovem la diferència de variances dintre dels grups: BARTLETT's
test--------------
bartlett.test(df$Age,df$Pclass)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df$Age and df$Pclass
## Bartlett's K-squared = 7.3627, df = 2, p-value = 0.02519
```

```
#---com que no podem assumir la igualtat de variances usem un test no
paramètric kruskal.test---
kruskal.test(df$Age,df$Pclass)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Age and df$Pclass
## Kruskal-Wallis chi-squared = 122.43, df = 2, p-value < 2.2e-16
```

Hi ha diferència d'edat per classe i supervivència?.

Ara el que ens interessa analitzar és si hi ha diferència significativa entre l'edat per classe i en funció de si van sobreviure.

Generem un gràfic interaction.plot per analitzar visualment les diferències.

Comencem fent una anàlisi paramètrica **ANOVA**. Tots dos coeficients de la regressió són significativament diferents a 0, per la qual cosa hi ha diferències entre les edats en

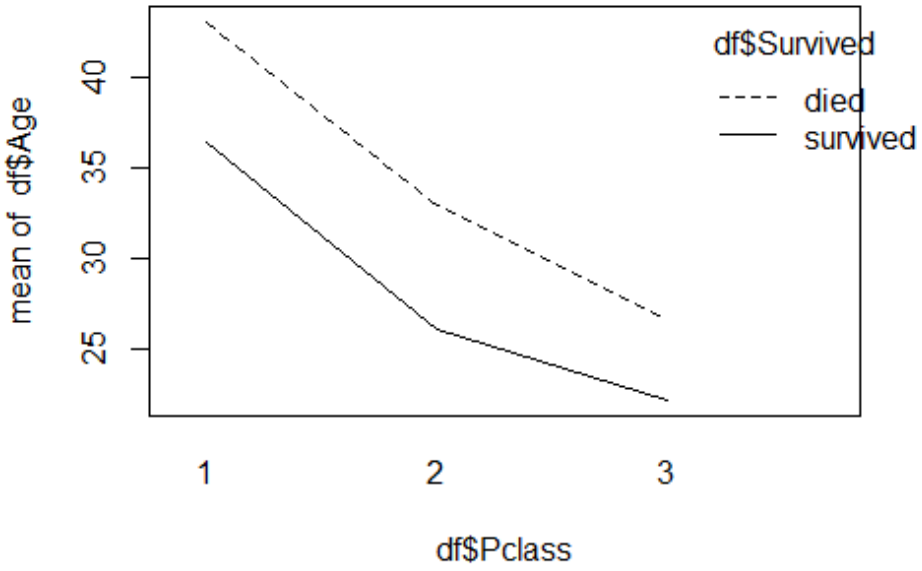
funció de la classe i la supervivència. Això seria així si poguéssim aplicar un test paramètric, però sabem que l'edat no es comporta com una Normal.

Com que no podem assumir la normalitat de l'edat, necessitem aplicar un test no paramètric **friedman.test**.

Primer hem de precalcular les dades. A la fórmula, a l'esquerra de '~' està la numèrica i a la dreta les variables factor. El resultat del p-valor de Friedman ens dóna 0.1353, més gran que alpha, ens hem de quedar h0: no hi ha diferència.

La conclusió: **"NO HI HA DIFERèNCIA SIGNIFICATIVA ENTRE L'EDAT EN FUNCIÓ DE LA CLASSE I LA SUPERVIVèNCIA"**.

```
#---TWO WAY ANALISIS DE LA VARIANÇA--------------------------
interaction.plot(df$Pclass, df$Survived, df$Age)
```




```
#---test parametric: ANOVA
anova(lm(Age~Pclass+Survived, data = df))

## Analysis of Variance Table
##
## Response: Age
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Pclass       2  26523 13261.5  75.984 < 2.2e-16 ***
## Survived     1   5887  5886.9  33.730 8.819e-09 ***
## Residuals 887 154809   174.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#---test no parametric: FRIEDMAN test
#---han de ser variables factor i s'han de precalcular les dades
agg<-aggregate(Age~Pclass+Survived, data = df, mean)
friedman.test(Age~Pclass|Survived, data = agg)

##
##  Friedman rank sum test
##
## data:  Age and Pclass and Survived
## Friedman chi-squared = 4, df = 2, p-value = 0.1353
```

Hi ha dependència entre la supervivència i els diferents sexes?.

Volem saber si la proporció entre el homes que es van salvar és diferent de la proporció de dones.

Generem primer una taula entre supervivència i sexe, es tracta d'una taula 2x2. Apliquem els tests de **fisher.test** i el de **chisq.test**. Tots dos tests ens donen un p-avlor de 0. Ens quedem h1: hi ha dependència.

La conclussio: **"HI HA DIFERENCIA ENTRE LA PROPORCIÓ DE PERSONES QUE ES SALVEN SI SE ES HOME O DONA"**.

