Pràctica 2: Neteja i validació de dades

Tipologia i cicle de vida de les dades

Jesús Marí i Víctor Boix

Data de presentació: 11/06/2019

Índex

---

1. Descripció del dataset

El dataset utilitzat en aquesta pràctica s'ha obtingut de la pàgina web de *Kaggle* (https://www.kaggle.com/c/titanic) i conté les dades dels passatgers del Titanic durant el desastre de 1912. L'objectiu d'aquest conjunt de dades és realitzar una predicció sobre la supervivència d'alguns passatgers a partir de les seves característiques. Per tant, es tracta d'un problema de classificació d'aprenentatge supervisat, on cal construir un model capaç de determinar el valor de l'atribut *Survived* (variable objectiu o dependent) a partir de la resta d'atributs (variables independents).

Les conjunt de dades està format per dos fitxers CSV:

- *Titanic_train.csv*. Conjunt d'entrenament que servirà per entrenar el model, per tant, conté l'atribut objectiu o classe *Survived*. Està format per 891 registres i 12 atributs.
- *Titanic_test.csv*. Conjunt de prova, conté els registres sobre els que cal realitzar la predicció. Està format per 418 registres i 11 atributs.

La informació que contenen els atributs és la següent:

- *PassengerId*. Identificador únic del viatger al dataset (nombre enter).
- *Survived*. Atribut que indica si el passatger va sobreviure a la catàstrofe (1: va sobreviure, 0: va morir). Només disponible al conjunt *train*.
- *Pclass*. Indica la classe en què viatjava el passatger (1: primera classe, 2: segona classe, 3: tercera classe).
- *Name*. Nom complet del viatger (cadena de text).
- *Sex*. Sexe del viatger (male: home, female: dona).
- *Age*. Edat del viatger en anys (nombre real).
- *SibSp*. Nombre de germans, germanes o marit/muller a bord del vaixell (nombre enter).
- *Parch*. Nombre de pares o fills a bord del vaixell (nombre enter).
- *Ticket*. Nombre del tiquet (cadena de text).
- *Fare*. Import pagat pel bitllet (nombre real).
- *Cabin*. Cabina on s'allotjava el passatger (cadena de text).
- *Embarked*. Port d'embarcament del viatger (C: Cherbourg, Q: Queenstown, S: Southampton).

2. Integració i selecció de les dades

Per tal de realitzar els processos de neteja i anàlisi de les dades carregarem els dos fitxers en dos dataframes que anomenarem *train* i *test*. Totes les operacions de neteja les realitzarem sobre els dos dataframes.

```
# Càrrega dels fitxers de dades
train <- read.csv('../data/Titanic_train.csv')
test <- read.csv('../data/Titanic_test.csv')
```

L'atribut *PassengerId* és un identificador numèric dels registres del dataframe que no ens aporta cap informació sobre els passatgers, per tant, l'eliminarem, ja que per identificar els registres ja disposem de l'índex.

```
# Eliminem l'atribut PassengerId a train i test
train <- train[,-1]
test <- test[,-1]
```

La resta d'atributs els mantindrem perquè ens donen informació sobre el viatger i poden resultar útils per al model. Podem fer una visualització dels primers registres amb la funció *head()*.

```
# Primers registres del dataframe train
head(train, 3)
```

```
##   Survived Pclass                                               Name
## 1        0      3                            Braund, Mr. Owen Harris
## 2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
## 3        1      3                             Heikkinen, Miss. Laina
##      Sex Age SibSp Parch           Ticket    Fare Cabin Embarked
## 1   male  22     1     0        A/5 21171  7.2500              S
## 2 female  38     1     0         PC 17599 71.2833   C85        C
## 3 female  26     0     0 STON/O2. 3101282  7.9250              S
```

3. Neteja de les dades

Comencem examinant el dataframe amb la funció *str()*, que ens mostra el tipus de dada i els valors dels primers registres per a cada atribut.

```
# Examinem el conjunt train
str(train)
```

```
## 'data.frame':    891 obs. of  11 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 ...
##  $ 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 ...
```

```
# Examinem el conjunt test
str(test)
```

```
## 'data.frame':    418 obs. of  10 variables:
##  $ 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 58 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 ...
```

L'atribut *Name* és de tipus factor, però té tants nivells com registres, per tant el convertirem a *character*.

```
# Conversió de Name a character
train[,'Name'] <- as.character(train[,'Name'])
test[,'Name'] <- as.character(test[,'Name'])
```

La variable *Survived* és l'atribut classe i, per tant, és de tipus categòric. En aquest cas farem la conversió a factor.

```
# Conversió de Survived a factor
train[,"Survived"] <- as.factor(train[,"Survived"])
```

Els tipus de dada per a cada atribut són:

- *Factor*: Sex, Ticket, Cabin, Embarked, Survived
- *int*: Pclass, SibSp, Parch
- *num*: Age, Fare
- *chr*: Name

Abans d'analitzar els elements buit i extrems pot resultar útil mostrar un resum de les dades amb la funció *summary()*.

```
# Resum de les dades train
summary(rbind(train[-1],test))
```

```
##      Pclass          Name               Sex           Age       
##  Min.   :1.000   Length:1309        female:466   Min.   : 0.17  
##  1st Qu.:2.000   Class :character   male  :843   1st Qu.:21.00  
##  Median :3.000   Mode  :character                Median :28.00  
##  Mean   :2.295                                   Mean   :29.88  
##  3rd Qu.:3.000                                   3rd Qu.:39.00
```

```
##  Max.   :3.000                                      Max.   :80.00
##                                                     NA's   :263
##      SibSp           Parch           Ticket         Fare
##  Min.   :0.0000   Min.   :0.000   CA. 2343:  11   Min.   :  0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1601    :   8   1st Qu.:  7.896
##  Median :0.0000   Median :0.000   CA 2144 :   8   Median : 14.454
##  Mean   :0.4989   Mean   :0.385   3101295 :   7   Mean   : 33.295
##  3rd Qu.:1.0000   3rd Qu.:0.000   347077  :   7   3rd Qu.: 31.275
##  Max.   :8.0000   Max.   :9.000   347082  :   7   Max.   :512.329
##                                   (Other) :1261   NA's   :1
##              Cabin      Embarked
##                 :1014    :  2
##  C23 C25 C27    :   6   C:270
##  B57 B59 B63 B66:   5   Q:123
##  G6             :   5   S:914
##  B96 B98        :   4
##  C22 C26        :   4
##  (Other)        : 271
```

3.1. Elements buits

Per començar examinem els elements buits a les columnes dels dos conjunts de dades. Podem veure que tenim molts valors NA a la columna *Age* dels dos dataframes i un a la columna *Fare* del conjunt de test.

```
# Nombre de valors NA per atribut
sapply(train, function(x) sum(is.na(x)))
```

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

```
sapply(test, function(x) sum(is.na(x)))
```

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

A més, els atributs *Cabin* i *Embarked* també contenen valors buits.

```
# Nombre de valors buits per atribut
sapply(train, function(x) nrow(train[x=='',]))
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch   Ticket
##        0        0        0        0      177        0        0        0
##     Fare    Cabin Embarked
##        0      687        2
```

```
sapply(test, function(x) nrow(test[x=='',]))
```

```
##   Pclass     Name      Sex      Age    SibSp    Parch   Ticket     Fare
##        0        0        0       86        0        0        0        1
##    Cabin Embarked
##      327        0
```

A continuació analitzarem els atributs que contenen valors NA o buit i els hi assignarem valors utilitzant diferents tècniques.

Fare

Per imputar el valor buit del preu del bitllet (*Fare*) podem utilitzar la informació de la classe del viatger (*Pclass*), ja que és lògic pensar que els bitllets de millor classe seran més cars. Això ho podem confirmar mostrant la mitjana de *Fare* per a cada valor de *Pclass*.

```
# Mitjana de Fare per cada valor de Pclass als conjunts train i test conjuntament
aggregate(Fare~Pclass, data=rbind(train[,-1],test), mean)
```

```
##   Pclass     Fare
## 1      1 87.50899
## 2      2 21.17920
## 3      3 13.30289
```

Com veiem, hi ha una diferència significativa entre la mitjana del preu de bitllet per a cada classe. Utilitzarem aquesta circumstància per imputar el valor NA de *Fare* amb la mitjana de valors per als viatgers de la mateixa classe.

```
# Imputem el valor NA de Fare amb la mitjana de valors de la mateixa classe
test[is.na(test$Fare),'Fare'] <- aggregate(Fare~Pclass, data=rbind(train[,-1],test),
                                           mean)[test[is.na(test$Fare),'Pclass'],'Fare']
```

Cabin

En el cas de l'atribut *Cabin*, substituirem els valors buits pel marcador 'NO'.

```
levels(train$Cabin)[levels(train$Cabin)==''] <- 'NO'
levels(test$Cabin)[levels(test$Cabin)==''] <- 'NO'
```

Embarked

Per imputar els valors buits del port d'embarcament (*Embarked*) podem aprofitar la informació del bitllet. Si consultem les dades dels viatgers que tenen un valor buit a *Embarked* veiem que tenen el mateix número de bitllet.

```
# Ticket dels viatgers amb valor buit a Embarked
train[train$Embarked=='',c("Embarked", "Ticket")]
```

```
##     Embarked Ticket
## 62          113572
## 830         113572
```

Si ampliem la cerca i mostrem el valor d'*Embarked* per a tots els viatgers amb un número de bitllet semblant (113XXX) veiem que la majoria han pujat a Southampton ('S').

```
# Viatgers amb un número de ticket semblant a l'anterior agrupats pel valor d'Embarked
 summary(train[grep("113.*",train$Ticket),"Embarked"])
```

```
##     C  Q  S
##  2  7  0 42
```

Per tant assignarem aquest valor als dos valors buits d'*Embarked*.

```
# Assignem el valor 'S' als valors buits d'Embarked
levels(train$Embarked) <- c("S", "C", "Q", "S")
```

Age

Finalment, per imputar el valor buit de l'edat (*Age*) farem servir el mètode basat en els k-veïns més propers (KNN). El que fa l'algorisme knn-imputation és identificar les k-observacions més properes i assignar la

mitjana ponderada a l'atribut amb valor NA. En el nostre cas utilitzarem k=3 i tindrem en compte tots els atributs excepte *Name*.

```
# Imputació de valors amb VIM
train$Age <- kNN(train[,-3], k=3)$Age
test$Age <- kNN(test[,-2], k=3)$Age
```

Per acabar, comprovem que ja no hi ha cap valor NA o buit.

```
# Nombre de valors NA per atribut
sapply(train, function(x) sum(is.na(x)))
```

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

```
sapply(test, function(x) sum(is.na(x)))
```

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

```
# Nombre de valors buits per atribut
sapply(train, function(x) nrow(train[x=='',]))
```

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

```
sapply(test, function(x) nrow(test[x=='',]))
```

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

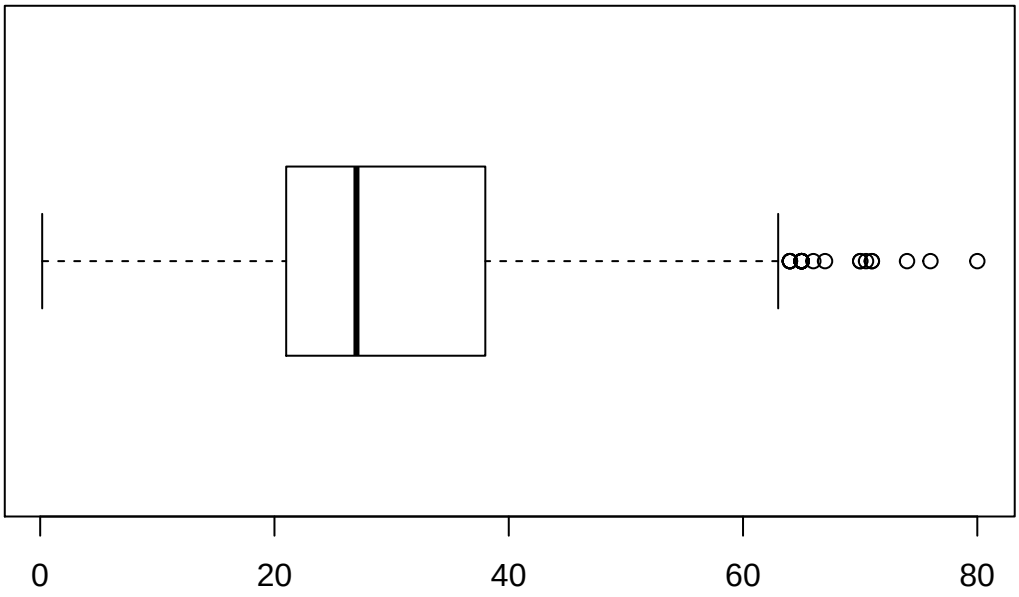
3.2. Valors extrems

Els valors extrems o *outliers* són aquells valors que es troben tan allunyats de la resta de valors que ens poden fer pensar que són erronis o tenen un origen diferent. Generalment es consideren *outliers* els valors que estan a més de 3 desviacions estàndards de la mitjana de la població. A continuació estudiarem els valors extrems per als atributs de tipus numèric, tant enters com reals.

