Pràctica 2: Tractament del dataset Titanic

Daniel Lopez Ramirez

7 de enero, 2020

Contents

1. Descripció del Dataset.

El dataset utilitzat és el que correspon al **Titanic: Machine Learning from Disaster** https://www.kaggle.com/c/titanic de kaggle. Aquest dataset conté tres fitxers csv, dos amb les mostres de train i test dels passatgers que hi havia al Titanic, i que es diferencien en que el fitxer de test no conté el valor de si la persona va sobreviure o no. I el tercer fitxer és una relació dels id's dels passatgers amb el valor de si va sobreviure o no per a la mostra de test.
Aquest dataset és important perquè permet estudiar quins passatgers van ser els més afectats per a l'incident del Titanic tenint en compte la classe en la que viatjaven, el sexe o d'altres variables, cosa que ens permet tenir més informació de com va succeir tot i intentar predir si el passatger va sobreviure o no, segons aquestes variables.

```
# Carreguem les dades dels fitxers
titanic_train <- read.csv("../csv/train.csv",header=TRUE, sep=",", na.strings="NA",
                          dec=".", strip.white=TRUE)
titanic_test <- read.csv("../csv/test.csv",header=TRUE, sep=",", na.strings="NA",
                         dec=".", strip.white=TRUE)
gender_submission <- read.csv("../csv/gender_submission.csv",header=TRUE, sep=",", na.strings="NA",
                              dec=".", strip.white=TRUE)
```

El dataset conté les següents dades:

Variable	Definició	Clau	Notes
PassengerId	Identificador del passatger		
Survived	Supervivent	0 = No, 1 = Yes	
Pclass	Classe de ticket	1 = 1st, 2 = 2nd, 3 = 3rd	A proxy for socio-economic status (SES)1st = Upper2nd = Middle3rd = Lower
Name	Nom del passatger		
Sex	Sexe		
Age	Edat en anys		Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
Sibsp	Nombre de germans/conjugues a bord		The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsisterSpouse = husband, wife (mistresses and fiancés were ignored)
Parch	Nombre de pares/fills a bord		The dataset defines family relations in this way... Parent = mother, fatherChild = daughter, son, stepdaughter, stepsonSome children travelled only with a nanny, therefore parch=0 for them.
Ticket	Número de ticket		
Fare	Tarifa		
Cabin	Número de cabina		
Embarked	Port d'embarcament		C = Cherbourg, Q = Queenstown, S = Southampton

2. Integració i selecció de les dades d'interes

Primer de tot farem unb primer anàlisi visual de les dades que contenen els datasets carregats.

```
# Revisem la informació del fitxer train.csv
summary(titanic_train)
```

```
##   PassengerId       Survived          Pclass
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000
##  Median :446.0   Median :0.0000   Median :3.000
##  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
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                   :  1   female:314   Min.   : 0.42
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   1st Qu.:20.12
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                Median :28.00
##  Abelson, Mr. Samuel                   :  1                Mean   :29.70
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
```

```
##  Adahl, Mr. Mauritz Nils Martin       :  1                Max.   :80.00
##  (Other)                              :885                NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
##                                   (Other) :852
##          Cabin     Embarked
##             :687    :  2
##  B96 B98    :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6         :  4   S:644
##  C22 C26    :  3
##  D          :  3
##  (Other)    :186
```

```
head(titanic_train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                  Name    Sex Age SibSp
## 1                             Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                              Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                            Allen, Mr. William Henry   male  35     0
## 6                                    Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500              S
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500              S
## 6     0           330877  8.4583              Q
```

```
#sapply(titanic_train, function(x)class(x))
str(titanic_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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
```

```
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
colSums(is.na(titanic_train))
```

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

```
colSums(titanic_train=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

El conjunt de training conté 891 observacions de 12 variables. D'aquestes 12 variables, hi ha algunes variables que el seu contingut no ens ajudarà per a la predicció, com són el número de ticket (Ticket) i el nom del passatger (Name). D'altra banda, hi ha variables que haurem de tractar i/o convertir, com:
* Convertir la variable *Survived* a factor.
* Convertir la variable *Pclass* a factor.
* Tractar la variable *Cabin* per extreure la coberta de la cabina i revisar els valors buits que conté.
* Tractar la variable *Age*, ja que hi ha força valors buits.
* Tractar la variable *Embarked*, ja que conté alguns valors buits.

```
# Revisem la informació del fitxer test.csv
summary(titanic_test)
```

```
##   PassengerId         Pclass
##  Min.   : 892.0   Min.   :1.000
##  1st Qu.: 996.2   1st Qu.:1.000
##  Median :1100.5   Median :3.000
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##                                          Name         Sex
##  Abbott, Master. Eugene Joseph            :  1   female:152
##  Abelseth, Miss. Karen Marie              :  1   male  :266
##  Abelseth, Mr. Olaus Jorgensen            :  1
##  Abrahamsson, Mr. Abraham August Johannes :  1
##  Abrahim, Mrs. Joseph (Sophie Halaut Easu):  1
##  Aks, Master. Philip Frank                :  1
##  (Other)                                  :412
##       Age            SibSp           Parch             Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   PC 17608:  5
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   113503  :  4
```

```
##  Median :27.00   Median :0.0000   Median :0.0000   CA. 2343:  4
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923   16966   :  3
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000   220845  :  3
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000   347077  :  3
##  NA's   :86                                        (Other) :396
##       Fare                   Cabin     Embarked
##  Min.   :  0.000                :327   C:102
##  1st Qu.:  7.896   B57 B59 B63 B66:  3   Q: 46
##  Median : 14.454   A34           :  2   S:270
##  Mean   : 35.627   B45           :  2
##  3rd Qu.: 31.500   C101          :  2
##  Max.   :512.329   C116          :  2
##  NA's   :1         (Other)       : 80
```

```
head(titanic_test)
```

```
##   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
## 1 34.5     0     0  330911  7.8292              Q
## 2 47.0     1     0  363272  7.0000              S
## 3 62.0     0     0  240276  9.6875              Q
## 4 27.0     0     0  315154  8.6625              S
## 5 22.0     1     1 3101298 12.2875              S
## 6 14.0     0     0    7538  9.2250              S
```

```
#sapply(titanic_test, function(x)class(x))
str(titanic_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       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 5
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ 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     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

```
colSums(is.na(titanic_test))
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0          86           0
##       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           1           0           0
```

```
colSums(titanic_test=="")
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0          NA           0
##       Parch      Ticket        Fare       Cabin    Embarked
##           0           0          NA         327           0
```

El conjunt de test conté 418 observacions de 11 variables. En aquest cas, no conté la variable *Survived* ja que és la que hem de predir. Tal com hem comentat amb el conjunt de training, eliminarem algunes variables com el número de ticket (Ticket) i el nom del passatger (Name). D'altra banda, hi ha variables que haurem de tractar i/o convertir, com:
* Convertir la variable *Pclass* a factor.
* Tractar la variable *Cabin* per extreure la coberta de la cabina. En aquest cas, veiem que hi han molts valors de la variable que estàn buits, i que haurem de tractar.
* Tractar la variable *Age*, ja que hi ha força valors buits.
* Tractar la variable *Fare*, ja que conté un valor buit.

```
# Revisem la informació del fitxer gender_submission.csv
summary(gender_submission)
```

```
##   PassengerId        Survived
##  Min.   : 892.0   Min.   :0.0000
##  1st Qu.: 996.2   1st Qu.:0.0000
##  Median :1100.5   Median :0.0000
##  Mean   :1100.5   Mean   :0.3636
##  3rd Qu.:1204.8   3rd Qu.:1.0000
##  Max.   :1309.0   Max.   :1.0000
```

```
head(gender_submission)
```

```
##   PassengerId Survived
## 1         892        0
## 2         893        1
## 3         894        0
## 4         895        0
## 5         896        1
## 6         897        0
```

```
#sapply(gender_submission, function(x)class(x))
str(gender_submission)
```

```
## 'data.frame':    418 obs. of  2 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
```

```
colSums(is.na(gender_submission))
```

```
## PassengerId    Survived
##           0           0
```

El conjunt de *gender_submission* conté els valors correctes de la variable *Survived* per al conjunt de test. L'unica tasca que haurem de realitzar és convertir la variable *Survived* a factor.

Un cop revisats els diversos conjunt de dades, anem a factoritzar les variables *Pclass* i *Survived*:

```
titanic_train$Pclass <- as.factor(titanic_train$Pclass)
titanic_test$Pclass <- as.factor(titanic_test$Pclass)
titanic_train$Survived<- as.factor(titanic_train$Survived)
```

3. Neteja de dades

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

Tal com hem comentat en l'apartat anterior algunes variables contenen zeros o elements buits. Per a poder gestionar aquests casos, combinarem els dos datasets. Per a combinar-los, afegirem la variable *Survived* al dataset de test per a després utilitzar *rbind* per a combinar-los.

```
titanic_test_survived <- titanic_test
titanic_test_survived$Survived <- NA
titanic <- rbind(titanic_train,titanic_test_survived)
```

Amb els dos datasets combinats, avaluem la nova informació:

```
summary(titanic)
```

```
##   PassengerId   Survived   Pclass                                    Name
##  Min.   :   1   0   :549   1:323   Connolly, Miss. Kate                 :   2
##  1st Qu.: 328   1   :342   2:277   Kelly, Mr. James                     :   2
##  Median : 655   NA's:418   3:709   Abbing, Mr. Anthony                  :   1
##  Mean   : 655                      Abbott, Mr. Rossmore Edward          :   1
##  3rd Qu.: 982                      Abbott, Mrs. Stanton (Rosa Hunt)     :   1
##  Max.   :1309                      Abelson, Mr. Samuel                  :   1
##                                    (Other)                              :1301
##      Sex           Age            SibSp           Parch
##  female:466   Min.   : 0.17   Min.   :0.0000   Min.   :0.000
##  male  :843   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
##               Median :28.00   Median :0.0000   Median :0.000
##               Mean   :29.88   Mean   :0.4989   Mean   :0.385
##               3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##               Max.   :80.00   Max.   :8.0000   Max.   :9.000
##               NA's   :263
##       Ticket          Fare                      Cabin      Embarked
##  CA. 2343:  11   Min.   :  0.000                  :1014    :  2
##  1601    :   8   1st Qu.:  7.896   C23 C25 C27    :   6   C:270
##  CA 2144 :   8   Median : 14.454   B57 B59 B63 B66:   5   Q:123
##  3101295 :   7   Mean   : 33.295   G6             :   5   S:914
##  347077  :   7   3rd Qu.: 31.275   B96 B98        :   4
##  347082  :   7   Max.   :512.329   C22 C26        :   4
##  (Other) :1261   NA's   :1         (Other)        : 271
```

```r
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                            Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                             Heikkinen, Miss. Laina female  26     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                           Allen, Mr. William Henry   male  35     0
## 6                                   Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500              S
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500              S
## 6     0           330877  8.4583              Q
```

```r
#sapply(titanic_test, function(x)class(x))
str(titanic)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 929 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```r
colSums(is.na(titanic))
```

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

```r
colSums(titanic=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
```

```
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
```

Com podem observar la variable *Cabin* conté molts valors buits (sobre un 80%), amb el que transformarem la variable en una nova variable, que indiqui si el passatger tenia cabina o no.

```
titanic$WithCabin <- ifelse(titanic$Cabin=="", "0","1")
titanic$WithCabin <- as.factor(titanic$WithCabin)
```

Pel que fa a la variable *Embarked*, conté dos valors buits. Anem a avaluar les possibles relacions de la variable Embarked amb les altres variables del dataset.

```
titanic[titanic$Embarked=="",]
```

```
##     PassengerId Survived Pclass                                      Name
## 62           62        1      1                      Icard, Miss. Amelie
## 830         830        1      1 Stone, Mrs. George Nelson (Martha Evelyn)
##        Sex Age SibSp Parch Ticket Fare Cabin Embarked WithCabin
## 62  female  38     0     0 113572   80   B28                 1
## 830 female  62     0     0 113572   80   B28                 1
```

Com podem veure, els dos passatgers que no ténen el port d'embarcament informat són dones que van sobreviure al naufragi i que anaven en primera classe. Tenint en compte aquesta informació, anem a avaluar quin és el valor més adient per aquests passatgers.

```
table(titanic$Embarked,titanic$Sex,titanic$Pclass,titanic$Survived)
```

```
## , ,  = 1,  = 0
##
##
##     female male
##          0    0
##   C      1   25
##   Q      0    1
##   S      2   51
##
## , ,  = 2,  = 0
##
##
##     female male
##          0    0
##   C      0    8
##   Q      0    1
##   S      6   82
##
## , ,  = 3,  = 0
##
##
##     female male
##          0    0
##   C      8   33
```

```
##   Q      9   36
##   S     55  231
##
## , ,  = 1,  = 1
##
##
##     female male
##          2    0
##   C     42   17
##   Q      1    0
##   S     46   28
##
## , ,  = 2,  = 1
##
##
##     female male
##          0    0
##   C      7    2
##   Q      2    0
##   S     61   15
##
## , ,  = 3,  = 1
##
##
##     female male
##          0    0
##   C     15   10
##   Q     24    3
##   S     33   34
```

