Práctica 2: Limpieza y Validación de datos

Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

Gabriel Viscarret Atienza

7 de enero de 2019

Contents

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Descripción de los datos

Los datos que vamos a manejar han sido suministrados desde el reto Titanic: Machine Learning from Disaster, publicado en la página https://www.kaggle.com/c/titanic/kernels?sortBy=relevance&group=everyone&

search=easy&page=1&pageSize=20&competitionId=3136 En este reto se suministra una colección de datos respecto a los viajeros del Titanic, separados en dos ficheros. Uno de entrenamiento que incluye la variable dependiente, Survived, que indica si el viajero sobrevivió al famoso accidente del Titanic. El otro, de test, incluye los mismos campos que el primero, salvo esta variable de supervivencia. Y es que el reto consiste en predecir si los viajeros del fichero test sobrevieron a la catástrofe. Indicar que se adjunta un tercer fichero con el resultado real de estos viajeros, para poder comparar la predicción con datos reales. El archivo de test y de entrenamiento contienen estos campos comunes:

- **PassengerID:** Es un identificador único del viajero dentro del conjunto de datos.
- **Pclass:** Indica la clase de tarifa contratada. Tiene los valores 1, 2 y 3, siendo el 1 la clase mas alta y 3 la mas sencilla.
- **Name:** Nombre del pasajero. Indicar que incluye en todos los registros el tratamiento referido al viajero (señor, señora, capitán, etc)
- **Sex:** Indica el género del viajero, pudiendo tomar los valores masculino (male) o femenino (female)
- **Age:** Edad del viajero en años. Será fraccional si es menor de un año. Además será también fraccional (con fracción .5) para los viajeros de los que se desconoce su edad, y se ha hecho una estimación.
- **SibSP:** Define la relación familiar, sumando el número de hermanos y hermanas o de cónyuges.
- **Parch:** Define la relación familiar, sumando tanto el número de ascendentes como el de descendentes.
- **Ticket:** Indica el número o código del billete.
- **Fare:** Indica el valor del billete que ha adquirido el viajero.
- **Cabin:** Indica el número de camarote.
- **Embarked:** Indica con un código la ciudad donde embarcó el viajero, siendo el código 'C' para Cherbourg, 'Q' para Queenstown y 'S' para la ciudad de Southampton.

Importancia y objetivo del análisis

El accidente del Titanic es de sobra conocido, con multitud de escritos, libros e incluso películas. También con multitud de leyendas sobre ese viaje. La principal es que al no haber medidas de rescate para todos los viajeros, se dio prioridad al acceso a los botes salvavidas a los pasajeros de primera clase. Vamos a intentar pues descubrir los factores mas influyentes que hicieron sobrevivir a los pasajeros, y si principalmente salvaron la vida los viajeros de primera clase. Además, se ha elegido este reto por verlo interesante (por su variedad, relación y deficiencias de los datos) para esta práctica que se basa en la limpieza e integración de los datos. Como objetivo final, trataremos ver como con estos datos, tras clasificarlos y limpiarlos, nos serán válidos para predecir la supervivencia de los pasajeros y personal del Titanic.

Integración y selección de los datos de interés a analizar

Comencemos incorporando los datos desde los ficheros

```
#Cargamos los ficheros
titanic.train=read.csv("dataset\\train.csv")
titanic.test=read.csv("dataset\\test.csv")
```

Para a continuación tratar y limpiar los datos, vamos a fusionar los dos datasets para hacerlo en conjunto.

```
#Para poder fusionarlos, tienen que tener las mismas columnas
titanic.test$Survived<-NA
titanic.full<-rbind(titanic.train,titanic.test)
```