```r
#-----TABULAR DATA-------------------------------------------
#---TWO INDEPENDENT PROPORTIONS-----------------

a<-table(df$Survived, df$Sex)#taula 2x2
fisher.test(a)

##
##  Fisher's Exact Test for Count Data
##
## data:  a
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0575310 0.1138011
## sample estimates:
## odds ratio
## 0.08128333

chisq.test(a)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  a
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Hi ha diferència en la classe en què es viatja segons el port d'embarcament?.

Volem saber si hi ha una diferència de proporció entre les classes en funció del port d'embarcament.

Creem una taula amb totes dues variables, és una taula 3x3. Calculem la **chisq.test** i el p-valor ens dóna 0. Ens quedem h1: hi ha dependència.

Si volem saber com contribueixen les variables a generar la dependència, determinem els valors esperats (esp) i els observats (obs) i calculem (obs-esp)^2/esp.

```
#--- R X C TABLES--------------------------

b<-table(df$Pclass,df$Embarked)
chisq.test(b)

##
##  Pearson's Chi-squared test
##
## data:  b
## X-squared = 127.01, df = 4, p-value < 2.2e-16

#per veure les diferencies---------------------
esp<-chisq.test(b)$expected
obs<-chisq.test(b)$observed
(obs-esp)^2/esp#els C en 1a tenen una gran contribució

##
##               C           Q           S
##   1 50.87168004 14.88095238  5.43196523
##   2  9.33868699 10.46722836  7.22965379
##   3  8.17935214 20.60376010  0.01002976
```

Correlacions.

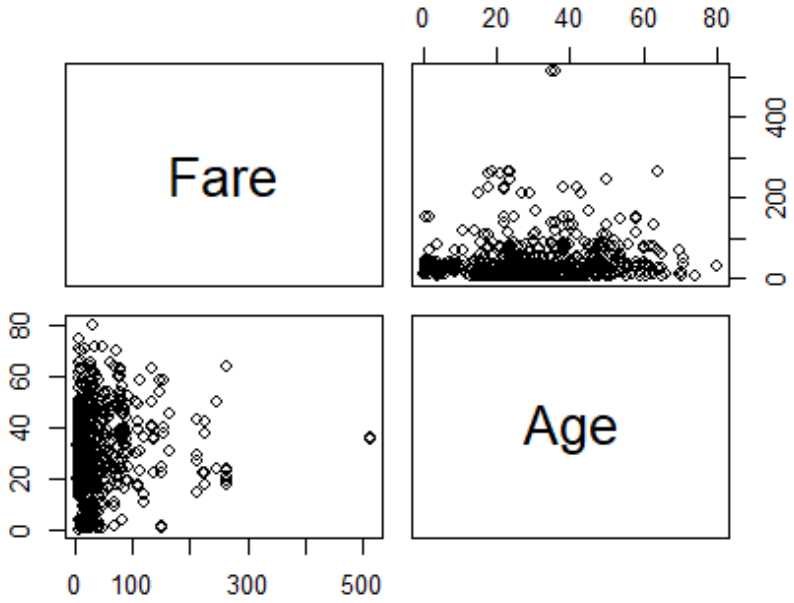
Quin tipus de correlació hi ha entre l'edat i l'import del passatge?.

Primer seleccionem les variables que són numèriques. Dibuixem un gràfic per analitzar la relació i calculem els coeficients de correlació.

Veiem que els nivells de correlació són baixos. Apliquem el test **cor** amb el métode de **'pearson'** per comprovar que són diferents de 0. El p-valor és de 0.004989. Ens quedem amb h1: la correlació és diferent de 0.