En aquest cas, ens interessa la taula on la *classe és 1 i els passatgers van sobreviure*:

	female	male
	2	0
C	42	17
Q	1	0
S	46	28

Segons la taula, el valor d'embarcament amb més freqüencia és S (Southampton), encara que el valor C (Cherbourg) també és força elevat, però si avaluem les dades tenint en compte els passatgers que van sobreviure, veurem que la majoria van embarcar a S. Per tant, als dos passatgers que no tenen el port d'embarcament els hi assignarem la S.

```
titanic$Embarked <- as.character(titanic$Embarked)
titanic$Embarked[titanic$Embarked==""] <- "S"
titanic$Embarked <- as.factor(titanic$Embarked)
```

Una altra variable que conté un valor buit, és la variable *Fare*. El registre conté les següents dades:

```
titanic[is.na(titanic$Fare) ,]
```

```
##      PassengerId Survived Pclass               Name  Sex  Age SibSp Parch
```

```
## 1044        1044     <NA>      3 Storey, Mr. Thomas male 60.5     0     0
##      Ticket Fare Cabin Embarked WithCabin
## 1044   3701   NA              S         0
```

Com podem veure, el passatger és un home, que va embarcar a Southampton i que era de tercera classe. Com que és només un registre el que hem de corregir, utilitzarem la mitjana del valor de *Fare* de tots els homes que van embarcar a Southampton a tercera classe:

```
titanic$Fare[is.na(titanic$Fare)] <- mean(titanic$Fare[titanic$Pclass=="3"
                                  & titanic$Embarked =="S" & titanic$Sex=="male"],na.rm=TRUE)
```

Finalment, hem de tractar els valors buits de la variable *Age*. Per a aquest tractament, utilitzarem la funció **missForest**, ja que és un mètode més robust per a corregir els valors buits. Per a poder utilizar-la, crearem un nou dataset, extraient variables que no utilitzarem posteriorment com *Name*, *Cabin*, *Ticket* i *Survived* (En aquest cas, la treiem per a que no calculi els valors buits de test).

```
titanic_1 <- subset(titanic, select = -c(Name,Ticket,Cabin,Survived))
titanic_mForest <- missForest(titanic_1,variablewise = TRUE)
```

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
```

```
colSums(is.na(titanic_mForest$ximp))
```

```
## PassengerId      Pclass         Sex         Age       SibSp       Parch
##           0           0           0           0           0           0
##        Fare    Embarked   WithCabin
##           0           0           0
```

```
titanic_mForest_data <- titanic_mForest$ximp
titanic_mForest_data$Survived <- titanic$Survived
```

Amb això ja tindriem les dades tractades:

```
str(titanic_mForest_data)
```

```
## 'data.frame':    1309 obs. of  10 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ WithCabin  : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
```

```r
colSums(is.na(titanic_mForest_data))
```

```
## PassengerId      Pclass         Sex         Age       SibSp       Parch
##           0           0           0           0           0           0
##        Fare    Embarked   WithCabin    Survived
##           0           0           0         418
```

```r
colSums(titanic_mForest_data=="")
```

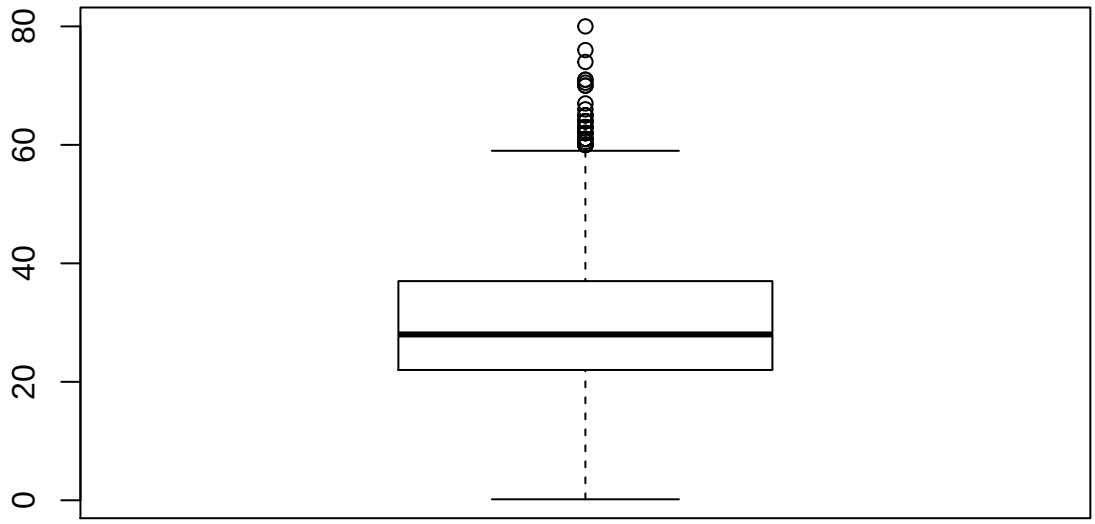
```
## PassengerId      Pclass         Sex         Age       SibSp       Parch
##           0           0           0           0           0           0
##        Fare    Embarked   WithCabin    Survived
##           0           0           0          NA
```

3.2. Identificació i tractament de valors extrems.

Per avaluar utilitzarem els gràfics **boxplot** sobre les variables de la mostra. No tindrem en compte les variables factoritzades per aquest anàlisi, ja que tots els seus valors estan dintre dels seus valors possibles (*Pclass*, *Sex*, *Embarked*, *WithCabin*).

Començarem per la variable *Age*:

```r
boxplot(titanic_mForest_data$Age)
```

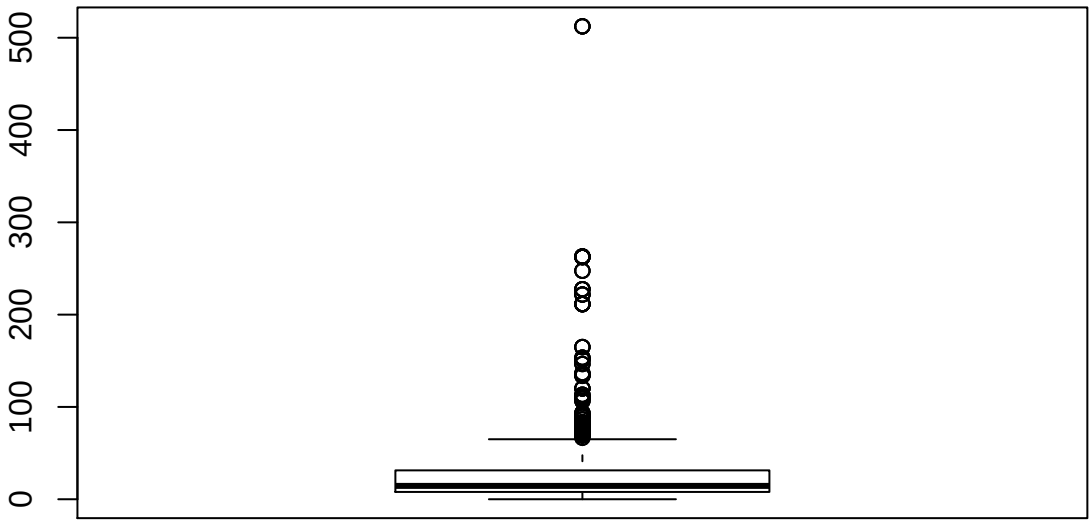


```
ggplot(data=titanic_mForest_data, aes(titanic_mForest_data$Survived, titanic_mForest_data$Age)) + geom_
```

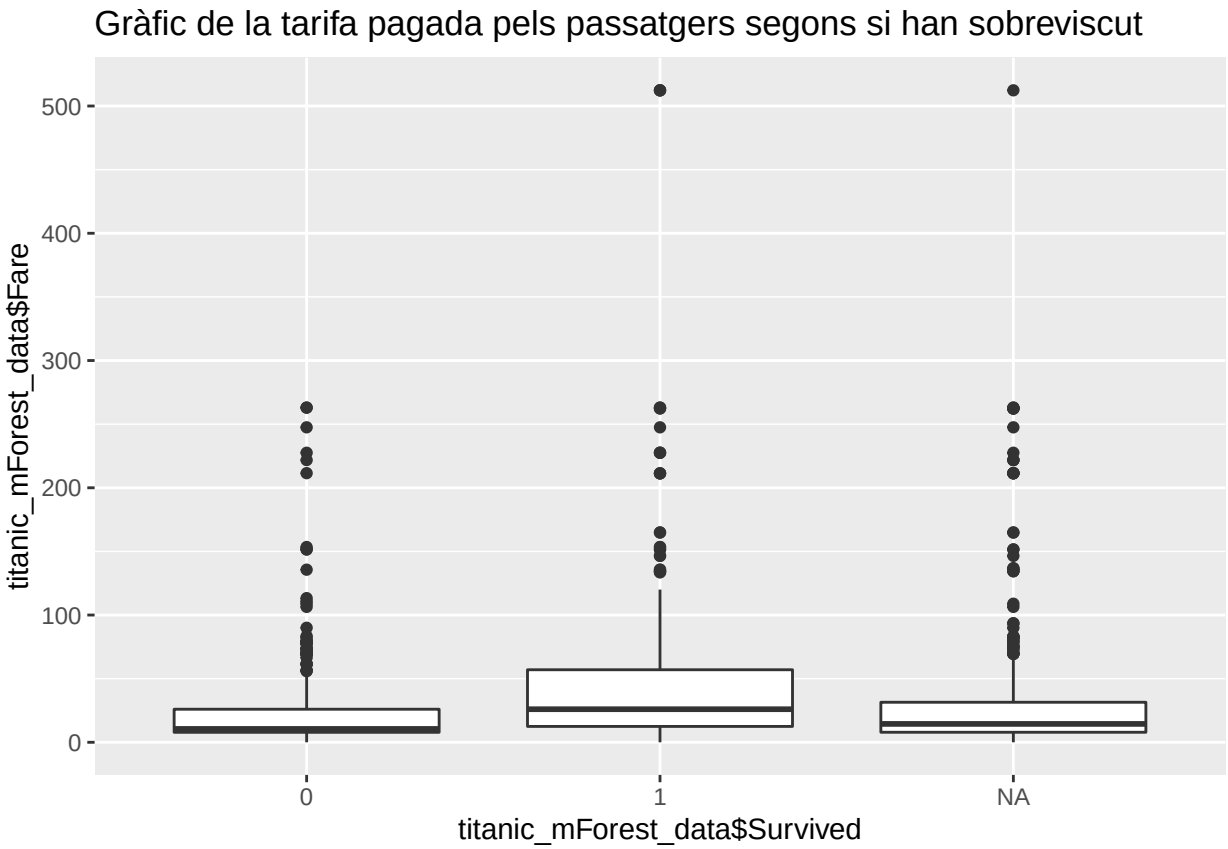
Com s'observa en el gràfic, podriem tenir diversos outliers a partir de 60 anys, però realment, podia haver-hi persones d'aquesta edat a la mostra. Per tant, donarem per vàlida la mostra i no aplicarem cap tractament als valors extrems de *Age*.

Analitzem ara la variable *Fare*:

```
boxplot(titanic_mForest_data$Fare)
```

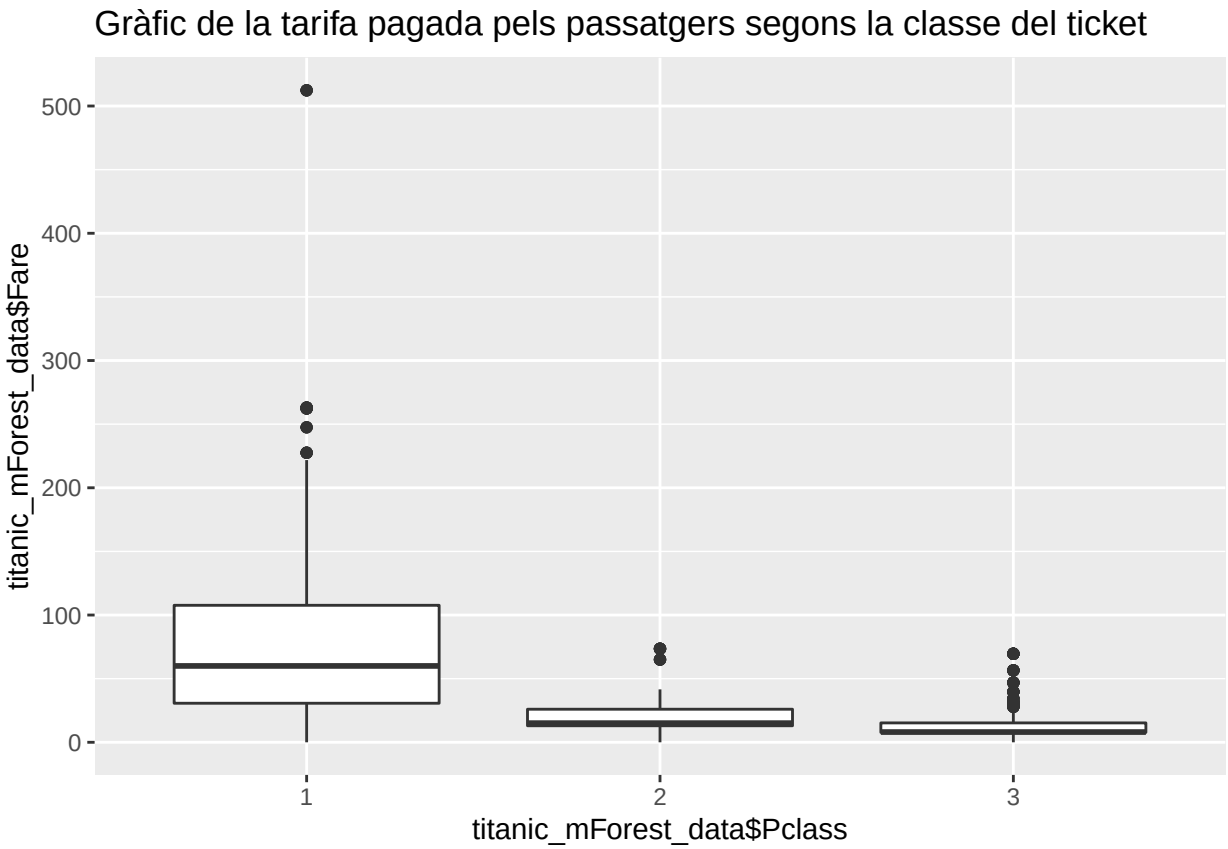



