Tipologia i cicle de vida de les dades

Autor: Lluís Calvo i Aleix Sicília

13-05-2021

Introducció

Aquesta pràctica consisteix en la PAC2 de l'assignatura de Tipologia i cicle de vida de les dades en el Màster de ciència de dades de la Universitat Oberta de Catalunya

---

Pràctica 2 - Cas Titanic

Descripció del dataset

(2) Aquest conjunt de dades forma part de la competició de Kaggle del Titanic. El 15 d'abril de 1912 va haver l'enfonsament del creuer Titànic. Desafortunadament, van morir 1502 de 2224 del total de tripulants (comptant passatgers i oficials). Encara que hi hagués un component de sort aleatòria de sobreviure al desastre, és àmpliament difós que alguns grups de persones tenien més probabilitats que d'altres. Aquest dataset s'extreu del portal de competició de projectes d'estadística Kaggle i tracta d'una mostra reduïda d'informació dels passatgers d'aquest tràgic viatge. En aquest projecte s'intentarà observar quins són els tipus de persones van tenir més probabilitats de supervivència? Al respondre aquesta pregunta podrem interpretar com va ser el succés i es poden extreure conclusions per tal d'evitar accidents de cara al futur.

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

Primerament instal·lem i carreguem les llibreries necessàries: * ggplot2 * dplyr * mlbench * MASS * pROC * randomForest * tidyverse * lsr

Primerament, instal·lem i carreguem les llibreries ggplot2 i dplry

```
# Llibreries utilitzades
packages <- c('ggplot2', 'dplyr', 'mlbench', 'MASS', 'pROC',
'randomForest', 'tidyverse', 'lsr', 'psych', 'ggthemes', 'tinytex',
'qqPlot')
# Instal·lació packets encara no instal·lats.
installed_packages <- packages %in% rownames(installed.packages())
```

```
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages], repos =
"http://cran.us.r-project.org")
}
library('ggplot2')
library('dplyr')
library('mlbench')
library('MASS')
library('pROC')
library('randomForest')
library('tidyverse')
library('lsr')
library('psych')
library('ggthemes')
library('tinytex')
```

Carreguem el fitxer de dades de train i test. Ens assegurarem de tenir mostres úniques.

```
train <- unique(read.csv('data/train.csv',stringsAsFactors = FALSE))
test <- unique(read.csv('data/test.csv',stringsAsFactors = FALSE))
filas=dim(train)[1]
filas=dim(test)[1]
```

Verifiquem l'estructura del joc de dades principal de train

```
str(train)
```

```
## '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)" "Heikkinen, Miss. Laina" "Futrelle,
Mrs. Jacques Heath (Lily May Peel)" ...
##  $ 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" ...
```

```
names(train)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

Verifiquem l'estructura del joc de dades principal de test

```
str(test)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen
Needs)" "Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

Veiem que tenim 891 registres de train que es corresponen als viatgers i tripulació del Titànic i 11 variables que els caracteritzen. En el model de test observem 418 observacions, tot i que no hi ha informació sobre si ha sobreviscut i és el que hem de predir. La resta de variables són iguals. No es poden integrar al ser conjunts de dades diferents (falta la columna "Survived"), per tant no es pot aplicar un merge ja que els conjunts de test ens falta informació.

Revisem la descripció de les variables contingudes al fitxer i si els tipus de variable es correspon al que hem carregat:

PassangerId (1) int amb el codi del passatger. No farà falta tenir-los en compte ja que no ens dona informació descriptiva sinó que es tractarà del codi de cada mostra

Survived (2) int, es tracta de si ha sobreviscut un tripulant o viatger o no. Al tractar-se d'una variable categòrica l'haurem de convertir a variable categòrica (factor).

Pblcass (3) int, indica la classe del passatger. Hi ha tres classes (1-3, sent 1 primera classe). Al tractar-se d'una variable categòrica l'haurem de convertir a variable categòrica (factor).

Name (4) string, indica el nom de cada passatger o tripulant. No és informació a tenir en compte de cara a l'anàlisis.

Sex (5) chr, indica el sexe de la persona (male/female). Al tractar-se d'una variable categòrica l'haurem de convertir a variable categòrica (factor).

Age (6) num Indica l'edat de cada persona. Hi ha edats amb decimals per tant no pot ser enter. Es discretitzarà per grups d'edat de múltiples de 10 anys per tal d'extreure informació més interpretable.

SibSp (7) int, indica si té germanes o esposes a bord.Al tractar-se d'una variable categòrica l'haurem de convertir a variable categòrica (factor).

Parch (8) int, indica si té pares o germanes a bord.Al tractar-se d'una variable categòrica l'haurem de convertir a variable categòrica (factor).

Ticket (9) chr, indica el codi de ticket que tenia cada passatger. No és informació rellevant per aquest cas.

Fare (10) num, indica el valor del ticket en dólars.

Cabin (11) chr, indica el número de cabina amb el que es va viatjar. No és informació rellevant per aquest cas.

Embarked (12) int, indica el port d'embarque (c=Cherbourg, Q=Queenstown, S=Southampton). Al tractar-se d'una variable categòrica l'haurem de convertir a variable categòrica (factor).

Per tal de corregir el dataset caldrà negligir certs atributs no necessaris com ara "cabin", "passangerId","Name", "Ticket", ja que donen informació massa general o no és informació rellevant.

```
df_titanic <- train[ -c(1,4,9,11) ]
```

Una vegada corregit, haurem de discretitzar certs camps. Per exemple, d'una manera senzilla haurem de categoritzar les columnes de "Survived", "Pblcass", "Sex", "SibSp", "Parch", "Embarked". També es podria discritzar "Age" per segments d'edat en múltiples de 10, però no es farà per tal de poder corregir els possibles valors buits a posteriori.

Abans de categoritzar, però tractarem els valors

```
summary(df_titanic)
```

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

```
df_titanic$Embarked[is.na(df_titanic$Embarked)] <- "A"
summary(df_titanic)
```

```
##     Survived          Pclass          Sex                 Age       
##  Min.   :0.0000   Min.   :1.000   Length:891         Min.   : 0.42  
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   1st Qu.:20.12
```

```
##  Median :0.0000   Median :3.000   Mode  :character   Median :28.00
##  Mean   :0.3838   Mean   :2.309                      Mean   :29.70
##  3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:38.00
##  Max.   :1.0000   Max.   :3.000                      Max.   :80.00
##                                                      NA's   :177
##      SibSp           Parch             Fare          Embarked
##  Min.   :0.000   Min.   :0.0000   Min.   :  0.00   Length:891
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:  7.91   Class :character
##  Median :0.000   Median :0.0000   Median : 14.45   Mode  :character
##  Mean   :0.523   Mean   :0.3816   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   Max.   :512.33
##
```

```
df_titanic$Embarked[df_titanic$Embarked == ""] <- "Desconegut"
df_titanic$Embarked[is.na(df_titanic$Embarked)] <- "Desconegut"
str(df_titanic)
```

```
## 'data.frame':    891 obs. of  8 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 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: chr  "S" "C" "S" "S" ...
```

```
summary(df_titanic)
```