Age

Si representem el diagrama de caixa de l'atribut *Age* veiem que la majoria de passatgers se situen entre els 20 i els 40 anys, però també apareixen alguns punts aïllats per sobre dels 60 anys.

```
# Diagrama de caixa de l'atribut Age
A.Age <- c(train$Age,test$Age)
boxplot(A.Age, horizontal = TRUE)
```



Si mostrem els valors extrems per a aquest atribut obtenim molts valors entre 60 i 80.

```
# Possibles outliers a l'atribut Age
boxplot.stats(A.Age)$out
```

```
##  [1] 66.0 65.0 71.0 70.5 65.0 65.0 65.0 65.0 65.0 65.0 64.0 65.0 65.0 71.0
## [15] 64.0 65.0 65.0 80.0 70.0 70.0 65.0 65.0 74.0 67.0 76.0 64.0 64.0 64.0
```

Si calculem els valors que es troben per sobre de 3 desviacions estàndard el nombre d'*outliers* disminueix a només 3.

```
# Edats superiors a 3 desviacions estàndard per sobre de la mitjana
A.Age[A.Age > (mean(A.Age) + 3 * sd(A.Age))]
```

```
## [1] 80 74 76
```

```
# Rang total de l'atribut Age
range(A.Age)
```

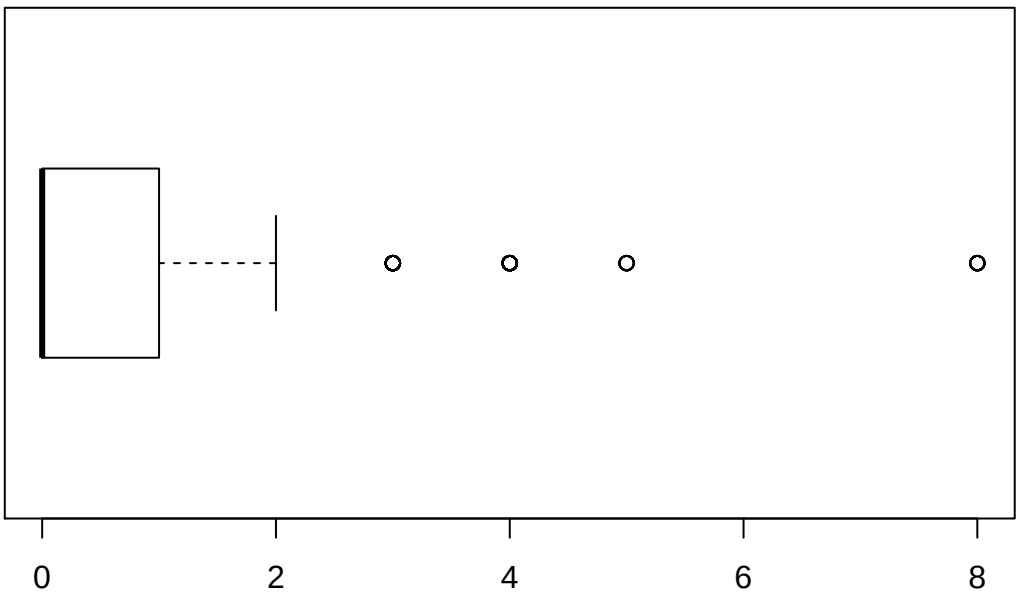
```
## [1]  0.17 80.00
```

Tot i que hem detectat 3 possibles outliers, tots els passatgers se situen entre els 0 i els 80 anys, que són edats perfectament possibles i per tant mantindrem tots els registres sense cap modificació.

SibSp

L'atribut *SibSp* ens indica el nombre de germans o marit/muller de cada passatger. Si mostrem la distribució de valors en una diagrama de caixa veiem que la majoria de valors se situen entre 0 i 1.

```
# Diagrama de caixa de l'atribut SibSp
A.SibSp <- c(train$SibSp, test$SibSp)
boxplot(A.SibSp, horizontal = TRUE)
```



Repetint les anàlisis anteriors veiem que la funció *boxplot* detecta com a extrems els valors superiors a 2, si calculem els valors superiors a 3 desviacions estàndard trobem els valors superiors a 3, mentre que el valor més alt de l'atribut és de 8. Tots aquests valors són perfectament possibles com a nombre de fills i per tant també els mantidrem.

```
# Possibles outliers a l'atribut SibSp
boxplot.stats(A.SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3
## [36] 5 4 3 4 8 4 3 4 8 4 8 3 4 5 3 4 8 4 8 4 3 3
```

```
# Valors per sobre de 3 SD
A.SibSp[A.SibSp > (mean(A.SibSp) + 3 * sd(A.SibSp))]
```

```
##  [1] 4 4 5 4 5 4 8 4 4 8 4 8 4 4 4 4 8 5 5 4 4 5 4 4 8 4 4 8 4 8 4 5 4 8 4
## [36] 8 4
```

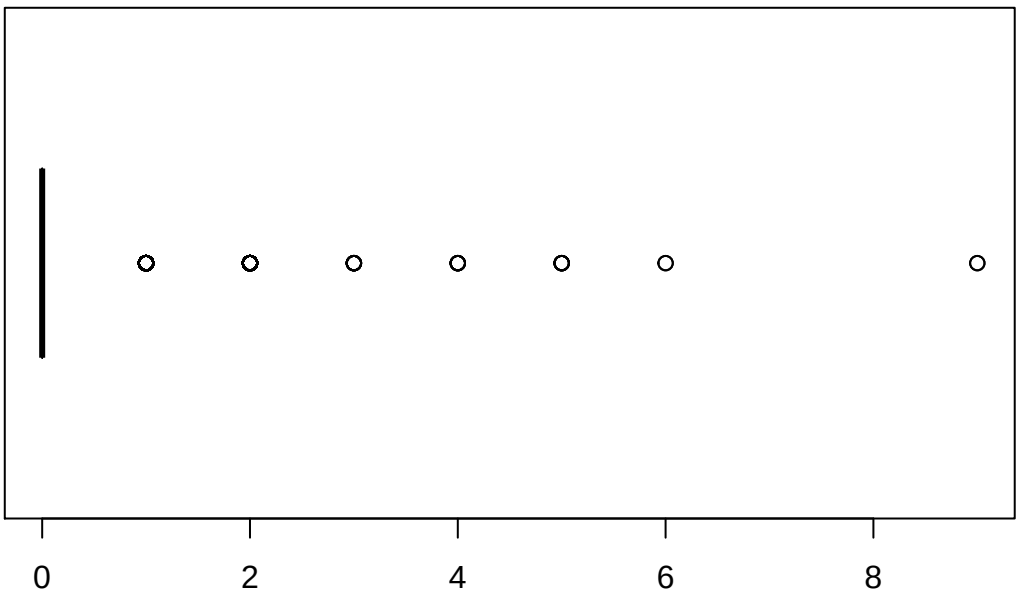
```
# Rang total de l'atribut
range(A.SibSp)
```

```
## [1] 0 8
```

Parch

L'atribut *Parch* indica el nombre de pares o fills al vaixell per a cada passatger. Si repetim els càlculs obtenim un resultat molt semblant a l'anterior. En aquest cas, la majoria de registres tenen un valor de 0, això fa que la mitjana sigui molt baixa i la resta de valors apareguin com a outliers. Malgrat tot, els rang de l'atribut se situa entre 0 i 9 fills, que són valors possibles i que encaixen amb els resultats de *SibSp*, per tant també els mantindrem.

```
# Diagrama de caixa de l'atribut Parch
A.Parch <- c(train$Parch, test$Parch)
boxplot(A.Parch, horizontal = TRUE)
```




```
# Possibles outliers a l'atribut Parch
boxplot.stats(A.Parch)$out
```

```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1
##  [36] 2 1 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1
##  [71] 1 2 1 2 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2
## [106] 2 3 4 1 2 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1
## [141] 2 1 1 2 5 2 1 1 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1
## [176] 1 2 1 2 3 1 2 1 2 2 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1
## [211] 1 5 2 1 1 1 1 3 1 2 2 1 2 1 2 1 2 4 1 1 2 1 1 1 4 6 2 3 1 1 2 2 2 1 1
## [246] 2 5 2 3 2 1 1 1 2 1 2 2 2 1 2 1 1 2 1 2 1 2 1 2 2 1 1 1 1 1 2 1 1 2 1
## [281] 1 1 2 1 2 9 1 1 1 2 2 2 1 9 1 1 2 2 1 1 2 1 1 1 1 1 1
```

```
# Valors per sobre de 3 SD
A.Parch[A.Parch > (mean(A.Parch) + 3 * sd(A.Parch))]
```

```
##  [1] 5 5 3 4 4 3 4 4 5 5 6 3 3 3 5 3 4 4 6 3 5 3 9 9
```

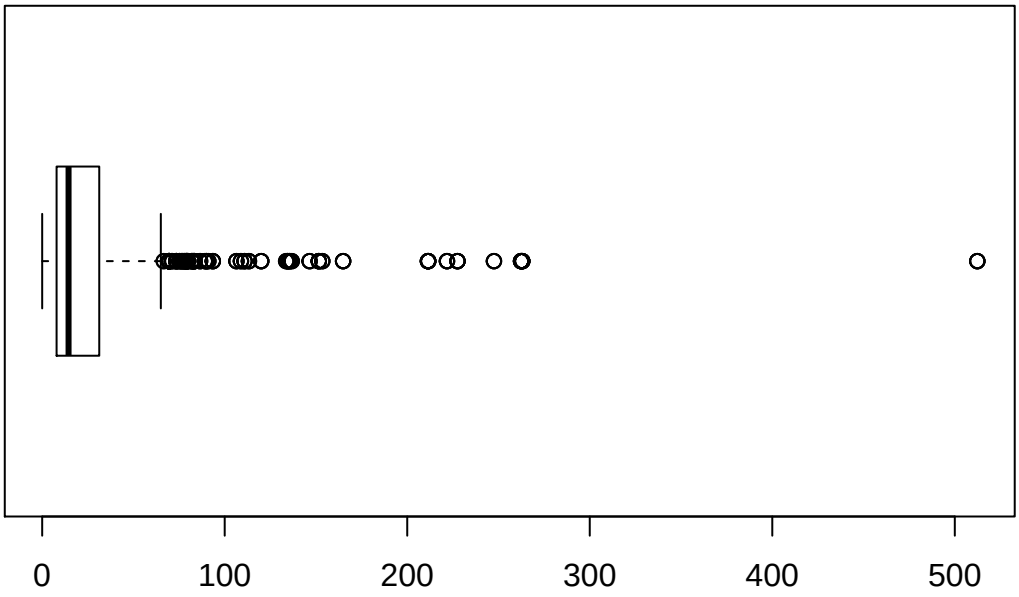
```
# Rang total de l'atribut
range(A.Parch)
```

```
## [1] 0 9
```

Fare

Per acabar estudiarem els valors extrem de l'atribut *Fare*, que indica el preu del bitllet de cada passatger. En aquest cas veiem que la majoria de bitllets se situen per sota de 50, tenim molts possibles *outliers* per sobre de 80 i un valor màxim superior als 500.

```
# Diagrama de caixa de l'atribut Fare
A.Fare <- c(train$Fare, test$Fare)
boxplot(A.Fare, horizontal = TRUE)
```




```
# Possibles outliers a l'atribut Fare
boxplot.stats(A.Fare)$out
```

```
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750
##   [8]  73.5000 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000
##  [15]  66.6000  69.5500  69.5500 146.5208  69.5500 113.2750  76.2917
##  [22]  90.0000  83.4750  90.0000  79.2000  86.5000 512.3292  79.6500
##  [29] 153.4625 135.6333  77.9583  78.8500  91.0792 151.5500 247.5208
##  [36] 151.5500 110.8833 108.9000  83.1583 262.3750 164.8667 134.5000
##  [43]  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000 263.0000
##  [50]  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
##  [57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042
##  [64]  91.0792  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000
##  [71] 221.7792 106.4250  71.0000 106.4250 110.8833 227.5250  79.6500
##  [78] 110.8833  79.6500  79.2000  78.2667 153.4625  77.9583  69.3000
##  [85]  76.7292  73.5000 113.2750 133.6500  73.5000 512.3292  76.7292
##  [92] 211.3375 110.8833 227.5250 151.5500 227.5250 211.3375 512.3292
##  [99]  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583 211.3375
## [106]  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
## [113]  89.1042 164.8667  69.5500  83.1583  82.2667 262.3750  76.2917
## [120] 263.0000 262.3750 262.3750 263.0000 211.5000 211.5000 221.7792
## [127]  78.8500 221.7792  75.2417 151.5500 262.3750  83.1583 221.7792
## [134]  83.1583  83.1583 247.5208  69.5500 134.5000 227.5250  73.5000
## [141] 164.8667 211.5000  71.2833  75.2500 106.4250 134.5000 136.7792
## [148]  75.2417 136.7792  82.2667  81.8583 151.5500  93.5000 135.6333
## [155] 146.5208 211.3375  79.2000  69.5500 512.3292  73.5000  69.5500
## [162]  69.5500 134.5000  81.8583 262.3750  93.5000  79.2000 164.8667
## [169] 211.5000  90.0000 108.9000
```

```
# Valors per sobre de 3 SD
A.Fare[A.Fare > (mean(A.Fare) + 3 * sd(A.Fare))]
```

```
##  [1] 263.0000 263.0000 247.5208 512.3292 247.5208 262.3750 263.0000
##  [8] 211.5000 227.5250 263.0000 221.7792 227.5250 512.3292 211.3375
## [15] 227.5250 227.5250 211.3375 512.3292 262.3750 211.3375 262.3750
## [22] 263.0000 262.3750 262.3750 263.0000 211.5000 211.5000 221.7792
```

```
## [29] 221.7792 262.3750 221.7792 247.5208 227.5250 211.5000 211.3375
## [36] 512.3292 262.3750 211.5000
```

```
# Rang total de l'atribut
range(A.Fare)
```

```
## [1]   0.0000 512.3292
```

Com en els casos anteriors, per a aquest atribut també hem decidit mantenir tots els valors.