Mostramos su resumen para una evaluación rápida

```
summary(titanic.full)
```

```
##   PassengerId      Survived          Pclass
```

```
##  Min.   :   1   Min.   :0.0000   Min.   :1.000
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000
##  Median : 655   Median :0.0000   Median :3.000
##  Mean   : 655   Mean   :0.3838   Mean   :2.295
##  3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1309   Max.   :1.0000   Max.   :3.000
##                 NA's   :418
##                                  Name          Sex           Age
##  Connolly, Miss. Kate              :   2   female:466   Min.   : 0.17
##  Kelly, Mr. James                  :   2   male  :843   1st Qu.:21.00
##  Abbing, Mr. Anthony               :   1                Median :28.00
##  Abbott, Mr. Rossmore Edward       :   1                Mean   :29.88
##  Abbott, Mrs. Stanton (Rosa Hunt):   1                3rd Qu.:39.00
##  Abelson, Mr. Samuel               :   1                Max.   :80.00
##  (Other)                           :1301                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
```

Después de haber descrito las columnas, y ver este breve resumen, indicamos la selección de datos que son de nuestro interés y los que no.

- PassengerID: Es un identificador del viajero añadido al dataset. Lo mantendremos por que nos ayudará a comprobar los resultados, pero por supuesto no entra en el análisis.
- Survived: Es la variable dependiente, que tendremos que predecir para el conjunto de test.
- Name: No tiene valor para el análisis, pero de momento no lo quitamos. Lo mantenemos para el proceso de limpieza ya que indica el tratamiento personal del viajero, que nos será útil. Sí que lo quitaremos en el análisis.
- Ticket: Es un código del billete, sin relación ni lógca que nos valga para el análisis ni para la limpieza, por lo que lo quitaremos.
- Cabin: Es un campo muy incompleto, con pocos datos, por lo que decidimos no incluirlo en el análisis. El resto de campos sí los usaremos en nuestro análisis.

Viendo los datos que queremos usar, vamos a comprobar primero el tipo de datos que R les ha asignado.

```
sapply(titanic.full,function (x) class(x))
```

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

Entendemos que Survived y Pclass deberían considerarse como factor. Además la edad la vamos a convertir a entero ya que no tiene mucho sentido trabajar con decimales. De esa manera unificamos todos los valores

menores de 1, y quitamos el indicativo de edad estimada.

```
titanic.full$Survived<-as.factor(titanic.full$Survived)
titanic.full$Pclass<-as.factor(titanic.full$Pclass)
titanic.full$Age<-as.integer(titanic.full$Age)
sapply(titanic.full,function (x) class(x))
```

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

Limpieza de los datos

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Observando el resumen de los datos, vemos columnas que contienen ceros o valores nulos.

Edad

Podemos ver como la edad tiene ceros y nulos. Los ceros son datos reales (en el dataset original había datos decimales para la edad menor de uno, que hemos transformado a entero. En el dataset original no había ningún cero) pero nos queda por resolver los nulos. Para ello vamos a meter la mediana de la edad en estos campos nulos. Y para afinar mas, vamos a agrupar los datos para calcular diferentes medias y asignarlas. En el campo nombre, se indica el tratamiento de cada persona (señor, señora, señorita, etc) que tienen cierta relación con la edad. Por eso vamos a extraer este tratamiento, agrupar por él y calcular la mediana para cada grupo.

Extraemos pues el tratamiento de cada viajero

```
titanic.full$tittle<-gsub('(.*, )|(\\..*)','',titanic.full$Name)
table(titanic.full$tittle)
```

```
##
##         Capt          Col          Don         Dona           Dr
##            1            4            1            1            8
##     Jonkheer         Lady        Major       Master         Miss
##            1            1            2           61          260
##         Mlle          Mme           Mr          Mrs           Ms
##            2            1          757          197            2
##          Rev          Sir the Countess
##            8            1            1
```

Para cada uno de estos grupos, calcularemos la mediana de la edad (sin tener en cuenta los nulos que sabemos que hay). Pero primero miramos cuales de los grupos tiene nulos.

```
titanic.NoAge=titanic.full[is.na(titanic.full$Age),]
table(titanic.NoAge$tittle)
```

```
##
##     Dr Master   Miss     Mr    Mrs     Ms
##      1      8     50    176     27      1
```