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

Després passem a discretitzar Age per diferents segments d'edat.

```
cols <- c("Survived", "Pclass", "Sex", "SibSp", "Parch", "Embarked")
df_titanic[cols] <- lapply(df_titanic[cols], factor)
```

```
summary(df_titanic)
```

```
##  Survived Pclass      Sex           Age        SibSp   Parch
Fare
##  0:549    1:216   female:314   Min.   : 0.42   0:608   0:678   Min.
:  0.00
##  1:342    2:184   male  :577   1st Qu.:20.12   1:209   1:118   1st
Qu.:  7.91
##           3:491                Median :28.00   2: 28   2: 80   Median
: 14.45
##                                Mean   :29.70   3: 16   3:  5   Mean
: 32.20
##                                3rd Qu.:38.00   4: 18   4:  4   3rd
Qu.: 31.00
##                                Max.   :80.00   5:  5   5:  5   Max.
:512.33
##                                NA's   :177     8:  7   6:  1
##        Embarked
##  C          :168
##  Desconegut:  2
##  Q          : 77
##  S          :644
##
##
##
```

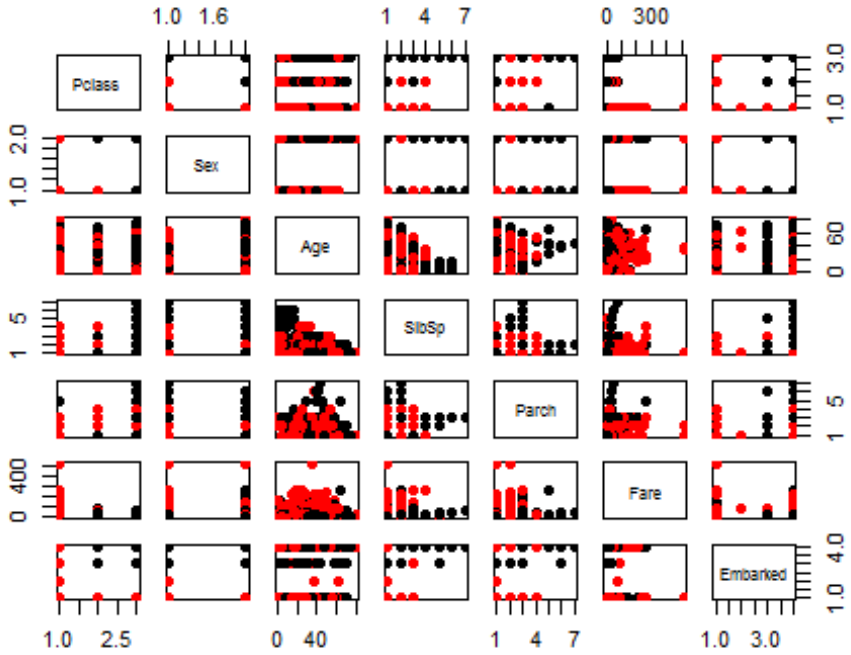
Observem les dades discretitzades

```
head(df_titanic)
```

```
##   Survived Pclass    Sex Age SibSp Parch    Fare Embarked
## 1        0      3   male  22     1     0  7.2500        S
## 2        1      1 female  38     1     0 71.2833        C
## 3        1      3 female  26     0     0  7.9250        S
## 4        1      1 female  35     1     0 53.1000        S
## 5        0      3   male  35     0     0  8.0500        S
## 6        0      3   male  NA     0     0  8.4583        Q
```

També es veurà un scatterplot per tal de veure si hi ha variables correlacionades. En vermell s'explicita si una persona ha sobreviscut o no (en vermell han sobreviscut).

```
# Basic Scatterplot Matrix
 pairs(~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, col =
factor(df_titanic$Survived), pch = 19, data = df_titanic)
```



Neteja de les dades.

Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

En primer lloc, hem de veure les estadístiques de valors buits tant desconegudes com sense valor.

```
# Estadístiques valors NA
colSums(is.na(df_titanic))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare
Embarked
##        0        0        0      177        0        0        0
0
```

```
# Estadístiques valors buits
colSums(df_titanic=="")
```

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

Al només haver dos observacions on no es té informació del port d'embarcament. Caldrà a tenir en compte que es conservaran els valors originals ja que realment no es

sabrà de quin port van sortir per falta d'informació. Per altra banda, s'assigna la mitjana per a valors buits de la variable "age"

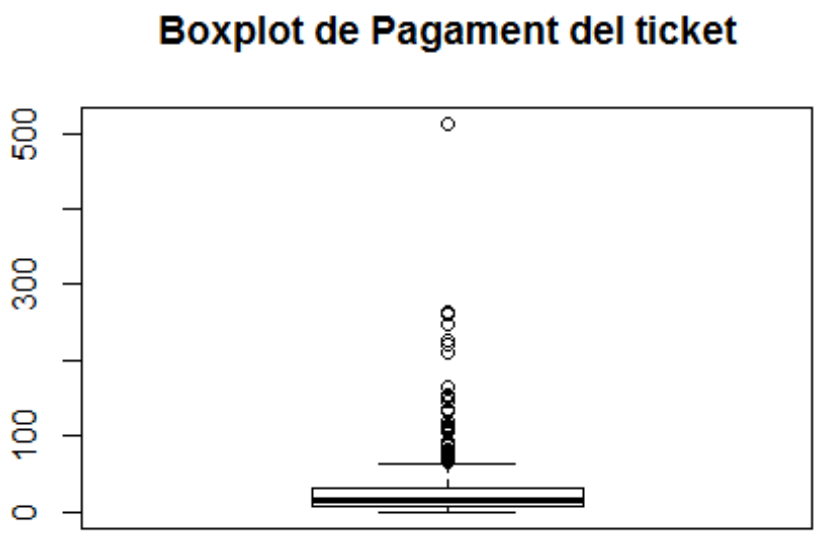
```
df_titanic$Age[is.na(df_titanic$Age)] <- mean(df_titanic$Age,na.rm=T)
```

Es pot observar com hi ha 177 observacions sense valors en el camp d'edat. És probable que molts viatgers no volguessin explicitar la seva edat. Es corregirà aquest camp mitjançant la mitjana de l'edat. També s'ha observat que hi ha dos observacions sense port d'origen.

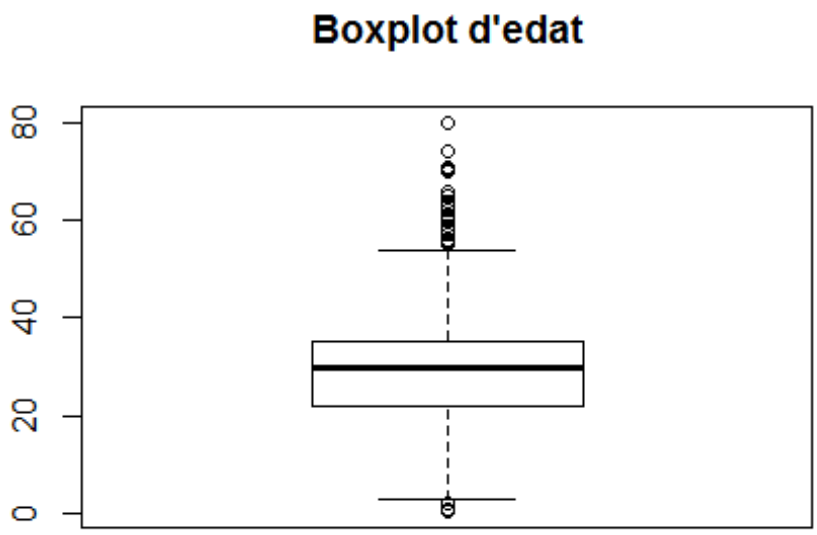
Identificació i tractament de valors extrems.