```
ggplot(data=titanic_mForest_data, aes(titanic_mForest_data$Survived, titanic_mForest_data$Fare)) + geom_
```

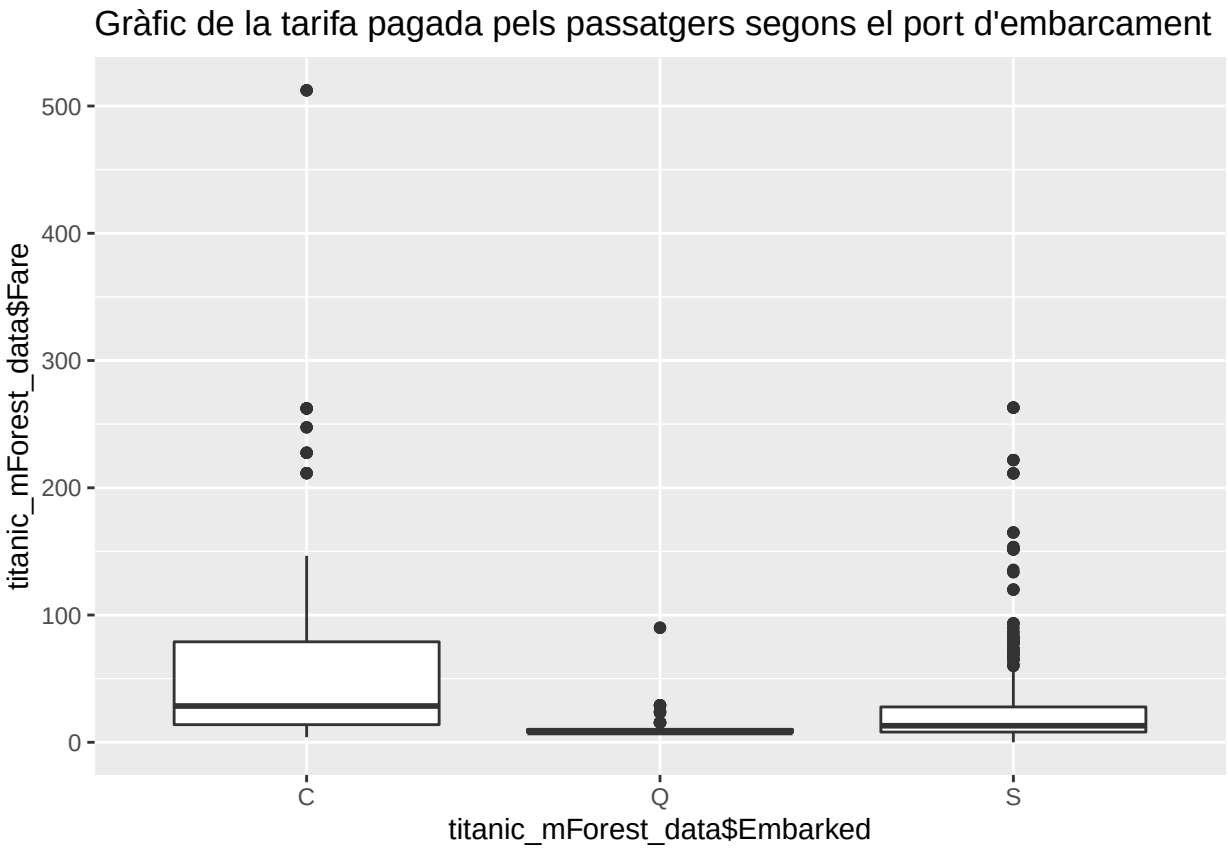


Observem que hi ha un outlier molt diferenciat de tots els altres (per sobre de 500$), però que apareix tant a la mostra de training com a la mostra de test. Per tant, hem de revisar aquests valors:

```
ggplot(data=titanic_mForest_data, aes(titanic_mForest_data$Pclass, titanic_mForest_data$Fare)) + geom_b
```



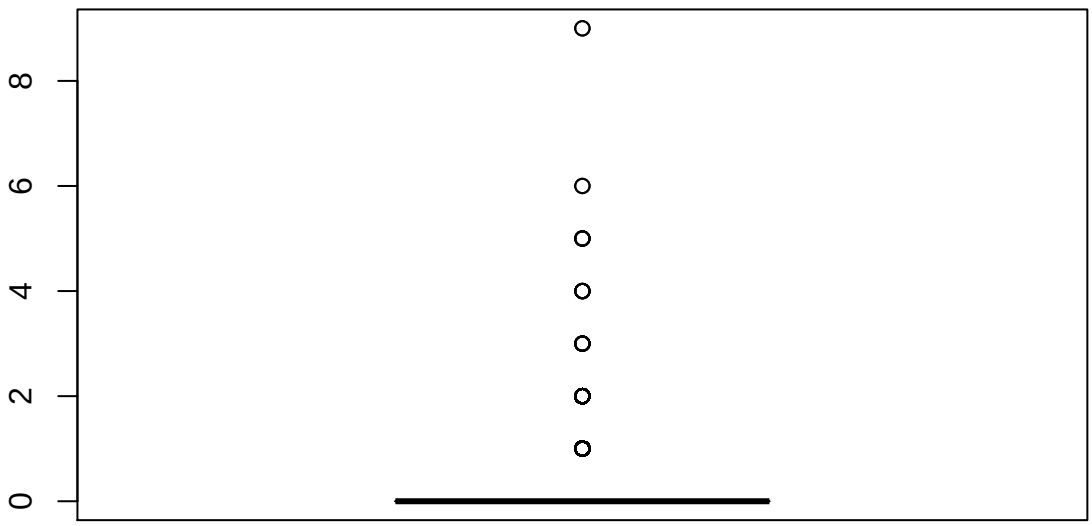

```
ggplot(data=titanic_mForest_data, aes(titanic_mForest_data$Embarked, titanic_mForest_data$Fare)) + geom_
```



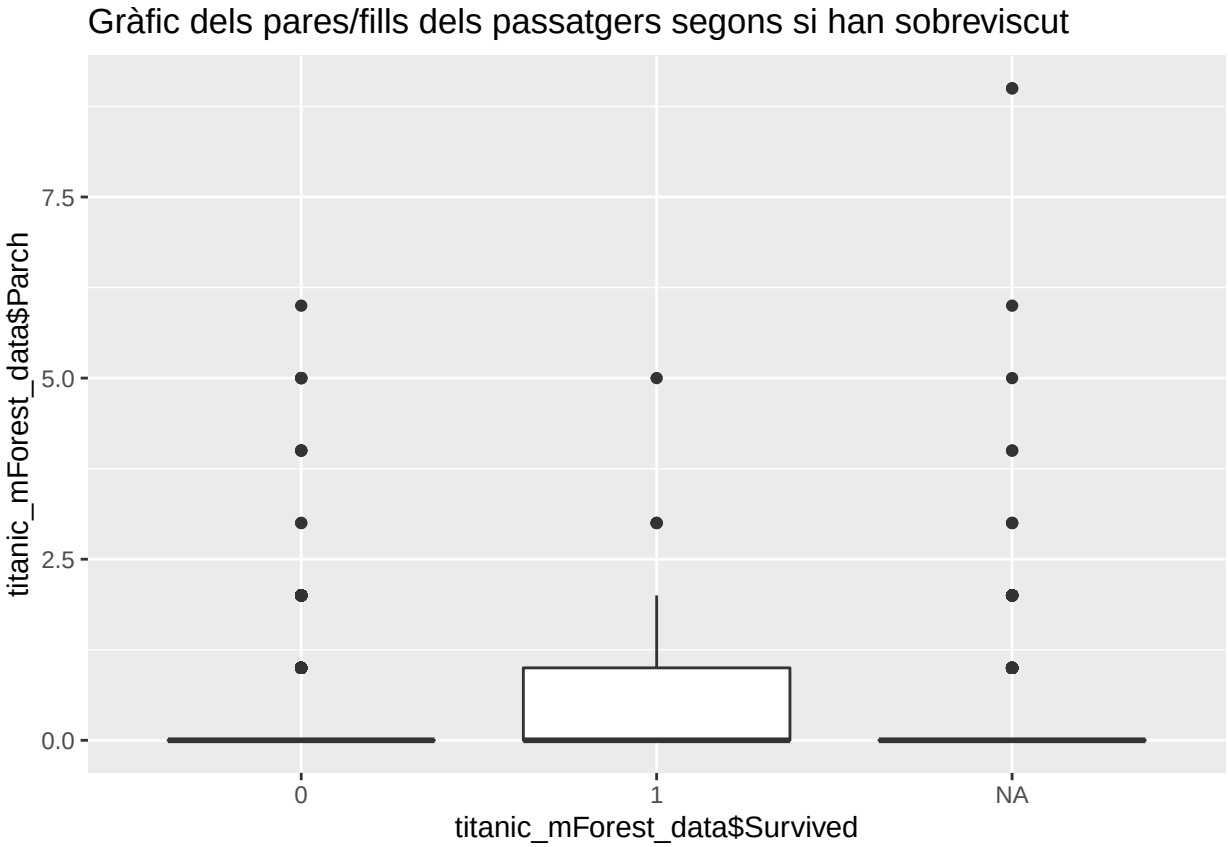
Tal com observem als gràfics, els passatgers que van pagar més de 500$ per un ticket, van embarcar al mateix port i anaven en primera classe. Per tant, és possible que aquests passatgers paguessin per un camarot molt exclussiu de primera classe. Per tant, donem per vàlida la mostra i no aplicarem cap tractament als valors extrems de *Fare*.

Finalment, tractarem els outliers de les variables *Parch* i *SibSp*:

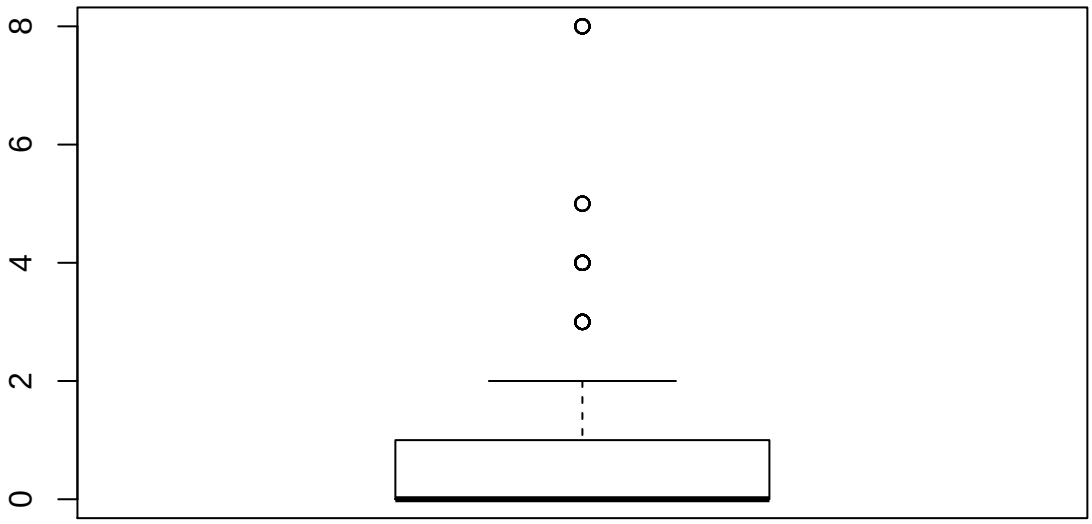
```
boxplot(titanic_mForest_data$Parch)
```



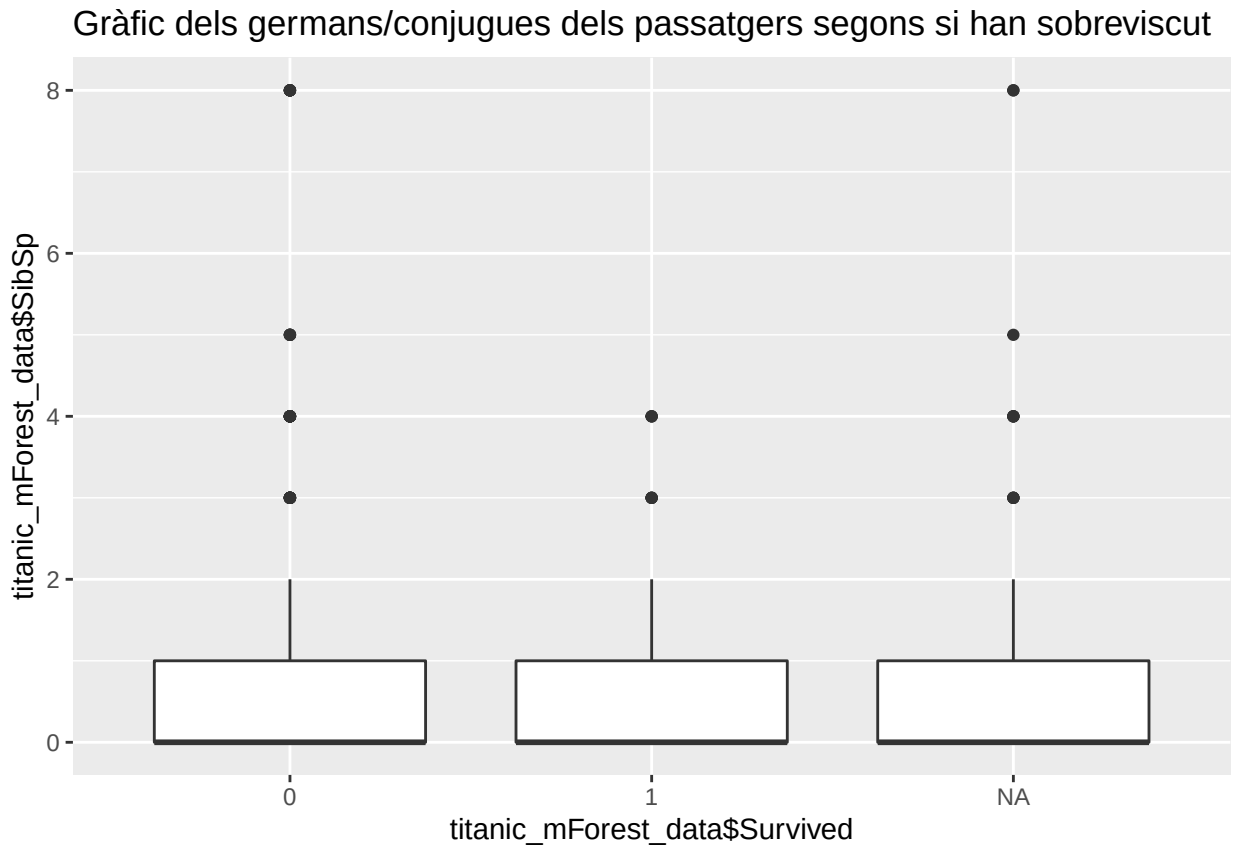