Calculamos pues la media de estos tipos de pasajeros.

```
titanic.Dr<-titanic.full[titanic.full$tittle=='Dr',]
titanic.Dr.median<-median(titanic.Dr$Age,na.rm=TRUE)

titanic.Master<-titanic.full[titanic.full$tittle=='Master',]
titanic.Master.median<-median(titanic.Master$Age,na.rm=TRUE)

titanic.Miss<-titanic.full[titanic.full$tittle=='Miss',]
titanic.Miss.median<-median(titanic.Miss$Age,na.rm=TRUE)

titanic.Mr<-titanic.full[titanic.full$tittle=='Mr',]
titanic.Mr.median<-median(titanic.Mr$Age,na.rm=TRUE)

titanic.Mrs<-titanic.full[titanic.full$tittle=='Mrs',]
titanic.Mrs.median<-median(titanic.Mrs$Age,na.rm=TRUE)

titanic.Ms<-titanic.full[titanic.full$tittle=='Ms',]
titanic.Ms.median<-median(titanic.Ms$Age,na.rm=TRUE)
```

Y asignamos estos cálculos a los nulos.

```
titanic.full[is.na(titanic.full$Age)&titanic.full$tittle=='Dr',"Age"]<-titanic.Dr.median
titanic.full[is.na(titanic.full$Age)&titanic.full$tittle=='Master',"Age"]<-titanic.Master.median
titanic.full[is.na(titanic.full$Age)&titanic.full$tittle=='Miss',"Age"]<-titanic.Miss.median
titanic.full[is.na(titanic.full$Age)&titanic.full$tittle=='Mr',"Age"]<-titanic.Mr.median
titanic.full[is.na(titanic.full$Age)&titanic.full$tittle=='Mrs',"Age"]<-titanic.Mrs.median
titanic.full[is.na(titanic.full$Age)&titanic.full$tittle=='Ms',"Age"]<-titanic.Ms.median
```

Con lo que tenmos este campo corregido. Si vemos su resumen, vemos sus nuevas estadísticas y que ya no tenemos nulos.

```
summary(titanic.full$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   22.00   29.00   29.41   35.50   80.00
```

Tarifa

Como hemos visto, la tarifa (Fare) también tiene nulos

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

```
##      PassengerId Survived Pclass               Name  Sex Age SibSp Parch
## 1044        1044     <NA>      3 Storey, Mr. Thomas male  60     0     0
##      Ticket Fare Cabin Embarked tittle
## 1044   3701   NA              S     Mr
```

Y también tiene ceros, 18 en concreto

```
titanic.full[titanic.full$Fare==0.0,]
```

```
##      PassengerId Survived Pclass                                Name
## 180          180        0      3                  Leonard, Mr. Lionel
## 264          264        0      1                 Harrison, Mr. William
## 272          272        1      3          Tornquist, Mr. William Henry
## 278          278        0      2           Parkes, Mr. Francis "Frank"
## 303          303        0      3       Johnson, Mr. William Cahoone Jr
## 414          414        0      2        Cunningham, Mr. Alfred Fleming
```

```
## 467          467        0      2                 Campbell, Mr. William
## 482          482        0      2      Frost, Mr. Anthony Wood "Archie"
## 598          598        0      3                   Johnson, Mr. Alfred
## 634          634        0      1         Parr, Mr. William Henry Marsh
## 675          675        0      2            Watson, Mr. Ennis Hastings
## 733          733        0      2                  Knight, Mr. Robert J
## 807          807        0      1                Andrews, Mr. Thomas Jr
## 816          816        0      1                     Fry, Mr. Richard
## 823          823        0      1       Reuchlin, Jonkheer. John George
## NA            NA     <NA>   <NA>                                  <NA>
## 1158        1158     <NA>      1 Chisholm, Mr. Roderick Robert Crispin
## 1264        1264     <NA>      1               Ismay, Mr. Joseph Bruce
##       Sex Age SibSp Parch Ticket Fare       Cabin Embarked   tittle
## 180  male  36     0     0   LINE    0                   S       Mr
## 264  male  40     0     0 112059    0         B94       S       Mr
## 272  male  25     0     0   LINE    0                   S       Mr
## 278  male  29     0     0 239853    0                   S       Mr
## 303  male  19     0     0   LINE    0                   S       Mr
## 414  male  29     0     0 239853    0                   S       Mr
## 467  male  29     0     0 239853    0                   S       Mr
## 482  male  29     0     0 239854    0                   S       Mr
## 598  male  49     0     0   LINE    0                   S       Mr
## 634  male  29     0     0 112052    0                   S       Mr
## 675  male  29     0     0 239856    0                   S       Mr
## 733  male  29     0     0 239855    0                   S       Mr
## 807  male  39     0     0 112050    0         A36       S       Mr
## 816  male  29     0     0 112058    0        B102       S       Mr
## 823  male  38     0     0  19972    0                   S Jonkheer
## NA   <NA>  NA    NA    NA   <NA>   NA        <NA>     <NA>     <NA>
## 1158 male  29     0     0 112051    0                   S       Mr
## 1264 male  49     0     0 112058    0 B52 B54 B56       S       Mr
```