A continuació, es detallarà un tractament d'outliers. Primerament veurem un boxplot de la variable "Fare" i "Age"

```
titanic.bp<-boxplot(df_titanic$Fare,main="Boxplot de Pagament del
ticket")
```




```
titanic.bp<-boxplot(df_titanic$Age,main="Boxplot d'edat")
```

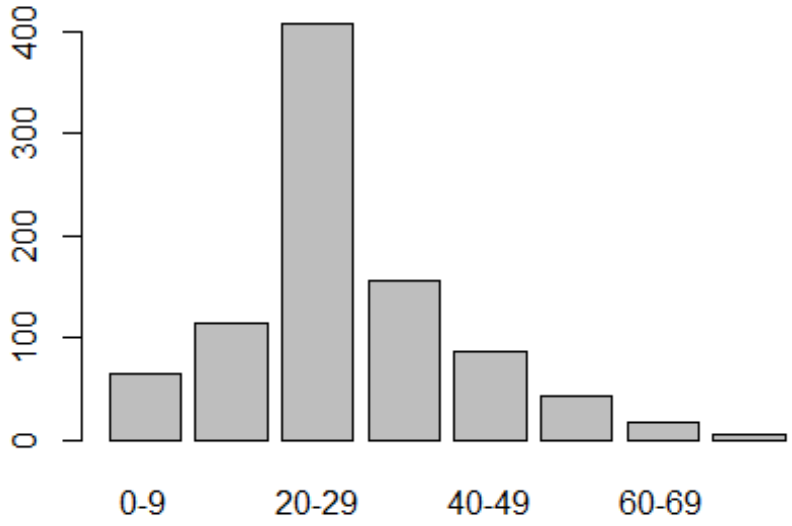


En el cas de la de l'edat podem discretitzar en rangs múltiples de 10

```
df_titanic["segment_edat"] <- cut(df_titanic$Age, breaks =
c(0,10,20,30,40,50,60,70,100), labels = c("0-9", "10-19", "20-29", "30-
39","40-49","50-59","60-69",">70"))
```

Veiem com s'agrupaven per segments d'edat i observem un predomini del grup d'edat entre 20 i 29 anys.

```
plot(df_titanic$segment_edat)
```



De la discretització de l'edat observem que realment la gent que viatjava era molt jove. El segment més gran erà de 20 a 29 anys. També veiem de la joventut de la tripulació. A continuació, caldrà tractar els valors extrems que van pagar unes taxes molt elevades. Ho analitzarem veient aquells valors que tenen una desviació estàndard superior a 3

```
# Criteri de 3 Desviacions estàndards (+/-3 SD)
Fare.outlier <- abs(scale(df_titanic$Fare)) > 3
colSums(Fare.outlier=="TRUE")

## [1] 20
```

En aquest cas hi ha 20 valors outliers. De moment, es conservaran en el model ja que estadísticament poden entrar (es tracta d'un 2,2% del total de les dades de train).

Anàlisis de les dades

Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

En primer lloc hem de veure un anàlisis descriptiu del conjunt de dades ja tractat.

```
summary(df_titanic)

##  Survived Pclass      Sex           Age        SibSp   Parch
Fare
##  0:549    1:216   female:314   Min.   : 0.42   0:608   0:678   Min.
```

```
:  0.00
##  1:342    2:184   male  :577   1st Qu.:22.00   1:209   1:118   1st
Qu.:  7.91
##           3:491                Median :29.70   2: 28   2: 80   Median
: 14.45
##                                Mean   :29.70   3: 16   3:  5   Mean
: 32.20
##                                3rd Qu.:35.00   4: 18   4:  4   3rd
Qu.: 31.00
##                                Max.   :80.00   5:  5   5:  5   Max.
:512.33
##                                                8:  7   6:  1
##        Embarked    segment_edat
##  C          :168   20-29  :407
##  Desconegut:  2   30-39  :155
##  Q          : 77   10-19  :115
##  S          :644   40-49  : 86
##                    0-9    : 64
##                    50-59  : 42
##                    (Other): 22
```

A continuació, s'observa un anàlisis de l'estructura composta de les dades.

```
str(df_titanic)
```

```
## 'data.frame':    891 obs. of  9 variables:
##  $ Survived    : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass      : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2
...
##  $ Sex         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1
1 ...
##  $ Age         : num  22 38 26 35 35 ...
##  $ SibSp       : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1
4 1 2 ...
##  $ Parch       : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1
2 3 1 ...
##  $ Fare        : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked    : Factor w/ 4 levels "C","Desconegut",..: 4 1 4 4 4 3 4
4 4 1 ...
##  $ segment_edat: Factor w/ 8 levels "0-9","10-19",..: 3 4 3 4 4 3 6 1
3 2 ...
```

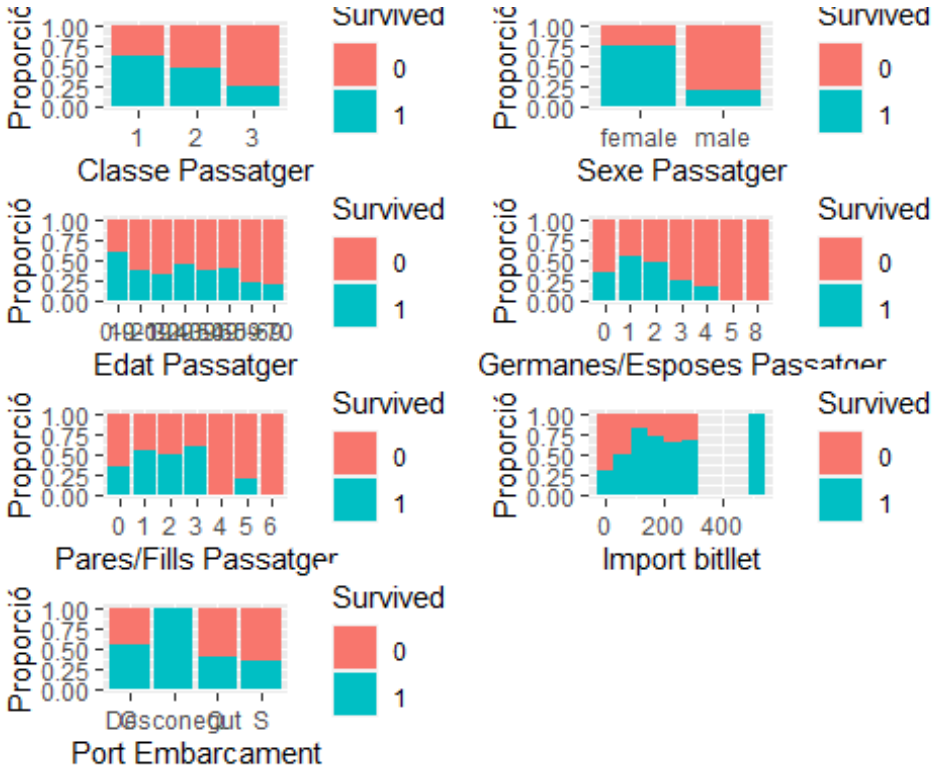
Podem observar com hi ha diferents variables ja tractades. Per exemple, es detalla que Survived i Sex tenen dos factors. En canvi, Pbclass en té 3, per 4 d'Embarked (si es té en compte les dos observacions amb port desconegut). També cal tenir en compte les 7 categories de SibSp i Parch i la nova variable de segment d'edat amb 8. Cal tenir en compte que la variable Age no tindrà sentit si usem segment d'edat per tal de tractar els valors outliers. Per la qual cosa aquesta variable pot ser negligible.

```
# Eliminarem la variable edat
df_titanic <- df_titanic[ -c(4) ]
```