```
ggplot(data=titanic_mForest_data, aes(titanic_mForest_data$Survived, titanic_mForest_data$Parch)) + geo
```




```
boxplot(titanic_mForest_data$SibSp)
```




```
ggplot(data=titanic_mForest_data, aes(titanic_mForest_data$Survived, titanic_mForest_data$SibSp)) + geo
```



Aquestes dues variables que estàn relacionades amb les families de passatgers, poden presentar algun outlier tenint en compte que tenen valors força elevats, però tenint en compte que aquestes variables contenen informació que no serà rellevant per al nostre estudi no tractarem aquesta informació. El que farem és crear una nova variable que indiqui si el passatger viatjava sol o amb familia, i la utilitzarem per al nostres estudi.

```
titanic_mForest_data$PassAlone <- ifelse(titanic_mForest_data$SibSp + titanic_mForest_data$Parch>0, 0,1)
titanic_mForest_data$PassAlone <- as.factor(titanic_mForest_data$PassAlone)
```

4. Anàlisi de les dades.

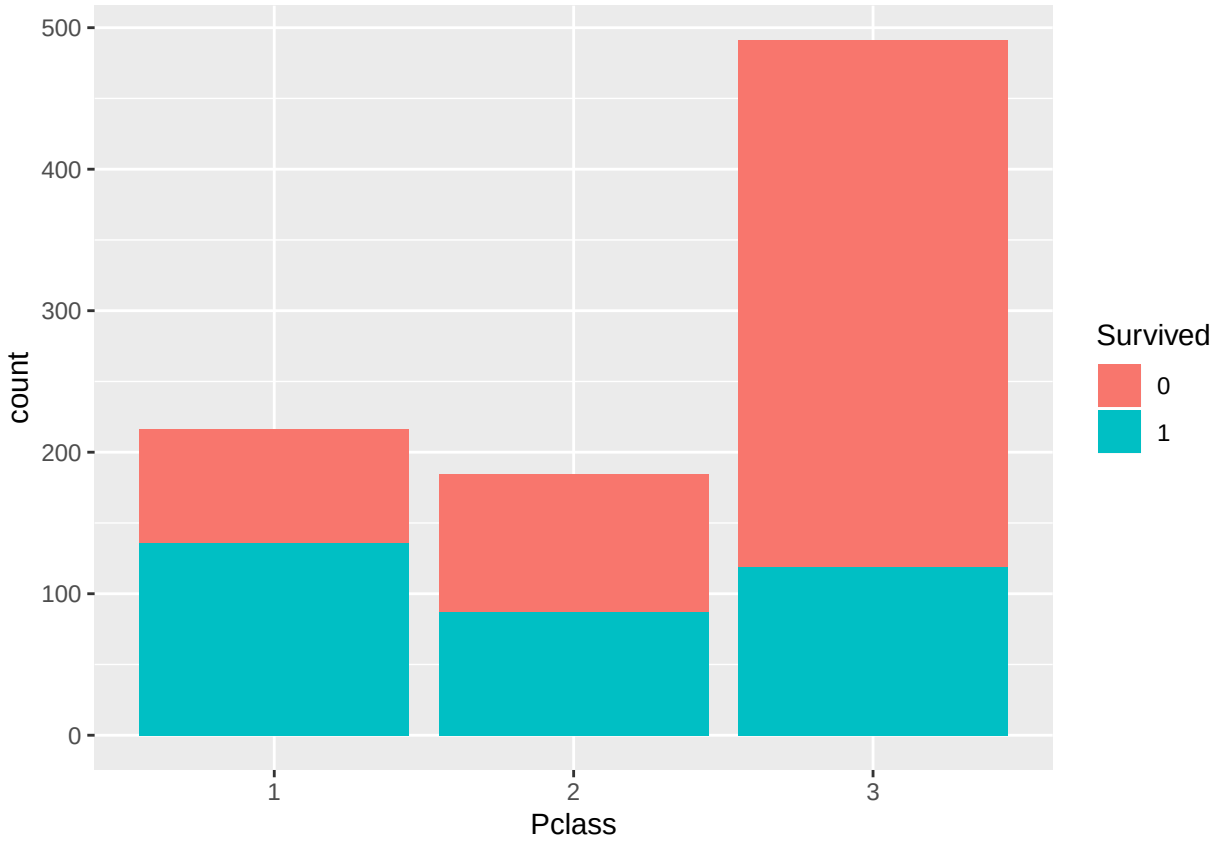
Un cop hem tractat les dades, eliminem les columnes que no utilitzarem (com *PassengerId*, *SibSp* i *Parch*),tornem a separar els datasets i els guardem com a fitxers CSV.

```
titanic_mForest_data <- subset(titanic_mForest_data,select = -c(SibSp,Parch))
titanic_train<-titanic_mForest_data[!is.na(titanic$Survived),]
titanic_test<-titanic_mForest_data[is.na(titanic$Survived),]
titanic_train<-subset(titanic_train,select = -PassengerId)
titanic_test <- subset(titanic_test,select = -Survived)
write.csv(titanic_train,"../csv/train_clean.csv")
write.csv(titanic_test,"../csv/test_clean.csv")
```

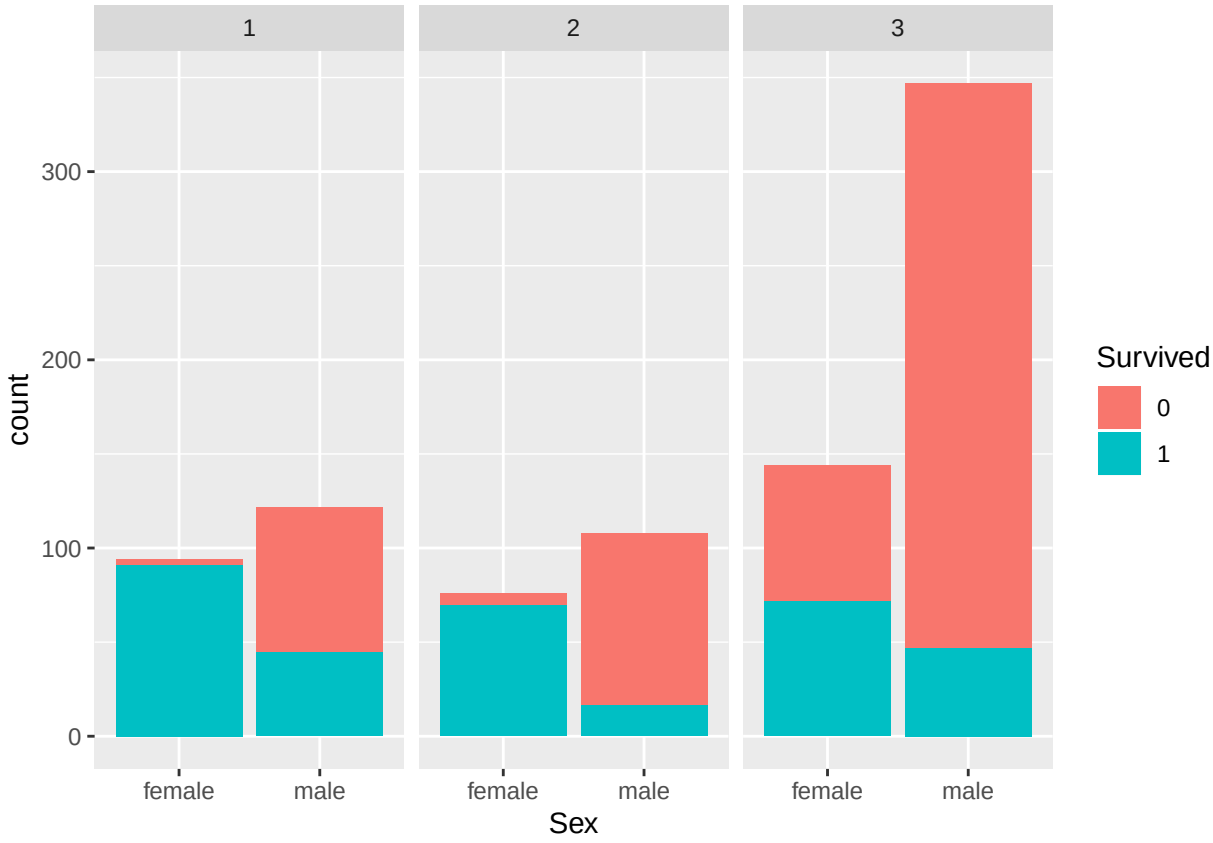
4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Abans de començar a seleccionar grups, anem a revisar la correlació entre les diverses variables de la mostra generant gràfics entre les diverses variables:

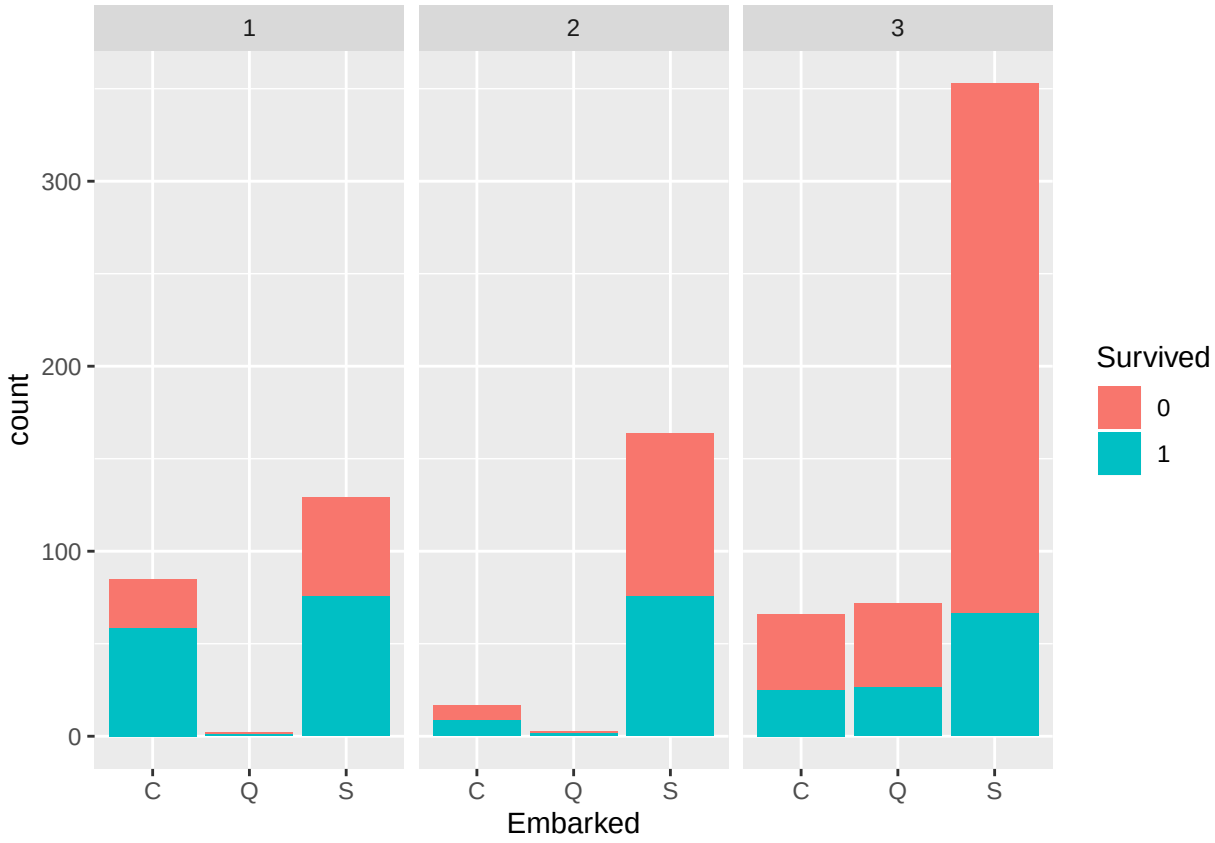
```
## Warning: Ignoring unknown aesthetics: position
```



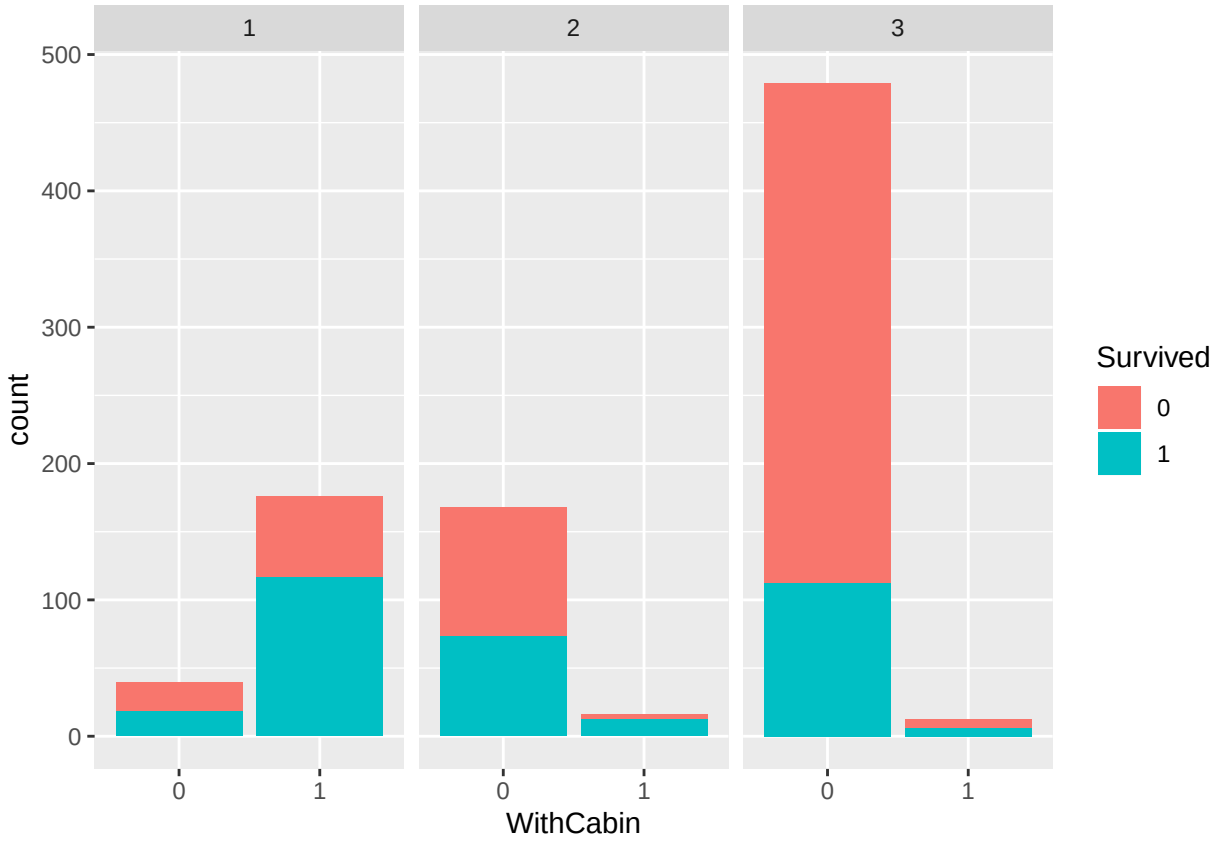