Para resolver el nulo, vamos a asignarle la mediana de un grupo similar a ese individuo.

```
titanic.fare.median<-median(titanic.full[titanic.full$Pclass==3&titanic.full$Sex=='male'&titanic.full$Em
titanic.full[is.na(titanic.full$Fare),"Fare"]<-titanic.fare.median
summary(titanic.full$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.896  14.454  33.276  31.275 512.329
```

Para los ceros, vamos calcular las medianas de los 18 individuos, clasificados en varios grupos. Son todos mujeres embarcadas en Southampton, de primera, segunda y tercera clase

```
titanic.fareP1.median<-median(titanic.full[titanic.full$Pclass==1&titanic.full$Sex=='male'&titanic.full$
titanic.fareP2.median<-median(titanic.full[titanic.full$Pclass==2&titanic.full$Sex=='male'&titanic.full$
titanic.fareP3.median<-median(titanic.full[titanic.full$Pclass==3&titanic.full$Sex=='male'&titanic.full$

titanic.full[titanic.full$Fare==0&titanic.full$Pclass==1,"Fare"]<-titanic.fareP1.median
titanic.full[titanic.full$Fare==0&titanic.full$Pclass==2,"Fare"]<-titanic.fareP2.median
titanic.full[titanic.full$Fare==0&titanic.full$Pclass==3,"Fare"]<-titanic.fareP3.median
summary(titanic.full$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.171   7.925  14.458  33.550  31.387 512.329
```

Embarque

Tiene dos valores nulos (asignados a una cadena vacía)

```
table(titanic.full$Embarked)
```

```
##
##       C   Q   S
##   2 270 123 914
```

Son sólo dos, por lo que simplemente les vamos a asignar el valor mayoritorio, S. Antes vamos a ver si separando en grupos se sigue cumpliendo que S es la mayoritaria.

```
table(titanic.full$Embarked,titanic.full$Pclass)
```

```
##
##       1   2   3
##       2   0   0
##   C 141  28 101
##   Q   3   7 113
##   S 177 242 495
```

```
table(titanic.full$Embarked,titanic.full$SibSp)
```

```
##
##       0   1   2   3   4   5   8
##       2   0   0   0   0   0   0
##   C 171  90   9   0   0   0   0
##   Q 100  14   4   0   5   0   0
##   S 618 215  29  20  17   6   9
```

Asignamos pues Southampton a estas dos variables. Siendo sólo dos no vamos a crear grandes distorsiones.

```
titanic.full[titanic.full$Embarked=="","Embarked"]<-"S"
```

Cabina

Hemos visto que esta columna tiene muchos campos nulos.

```
sum(is.na(titanic.full$Cabin))
```

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

```
sum(titanic.full$Cabin=="")
```

```
## [1] 1014
```

1014 de 1309 observaciones son nulas, por lo que no pueden tener validez en el estudio y la vamos a descartar.