Realitzarem un analisi exploratori de les dades mitjançant histogrames de freqüència dels següents atributs

- Pclass: Classe en que viatja el passatger
- Sex: Sexe del passatger
- segment_Edat: Segment d'edat del passatger
- SibSp: Geramnes/Esposses del passatger
- Parch: Pares/Fills del passatger
- Fare: Import bitllet
- Embarked: Port d'embarcament

```
require(gridExtra)

p1 <- ggplot(df_titanic, aes(x=factor(Pclass), fill = Survived)) +
geom_bar(position = "fill") + xlab("Classe Passatger") +
ylab("Proporció")

p2 <- ggplot(df_titanic, aes(x=factor(Sex), fill = Survived)) +
geom_bar(position = "fill") + xlab("Sexe Passatger") + ylab("Proporció")

p3 <- ggplot(df_titanic, aes(x=factor(segment_edat), fill = Survived)) +
geom_bar(position = "fill") + xlab("Edat Passatger") + ylab("Proporció")

p4 <- ggplot(df_titanic, aes(x=factor(SibSp), fill = Survived)) +
geom_bar(position = "fill") + xlab("Germanes/Esposes Passatger") +
ylab("Proporció")

p5 <- ggplot(df_titanic, aes(x=factor(Parch), fill = Survived)) +
geom_bar(position = "fill") + xlab("Pares/Fills Passatger") +
ylab("Proporció")

p6 <- ggplot(df_titanic, aes(x=Fare, fill = Survived)) +
geom_histogram(bins=10, position = "fill") + xlab("Import bitllet") +
ylab("Proporció")

p7 <- ggplot(df_titanic, aes(x=factor(Embarked), fill = Survived)) +
geom_bar(position = "fill") + xlab("Port Embarcament") +
ylab("Proporció")

grid.arrange(p1, p2, p3, p4, p5, p6, p7, nrow = 4)
```



D'un analisi visual, es pot apreciar que per certs atributs, la proporció de supervivent/no supervivents es molt diferent:

- Classe: Es pot apreciar que hi ha haver més supervivents dins de les classes més elevades.
- Sexe: Es pot apreciar que entre les dones la supervivència va ser al volant del 75% i en canvi entre els homes va ser al voltant del 25%
- Import bitllet: Sembla que com més car era el bitllet, més possibilitats hi havia de sobrevieure. Deduïm (Encara que caldria contrastar-ho) que els que no han pagat billet, son els membres de la tripulació, i en aquest sector es pot veure que la supervivència es molt inferiror a la resta.
- Port embarcament: Sembla que que les persones que no es coneix el port d'embarcament, tenien moltes ménys possibilitats de sobreviure, potser aquestes persones son part de la tripulació. S'hauria d'estudiar en un analisi més detallat.

Comprovació de la normalitat i homogeneïtat de la variància

Comprovació de la normalitat

En primer lloc, hem de veure la normalitat del model. Això ho podem fer mitjançant dos tests, el de Kolmogorov-Smirnov i el de Shapiro-Wilk sobre la variable "Fare".

```
ks.test(df_titanic$Fare, pnorm, mean(df_titanic$Fare),
sd(df_titanic$Fare))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df_titanic$Fare
## D = 0.28185, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

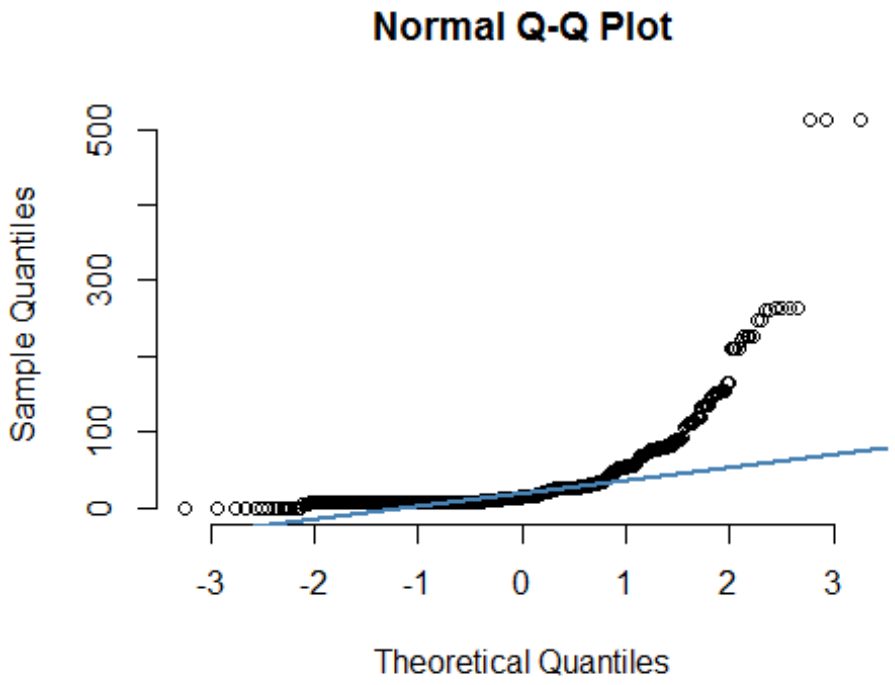
```
shapiro.test(df_titanic$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_titanic$Fare
## W = 0.52189, p-value < 2.2e-16
```

En ambdues proves, el p-valor es més petit que el nivell de significació, que gereralment es pren com a alpha = 0,05. En ambdos casos el valor es molt més petit, per tant, assumirem que la mostra no es una distribució normal.

Realitarem un qqplot (Quantile-Quantile plot) per veure la correlació entre la nostra distribució i una distribució normal, per confirmar visualment el resultat anteriroment, és a dir, que la distribució no és normal.