3.3. Conversió de dades

A continuació transformarem alguns atributs per tal de reduir-ne el nombre de categories o obtenir-ne informació que pugui resultat interessant per a la fase d'anàlisi.

Name

A l'atribut *Name* n'extraurem el títol de tractament, ja que indica l'estatus del viatger i pot influir en la seva supervivencia.

```
# Extraiem el títol de l'atribut Name
train$Title <- sub(".*\\s([A-Za-z]+)\\..*", "\\1", train$Name)
test$Title <- sub(".*\\s([A-Za-z]+)\\..*", "\\1", test$Name)

# Nombre de registres per categoria de Title
table(c(train$Title, test$Title))
```

```
##
##     Capt      Col Countess      Don     Dona       Dr Jonkheer        L
##        1        4        1        1        1        8        1        1
##     Lady    Major   Master     Miss     Mlle      Mme       Mr      Mrs
##        1        2       61      260        2        1      757      196
##       Ms      Rev      Sir
##        2        8        1
```

A continuació, com que el nombre de categories és força gran i algunes tenen molt pocs registres, unirem els títols poc freqüents ens una mateixa categoria 'other'. Per acabar, convertim l'atribut a tipus factor.

```
# Marquem com a 'other' els títols menys freqüents
others <- c("Capt","Countess","Don","Dona","Jonkheer","L","Lady","Mme","Sir",
            "Major","Mlle","Ms","Col","Dr","Rev")
train[train$Title %in% others, 'Title'] = 'other'
test[test$Title %in% others, 'Title'] = 'other'

# Convertim Title a factor
train$Title <- as.factor(train$Title)
test$Title <- as.factor(test$Title)
```

Cabin

L'atribut cabina conté més de 100 nivells. Una manera de reduir-los pot ser eliminar els números i conservar únicament la lletra, ja que segurament indica el tipus de cabina o la zona on s'allotjaven els passatgers i pot resultar significativa.

```
# Extraiem la lletra de l'atribut Cabin
levels(train$Cabin)[-1] <- sub("([A-T]).*","\\1", levels(train$Cabin)[-1])
levels(test$Cabin)[-1] <- sub("([A-T]).*","\\1", levels(test$Cabin)[-1])
```

```r
# Freqüència de supervivents per cabina
table(train[,c("Survived","Cabin")])
```

```
##          Cabin
## Survived  NO   A   B   C   D   E   F   G   T
##        0 481   8  12  24   8   8   5   2   1
##        1 206   7  35  35  25  24   8   2   0
```

Com veiem, les cabines A, F, G i T tenen molt pocs passatgers i un percentatge de supervivents força semblant. Per evitar tenir cabines amb pocs registres i millorar l'eficàcia de les anàlisis unirem aquestes cabines en una mateixa categoria que anomenarem X.

```r
# Reanomenem a X les cabines A, F, G i T
levels(train$Cabin) <- c("NO", "X", "B", "C", "D", "E", "X", "X", "X")
levels(test$Cabin) <- c("NO", "X", "B", "C", "D", "E", "X", "X", "X")
```

Ticket

Finalment, tot i que l'atribut *Ticket* també conté molts nivells i sembla poc útil, podem provar d'extreure'n alguna informació interessant. A continuació mostrem les dades d'alguns passatgers que comparteixen el valor de *Ticket*.

```r
# Valors més freqüents de Ticket
sort(table(train$Ticket), decreasing = TRUE)[c(1:5)]
```

```
##
##     1601   347082 CA. 2343  3101295   347088
##        7        7        7        6        6
```

```r
# Passatgers amb el Ticket més freqüent
train[train$Ticket=='1601',]
```

```
##     Survived Pclass            Name  Sex Age SibSp Parch Ticket    Fare
## 75         1      3   Bing, Mr. Lee male  32     0     0   1601 56.4958
## 170        0      3   Ling, Mr. Lee male  28     0     0   1601 56.4958
## 510        1      3  Lang, Mr. Fang male  26     0     0   1601 56.4958
## 644        1      3 Foo, Mr. Choong male  32     0     0   1601 56.4958
## 693        1      3    Lam, Mr. Ali male  32     0     0   1601 56.4958
## 827        0      3    Lam, Mr. Len male  28     0     0   1601 56.4958
## 839        1      3 Chip, Mr. Chang male  32     0     0   1601 56.4958
##     Cabin Embarked Title
## 75     NO        S    Mr
## 170    NO        S    Mr
## 510    NO        S    Mr
## 644    NO        S    Mr
## 693    NO        S    Mr
## 827    NO        S    Mr
## 839    NO        S    Mr
```

Com veiem, els passatgers amb un mateix valor per a *Ticket* també comparteixen moltes característiques, perquè segurament formaven una família o viatjaven junts. Això ho aprofitarem per calcular la mida del grup de persones que viatjaven conjuntament i crear l'atribut *Group*.

```r
# Creem una taula de freqüències per als tickets
tickets <- as.data.frame(table(rbind(train["Ticket"],test["Ticket"])))
colnames(tickets) <- c("Ticket","Group")

# Assignem el valor de la freqüència a un nou atribut "Group"
```

```
train <- join(train, tickets, by="Ticket")
test <- join(test, tickets, by="Ticket")
```

Fare

Finalment, com que el preu del bitllet sembla ser el mateix per a cada ticket, és a dir, sembla tractar-se de l'import per a tot el grup, dividirem l'import del passatge (*Fare*) per les dimensions del grup (*Group*) per tal d'obtenir el preu individual del bitllet.

```
# Càlcul del preu individual del bitllet
train$Fare <- train$Fare / train$Group
test$Fare <- test$Fare / test$Group
```

Per acabar, eliminem els atributs *Name* i *Ticket* per evitar la redundància de dades, ja que els hem utilitzat per generar els atributs derivats *Title* i *Group*.

```
# Eliminem l'atribut Name i Ticket
train <- train[,-c(3,8)]
test <- test[,-c(2,7)]
```

3.4. Exportació de dades

Una vegada hem netejat les dades les exportarem en un nou fitxer.

```
# Exportació dels conjunt train i test nets
write.csv(train, '../data/Titanic_train_clean.csv', row.names = FALSE)
write.csv(test, '../data/Titanic_test_clean.csv', row.names = FALSE)
```

4. Anàlisi de dades

4.1. Selecció de dades

La fase d'anàlisi s'enfocarà en estudiar les dades des de diferents punts de vista i intentant respondre diferents preguntes:

1. Com és la distribució de valors per a les variables numèriques contínues? La mitjana de valors és significativament diferent per als supervivents i no supervivents?
2. Hi ha correlacions entre variables numèriques que permetin prescindir d'algun atribut?
3. La probabilitat de sobreviure és significativament diferent entre les diferents categories?
4. Quina és la probabilitat de sobreviure per als passatgers del conjunt de test?

La primera pregunta estudia la distribució de les variables numèriques contínues (*Age* i *Fare*). Començarem estudiant si els seus valors segueixen una distribució normal, a continuació compararem la variància en funció del valors de la classe (*Survived*) i, finalment, aprofitarem aquest resultat per comparar la mitjana de valors en funció del valor de *Survived*. Això ens permetrà saber si la mitjana d'edat i del preu del bitllet són significativament diferents entre els supervivents i els no supervivents.

Per a respondre la segona pregunta estudiarem la correlació entre les variables quantitatives (*Pclass*, *Age*, *SibSp*, *Parch*, *Fare* i *Group*). En principi es tracta de les variables independents del conjunt de dades, per aquest motiu, si detectem alguna dependència important significarà que hi ha redundància en les dades i pot ser necessari eliminar alguna columna. Per fer aquesta anàlisi treballarem amb els dos conjunts de dades: *train* i *test*.

La tercera pregunta se centra en les variables qualitatives (*Sex*, *Cabin*, *Embarked* i *Title*) on avaluarem si hi ha diferències significatives en el percentatge de supervivents entre els diferents grups definits per aquestes variables. Per fer aquesta anàlisi utilitzarem el test de χ^2 amb la funció *chisq.test* sobre les dades de *train*.

Finalment, la darrera pregunta és la més important perquè recull els resultats de les tres anteriors i planteja el problema principal de l'estudi: la creació d'un model predictiu per a la classe *Survived*. Treballarem amb el conjunt *train* per crear un model de regressió logística a partir de totes les variables, tant les numèriques com les categòriques, i l'utilitzarem per efectuar prediccions sobre el conjunt de *test*.

```
# Variables numèriques
var_num <- c("Pclass","Age","SibSp","Parch","Fare","Group")

# Variables categòriques
var_cat <- c("Sex","Cabin","Embarked","Title")
```

4.2. Anàlisi de la distribució de variables contínues

Per fer les anàlisis de normalitat i homogeneïtat de la variància ens centrarem en els atributs de tipus *numeric*, que són *Age* i *Fare*. En primer lloc unirem els dos conjunts de dades en un mateix dataframe per facilitar-ne l'estudi.

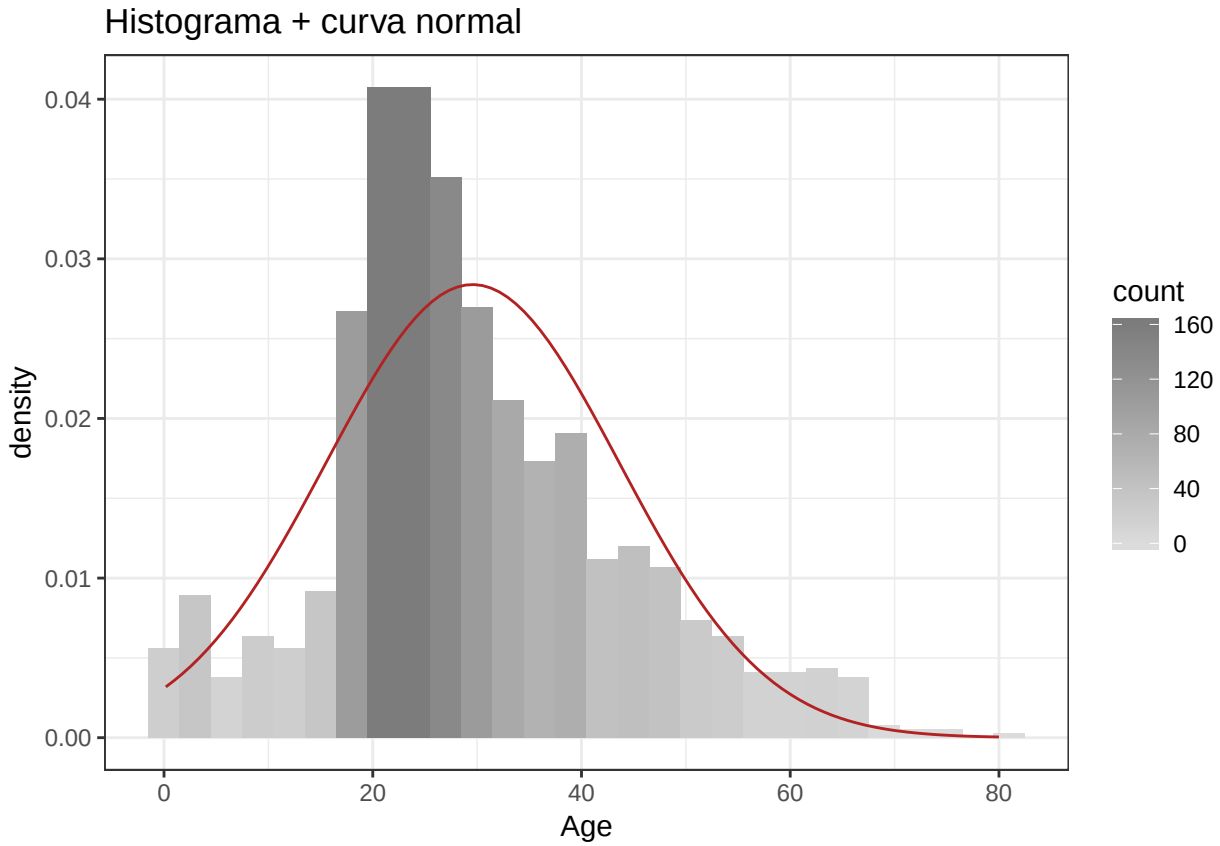
```
# Unim les dades en un dataframe
test_and_train <- rbind(train[c("Age","Fare")],test[c("Age","Fare")])
```