Però com que sabem que cap de les dues variables es comporten com una normal, hem d'aplicar proves no paramètriques: **'spearman'** i **'kendall'**. Tots dos p-valors són pròxims a 0, així que validem la h1: la correlació entre les dues variables és diferent de 0.

```r
df.numeric<-select_if(df, is.numeric)#selecció de les columnes que són
numeriques
pairs(df.numeric)
```




```r
#matriu de correlacions--------------------------------------
cor(df.numeric)#els nivells de correlació són molt baixos
```

```
##            Fare       Age
## Fare 1.0000000 0.1012841
## Age  0.1012841 1.0000000
```

```r
cor.test(df$Fare,df$Age, method = 'pearson')#p-valor=0--->h1: la
correlació és diferent de 0
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$Fare and df$Age
## t = 3.0355, df = 889, p-value = 0.002471
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03584518 0.16585824
## sample estimates:
##       cor
## 0.1012841
```

```r
#variants no parametriques-------------------------------------
cor.test(df$Fare,df$Age, method = 'spearman')#p-valor=0--->h1: la
correlació és diferent de 0
```

```
## Warning in cor.test.default(df$Fare, df$Age, method = "spearman"):
Cannot
## compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  df$Fare and df$Age
## S = 100490000, p-value = 9.652e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1476316
```

```r
cor.test(df$Fare,df$Age, method = 'kendall')#p-valor=0--->h1: la
correlació és diferent de 0
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$Fare and df$Age
## z = 4.4858, p-value = 7.262e-06
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1021208
```

Mirem si es pot generar una regressió entre totes dues variables.

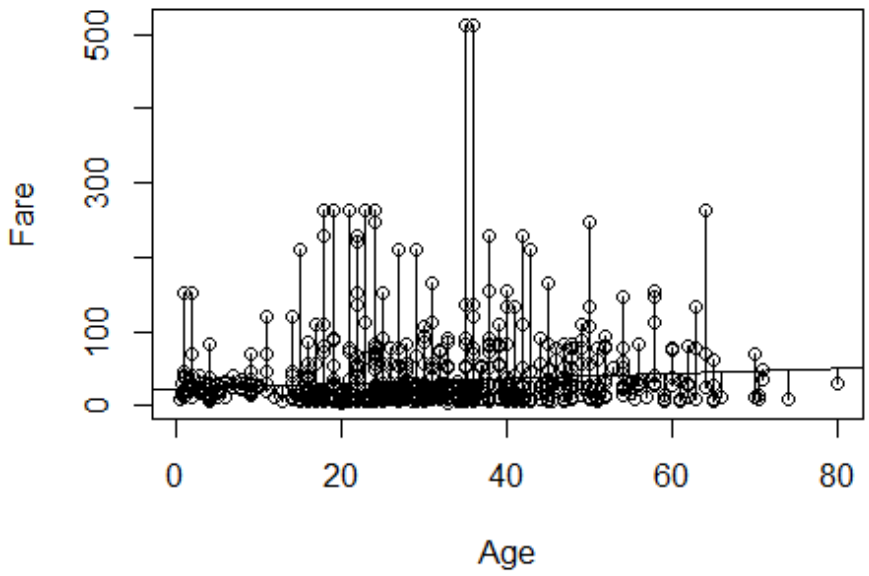
Primer generem la regressió entre les dues variables i analitzem els resultats amb la funció summary.

La conclusió és que els coeficients són diferents de 0 però la R^2 és molt baixa 0.008834. Ho veiem en el gràfic.

```r
#---SIMPLE LINEAR REGRESSION---------------
attach(df)
lmFareAge<-lm(Fare ~ Age, data=df)
summary(lmFareAge)#els coeficients són significatius però r2
ajustada=0.03
```

```
##
## Call:
## lm(formula = Fare ~ Age, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -40.50 -22.76 -16.79   2.00 477.58
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.5995     3.7695   5.995 2.95e-09 ***
## Age           0.3470     0.1143   3.036  0.00247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.46 on 889 degrees of freedom
## Multiple R-squared:  0.01026,    Adjusted R-squared:  0.009145
## F-statistic: 9.214 on 1 and 889 DF,  p-value: 0.002471

#---generem el gràfic-------------
plot(Fare~Age, data=df)
abline(lmFareAge)
segments(df$Age,fitted(lmFareAge),df$Age,df$Fare)
```



Models lineals generalitzats.

Volem saber quines variables influeixen més en la supervivència.

Apliquem el **model lineal generatitzat** per comprovar quines variables / valors tenen més pes (coeficients) a l'hora de determinar la supervivència.