```
## Warning: Ignoring unknown aesthetics: position
```



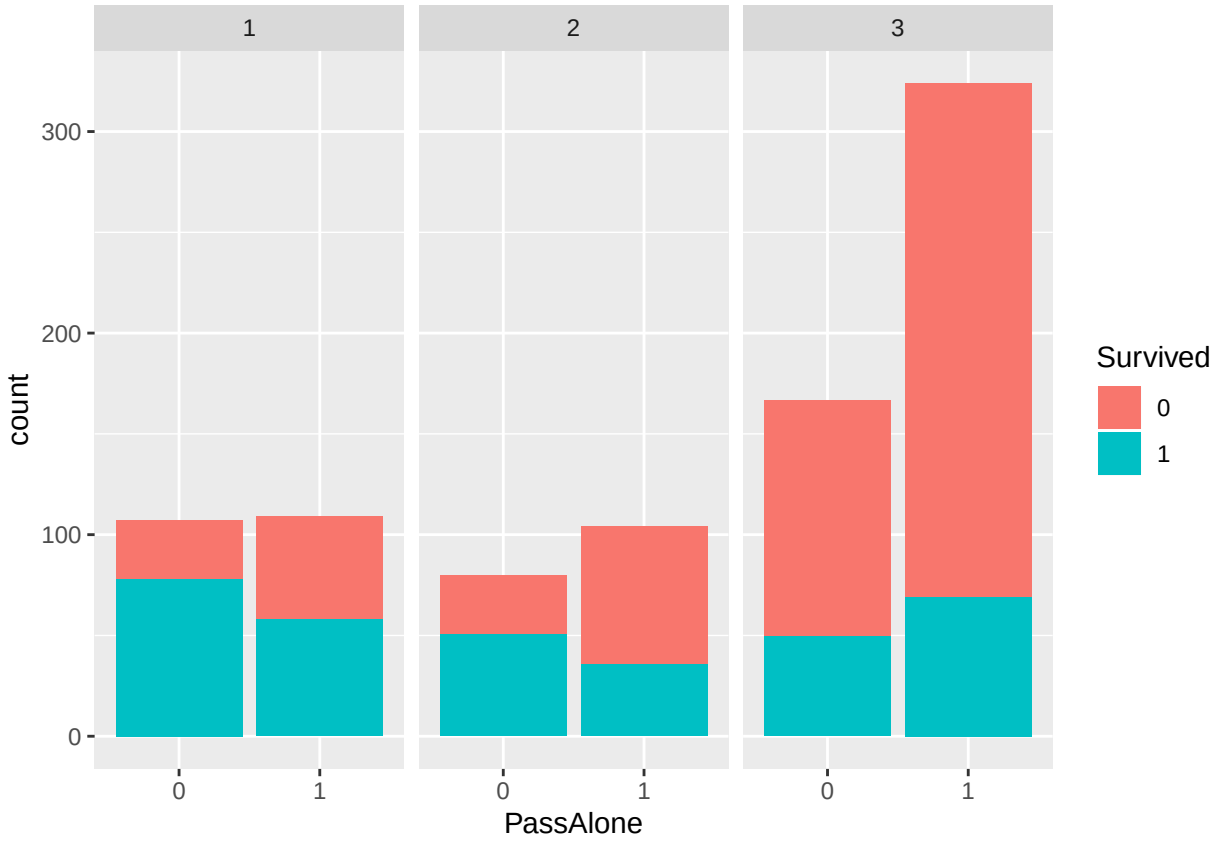

```
## Warning: Ignoring unknown aesthetics: position
```



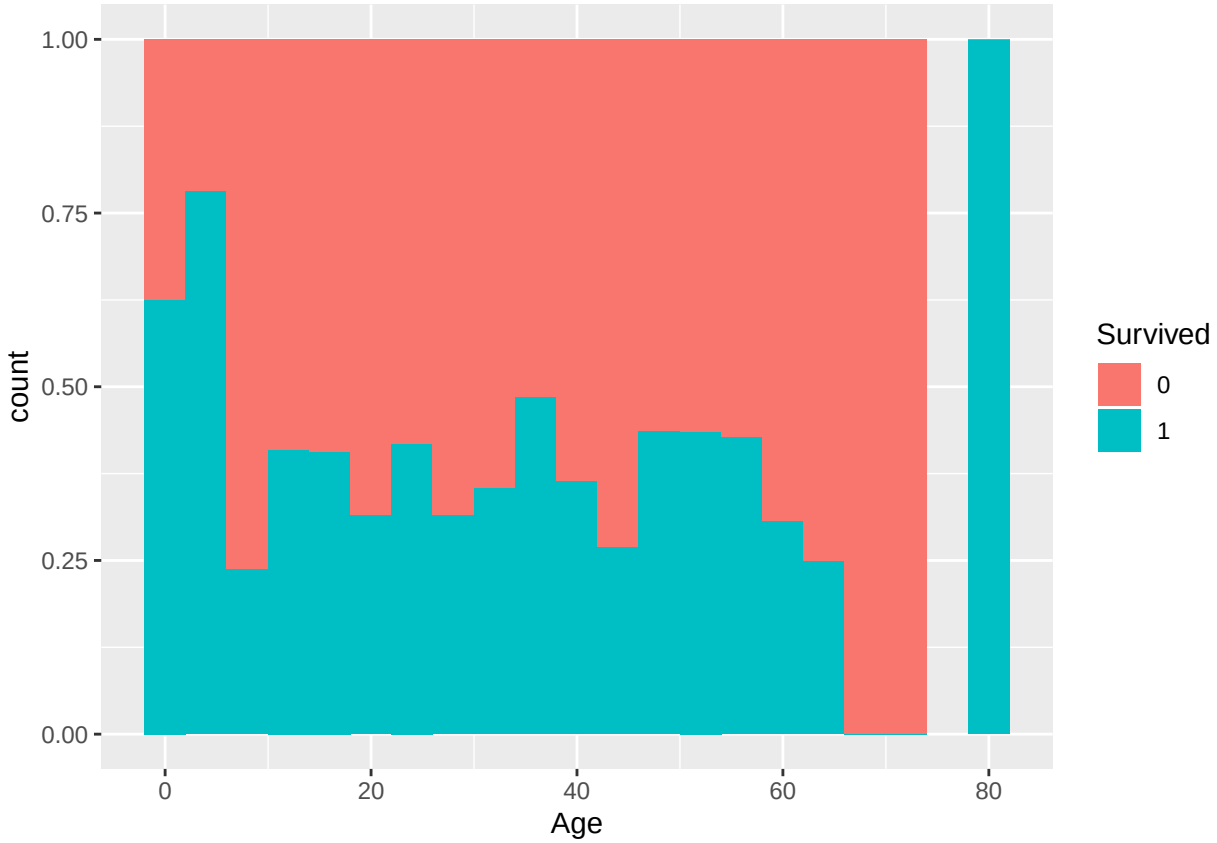