4.2.1 Normalitat

Age

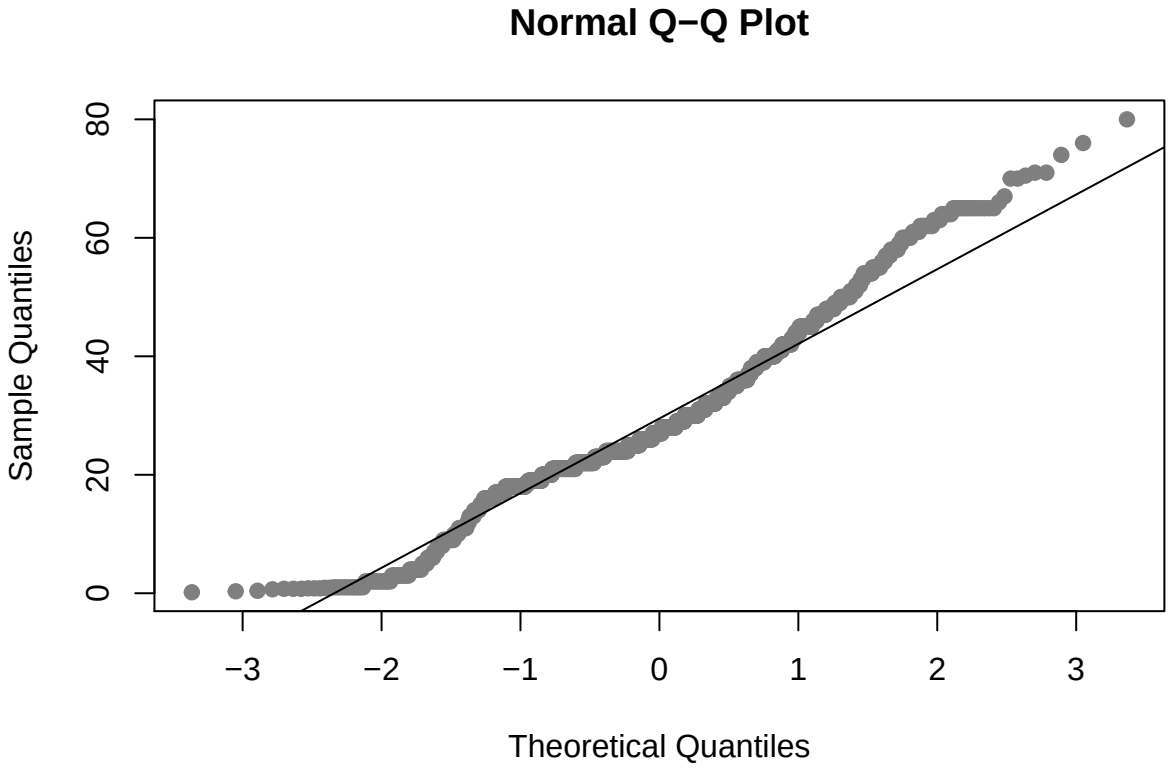
Començarem analitzant la normalitat de la distribució de l'atribut *Age* de manera gràfica, representant les dades en un histograma i utilitzant la gràfica Q-Q.

```
# Representem les dades en un histograma
ggplot(data = test_and_train, aes(x = Age)) +
  geom_histogram(aes(y = ..density.., fill = ..count..), binwidth = 3) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(test_and_train$Age),
                            sd = sd(test_and_train$Age))) +
  ggtitle("Histograma + curva normal") +
  theme_bw()
```



Tot i que la majoria de franges d'edat s'acosten a una distribució normal, l'interval aproximat entre 18 i 25 anys és molt més freqüent que la resta i s'allunya molt d'aquesta distribució. Com veiem a continuació, la gràfica Q-Q segueix la línia de normalitat a la regió central però s'allunya als extrems.

```
# Representació de la gràfica Q-Q d'Age
qqnorm(test_and_train$Age, pch = 19, col = "gray50")
qqline(test_and_train$Age)
```



Per tal d'avaluar la normalitat de la distribució de manera numèrica utilitzarem el contrast d'hipòtesis. En primer lloc farem el test de Shapiro-Wilk.

```
# Test de normalitat de Shapiro-Wilk
shapiro.test(test_and_train$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  test_and_train$Age
## W = 0.9699, p-value = 7.154e-16
```

Prenent com a hipòtesi nul · la que la distribució de valors d'*Age* és normal i fixant un nivell de significació del 5%, veiem com el càlcul del p-valor és molt inferior al nivell de significació i, per tant, ens veiem obligats a rebutjar la hipòtesi nul · la; és a dir, no podem afirmar que la distribució d'*Age* sigui normal.

Com que es tracta d'una mostra relativament gran ($|T| > 50$), verificarem els resultats amb el test d'Anderson-Darling.

```
# Test de normalitat d'Anderson-Darling
ad.test(test_and_train$Age)
```

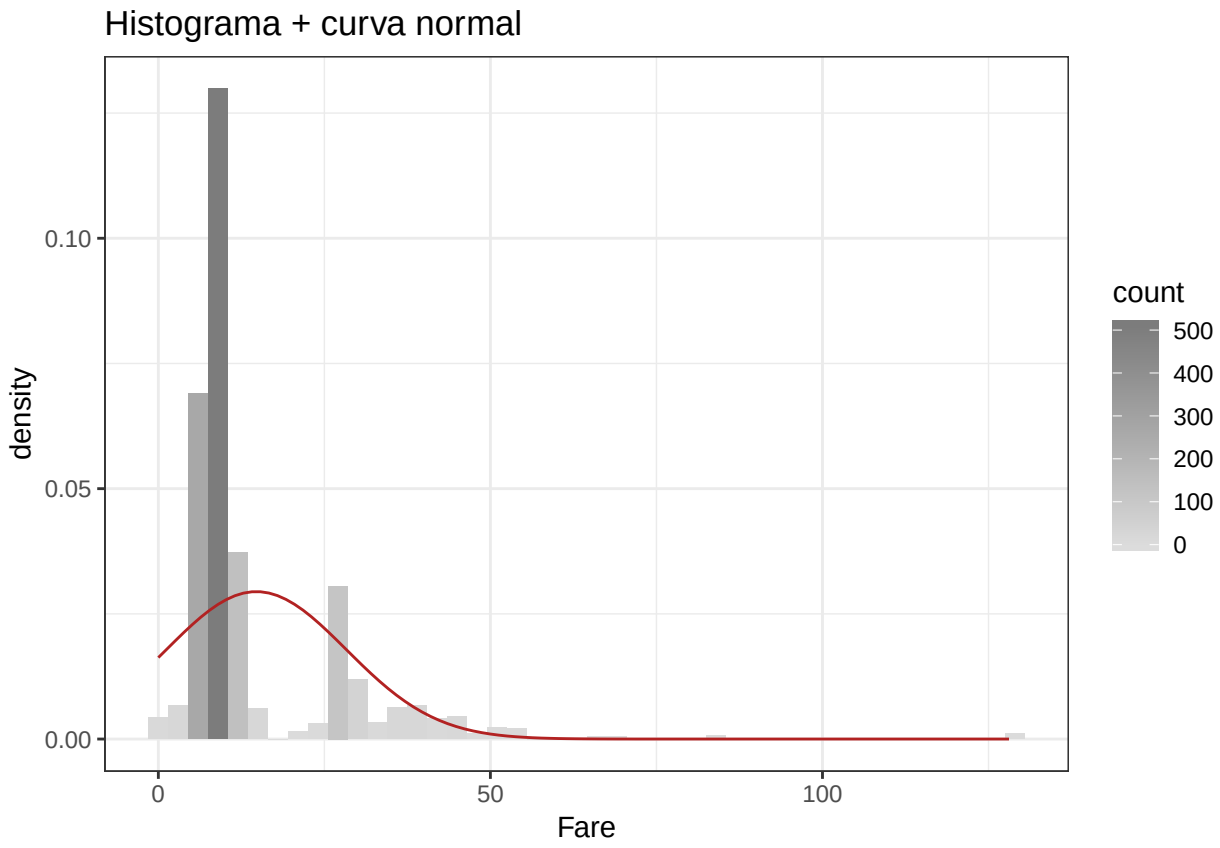
```
##
##  Anderson-Darling normality test
##
## data:  test_and_train$Age
## A = 14.398, p-value < 2.2e-16
```

Els resultats també ens diuen que no podem considerar que la distribució d'edats segueixi una corba normal.

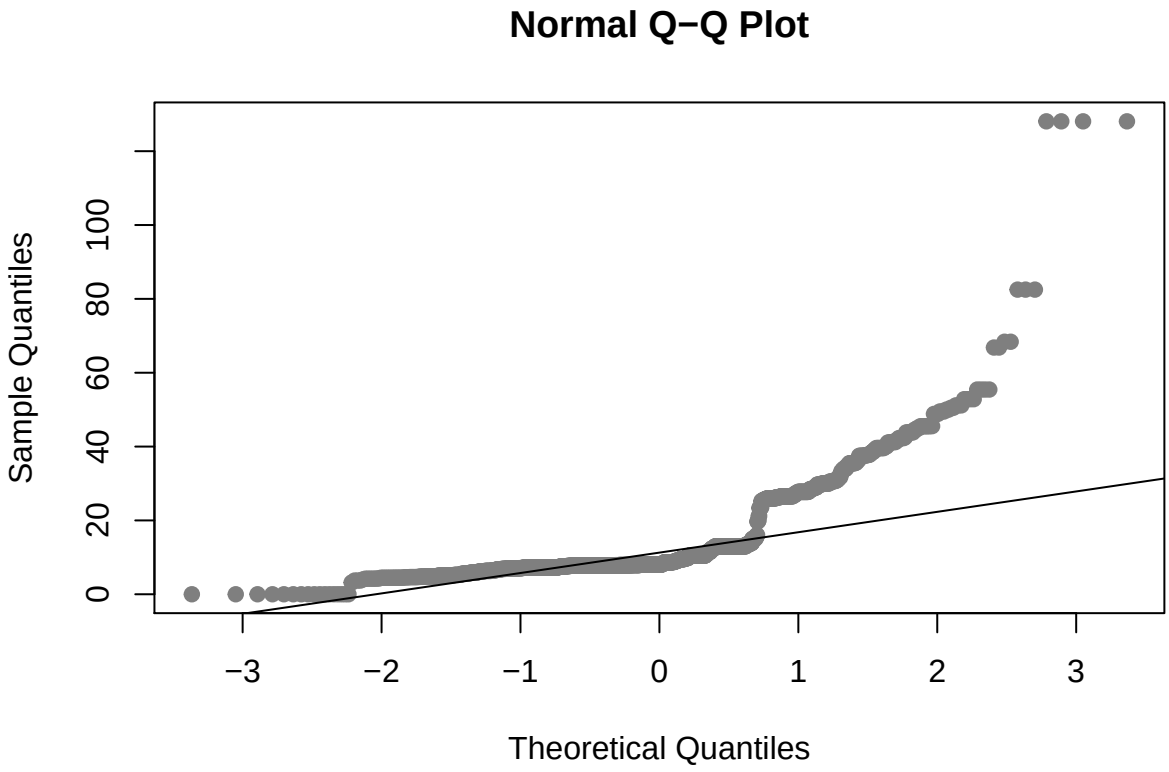
Fare

A continuació repetim les gràfiques i els càlculs per avaluar la normalitat de l'atribut *Fare*.

```r
# Representem les dades en un histograma
ggplot(data = test_and_train, aes(x = Fare)) +
  geom_histogram(aes(y = ..density.., fill = ..count..), binwidth = 3) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(test_and_train$Fare),
                            sd = sd(test_and_train$Fare))) +
  ggtitle("Histograma + curva normal") +
  theme_bw()
```




```r
# Representació de la gràfica Q-Q per a Fare
qqnorm(test_and_train$Fare, pch = 19, col = "gray50")
qqline(test_and_train$Fare)
```



A nivell gràfic, els valors de *Fare* tampoc semblen seguir una distribució normal. Això ho confirmem amb els tests de Shapiro-Wilk i d'Anderson-Darling.

```
# Test de normalitat de Shapiro-Wilk
shapiro.test(test_and_train$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  test_and_train$Fare
## W = 0.67421, p-value < 2.2e-16
```

```
# Test de normalitat d'Anderson-Darling
ad.test(test_and_train$Fare)
```

```
##
##  Anderson-Darling normality test
##
## data:  test_and_train$Fare
## A = 139.02, p-value < 2.2e-16
```

En tots dos casos el càlcul del p-valor resulta molt inferior al nivell de significació i, per tant, no podem afirmar que els valors de l'atribut *Fare* segueixin una distribució normal.

4.2.2. Homoscedasticitat

A continuació estudiarem si la variància de les variables numèriques *Age* i *Fare* és constant als grups de les persones que sobreviuen respecte a les que no ho fan. Com que cap de les dues de les variables té una distribució normal utiltizarem el test Fligner-Killeen. En aquest test la hipòtesi nul · la contempla que les variàncies entre els grups són iguals.

Age

```
# Apliquem el test Fligner-Killeen a Age respecte als valors de Survived
fligner.test(Age ~ Survived, data=train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 0.0020217, df = 1, p-value =
## 0.9641
```

Com el p-valor resultat és més gran que 0.05 no hi ha una justificació per rebutjar la hipòtesi nul · la. Això ens indica que la variància de l'edat és similar en ambdós grups.