```
qqnorm(df_titanic$Fare, pch = 1, frame = FALSE)
qqline(df_titanic$Fare, col = "steelblue", lwd = 2)
```



Es pot apreciar clarament la diferència de la distribució entre la nostra mostra i la d'una mostra normal, per tant tal i com s'havia observat no hi ha normalitat.

Comprovació de la homogeneïtat de la variància

A continuació, es veurà la variància dels errors a partir d'un anàlisis de l'homoscedasticitat. Com en el punt anterior hem observat que les dades no segueixen una distribució normal, utilitzarem el test Fligner-Killeen, ja que el tests de Levene únicament per mostres amb una distribució noraml

```
res <- bartlett.test(Fare ~ Survived, data = df_titanic)
res

##
##  Bartlett test of homogeneity of variances
##
## data:  Fare by Survived
## Bartlett's K-squared = 243.67, df = 1, p-value < 2.2e-16
```

Com p-value és inferior a 0,05 es rebutja la hipòtesis d'homogeneïtat en la variància segons el test de Barlett.

```
fligner.test(Fare ~ Survived, data = df_titanic)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 96.253, df = 1, p-value < 2.2e-16
```

La prova de Fligner-Killeen dona un p-valor inferior al nivell de significació (<0,005), per tant la variable "Fare" presenta variables estadísticament diferents per als diferents grups de "Survived".

Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Comparació entre dos grups de dades

A l'apartat anterior hem conclós que la normalitat i l'homoscedasticitat no es compleixen, per tant, harem d'aplicar proves no paràmetriques de contrast d'hipotèsis, com Wilcoxon i Mann-Whitney. Compararem la relació entre "Fare" i "Survived" per els grups d'edat "20-29" i "50-59".

```
wilcox.test(Fare ~ Survived, data = df_titanic, subset = segment_edat
%in% c("20-29", "50-59"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Fare by Survived
## W = 14676, p-value = 9.551e-10
## alternative hypothesis: true location shift is not equal to 0
```

No s'observen diferències estadísticament significatives en la supervivència entre els segments d'edat "20-29" i "50-59".

Comparació entre més de dos grups de dades

Utilitzarem el test de Kruskal-Wallis com a test no paramètric.

```
kruskal.test(Fare ~ Survived, data = df_titanic)
```

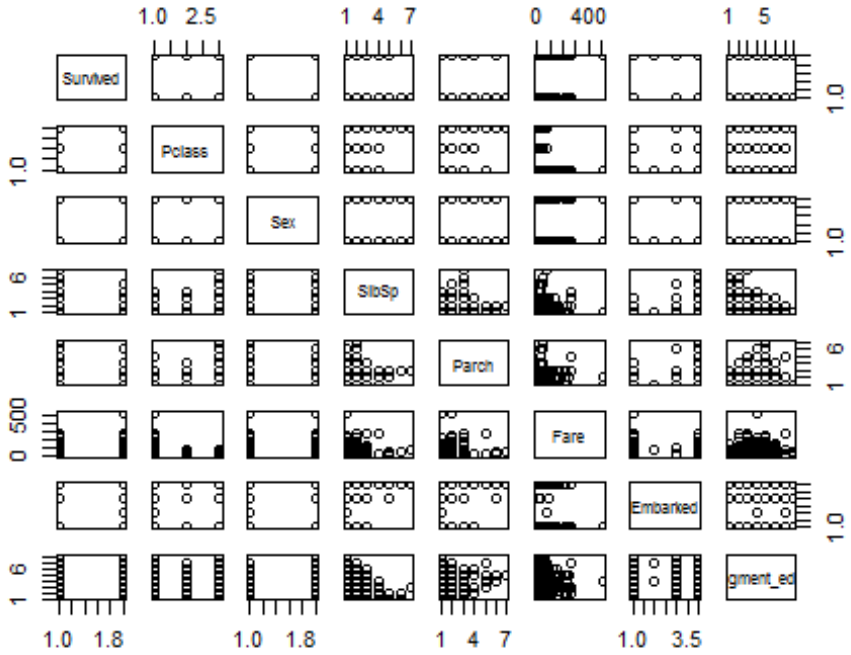
```
##
##  Kruskal-Wallis rank sum test
##
## data:  Fare by Survived
## Kruskal-Wallis chi-squared = 93.277, df = 1, p-value < 2.2e-16
```

Com que el p-valor obtingut és inferior que el nivell de significació, podem concloure que la supervivència mostra diferències significatives segons el preu pagat per el bitllet.

Regressió

Representació del conjunt de dades per parelles de variables.

```
plot(df_titanic)
```



Anem a analitzar la relació entre el preu pagat pel bitllet i la supervivència, després de veure certa relació entre aquest atributs.

```
logit_1 <- glm(Survived~df_titanic$Fare, family = binomial,data =
df_titanic)
summary(logit_1)
```

```
##
## Call:
## glm(formula = Survived ~ df_titanic$Fare, family = binomial,
##     data = df_titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4906  -0.8878  -0.8531   1.3429   1.5942
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.941330   0.095129  -9.895  < 2e-16 ***
## df_titanic$Fare  0.015197   0.002232   6.810 9.79e-12 ***
## ---
## 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: 1117.6  on 889  degrees of freedom
```

```
## AIC: 1121.6
##
## Number of Fisher Scoring iterations: 4
```

Podem observar que es tracta d'una regressió lineal binària, ja que la variable de decisió és binària (sobreviu o no). Per tant, podem observar com el valor de Z>3 es tracta d'una variable a tenir en compte. No obstant, només estem definint el model amb una variable, pel qual hem d'incloure, més variables.

```
logit_1 <- glm(Survived~df_titanic$Fare+df_titanic$Sex+df_titanic$Sex,
family = binomial,data = df_titanic)
summary(logit_1)

##
## Call:
## glm(formula = Survived ~ df_titanic$Fare + df_titanic$Sex +
df_titanic$Sex,
##     family = binomial, data = df_titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2082  -0.6208  -0.5824   0.8126   1.9658
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.647100   0.148502   4.358 1.32e-05 ***
## df_titanic$Fare      0.011214   0.002295   4.886 1.03e-06 ***
## df_titanic$Sexmale -2.422760   0.170515 -14.208  < 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:  884.31  on 888  degrees of freedom
## AIC: 890.31
##
## Number of Fisher Scoring iterations: 5

anova(logit_1)

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev
```

```
## NULL                              890    1186.66
## df_titanic$Fare  1   69.086       889    1117.57
## df_titanic$Sex   1  233.259       888     884.31
```

A continuació, podem observar com la variable de sexe masculí provocava una disminució de la probabilitat de supervivència. Per tant, és una de les variables que més influeix en la supervivència del titànic.

En el següent, cas aplicarem una minimització del valor AIC per tal de veure les variables que descriuen millor el model.