```
## Warning: Ignoring unknown aesthetics: position
```



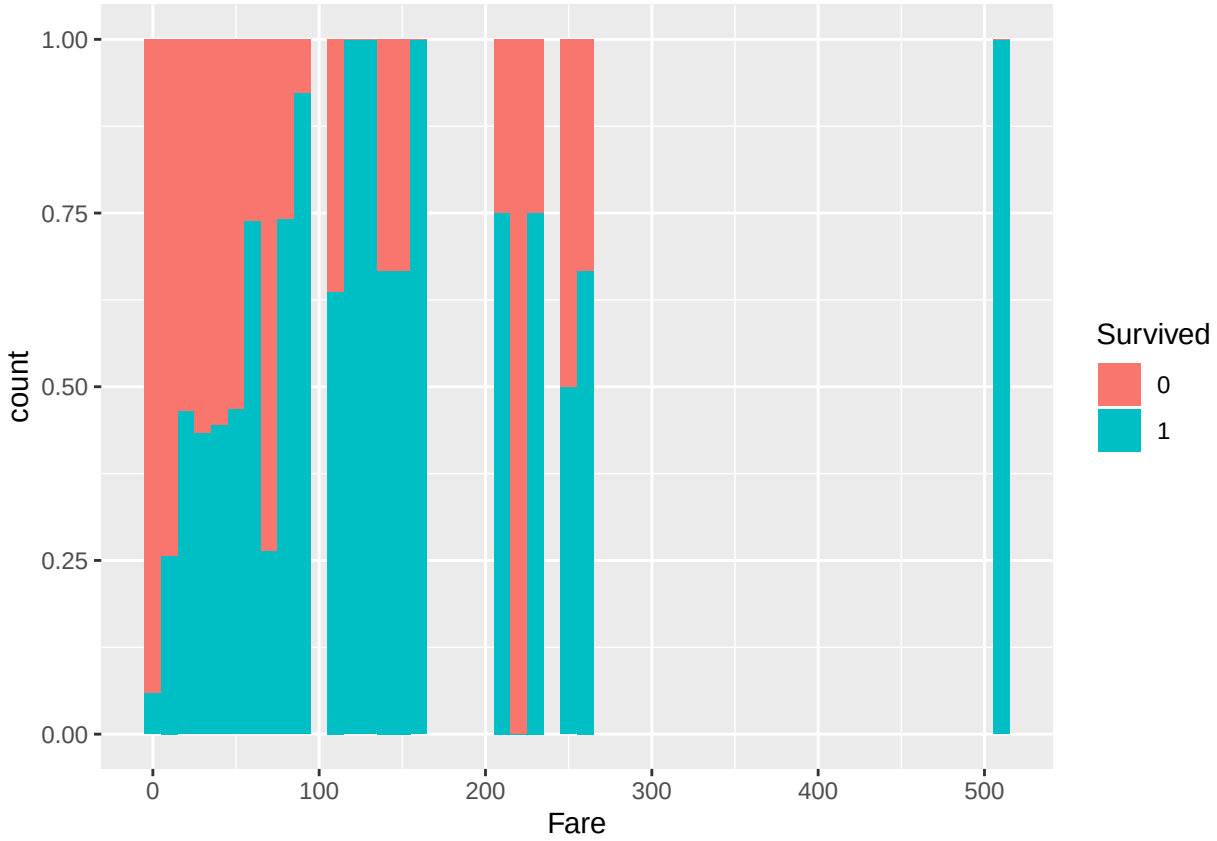

```
## Warning: Ignoring unknown aesthetics: position
```




```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
## Warning: Removed 60 rows containing missing values (geom_bar).
```



Com es pot observar amb els gràfics generats:

- El major nombre de passatgers que van sobreviure van ser de primera classe.
- Les dones van ser les que van sobreviure més del naufragi, en vers dels homes (a prod d'un 50% més).
- En quant a l'embarcament, van sobreviure més passatgers que van embarcar a Southampton, ja que va ser el port on va embarcar més gent, però si ho avaluem pel ratio dels passatgers embarcats i passatgers que van sobreviure, Cherbourg té un millor ratio, i per tant va sobreviure més gent de la que va embarcar a Cherbourg.
- Els passatgers amb cabina, van sobreviure més, en proporció, que els passatgers sense cabina.
- La variable que indica si els passatgers tenien familia o no, no sembla tenir gaire relació amb si els passatgers han sobreviscut o no.
- En quant a l'edat dels passatgers, la major mortalitat es registra entre els 8-10 anys i els 40-45 anys, tenint una mortalitat total sobre els 65 anys.
- La tarifa ens indica que contra més baixa era la tarifa més mortalitat hi va haver, encara que podem observar algunes excepcions.

Per tant, podem dir que les variables que poden tenir relació sobre la supervivència poden ser: *Pclass*, *Sex*, *Embarked*, *WithCabin*, *Age* i *Fare*.

```
titanic_train_classe1 <- titanic_train[titanic_train$Pclass==1,]
titanic_train_classe2 <- titanic_train[titanic_train$Pclass==2,]
titanic_train_classe3 <- titanic_train[titanic_train$Pclass==3,]
titanic_train_dona<- titanic_train[titanic_train$Sex==0,]
titanic_train_home <- titanic_train[titanic_train$Sex==1,]
titanic_train_classe1 <- titanic_train[titanic_train$Pclass==1,]
titanic_train_EmbC <- titanic_train[titanic_train$Embarked=="C",]
titanic_train_EmbQ <- titanic_train[titanic_train$Embarked=="Q",]
titanic_train_EmbS <- titanic_train[titanic_train$Embarked=="S",]
# Desfactoritzem les variables necessàries
titanic_train$WithCabin <- as.numeric(as.character(titanic_train$WithCabin))
titanic_train$Sex <- as.numeric(titanic_train$Sex)
titanic_train$Pclass <- as.numeric(as.character(titanic_train$Pclass))
titanic_train$PassAlone <- as.numeric(as.character(titanic_train$PassAlone))
titanic_train$Survived <-as.numeric(as.character(titanic_train$Survived))
titanic_test$WithCabin <- as.numeric(as.character(titanic_test$WithCabin))
titanic_test$Sex <- as.numeric(titanic_test$Sex)
titanic_test$Pclass <- as.numeric(as.character(titanic_test$Pclass))
titanic_test$PassAlone <- as.numeric(as.character(titanic_test$PassAlone))
```

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Per a comprovar la normalitat utilitzarem el test de *Shapiro-Wilk* sobre les variables de la mostra que son numèriques. D'altra banda al ser una mostra amb més de 30 registres, podem comsiderar el *Teorema del Limit Central* per assegurar que la mostra segueix una distribució normal.

```
shapiro.test(titanic_train[,"Age"])
```

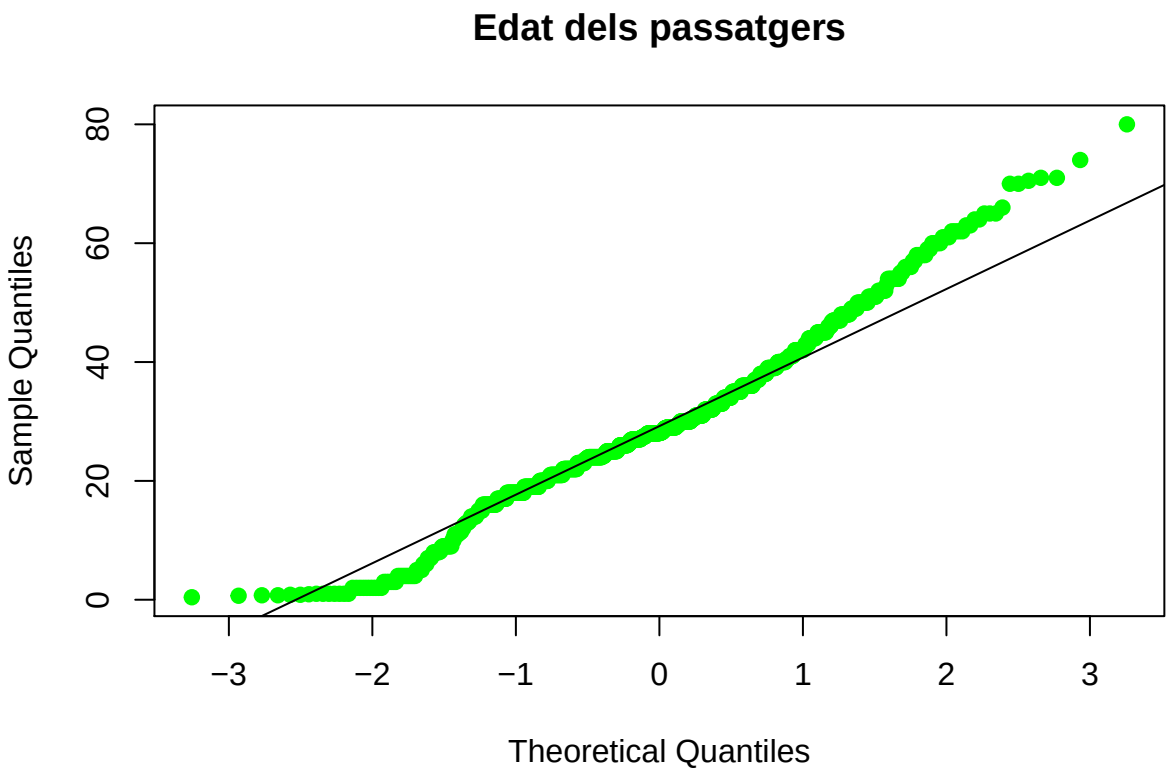
```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_train[, "Age"]
## W = 0.9802, p-value = 1.229e-09
```

```
shapiro.test(titanic_train[,"Fare"])
```

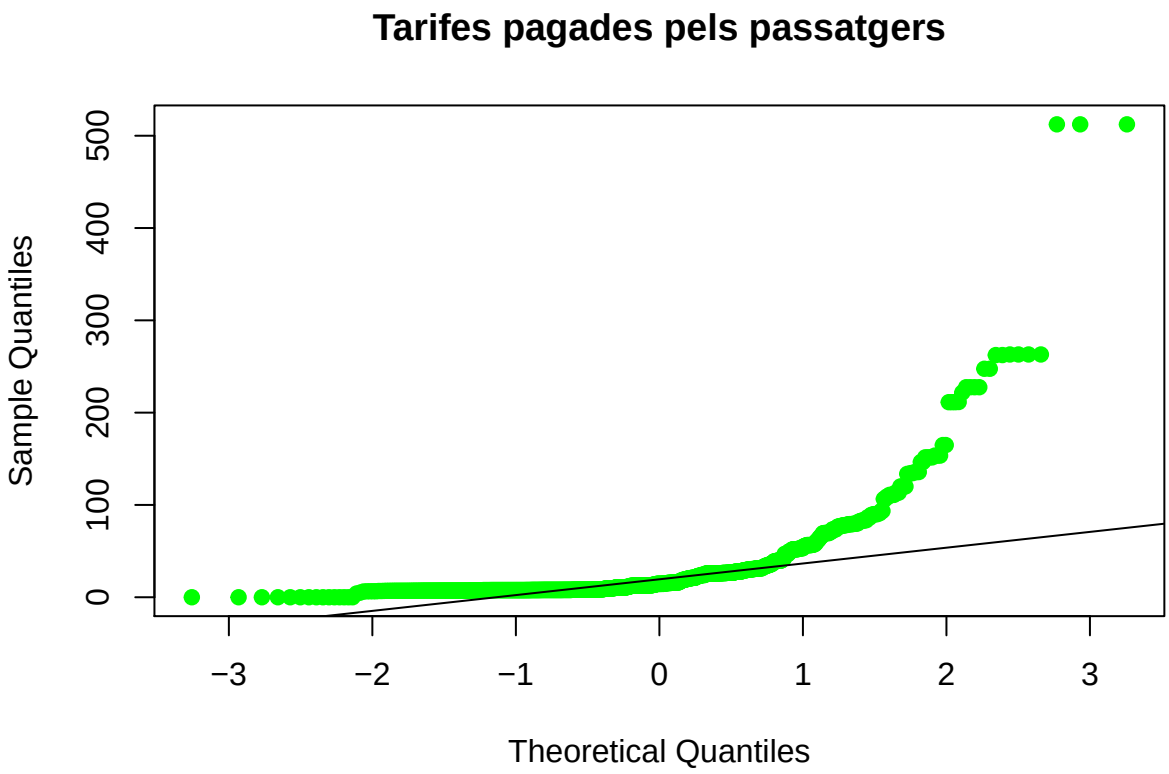
```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_train[, "Fare"]
## W = 0.52189, p-value < 2.2e-16
```

Com podem observar amb el test realitzat, el p-value és menor que el nivell de significació, i per tant no podem assegurar que segueixi una distribució normal. Si realitzem els *Q-Qplot* de les dues variables:

```
qqnorm(titanic_train$Age, pch = 19, col = "green",main="Edat dels passatgers")
qqline(titanic_train$Age)
```




```
qqnorm(titanic_train$Fare, pch = 19, col = "green",main="Tarifes pagades pels passatgers")
qqline(titanic_train$Fare)
```



Es veu com la distribució de l'edat dels passatgers s'assembla a una distribució normal, en canvi la tarifa, clarament, no segueix una distribució normal.

Els diversos tests realitzats ens indiquen que la mostra no segueix una distribució normal, però com la mostra (tant de training com de test) és suficientment elevada, segons el teorema del limit central, aquesta mostra seguirà una distribució normal.

Per estudiar la homogeneitat de les variances utilitzarem el test no paramètric de *Fligner-Killen*, ja que com hem comprobat anteriorment les variables no segueixen una distribució normal.

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 3.1669, df = 1, p-value =
## 0.07514
```

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

```
##
##  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
```

Tal com indiquen els tests, la variable *Age* al tenir un *p-value* superior a 0.05 podem dir que les variancies de les mostres son homogènies. En canvi la variable *Fare* té un *p-value* menor a 0.05 i per tant, les variancies de les mostres no son homogènies.

4.3. 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.