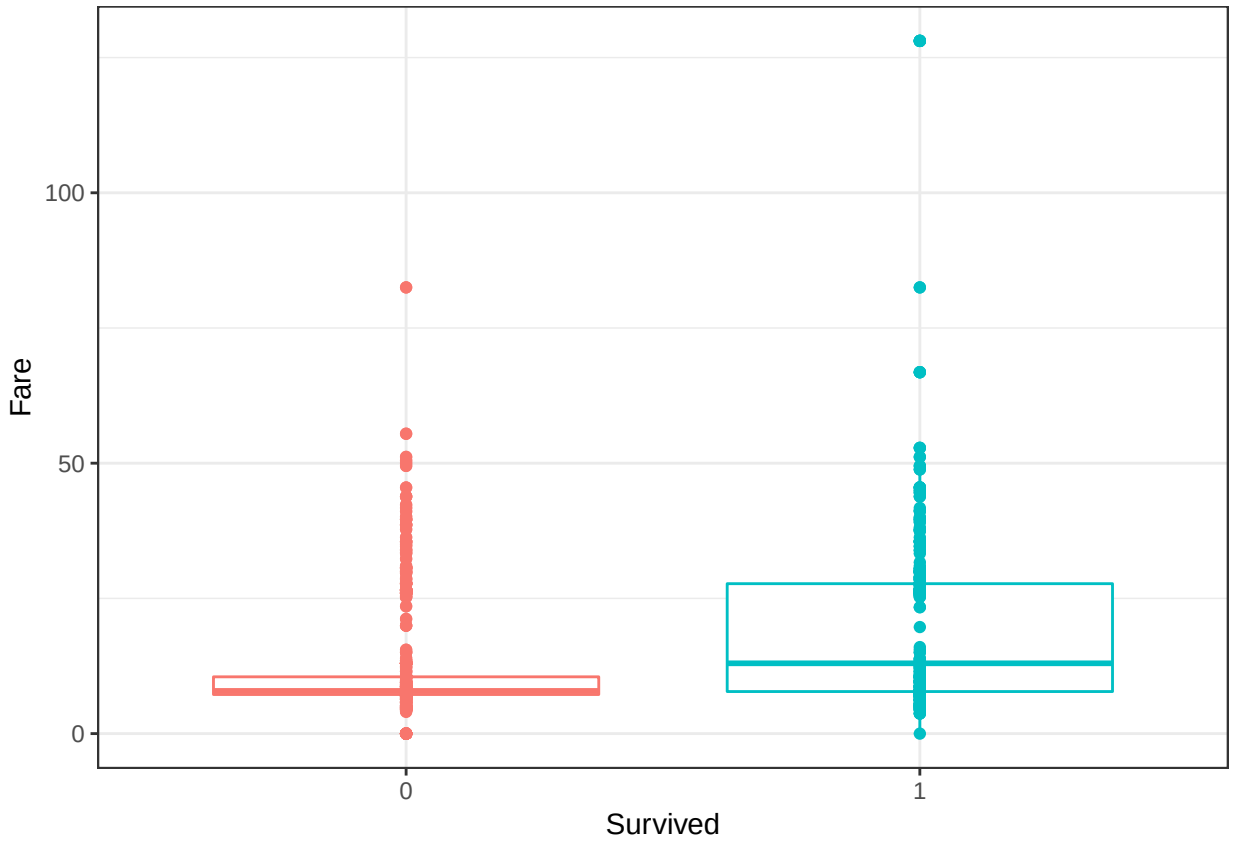
Fare

```
# Apliquem el test Fligner-Killeen a Fare respecte als valors de Survived
fligner.test(Fare ~ Survived, data=train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 126.45, df = 1, p-value <
## 2.2e-16
```

Com el p-valor resultant és pràcticament 0, podem rebutjar la hipòtesi nul · la. Això ens indica que la variància de la variable *Fare* és significativament diferent en ambdós grups. Aquest resultat el podem comprovar gràficament mostrant la distribució de valors per a les dues classes en un diagrama de caixa.

```
# Diagrama de caixa de Fare per a les classes de Survived
ggplot(data = train, aes(x = Survived, y = Fare, colour = train$Survived)) +
  geom_boxplot() +
  geom_point() +
  theme_bw() +
  theme(legend.position = "none")
```



4.2.3. Comparació de les mitjanes d'*Age* i *Fare*

Les variables *Age* i *Fare* presenten una important desviació respecte de la normal, per tant, emprarem el test no paramètric de Kruskal-Wallis per analitzar la diferència de les seves mitjanes. En aquest test la hipòtesi nul · la contempla que les distribucions entre els grups són iguals.

```
# Test de Kruskal-Wallis aplicat a Age respecte Survived
kruskal.test(Age  ~ Survived, data = train )
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Age by Survived
## Kruskal-Wallis chi-squared = 2.6638, df = 1, p-value = 0.1027
```

```
# Test de Kruskal-Wallis aplicat a Fare respecte Survived
kruskal.test(Fare  ~ Survived, data = train )
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Fare by Survived
## Kruskal-Wallis chi-squared = 67.238, df = 1, p-value = 2.407e-16
```

Com podem veure, el test de Kruskal-Wallis per a l'atribut *Age* ens dóna un p-valor superior a 0.05, per tant, no podem rebutjar la hipòtesi nul · la i hem de considerar que les dues mostres provenen de la mateixa distribució, és a dir, que la mitjana d'edat per a supervivents i no supervivents és la mateixa. A partir d'aquests resultats, ens plantejarem la possibilitat d'excloure aquest atribut del model.

D'altra banda, l'atribut *Fare* ens mostra un p-valor molt petit, això vol dir que podem afirmar que existeix una diferència significativa entre la mitjana del preu del bitllet per a les dues mostres.

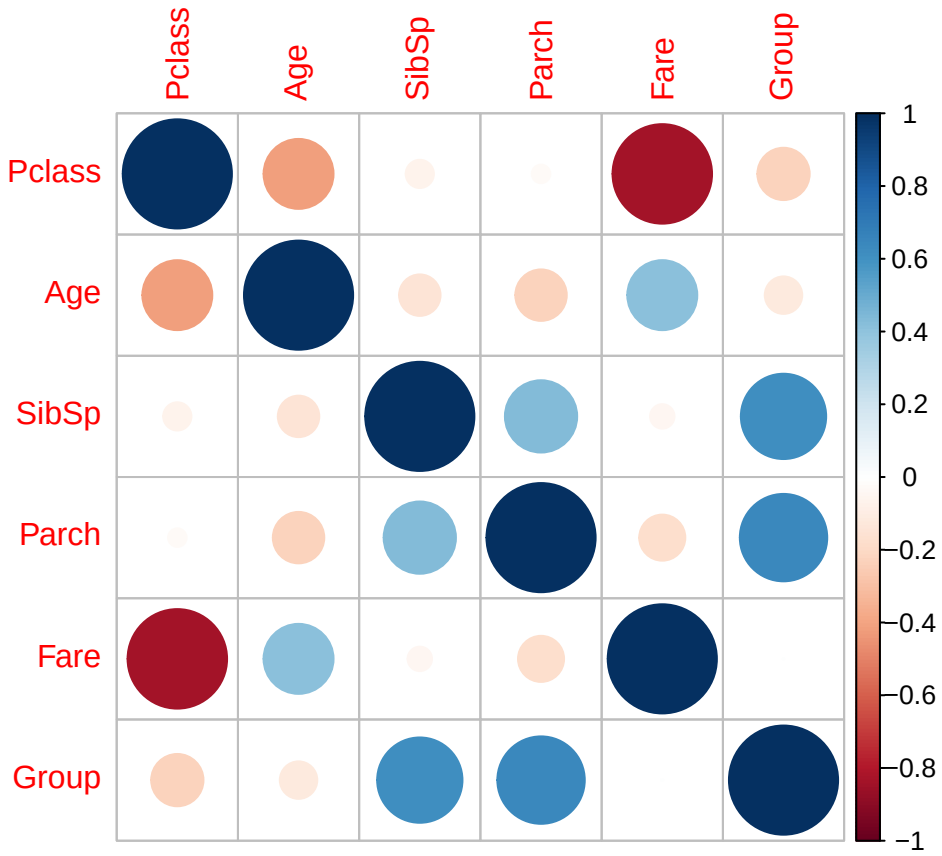
4.3. Anàlisi de correlació

En aquest apartat ens preguntem si hi ha correlacions entre variables numèriques que permetin prescindir d'algun atribut. Per analitzar les possibles dependències generarem una matriu de correlacions entre totes les variables quantitatives dels conjunts de dades *train* i *test* conjuntament; a més, utilitzarem la correlació de *Spearman*, que no suposa cap tipus de distribució entre les dades, perquè com hem vist hi ha variables numèriques que no segueixen una distribució normal.

```
# Correlació entre variables numèriques
corr <- cor(rbind(train[var_num], test[var_num]), method="spearman")
corr
```

```
##             Pclass        Age       SibSp       Parch         Fare
## Pclass  1.00000000 -0.4105725 -0.06667944 -0.02875173 -0.831008732
## Age    -0.41057250  1.0000000 -0.14426340 -0.22199667  0.411368607
## SibSp  -0.06667944 -0.1442634  1.00000000  0.43837300 -0.049973736
## Parch  -0.02875173 -0.2219967  0.43837300  1.00000000 -0.178033866
## Fare   -0.83100873  0.4113686 -0.04997374 -0.17803387  1.000000000
## Group  -0.22951316 -0.1177524  0.61096280  0.64137006  0.000211549
##               Group
## Pclass -0.229513158
## Age    -0.117752436
## SibSp   0.610962797
## Parch   0.641370061
## Fare    0.000211549
## Group   1.000000000
```

```
# Representació gràfica de la matriu de correlacions
corrplot(corr, method="circle")
```



Com era d'esperar, hi ha una forta correlació entre els atributs *Fare* i *Pclass*; el coeficient de *Spearman* és de -0.83, força proper a -1, que vol dir que per a classes més altes (valors més petits) augmenta el preu del bitllet i a l'inversa. A més, també s'aprecia una correlació positiva entre *Group-Parch* (0.64) i *Group-SibSp* (0.61) perquè tots aquests atributs informen de la mida de la família o grup de gent que viatjava conjuntament. Altres correlacions menys importants són *Age-Pclass* (-0,41), classes més altes (valors petits de Pclass) tenen edats més grans, i *SibSp-Parch* (0.43).

A partir d'aquests resultats ens plantejarem la possibilitat d'excloure atributs redundants en el nostre model, com *Pclass* o *Fare*, en funció del nivell de significació que tinguin.

4.4. Contrast d'hipòtesis

A la tercera pregunta ens qüestionem si les diferents categories de les variables qualitatives (*Sex*, *Cabin*, *Embarked* i *Title*) influeixen en la probabilitat de sobreviure. Atès que la variable *Survived* també és qualitativa, emprarem el test χ^2 amb la hipotesi nul · la que diu que les variables examinades no són independents. Primer calcularem les freqüències de *Survived* per a cada categoria i després aplicarem la funció *chisq.test*.

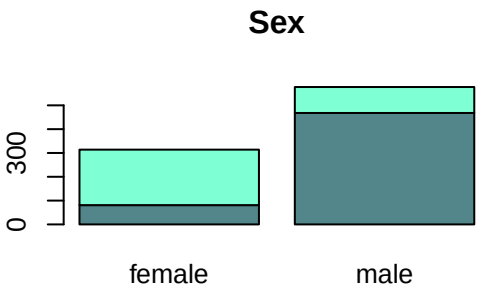
```
# Test chi quadrat per a totes les variables qualitatives
for (var in var_cat){
  # Creem una taula de freqüències
  taula = table(train[, c("Survived", var)])
  print(taula)
  # Apliquem el test
  print(chisq.test(taula))
}
```

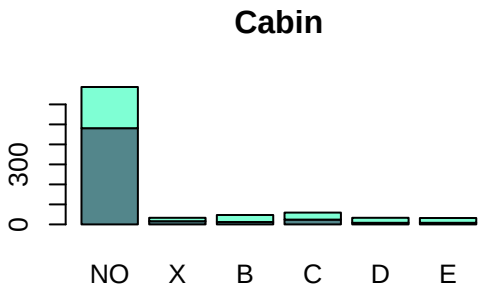
```
##          Sex
```

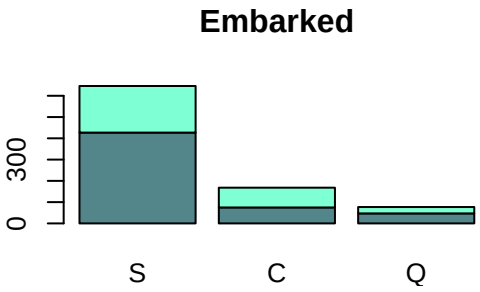
```
## Survived female male
##        0     81  468
##        1    233  109
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  taula
## X-squared = 260.72, df = 1, p-value < 2.2e-16
##
##         Cabin
## Survived  NO   X   B   C   D   E
##        0 481  16  12  24   8   8
##        1 206  17  35  35  25  24
##
##  Pearson's Chi-squared test
##
## data:  taula
## X-squared = 97.337, df = 5, p-value < 2.2e-16
##
##         Embarked
## Survived   S   C   Q
##        0 427  75  47
##        1 219  93  30
##
##  Pearson's Chi-squared test
##
## data:  taula
## X-squared = 25.964, df = 2, p-value = 2.301e-06
##
##         Title
## Survived Master Miss  Mr Mrs other
##        0     17   55 436  26    15
##        1     23  127  81  98    13
##
##  Pearson's Chi-squared test
##
## data:  taula
## X-squared = 282.24, df = 4, p-value < 2.2e-16
```

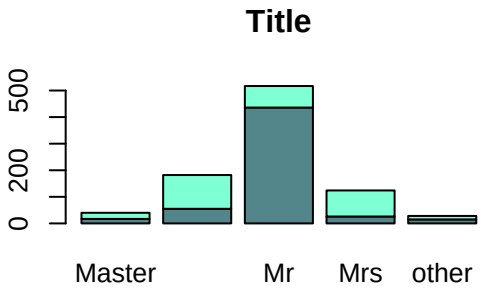
A L'execució podem veure com el p-valor és inferior al 0.05 en tots els casos, per tant, rebutgem la hipòtesi nul · la. Això vol dir que podem afirmar que hi ha diferències significatives entre el nombre de supervivents per a les diferents categories de *Sex*, *Cabin*, *Embarked* i *Title*. Aquestes diferències es poden apreciar gràficament si mostrem un diagrama de barres per a cada categoria.

```
# Representació dels supervivents i no supervivents per categories
layout(matrix(c(1,2,3,4,5,5)))
par(mfrow=c(2,2))
for (var in var_cat){
  barplot(table(train[,c("Survived", var)]), main = var,
          col = c("cadetblue4","aquamarine"))
}
```

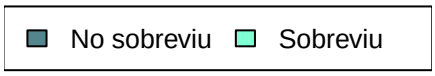









```
plot.new()
legend("center", ncol=2, legend=c("No sobreviu","Sobreviu"),
       fill = c("cadetblue4","aquamarine"))
```



4.5. Regressió logística

Finalment, calcularem la probabilitat de supervivència per als passatgers del conjunt *test*. En el nostre cas, com que la variable a predir (dependent) és dicotòmica, utilitzarem una regressió logística.