```
# Regressió binomial
logit_1 <- glm(Survived~., family = binomial,data = df_titanic)
summary(logit_1)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = df_titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0886  -0.5667  -0.4118   0.5986   2.3905
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          4.793e+00  6.401e-01   7.487 7.07e-14 ***
## Pclass2             -9.100e-01  3.064e-01  -2.970 0.002982 **
## Pclass3             -1.859e+00  3.056e-01  -6.083 1.18e-09 ***
## Sexmale             -2.752e+00  2.087e-01 -13.185  < 2e-16 ***
## SibSp1               4.048e-02  2.280e-01   0.178 0.859107
## SibSp2              -3.605e-01  5.657e-01  -0.637 0.523924
## SibSp3              -2.583e+00  7.445e-01  -3.469 0.000523 ***
## SibSp4              -2.488e+00  8.418e-01  -2.955 0.003127 **
## SibSp5              -1.591e+01  9.498e+02  -0.017 0.986633
## SibSp8              -1.573e+01  7.541e+02  -0.021 0.983354
## Parch1               5.031e-03  3.153e-01   0.016 0.987268
## Parch2              -2.405e-01  3.967e-01  -0.606 0.544321
## Parch3               2.992e-01  1.066e+00   0.281 0.778997
## Parch4              -1.614e+01  1.083e+03  -0.015 0.988112
## Parch5              -1.813e+00  1.192e+00  -1.521 0.128334
## Parch6              -1.673e+01  2.400e+03  -0.007 0.994437
## Fare                 3.499e-03  2.792e-03   1.253 0.210168
## EmbarkedDesconegut   1.426e+01  1.628e+03   0.009 0.993011
## EmbarkedQ            5.257e-02  3.883e-01   0.135 0.892318
## EmbarkedS           -3.614e-01  2.482e-01  -1.456 0.145363
## segment_edat10-19   -2.044e+00  5.357e-01  -3.815 0.000136 ***
## segment_edat20-29   -2.273e+00  4.973e-01  -4.570 4.88e-06 ***
## segment_edat30-39   -2.004e+00  5.206e-01  -3.849 0.000118 ***
## segment_edat40-49   -2.613e+00  5.658e-01  -4.618 3.88e-06 ***
## segment_edat50-59   -2.915e+00  6.446e-01  -4.522 6.12e-06 ***
## segment_edat60-69   -3.297e+00  8.793e-01  -3.750 0.000177 ***
```

```
## segment_edat>70    -2.923e+00  1.290e+00  -2.266 0.023456 *
## ---
## 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:  753.05  on 864  degrees of freedom
## AIC: 807.05
##
## Number of Fisher Scoring iterations: 15
```

```
logit_2 <- stepAIC(logit_1)
```

```
## Start:  AIC=807.05
## Survived ~ Pclass + Sex + SibSp + Parch + Fare + Embarked +
segment_edat
##
##                Df Deviance     AIC
## - Parch         6   761.58  803.58
## - Embarked      3   756.70  804.70
## - Fare          1   754.87  806.87
## <none>              753.05  807.05
## - SibSp         6   781.07  823.07
## - segment_edat  7   785.79  825.79
## - Pclass        2   791.88  841.88
## - Sex           1   971.96 1023.96
##
## Step:  AIC=803.58
## Survived ~ Pclass + Sex + SibSp + Fare + Embarked + segment_edat
##
##                Df Deviance     AIC
## - Embarked      3   766.11  802.11
## - Fare          1   762.62  802.62
## <none>              761.58  803.58
## - SibSp         6   792.11  822.11
## - segment_edat  7   801.26  829.26
## - Pclass        2   809.94  847.94
## - Sex           1   982.64 1022.64
##
## Step:  AIC=802.11
## Survived ~ Pclass + Sex + SibSp + Fare + segment_edat
##
##                Df Deviance     AIC
## - Fare          1   767.80  801.80
## <none>              766.11  802.11
## - SibSp         6   801.02  825.02
## - segment_edat  7   805.91  827.91
## - Pclass        2   815.57  847.57
## - Sex           1  1001.70 1035.70
```

```
##
## Step:  AIC=801.8
## Survived ~ Pclass + Sex + SibSp + segment_edat
##
##                Df Deviance     AIC
## <none>              767.80  801.80
## - SibSp         6   801.09  823.09
## - segment_edat  7   808.21  828.21
## - Pclass        2   857.75  887.75
## - Sex           1  1010.62 1042.62
```

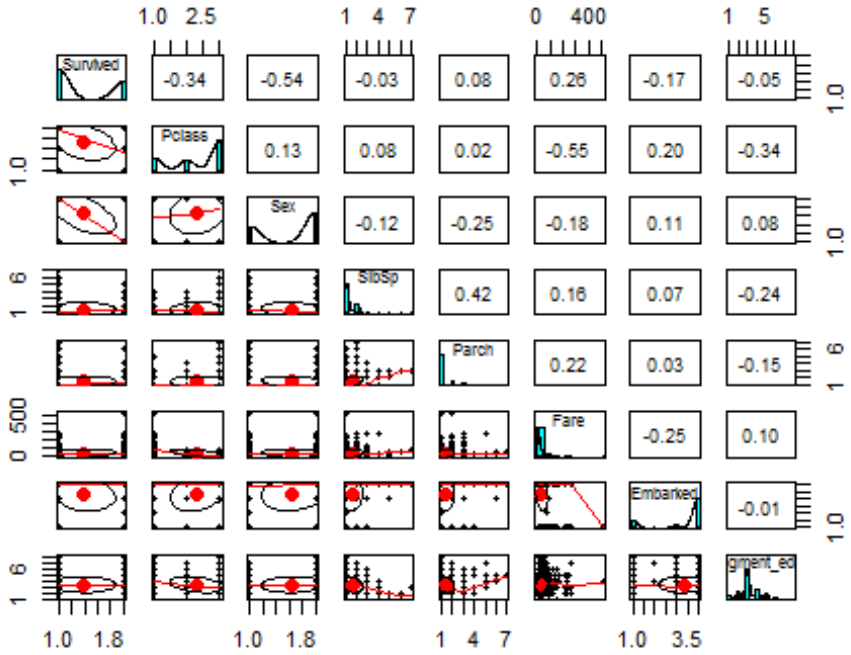
En aquest anàlisis es pot observar com les variables més rellevants és el sexe (Z<-13,15). Això ens indica que es van complir els procediments típics del codi mariner, on en aquella època tenien prioritat el sexe femení en cas d'abordatge. També tenies menys probabilitats de sobreviure si eres de tercera classe o segona. Altres variables rellevants són els diferents grups d'edat. Els menors de 20 anys i majors de 70 van tenir més possibilitats de supervivència.

Analisi de correlació

```
summary(df_titanic)
```

```
##  Survived Pclass      Sex      SibSp   Parch        Fare
##  0:549    1:216   female:314   0:608   0:678   Min.   :  0.00
##  1:342    2:184   male  :577   1:209   1:118   1st Qu.:  7.91
##           3:491                2: 28   2: 80   Median : 14.45
##                                3: 16   3:  5   Mean   : 32.20
##                                4: 18   4:  4   3rd Qu.: 31.00
##                                5:  5   5:  5   Max.   :512.33
##                                8:  7   6:  1
##        Embarked    segment_edat
##  C          :168   20-29  :407
##  Desconegut:  2   30-39  :155
##  Q          : 77   10-19  :115
##  S          :644   40-49  : 86
##                    0-9    : 64
##                    50-59  : 42
##                    (Other): 22
```