En aquest cas, aplicarem les següents proves estadístiques als grups de dades:

- Contrast d'Hipòtesis
- Anova
- Regressió Lineal Multivariable
- RandomForest

4.3.1. Contrast d'Hipòtesis

4.3.1.1 Contrast d'Hipòtesis variable *WithCabin*

En aquest primer contrast avaluarem si la mitjana de passatgers amb cabina que van sobreviure és igual a la mitjana de passatgers sense cabina que van sobreviure, o bé la mitjana de passatgers amb cabina que van sobreviure és menor que la mitjana de passatgers sense cabina que van sobreviure.

$$\begin{cases} H_0: & \mu_r = \mu_u \\ H_1: & \mu_r < \mu_u \end{cases}$$

Primer de tot, creem totes les variables necessàries per al test:

```
titanic_train_wcabin <- titanic_train$Survived[titanic_train$WithCabin==0]
titanic_train_wocabin <- titanic_train$Survived[titanic_train$WithCabin==1]
```

Amb aquestes variables avaluem si les variances són iguals:

```
var.test(titanic_train_wcabin, titanic_train_wocabin, conf.level=.95,alternative = "less")
```

```
##
##  F test to compare two variances
##
## data:  titanic_train_wcabin and titanic_train_wocabin
## F = 0.94148, num df = 686, denom df = 203, p-value = 0.2889
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.000000 1.128002
## sample estimates:
## ratio of variances
##          0.9414772
```

El *p-valor* és superior a 0.05, per tant no hi ha una diferència significativa entre les dues variances i podem utilitzar el mètode paramètric de variances desconegudes però iguals.

```
titanic_tstud_Cab<-t.test(titanic_train_wcabin, titanic_train_wocabin, var.equal = TRUE,alternative = "
print(titanic_tstud_Cab)
```

```
##
##  Two Sample t-test
##
## data:  titanic_train_wcabin and titanic_train_wocabin
## t = -9.9626, df = 889, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3061872
## sample estimates:
## mean of x mean of y
## 0.2998544 0.6666667
```

Aplicant la *T d'Student*, observem que el *p-value* és més petit que 0.05 i per tant podem rebutjar la hipòtesi nul · la de que la mitjana de passatgers amb cabina que van sobreviure és igual a la mitjana de passatgers sense cabina que van sobreviure. Podem afirmar que es compleix la hipòtesi alternativa de que la mitjana de passatgers amb cabina que van sobreviure és més gran que la mitjana de passatgers sense cabina que van sobreviure.

4.3.1.2 Contrast d'Hipòtesis variable *Sex*

En aquest primer contrast avaluarem si la mitjana de dones que van sobreviure és igual a la mitjana d'homes que van sobreviure, o bé la mitjana de dones que van sobreviure és més gran que la mitjana d'homes que van sobreviure.

$$\begin{cases} H_0: & \mu_r = \mu_u \\ H_1: & \mu_r > \mu_u \end{cases}$$

Primer de tot, creem totes les variables necessàries per al test:

```
titanic_train_dones <- titanic_train$Survived[titanic_train$Sex==1]
titanic_train_homes <- titanic_train$Survived[titanic_train$Sex==2]
```

Amb aquestes variables avaluem si les variances són iguals:

```
var.test(titanic_train_dones, titanic_train_homes, conf.level=.95,alternative="greater")
```

```
##
##  F test to compare two variances
##
## data:  titanic_train_dones and titanic_train_homes
## F = 1.2511, num df = 313, denom df = 576, p-value = 0.01109
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  1.064778      Inf
## sample estimates:
## ratio of variances
##           1.251103
```

El *p-valor* és inferior a 0.05, per tant hi ha una diferència significativa entre les dues variances.

```
titanic_tstud_Sex<-t.test(titanic_train_dones, titanic_train_homes, var.equal = FALSE,alternative = "gr
print(titanic_tstud_Sex)
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic_train_dones and titanic_train_homes
## t = 18.672, df = 584.43, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5043259       Inf
## sample estimates:
## mean of x mean of y
## 0.7420382 0.1889081
```

Aplicant la *T d'Student*, observem que el *p-value* és menor que 0.05 i per tant podem rebutjar la hipòtesi nul · la de que la mitjana de dones que van sobreviure és igual a la mitjana d'homes que van sobreviure i per tant podem afirmar que es compleix la hipòtesi alternativan de que la mitjana de dones que van sobreviure és més gran que la mitjana d'homes que van sobreviure.

4.3.2. Anova multifactorial

En aquest cas volem contrastar la hipòtesi de diverses variables de la mostra on la hipòtesi nul · la és que totes les mitjanes poblacionals de la mostra són iguals, i la hipòtesi alternativa, que no totes les mitjanes poblacionals són iguals. Utilitzarem les variables *Pclass*, *Sex* i *Age*.

```
titanic_train_aov <- aov(Survived~Pclass+Sex+Age+Pclass:Sex+Pclass:Age+Sex:Age,data=titanic_train)
summary(titanic_train_aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Pclass        1  24.14   24.14 168.645  < 2e-16 ***
## Sex           1  53.34   53.34 372.577  < 2e-16 ***
## Age           1   3.31    3.31  23.091 1.82e-06 ***
## Pclass:Sex    1   2.62    2.62  18.292 2.10e-05 ***
## Pclass:Age    1   0.01    0.01   0.038    0.845
## Sex:Age       1   0.77    0.77   5.349    0.021 *
## Residuals   884 126.55    0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tal com es pot observar, la major part dels p-valors són menors que 0.05, però hi ha dos que els valors si que són més grans. Tot i això podem rebutjar la hipòtesi nul · la i assegurar que no totes les mitjanes de les variables són iguals. Si avaluem el valor F, el factor *Sex* té el valor més alt, seguit del factor *Pclass* per tant aquests dos factors són més significatius que els altres.

4.3.3. Regressió Lineal Multifactorial

Un cop hem avaluat les diverses variables de la mostra per a veure si poden ser importants en el model o no, anem a utilitzar un model de regressió lineal per a realitzar les prediccions.

```r
titanic_train$Survived <- as.numeric(titanic_train$Survived)
titanic_lr <- lm(formula=Survived ~ Pclass + Sex + Age + Fare + Embarked + WithCabin + PassAlone,data =
summary(titanic_lr)
```

```
##
## Call:
## lm(formula = Survived ~ Pclass + Sex + Age + Fare + Embarked +
##     WithCabin + PassAlone, data = titanic_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08243 -0.19564 -0.07624  0.24220  1.00946
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7209297  0.0952387  18.070  < 2e-16 ***
## Pclass      -0.1563907  0.0256306  -6.102 1.57e-09 ***
## Sex         -0.4871768  0.0286876 -16.982  < 2e-16 ***
## Age         -0.0052080  0.0010929  -4.765 2.20e-06 ***
## Fare        -0.0001491  0.0003248  -0.459   0.6462
## EmbarkedQ   -0.0022107  0.0556489  -0.040   0.9683
## EmbarkedS   -0.0737710  0.0343627  -2.147   0.0321 *
## WithCabin    0.1118530  0.0447571   2.499   0.0126 *
## PassAlone    0.0224715  0.0290631   0.773   0.4396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3809 on 882 degrees of freedom
## Multiple R-squared:  0.3928, Adjusted R-squared:  0.3873
## F-statistic: 71.33 on 8 and 882 DF,  p-value: < 2.2e-16
```

Revisant els regressors del model de regressió resultant individualment, es pot observar que tots els valors són menors de 0.05 excepte *Fare*, *EmbarkedQ* i *PassAlone* i per tant en tots aquests valors que són menors, podem rebutjar la hipòtesi nul · la i dictaminar que aquests regressors són vàlids per a predir la qualitat del son. En canvi, els altres regressors tenen un valor superior a 0.05 i per tant, no podem rebutjar la hipòtesi nul · la de que no és un regressor vàlid per a predir la supervivència del passatger.

Segons els valors d'R2 i p-value obtinguts al model, estem davant d'un model poc precís, ja que explica el **39.28%** de la variabilitat de la supervivència del passatger.

Amb el model creat, anem a predir la supervivència dels passatgers de la mostra de test:

```r
titanic_lr_pred <- data.frame(predict.lm(titanic_lr,newdata=titanic_test,interval="prediction"))
titanic_lr_pred_df<- data.frame(ifelse(titanic_lr_pred$fit<0.5,0,1))
colnames(titanic_lr_pred_df) <- c("Survived")
titanic_lr_cm <- with(gender_submission,table(titanic_lr_pred_df$Survived, Survived))
titanic_lr_pred_error <- 100 * titanic_lr_cm[2] / sum(titanic_lr_cm)
Output_lr<- data.frame(PassengerID = gender_submission$PassengerId, Survived = titanic_lr_pred_df$Survi
write.csv(Output_lr, file = "../csv/test_lr_pred.csv")
```

Un cop realitzada la predicció, veiem que el ratio d'encert és del **36.36%**, un ratio de predicció molt baix, que ja ens esperàvem després d'haver vist els resultats del model.

4.3.4. RandomForest

Finalment, aplicarem un arbre de classificació de tipus *RandomForest*, ja que aquests tipus d'arbres milloren la tasa de classificació ja que combinen el resultat de multiples Arbres de Decisió en diferents mostres per reduïr la variació en les prediccions, i així minimitzar l'Over-fitting que es produeix amb els Arbres de Decisió normals.

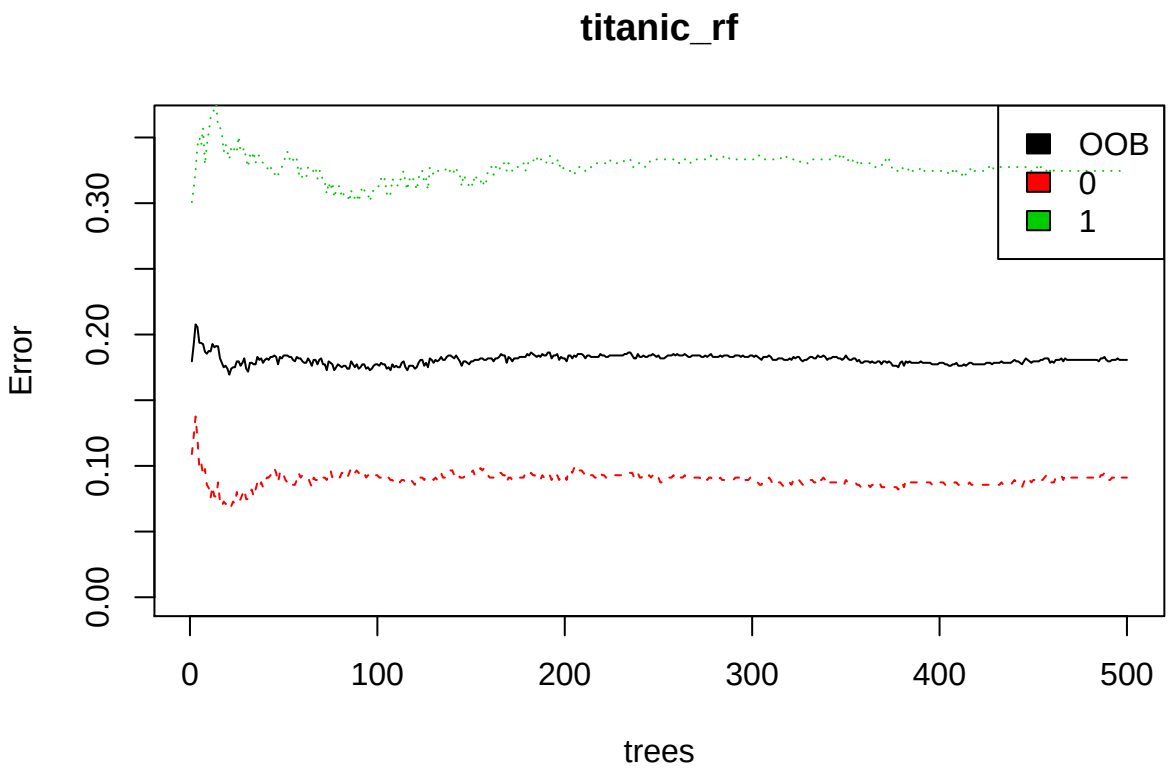
Executem el *RandomForest* sobre totes les variables que tenim en la mostra de training:

```
titanic_train$Survived <- as.factor(titanic_train$Survived)
titanic_rf <- randomForest(Survived ~ Pclass + Sex + Age + Fare + Embarked + WithCabin + PassAlone,data
titanic_rf
```

```
##
## Call:
##  randomForest(formula = Survived ~ Pclass + Sex + Age + Fare +      Embarked + WithCabin + PassAlone
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 18.07%
## Confusion matrix:
##     0   1 class.error
## 0 499  50  0.09107468
## 1 111 231  0.32456140
```

L'estimació de l'error és d'un 18% amb aquest mètode. Generem la gràfica amb els errors de predicció:

```
plot(titanic_rf, ylim=c(0,0.36))
legend('topright', colnames(titanic_rf$err.rate), col=1:3, fill=1:3)
```



En la gràfica podem observar els errors en la predicció tant de passatgers vius com morts, com la mitjanan entre els dos valors. Podem observar que és més fàcil predir els passatgers que moren que els que viuen.