4.5.1. Creació del model

Per generar el model de regressió logística utilitzarem la funció *regLog* amb totes les variables disponibles, tant les categòriques com les numèriques. Les variables numèriques no requereixen cap tipus de transformació, però les categòriques s'hauran de binaritzar, és a dir, haurem de crear un atribut de tipus binari per a cada categoria; malgrat tot, la funció *regLog* ja realitza aquesta transformació i per tant no serà necessari modificar cap atribut.

Per tal d'avaluar el model generat serà necessari dividir les dades de *train* en un conjunt d'entrenament i un altre de test. Malgrat tot, com que el nombre de dades no és gaire gran utilitzarem la validació creuada o *cross validaton* per poder generar un model a partir de totes les dades. En el nostre cas utilitzarem 4 particions (4-fold).

```
# Creem les particions de la cross validation
folds <- trainControl(method = "cv", number=4)
# Entrenem el model
regLog <- train(Survived ~ Pclass + Age + SibSp + Parch + Fare + Group + Sex +
                  Cabin + Embarked + Title,
```

```
              data = train, trControl = folds, method = "glm",
              family = binomial(link="logit"))
# Mostrem els coeficients
summary(regLog)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3301  -0.5404  -0.3573   0.5580   2.5177
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  20.055331 480.076204   0.042 0.966678
## Pclass       -0.869974   0.213987  -4.066 4.79e-05 ***
## Age          -0.037018   0.009232  -4.010 6.08e-05 ***
## SibSp        -0.646294   0.151590  -4.263 2.01e-05 ***
## Parch        -0.409425   0.163344  -2.507 0.012193 *
## Fare          0.014962   0.012562   1.191 0.233643
## Group         0.093674   0.100388   0.933 0.350755
## Sexmale     -15.845286 480.075440  -0.033 0.973670
## CabinX        0.382084   0.478797   0.798 0.424865
## CabinB        0.486244   0.556229   0.874 0.382020
## CabinC        0.136883   0.474432   0.289 0.772949
## CabinD        1.105499   0.573031   1.929 0.053704 .
## CabinE        1.467089   0.558422   2.627 0.008609 **
## EmbarkedC     0.410740   0.257764   1.593 0.111054
## EmbarkedQ     0.414360   0.348238   1.190 0.234095
## TitleMiss   -16.247682 480.075715  -0.034 0.973002
## TitleMr      -3.274970   0.541643  -6.046 1.48e-09 ***
## TitleMrs    -15.292186 480.075787  -0.032 0.974589
## Titleother   -3.037188   0.792428  -3.833 0.000127 ***
## ---
## 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:  705.96  on 872  degrees of freedom
## AIC: 743.96
##
## Number of Fisher Scoring iterations: 14
```

El model generat ens dóna una precisió de 0.828, que vol dir que més d'un 80% dels registres s'han classificat correctament.

Si analitzem els coeficients de la regressió i el p-valor calculat per a cada un, veiem que els atributs *Fare*, *Group*, *Sex* i *Embarked* resulten poc rellevants per a la regressió. Això s'explica perquè, sota la hipòtesi nul · la que el coeficient és 0, el seu p-valor és força gran i ens obliga a acceptar-la, que vol dir que podem prescindir d'aquests atributs. En canvi *Age*, a diferència del que ens mostrava el test de Kruskal-Wallis, sí que resulta important per al model. A partir d'aquests resultats tindrem en compte:

- **Fare**. A l'anàlisi de correlació hem vist que hi ha un lligam fort entre *Pclass* i *Fare*, per tant, resulta

lògic mantenir *Pclass* a la regressió (molt més rellevant) i excloure *Fare*.

- **Group**. L'anàlisi de correlació també ens ha mostrat un lligam important de *Group* amb *SibSp* i *Parch*. Com que aquests dos també resulten rellevants per a la regressió optarem per excloure *Group*.
- **Sex**. Tot i que resulta sorprenent que *Sex* tingui tan poca importància, cal tenir en compte que segurament hi ha redundància de dades amb *TitleMr* i aquest atribut sí que resulta molt significatiu. Per tant, també exclourem *Sex* de la regressió.
- **Embarked**. La informació d'aquest atribut no la trobem en cap altre variable i el p-valor obtingut no és tant alt com a la resta d'atributs exclosos. En aquest cas optarem per mantenir aquesta variable perquè en el nou model pot ser que prengui més importància.

A continuació crearem un nou model excloent les variables comentades.

```
# Entrenem el model
regLog <- train(Survived ~ Pclass + Age + SibSp + Parch + Cabin + Embarked + Title,
                data = train, trControl = folds, method = "glm",
                family = binomial(link="logit"))
# Mostrem els coeficients
summary(regLog)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3517  -0.5535  -0.3590   0.5642   2.5340
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.826894   0.755552   6.389 1.67e-10 ***
## Pclass      -1.011188   0.184951  -5.467 4.57e-08 ***
## Age         -0.038311   0.009065  -4.226 2.38e-05 ***
## SibSp       -0.560149   0.124992  -4.481 7.41e-06 ***
## Parch       -0.332872   0.134017  -2.484  0.01300 *
## CabinX       0.444359   0.468356   0.949  0.34274
## CabinB       0.733070   0.503920   1.455  0.14574
## CabinC       0.289978   0.447635   0.648  0.51711
## CabinD       1.242022   0.554660   2.239  0.02514 *
## CabinE       1.517975   0.555760   2.731  0.00631 **
## EmbarkedC    0.528518   0.246707   2.142  0.03217 *
## EmbarkedQ    0.417723   0.345704   1.208  0.22692
## TitleMiss   -0.381545   0.505025  -0.755  0.44995
## TitleMr     -3.238826   0.545273  -5.940 2.85e-09 ***
## TitleMrs     0.561431   0.561977   0.999  0.31778
## Titleother  -2.300479   0.707178  -3.253  0.00114 **
## ---
## 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:  718.54  on 875  degrees of freedom
## AIC: 750.54
##
## Number of Fisher Scoring iterations: 5
```

Com veiem, en aquest nou model hi ha més variables que resulten significatives i també millora la importància d'*Embarked*. Si calculem novament la precisió veiem que gairebé no ha variat.

```
# Calculem la precisió del model
regLog$results['Accuracy']
```

```
##    Accuracy
## 1 0.8293843
```

4.5.2. Corba ROC

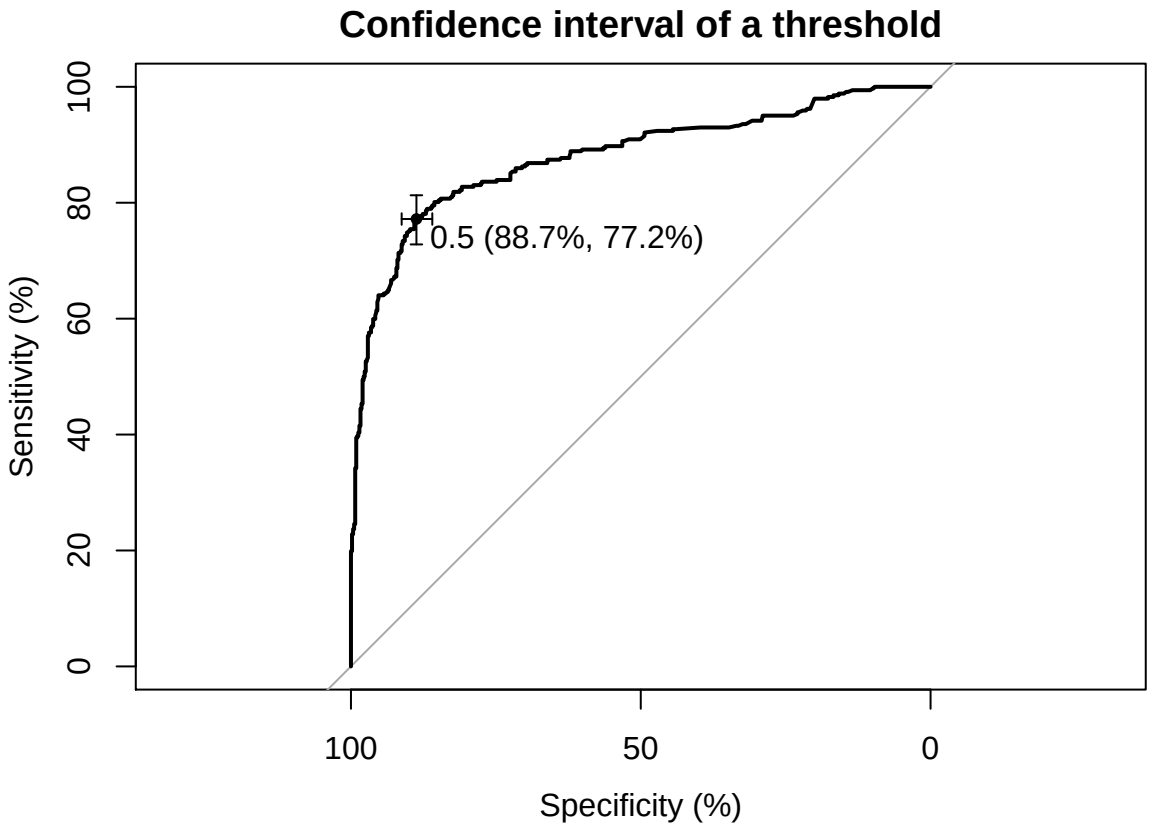
El model generat ens permet obtenir tant la probabilitat de sobreviure per a cada viatger com realitzar una predicció de *Survived*. A continuació calcularem la probabilitat de supervivència sobre el conjunt *train*, representarem la corba ROC i trobarem el llindar òptim per a determinar la supervivència de cada viatger.

```
# Calculem la probabilitat de supervivència dels passatgers del conjunt train
prob_Survived <- predict(regLog, train, type="prob")['1']
colnames(prob_Survived) <- "prob_Survived"
# Unim les dades al conjunt train
y_train <- cbind(train, prob_Survived)
head(y_train,5)
```

```
##   Survived Pclass    Sex Age SibSp Parch     Fare Cabin Embarked Title
## 1        0      3   male  22     1     0  7.25000    NO        S    Mr
## 2        1      1 female  38     1     0 35.64165     C        C   Mrs
## 3        1      3 female  26     0     0  7.92500    NO        S  Miss
## 4        1      1 female  35     1     0 26.55000     C        S   Mrs
## 5        0      3   male  35     0     0  8.05000    NO        S    Mr
##   Group prob_Survived
## 1     1    0.05475947
## 2     2    0.96006168
## 3     1    0.60245284
## 4     2    0.94081491
## 5     1    0.05806431
```

Per a obtenir el llindar més adequat de probabilitat representarem la corba ROC.

```
# Corba ROC del conjunt train
plot.roc(y_train$Survived, y_train$prob_Survived,
         main="Confidence interval of a threshold",
         percent=TRUE,  ci=TRUE, of="thresholds",
         thresholds="best",
         print.thres="best")
```



Els càlculs mostren que el llindar òptim per a la classe *Survived* és de 0.5; això vol dir que considerarem com a supervivents els passatgers que obtinguin una probabilitat superior a 0.5 segons el model. A partir d'aquest resultat podem fer una predicció del valor de *Survived* i mostrar la matriu de confusió.

```
# Assignem com supervivents les probabilitats superiors al 50%
y_train$Survived_pred <- ifelse(y_train$prob_Survived>0.5, 1, 0)
# Convertim les prediccions atipus factor
y_train$Survived_pred <- as.factor(y_train$Survived_pred)
# Mostrem la matriu de confusió
gmodels::CrossTable(y_train$Survived, y_train$Survived_pred)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  891
##
##
##                  | y_train$Survived_pred
## y_train$Survived |         0 |         1 | Row Total |
```

```
## -----------------|-----------|-----------|-----------|
##                0 |       481 |        68 |       549 |
##                  |    55.923 |    92.816 |           |
##                  |     0.876 |     0.124 |     0.616 |
##                  |     0.865 |     0.203 |           |
##                  |     0.540 |     0.076 |           |
## -----------------|-----------|-----------|-----------|
##                1 |        75 |       267 |       342 |
##                  |    89.771 |   148.994 |           |
##                  |     0.219 |     0.781 |     0.384 |
##                  |     0.135 |     0.797 |           |
##                  |     0.084 |     0.300 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |       556 |       335 |       891 |
##                  |     0.624 |     0.376 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

Com veiem, dels 891 registres del conjunt *train*, tenim:

- *Verdaders negatius*: 481 no supervivents s'han classificat correctament.
- *Falsos positius*: 68 no supervivents s'han classificat com a supervivients.
- *Falsos negatius*: 75 supervivents s'han classificat com a no supervivents.
- *Verdaders positius*: 267 supervivents s'han classificat correctament.

4.5.3. Predicció del conjunt test

A partir del model de regressió logística generat als apartats anteriors, podem fer una predicció de la probabilitat de supervivència del conjunt *test* amb la funció *predict*.