```
# Correlacions entre variables
pairs.panels(df_titanic)
```



En l'anterior scatterplot, com la probabilitat de sobreviure està molt influenciada perquè sigui un sexe femení. També una de les variables més correlacionades amb la supervivència és la classe. Si és de classe 1 és més probable de sobreviure a una de classe 3. També estan molt relacionades les classes amb lex taxes (55%). També hi ha una certa relació entre el segment d'edat i la classe d'embarcament

Aplicació de model predictiu (Random Forest)

Aplicarem un model d'arbres de decisió (Random Forest), que per les condicions del conjunt de dades sembla més adient. No obstant, caldria comparar els diferents errors entre models i també observar si es pot usar models predictius en paral·lel amb diferents classificadors.

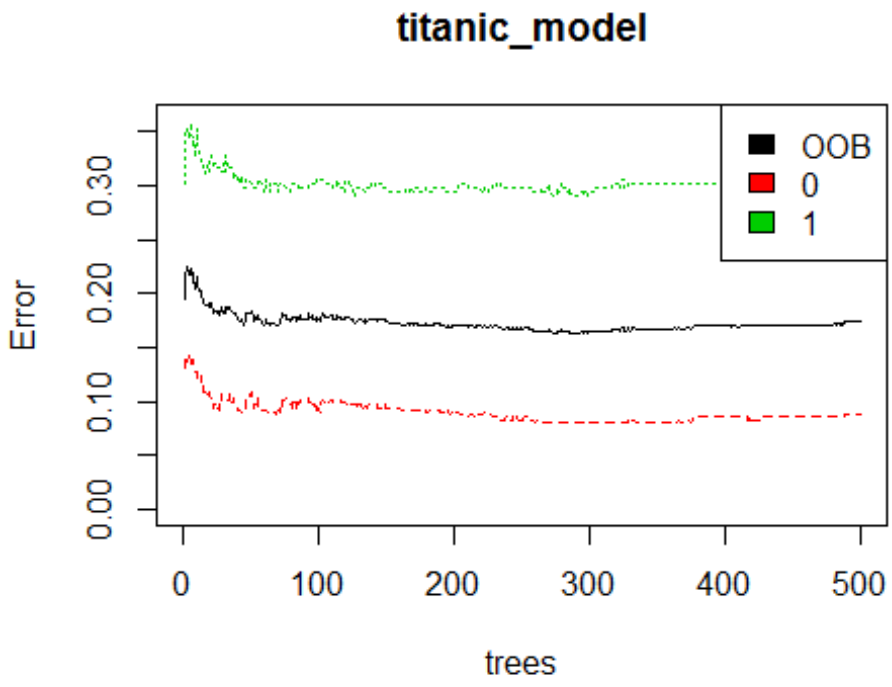
```
# Obtindrem la llibreria corresponent
library('randomForest')
train <- df_titanic[1:712,]
test <- df_titanic[713:891,]
# Set a random seed
set.seed(754)
# Construim un model de RandomForest amb les variables ja tractades
(Sexe,classe d'embarcament, port d'embarcament, sexe, situació familiar,
preu del ticket i segment d'edat)
titanic_model <- randomForest(Survived ~.,
                              data = train)
```

A continuació, representarem els resultats del model predictiu Random Forest. Es detallaran els errors per número d'estimadors (arbres de decisió en aquest cas), un ranking d'importàncies relatives i les respectives prediccions.

Representació dels resultats a partir de taules i gràfiques

En primer lloc, observarem la precisió del model resultant per Random Forest. Tot i que aquest model és de més difícil interpretació gràfica que altres mètodes com ara una regressió lineal, s'intentarà extreure tota la informació possible.

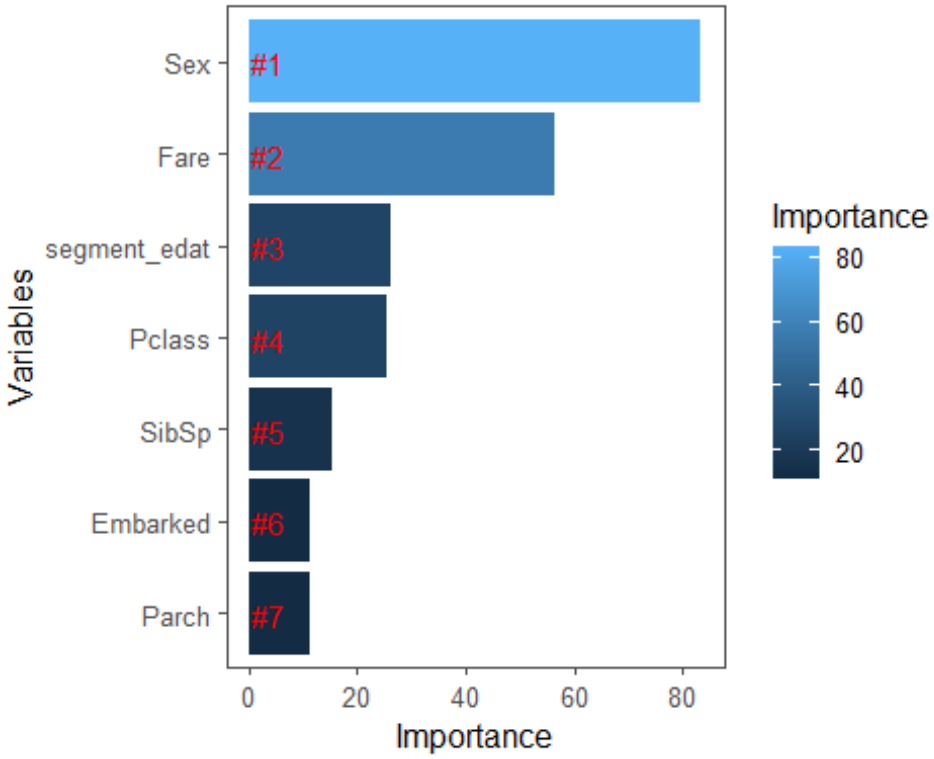
```
# Mostrem l'error del model
plot(titanic_model, ylim=c(0,0.36))
legend('topright', colnames(titanic_model$err.rate), col=1:3, fill=1:3)
```



En l'anterior punt vam veure com el model ens dona una precisió aproximada del 80%. Tindríem una precisió al voltant del 90% pels no supervivents i d'un 70% pels supervivents, en training set. Un cop aplicat el Random Forest veurem la importància de cada variable en el model per tal d'interpretar el model. Al ser un model més complicat que altres mètodes de graficar degut a la seva estructura de "black box" hem optat per representar aquella estrucutra de múltiples arbres com la importància relativa, que mostra les variables de decisió més importants que pren el random Forest

```
# Treiem les variables amb més importància relativa
importance    <- importance(titanic_model)
varImportance <- data.frame(Variables = row.names(importance),
```

```
                            Importance = round(importance[
,'MeanDecreaseGini'],2))
# Creem un ranking basat en importància relativa
rankImportance <- varImportance %>%
  mutate(Rank = paste0('#',dense_rank(desc(Importance))))
# Usem ggplot2 per visualitzar la importància relativa
ggplot(rankImportance, aes(x = reorder(Variables, Importance),
                           y = Importance, fill = Importance)) +
  geom_bar(stat='identity') +
  geom_text(aes(x = Variables, y = 0.5, label = Rank),
            hjust=0, vjust=0.55, size = 4, colour = 'red') +
  labs(x = 'Variables') +
  coord_flip() +
  theme_few()
```



Tal i com hem vist anteriorment, podem veure les variables que més importen en la decisió de l'algoritme de RandomForest. Podem observar com les variables més rellevants són Sexe, seguit de les taxes (Fare), el segment de l'edat i la classe. Com les taxes pagades estan correlacionades amb les classes té sentit. A continuació, es detallarà les prediccions del nostre model