Conclusió: **"EL FET DE SER HOME I VIATJAR EN 2a I 3a CLASSE AFAVOREIX LA NO SUPERVIVÈNCIA"**.

```
rl<-glm(Survived~Pclass+Sex+Age+Fare+Embarked,family=binomial('logit'))
summary(rl)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Fare + Embarked,
##     family = binomial("logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4575  -0.6785  -0.4028   0.6464   2.4495
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.7155560  0.4486886   8.281  < 2e-16 ***
## Pclass2     -1.0095740  0.2968950  -3.400 0.000673 ***
## Pclass3     -2.3434072  0.3013591  -7.776 7.48e-15 ***
## Sexmale     -2.5612592  0.1880918 -13.617  < 2e-16 ***
## Age         -0.0265561  0.0068230  -3.892 9.94e-05 ***
## Fare        -0.0008931  0.0021429  -0.417 0.676840
## EmbarkedQ    0.0287383  0.3701369   0.078 0.938113
## EmbarkedS   -0.5575369  0.2334985  -2.388 0.016952 *
## ---
## 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:  802.99  on 883  degrees of freedom
## AIC: 818.99
##
## Number of Fisher Scoring iterations: 5
```

Arbres de decisió.

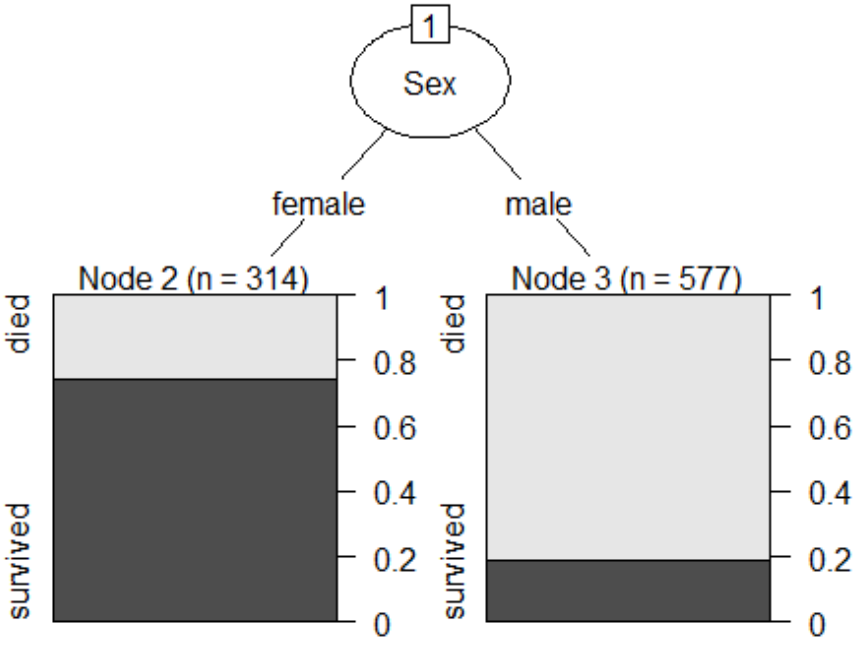
La variable més important per al model és la variable 'Sex'. El model només utilitza aquesta variable de les proposades i amb una única variable comet només un error del 21%.

Conclussió: **"EL SEXE ES DETERMINANT A LA HORA DE SOBREVIURE"**.

```
#--------------arbres de classificació----------------------------

model<-C5.0(Survived~Pclass+Sex, data=df, rules=F)
summary(model)
```

```
##
## Call:
## C5.0.formula(formula = Survived ~ Pclass + Sex, data = df, rules = F)
##
##
```

```
## C5.0 [Release 2.07 GPL Edition]      Sun Jun 10 21:29:05 2018
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 891 cases (3 attributes) from undefined.data
##
## Decision tree:
##
## Sex = female: survived (314/81)
## Sex = male: died (577/109)
##
##
## Evaluation on training data (891 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       2  190(21.3%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     468    81    (a): class died
##     109   233    (b): class survived
##
##
##  Attribute usage:
##
##  100.00% Sex
##
##
## Time: 0.0 secs

plot(model)
```



5. CONCLUSIÓ i RESOLUCIÓ DEL PROBLEMA.

Amb aquesta anàlisi hem pogut comprovar la dita de **"EN CAS DE NAUFRAGI, LES DONES (I ELS NENS) PRIMER**.

6. RECURSOS

1. Dalgaard, P. (2008). *Introductory statistics with R.* Springer Science & Business Media.
2. Megan Squire (2015). *Clean Data.* Packt Publishing Ltd. Chapter 1&2.