```
# Calculem la probabilitat de supervivència dels passatgers del conjunt test
prob_Survived <- predict(regLog, test, type="prob")['1']
colnames(prob_Survived) <- "prob_Survived"

# Afegim la predicció al conjunt de dades
solution <- cbind(test, prob_Survived)
# Assignem com a supervivents els passatgers amb probabilitats superiors al 50%
solution$Survived_pred <- ifelse(solution$prob_Survived>0.5, 1,0)
# Convertim la predicció a tipus factor
solution$Survived_pred<-as.factor(solution$Survived_pred)

# Mostrem els resultats per als primers registres
head(solution,5)
```

```
##   Pclass    Sex  Age SibSp Parch    Fare Cabin Embarked Title Group
## 1      3   male 34.5     0     0 7.82920    NO        Q    Mr     1
## 2      3 female 47.0     1     0 7.00000    NO        S   Mrs     1
## 3      2   male 62.0     0     0 9.68750    NO        Q    Mr     1
## 4      3   male 27.0     0     0 8.66250    NO        S    Mr     1
## 5      3 female 22.0     1     1 6.14375    NO        S   Mrs     2
##   prob_Survived Survived_pred
## 1    0.08710497             0
## 2    0.49849730             0
## 3    0.08379424             0
```

```
## 4    0.07727959            0
## 5    0.64996758            1
```

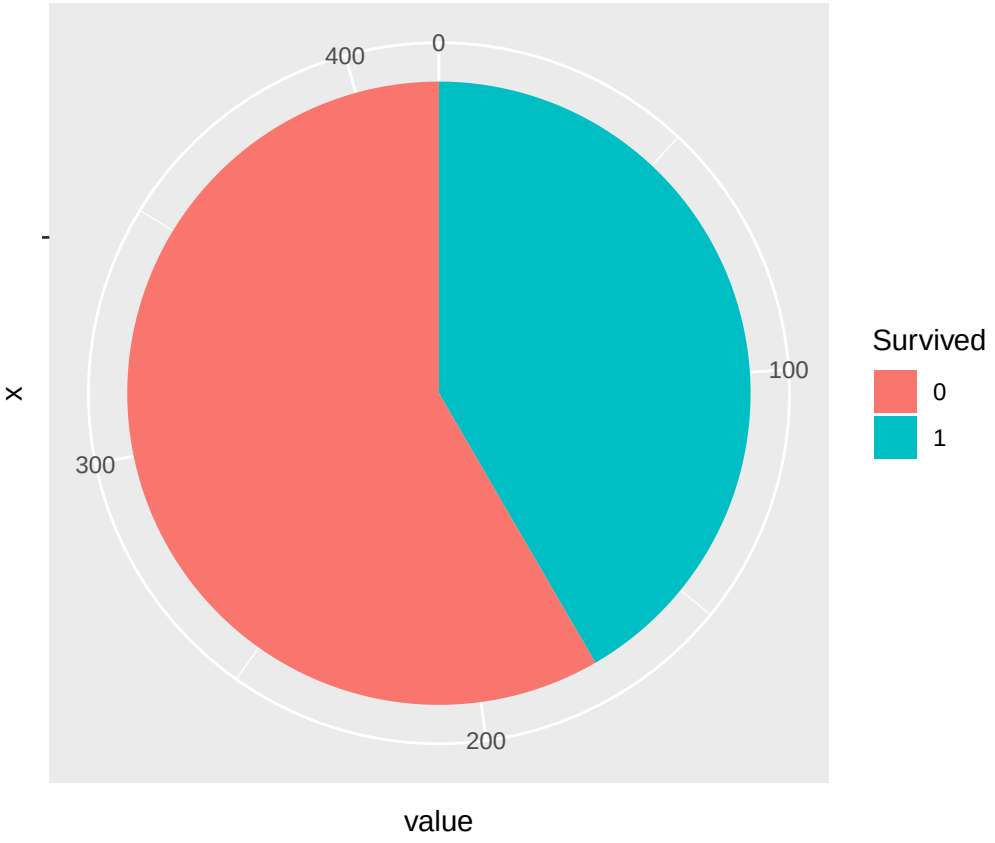
5. Representació de dades

En els apartats anteriors hem anat complementant les diferents anàlisis amb gràfiques i taules que ens mostraven els resultats obtinguts. En aquest apartat ens centrarem en representar els resultats de les prediccions sobre el conjunt *test*.

Diagrama de sectors

Podem representar el resultat total de supervivents i no supervivents amb un diagrama de sectors.

```
# Diagrama de sectors de Survived
taula <- data.frame(table(solution$Survived_pred))
colnames(taula) <- c("Survived", "value")

ggplot(taula, aes(x="", y=value, fill=Survived)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0)
```



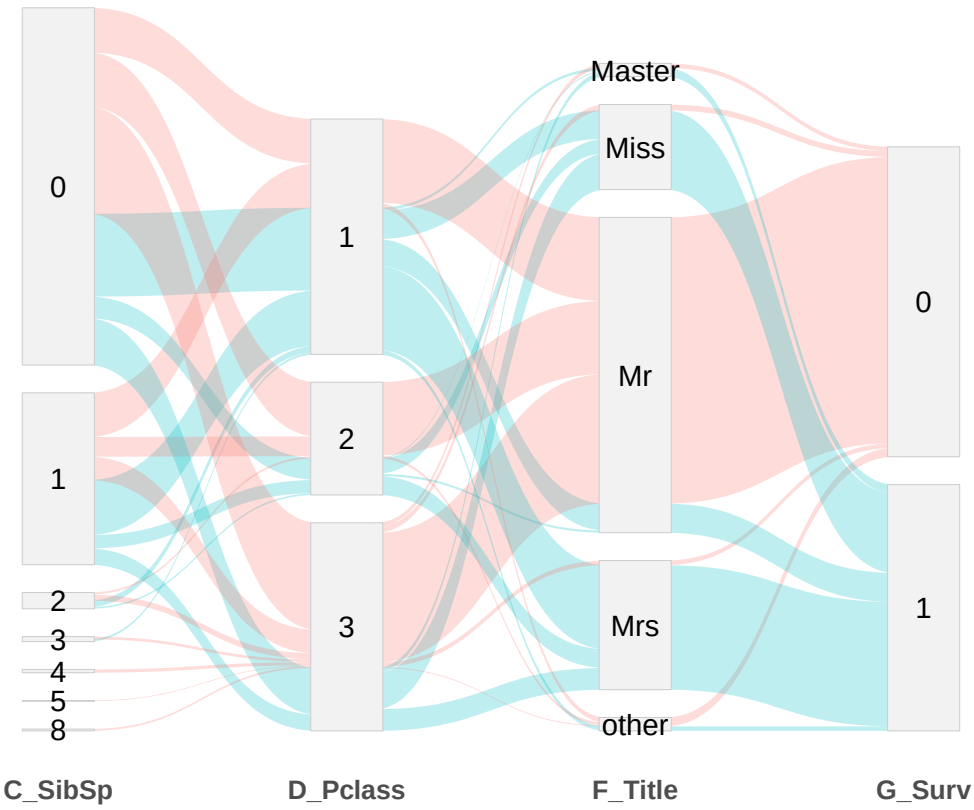
Gràfic de conjunts paral · lels

Per tal de tenir una visió general de la solució al problema crearem un gràfic de conjunts paral · lels que mostri les variables enteres i categòriques que han resultat més significatives per al model (*Pclass*, *SibSp* i *Title*) juntament amb la classe *Survived*.

```r
# Generem un gràfic de conjunts paral·lels
solution_temp <-solution
names(solution_temp)[names(solution_temp) == "Title"] <- "F_Title"
names(solution_temp)[names(solution_temp) == "SibSp"] <- "C_SibSp"
names(solution_temp)[names(solution_temp) == "Pclass"] <- "D_Pclass"
names(solution_temp)[names(solution_temp) == "Survived_pred"] <- "G_Surv"

solution_temp$C_SibSp<- as.factor(solution_temp$C_SibSp)
solution_temp$D_Pclass<- as.factor(solution_temp$D_Pclass)

data <- reshape2::melt(solution_temp)
```

```
## Using D_Pclass, Sex, C_SibSp, Cabin, Embarked, F_Title, G_Surv as id variables
```

```r
data <- gather_set_data(data, c("F_Title", "C_SibSp",  "D_Pclass", "G_Surv"))

ggplot(data, aes(x, id = id, split = y, value = value)) +
  geom_parallel_sets(aes(fill = G_Surv), alpha = 0.25, axis.width = 0.2,
                     n=100, strength = 0.5) +
  geom_parallel_sets_axes(axis.width = 0.25, fill = "gray95",
                          color = "gray80", size = 0.15) +
  geom_parallel_sets_labels(colour = 'black',angle = 0 ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text.y = element_blank(),
    axis.text.x = element_text(size = 10, face = "bold"),
    axis.title.x  = element_blank()
    )
```



El diagrama mostra clarament que la majoria de no supervivents són homes adults (títol *Mr*), probablement perquè els primers a evacuar el vaixell foren dones i nens. A més, veiem com els pocs *Mr* supervivents són gairebé tots de primera classe. També podem observar com el percentatge de supervivents entre els passatgers de primera classe resulta superior a la resta.

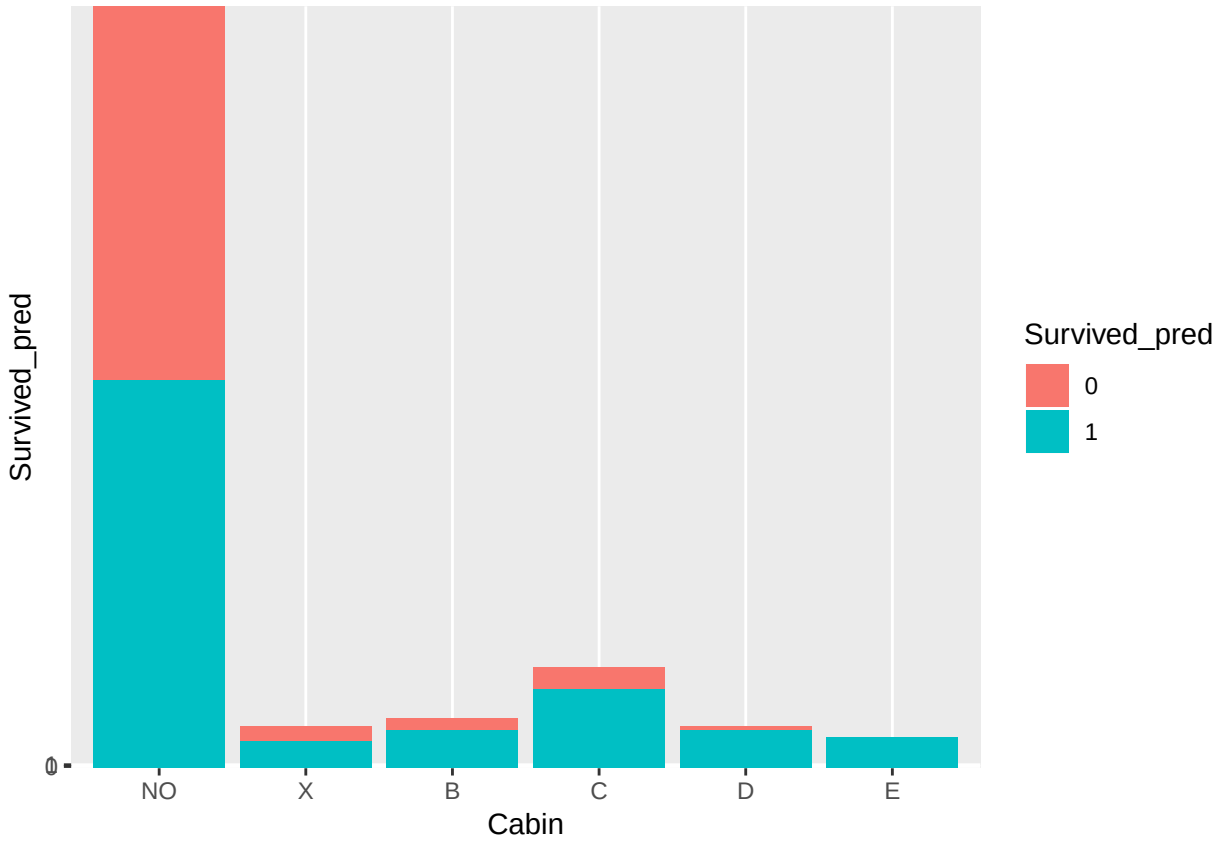
Diagrama de barres

Per observar les diferències entre la resta de variables categòriques (*Cabin* i *Embarked*) podem fer un diagrama de barres.