```
titanic_rf$importance
```

```
##             MeanDecreaseGini
## Pclass             29.571521
## Sex                97.505422
## Age                54.496134
## Fare               57.177902
## Embarked            9.930890
## WithCabin          16.680179
## PassAlone           7.443058
```

La importància dels diversos paràmetres en la classificació queda palesa en la taula anterior, on es pot observar que els paràmetres *Sex*, *Fare*, *Age* i *Pclass* són els més importants per a la classificació. Amb el model generat, anem a generar la predicció de la mostra de test:

```
titanic_rf_pred <- predict(titanic_rf,titanic_test)
(titanic_rf_cm <- with(gender_submission,table(titanic_rf_pred, Survived)))
```

```
##                Survived
## titanic_rf_pred   0   1
##               0 247  41
##               1  19 111
```

```r
titanic_rf_pred_error <- 100 * sum(diag(titanic_rf_cm)) / sum(titanic_rf_cm)
print(titanic_rf_pred_error)
```

```
## [1] 85.64593
```

```r
Output_rf<- data.frame(PassengerID = gender_submission$PassengerId, Survived = titanic_rf_pred)
write.csv(Output_rf, file = "../csv/test_rf_pred.csv")
```

Com podem observar, el model ha predit correctament un **85.65%** dels casos de la mostra de test.

5. Representació dels resultats a partir de taules i gràfiques.

En tot l'estudi que estem realitzant, a part de realitzar la neteja de les dades, també hem aplicat proves estadístiques sobre les dades, primer, per obtenir més informació sobre les variables de la mostra a l'hora d'utilitzar-les en models de predicció així com els models de predicció que s'han generat i testejat amb les mostres.

Primer de tot hem realitzat dos contrast d'hipòtesis sobre les variables *WithCabin* i *Sex*:

```r
print(titanic_tstud_Cab)
```

```
##
##  Two Sample t-test
##
## data:  titanic_train_wcabin and titanic_train_wocabin
## t = -9.9626, df = 889, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3061872
## sample estimates:
## mean of x mean of y
## 0.2998544 0.6666667
```

```r
print(titanic_tstud_Sex)
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic_train_dones and titanic_train_homes
## t = 18.672, df = 584.43, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5043259       Inf
## sample estimates:
## mean of x mean of y
## 0.7420382 0.1889081
```

Per a la variable *WithCabin*, podem afirmar que es compleix la hipòtesi alternativa de que la mitjana de passatgers amb cabina que van sobreviure és més gran que la mitjana de passatgers sense cabina que van sobreviure.

Per a la variable *Sex*, podem afirmar que es compleix la hipòtesi alternativan de que la mitjana de dones que van sobreviure és més gran que la mitjana d'homes que van sobreviure.

Seguidament hem realitzat una ANOVA Multifactorial per contrastar la hipòtesi de diverses variables de la mostra on la hipòtesi nul · la és que totes les mitjanes poblacionals de la mostra són iguals, i la hipòtesi alternativa, que no totes les mitjanes poblacionals són iguals. Utilitzarem les variables Pclass, Sex i Age.

```
summary(titanic_train_aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Pclass        1  24.14   24.14 168.645  < 2e-16 ***
## Sex           1  53.34   53.34 372.577  < 2e-16 ***
## Age           1   3.31    3.31  23.091 1.82e-06 ***
## Pclass:Sex    1   2.62    2.62  18.292 2.10e-05 ***
## Pclass:Age    1   0.01    0.01   0.038    0.845
## Sex:Age       1   0.77    0.77   5.349    0.021 *
## Residuals   884 126.55    0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultat obtingut indica que la major part dels p-valors són menors que 0.05, però hi ha dos que els valors si que són més grans. Tot i això podem rebutjar la hipòtesi nul · la i assegurar que no totes les mitjanes de les variables són iguals. Si avaluem el valor F, el factor *Sex* té el valor més alt, seguit del factor *Pclass* per tant aquests dos factors són més significatius que els altres.

Un cop hem obtingut informació sobre les dades, passem a aplicar models de predicció. Primer, s'ha crear un model de regressió lineal:

```
print(titanic_lr_cm)
```

```
##    Survived
##       0   1
##   0 256   5
##   1  10 147
```

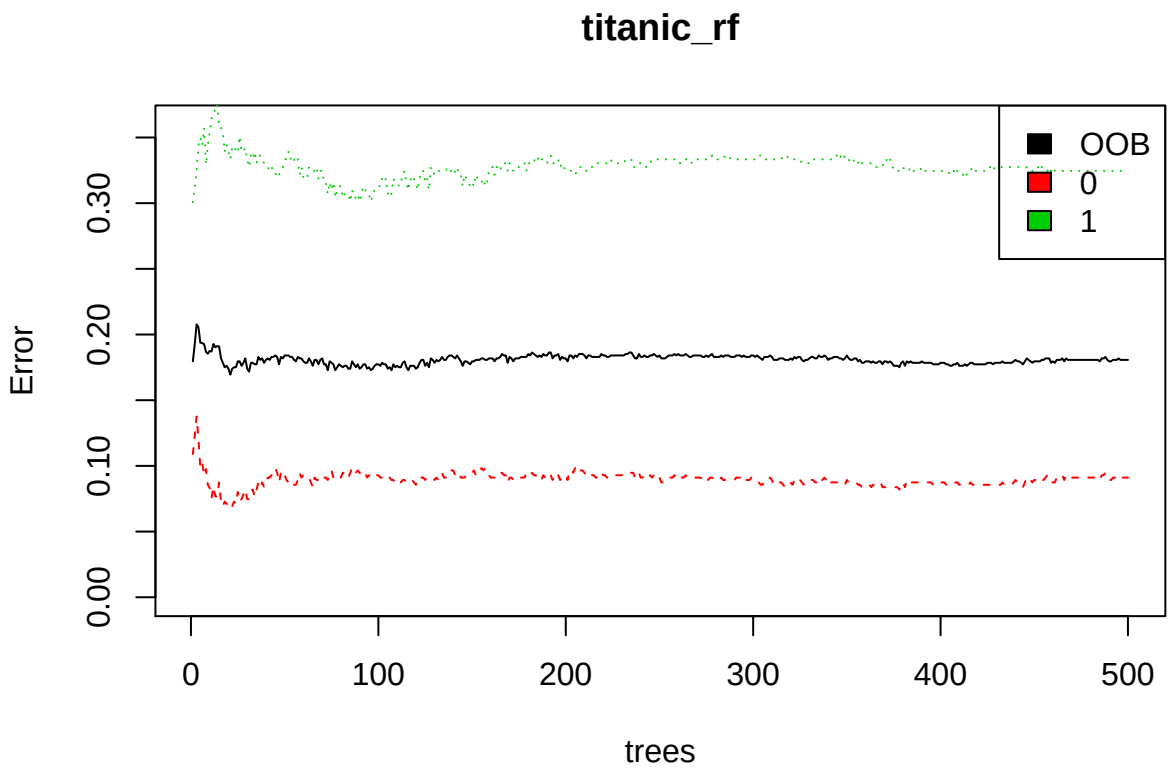
```
print(titanic_lr_pred_error)
```

```
## [1] 2.392344
```

Segons els valors d'R2 i p-value obtinguts al model,aquest model és poc precís, ja que explica el **39.28%** de la variabilitat de la supervivència del passatger. Aquests valors del model s'han traduït en un ratio d'encert de predicció molt baix, **36.36%**, i per tant, aquest model no seria un bon model per a la predicció dels supervivents del Titanic.

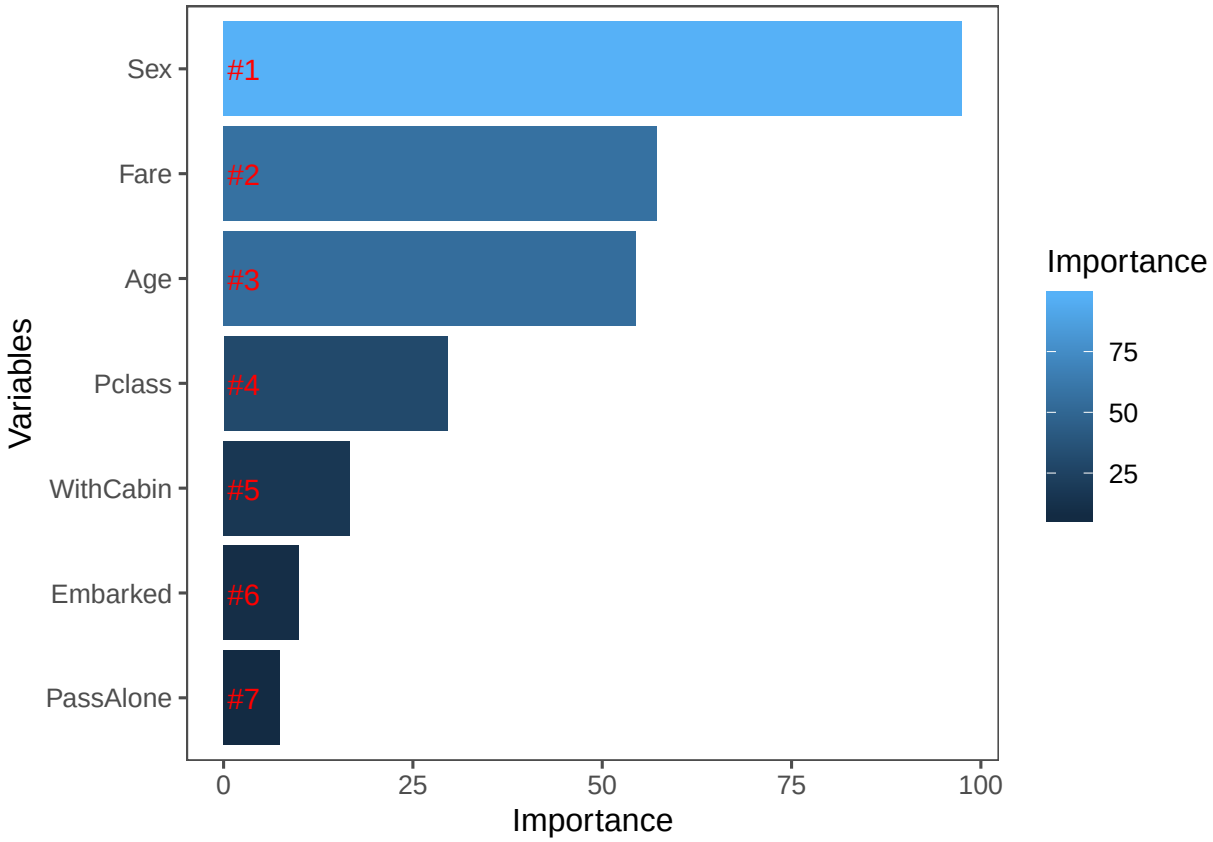
Un cop avaluat el model de regressió lineal, hem provat amb un model *RandomForest*:

```
plot(titanic_rf, ylim=c(0,0.36))
legend('topright', colnames(titanic_rf$err.rate), col=1:3, fill=1:3)
```



En la gràfica podem observar els errors en la predicció tant de passatgers vius com morts, com la mitjanan entre els dos valors. Podem observar que és més fàcil predir els passatgers que moren que els que viuen. L'estimació de l'error és d'un **18%**.

```
importance    <- importance(titanic_rf)
varImportance <- data.frame(Variables = row.names(importance),
                            Importance = round(importance[ ,'MeanDecreaseGini'],2))
rankImportance <- varImportance %>% mutate(Rank = paste0('#',dense_rank(desc(Importance))))
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()
```



Aquest gràfic ens explica de manera visual la importància dels diversos paràmetres del model, on es pot observar que els paràmetres *Sex*, *Fare*, *Age* i *Pclass* són els més importants per a la classificació.

```
print(titanic_rf_cm)
```

```
##                 Survived
## titanic_rf_pred   0   1
##               0 247  41
##               1  19 111
```

```
print(titanic_rf_pred_error)
```

```
## [1] 85.64593
```

Com podem observar, el model ha predit correctament un **85.65%** dels casos de la mostra de test.

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

Primer de tot, hem carregat les dades del dataset **Titanic: Machine Learning from Disaster**. hem realitzat una revissió de les dades per a netejar, corregir i crear variables en el dataset, per a després poder aplicar diversos mètodes estadístics.

La idea era estudiar aquest dataset amb diversos models per a poder predir la supervivència o no d'un passatger del Titanic. Per això, el primer que hem fet és revisar la interelació entre les diverses variables del dataset i ens ha generat un nou subset de dades on les variables que podien tenir relació amb la supervivència eren *Pclass*, *Sex*, *Embarked*, *WithCabin*, *Age* i *Fare*. Amb aquest primer cribatge hem analitzat la relació entre elles, observant que *Sex* i *Pclass* podien ser els factors més significatius de la mostra a l'hora de realitzar prediccions.

Per a fer proves de prediccions, hem seleccionat dos models: regressió lineal i random forest. El primer model no s'adaptava gaire bé a les dades, i per tant no és un bon model per a utilitzar en prediccions d'aquest dataset. En canvi el segon model, al ser un model que minimnitza l'over-fitting de les dades, ens ha donat uns resultats força bons, ja que la predicció ha arribat quasi a un **86%**.

Tot i que a la mostra tenim força variables i diverses, els mètodes emprats, ens han demostrat, que és un subconjunt més reduït d'aquestes variables, les que permeten una millor predicció del model, com són *Pclass*, *Sex* i *Age*.

Amb aquests resultats, i afinant una mica més les diverses variables de la mostra podriem arribar a aconseguir un model amb una predicció millor, però tampoc seria gaire millor. Per tant, aquest resultat obtingut pel procés crec que és un bon resultat.