```
# Predim usant testing set
prediction <- predict(titanic_model, test)
prediction
```

```
## 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729
730 731 732
```

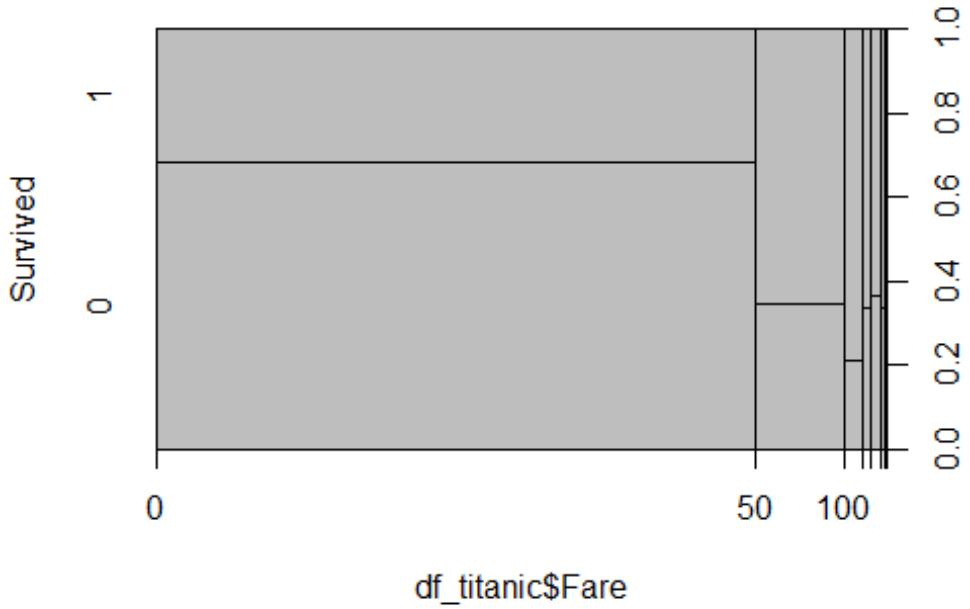
```
##   0   0   0   0   1   1   0   0   1   0   0   0   0   0   1   1   0
1   1   0
## 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749
750 751 752
##   0   0   0   0   0   1   0   0   0   1   1   0   0   0   0   1   0
0   1   0
## 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769
770 771 772
##   0   0   1   1   0   0   0   1   0   0   0   1   0   1   0   1   0
0   0   0
## 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789
790 791 792
##   1   0   1   0   0   1   0   1   1   1   0   0   0   0   0   0   1
0   0   0
## 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809
810 811 812
##   0   0   0   0   1   0   0   0   0   1   0   0   0   0   0   0   0
1   0   0
## 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829
830 831 832
##   0   0   0   0   1   0   0   0   1   0   0   1   0   0   1   1   0
1   0   1
## 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849
850 851 852
##   0   0   0   1   0   0   1   0   0   0   1   0   0   0   0   0   0
1   0   0
## 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869
870 871 872
##   1   1   1   0   1   0   1   0   0   0   1   0   0   1   1   0   0
1   0   1
## 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889
890 891
##   0   0   1   1   0   0   0   1   1   0   1   0   0   0   0   1   0
0   0
## Levels: 0 1
```

```
write.csv(prediction, file =
'Output/sicilia_calvo_titanic_predictions.csv', row.names = F)
write.csv(df_titanic, file =
'Output/sicilia_calvo_titanic_df_titanic.csv')
```

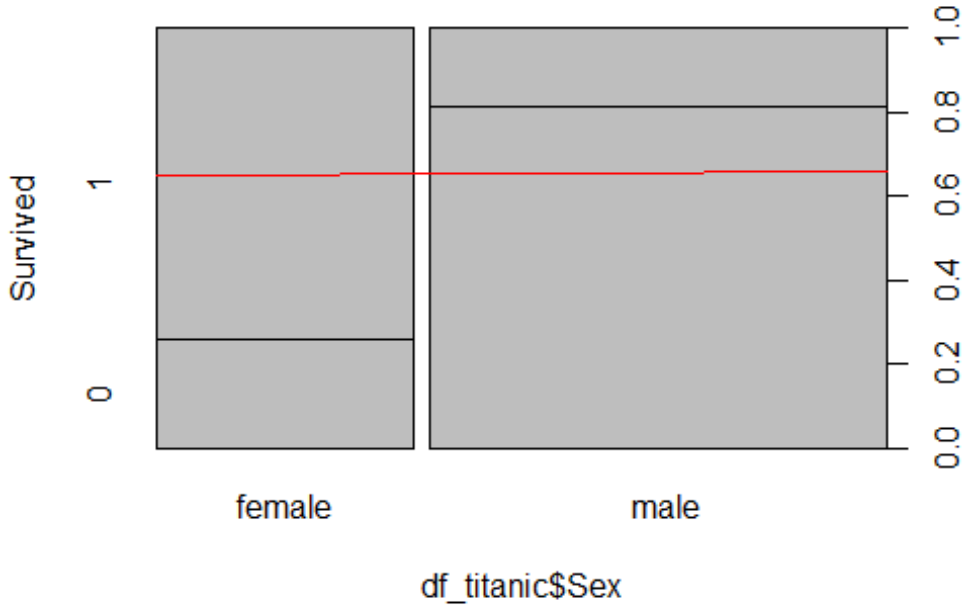
Representació del model de regressió binomial

En aquest apartat es podrà observar com es comporta la regressió lineal del nostre model.

```
logit_1 <- glm(Survived~df_titanic$Fare+df_titanic$Sex, family =
binomial,data = df_titanic)
plot(Survived~df_titanic$Fare+df_titanic$Sex,
data=df_titanic,type="l",lty=1,col=16,ylim=c(0,1),xlim=c(0,7))
```




```r
abline(logit_1, col=2)
```



En aquest gràfic es pot representar la regressió binomial com la probabilitat de que una persona sobrevisqui. Es pot apreciar com segons la taxa de pagamaent entre 0 i 50 la probabilitat era molt més alta que en les altres taxes. Una altra probabilitat de sobreviure era ser dona, en comparació a ser home. Tal i com es pot observar el logit_1 només tenint en compte sex i Fare no pot fer una bona delimitació del model ja que està subajustat al model de dades i no extreu un model correcte.

Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Les variables més rellevants a l'hora de determinar la supervivència d'un passatger són el sexe seguit de variables relacionades amb el seu nivell adquisitiu (preu del ticket o classe) i la seva edat. També és cert que hi ha una certa relació entre el preu del ticket, la classe d'embarcament i el segment d'edat.

Les dades tenen una qualitat correcta i estan majoritàriament ben informades. Disposen d'una variable de classe "survived" que les fa aptes per un classificador.

En resum, aconseguim obtenir una predicció de cada classe segons el tipo d'observació amb un model de Random Forest amb una precisió al voltant d'un 80%.

Codi

El codi s'ha anat adjuntant en la resolució de cada apartat. Per tant, ja no faria falta aquest apartat.