```
# Diagrama de barres de Cabin
taula <- melt(solution[,c("Survived_pred", "Cabin")])
```

```
## Using Survived_pred, Cabin as id variables
```

```
ggplot(solution, aes(x=Cabin, y=Survived_pred, fill=Survived_pred)) +
    geom_bar(stat="identity")
```

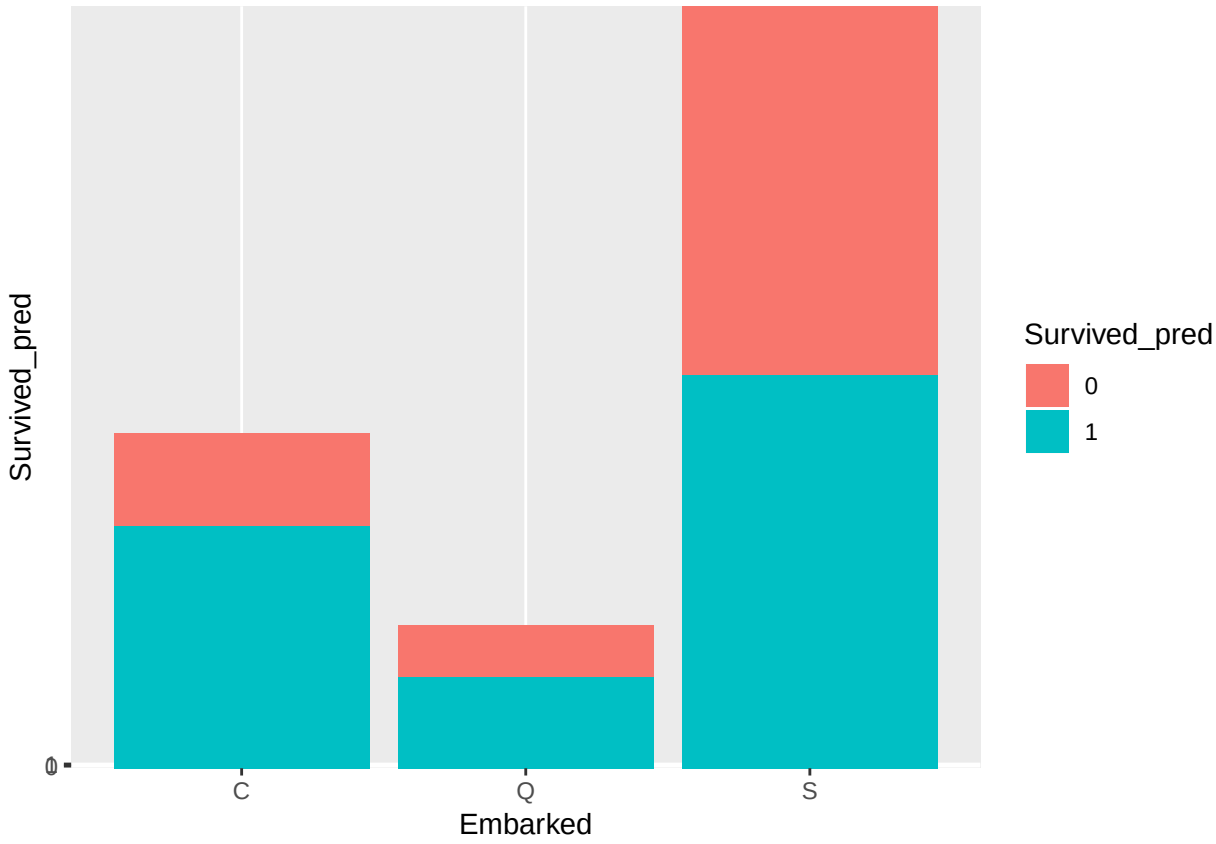


Com veiem, hi ha una diferència important entre els viatgers que no tenen cabina i la resta.

```
# Diagrama de barres d'Embarked
taula <- melt(solution[,c("Survived_pred", "Embarked")])
```

```
## Using Survived_pred, Embarked as id variables
```

```
ggplot(solution, aes(x=Embarked, y=Survived_pred, fill=Survived_pred)) +
    geom_bar(stat="identity")
```

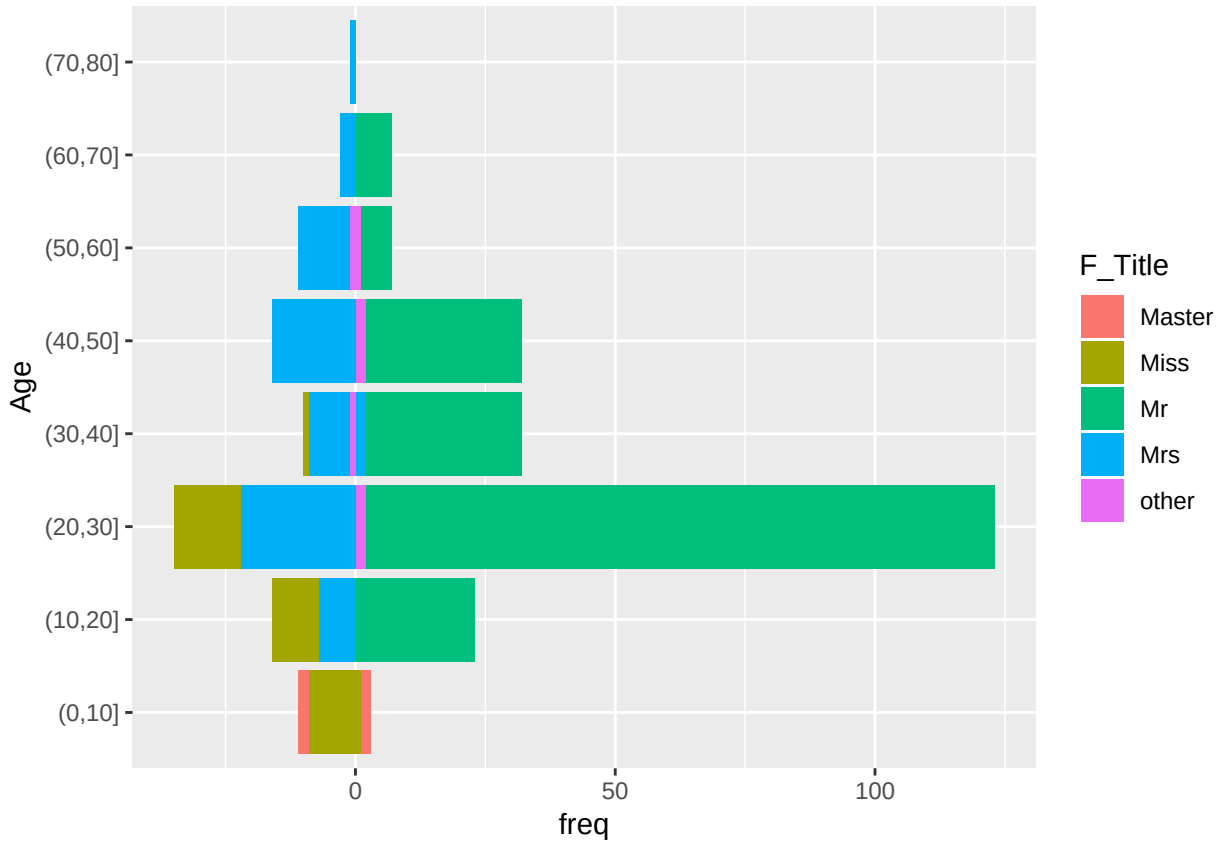


En aquest cas veiem que la probabilitat de sobreviure és inferior entre el passatgers embarcats a Southampton.

Piràmide segons Age

A continuació estudiarem els resultats en funció del valor de la variable numèrica *Age*. Per fer-ho crearem una piràmide que agrupi els passatgers per edats, representant les persones que sobreviuen a l'esquerra del 0 i els que no ho fan a la dreta. A més, mostrarem en diferents colors la categoria *Title*.

```
# Representació de la piràmide en funció d'Age
piramid_solution <- solution_temp[, c("Age","G_Surv","F_Title")]
piramid_solution$Age <- cut(piramid_solution$Age, seq(-0,100,by=10))
piramid_solution <- count(piramid_solution, c("Age","G_Surv","F_Title"))


ggplot(data = piramid_solution, aes(x = Age, y = freq, fill = F_Title)) +

  geom_bar(data = subset(piramid_solution, G_Surv == 0),
           stat = "identity",
           position = "identity") +
  geom_bar(data = subset(piramid_solution, G_Surv == 1) ,
           stat = "identity",
           position = "identity",
           mapping = aes(y = -freq)) +
        coord_flip() +
  scale_y_continuous(labels = abs)
```



Al gràfic veiem clarament com l'interval d'edats (20,30] és el més nombrós i també com els homes adults (títol de Mr) són els que majoritàriament no sobreviuen. Destaquem un petit grup amb titol "Master" (es refereix als nens en anglés) de (0,10] anys en què sobreviuen aproximadament la meitat.

6. Resolució del problema

A la fase d'anàlisi ens hem plantejat quatre preguntes per tal d'estudiar les dades. Les tres primeres preguntes estaven enfocades a estudiar la distribució de valors i la seva possible importància per a la construcció del model, mentre que la darrera pregunta plantejava l'objectiu final del problema.

Tant la representació gràfica (histograma i gràfica Q-Q) com els test estadístics (Shapiro-Wilk i Anderson-Darling) ens han mostrat que les variables numèriques contínues (*Age* i *Fare*) no segueixen una distribució normal. D'altra banda, si separem les passatgers segons el valor de la classe *Survived*, tant la mitjana com la variància dels valors d'*Age* poden considerar-se d'una mateixa població i, per tant, iguals per als dos grups. En canvi, la mitjana i la variància de *Fare* sí que mostren diferencies significatives en funció del valor de *Survived*. Aquest resultat ens indica que *Fare* pot tenir més importància per construir el model predictiu.

Utilitzant el coeficient de correlació d'Spearman hem obtingut una correlació important entre *Pclass-Fare*, i dues correlacions notables entre *Group-SibSp* i *Group-Parch*. Això ens ha plantejat la possibilitat de prescindir d'alguns d'aquests atributs al model.

Per estudiar la importància de les variables categòriques en el valor de *Survived* hem utilitzat el test *chi quadrat* (χ^2). Els resultats ens han mostrat que hi ha diferències significatives en el nombre de supervivents per a cada categoria; per tant, que les variables categòriques *Sex*, *Cabin*, *Embarked* i *Title* poden resultar importants per al model.

Finalment, hem construït un model de regressió logística capaç d'efectuar prediccions sobre el valor de la

classe *Survived* a partir de tots els atributs. La informació d'aquest model i els resultats de les proves anteriors els hem utilitzat per prescindir d'atributs redundants o poc significatius per al model (*Fare*, *Group* i *Sex*). El model final l'hem construït amb la resta de variables (*Pclass*, *Age*, *SibSp*, *Parch*, *Cabin*, *Embarked* i *Title*), l'hem avaluat amb una validació creuada en 4 particions (4-fold) i ens ha donat una precisió superior al 80%. Aquest model l'hem utilitzat per calcular la probabilitat de sobreviure per als passatgers del conjunt *test* i efectuar una predicció per al valor de *Survived*.

Per acabar, podem avaluar els resultats obtinguts guardant les prediccions de *test* en un fitxer (*submission.csv*) i enviant les dades a *Kaggle*. Si ho fem obtenim una precisió de 0.75598, que vol dir que hem classificat correctament més del 75% dels registres del conjunt *test*.

```
# Creem un dataframe amb l'identificador del viatger i la nostra predicció
submission <- data.frame(seq(892,1309),solution$Survived_pred)
# Assignem els noms correctes a les columnes
colnames(submission) <- c("PassengerId", "Survived")
# Guardem el dataframe en un fitxer
write.csv(submission, '../data/submission.csv', row.names = FALSE)
```

7. Recursos

- Calvo, M.; Subirats, L.; Pérez, D. (2019). *Introducción a la limpieza y análisis de los datos*. Editorial UOC.
- Squire, Megan (2015). *Clean Data*. Packt Publishing Ltd.
- Dalgaard, Peter (2008). *Introductory statistics with R*. Springer Science & Business Media.
- *Data visualization catalogue*. https://datavizcatalogue.com.

8. Taula de contribucions al treball

- **Recerca prèvia:** Jesús Marí, Víctor Boix
- **Redacció de les respostes:** Jesús Marí, Víctor Boix
- **Desenvolupament del codi:** Jesús Marí, Víctor Boix