Identificación y tratamiento de valores extremos

Volvemos a mostrar el resumen del dataset

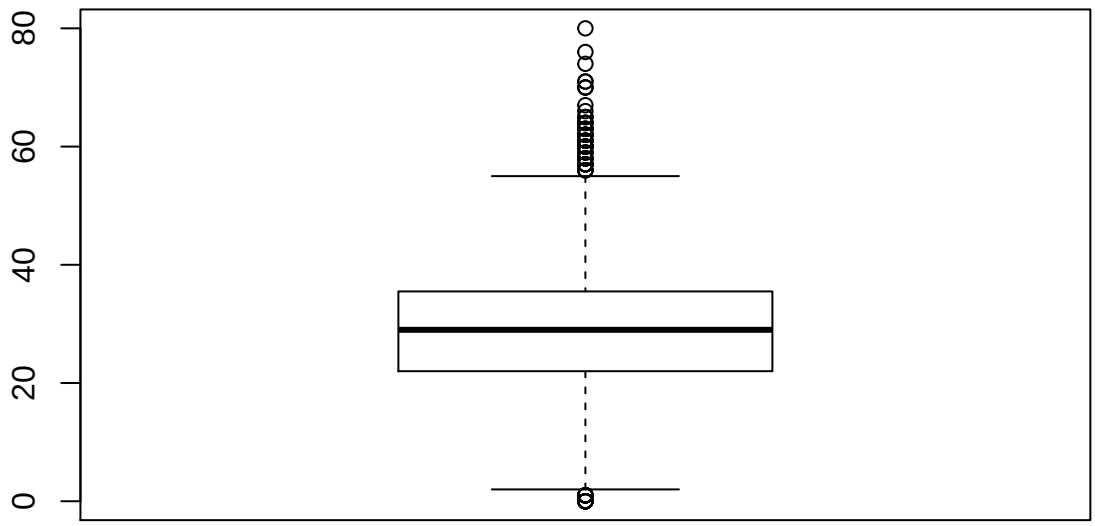
```
summary(titanic.full)
```

```
##   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.   : 0.00   Min.   :0.0000   Min.   :0.000
##  male  :843   1st Qu.:22.00   1st Qu.:0.0000   1st Qu.:0.000
##               Median :29.00   Median :0.0000   Median :0.000
##               Mean   :29.41   Mean   :0.4989   Mean   :0.385
##               3rd Qu.:35.50   3rd Qu.:1.0000   3rd Qu.:0.000
##               Max.   :80.00   Max.   :8.0000   Max.   :9.000
##
##       Ticket          Fare                     Cabin      Embarked
##  CA. 2343:  11   Min.   :  3.171                  :1014    :  0
##  1601    :   8   1st Qu.:  7.925   C23 C25 C27    :   6   C:270
##  CA 2144 :   8   Median : 14.458   B57 B59 B63 B66:   5   Q:123
##  3101295 :   7   Mean   : 33.550   G6             :   5   S:916
##  347077  :   7   3rd Qu.: 31.387   B96 B98        :   4
##  347082  :   7   Max.   :512.329   C22 C26        :   4
##  (Other) :1261                     (Other)        : 271
##     tittle
##  Length:1309
##  Class :character
##  Mode  :character
##
##
##
##
```

- **Pasanger id:** Es un identificador, que podremos usar para indexar, pero no forma parte del análisis
- **Survived:** Es la variable dependiente, que toma 0 o 1 (o nulo para los del test)
- **Name** Son textos con los nombres, por lo que no tiene sentido analizar valores extremos
- **Sex:** Es un factor con dos valores, male, female
- **Age:** Esta variable si que entra en el análisis, y es de tipo entero

```
boxplot(titanic.full$Age)
```



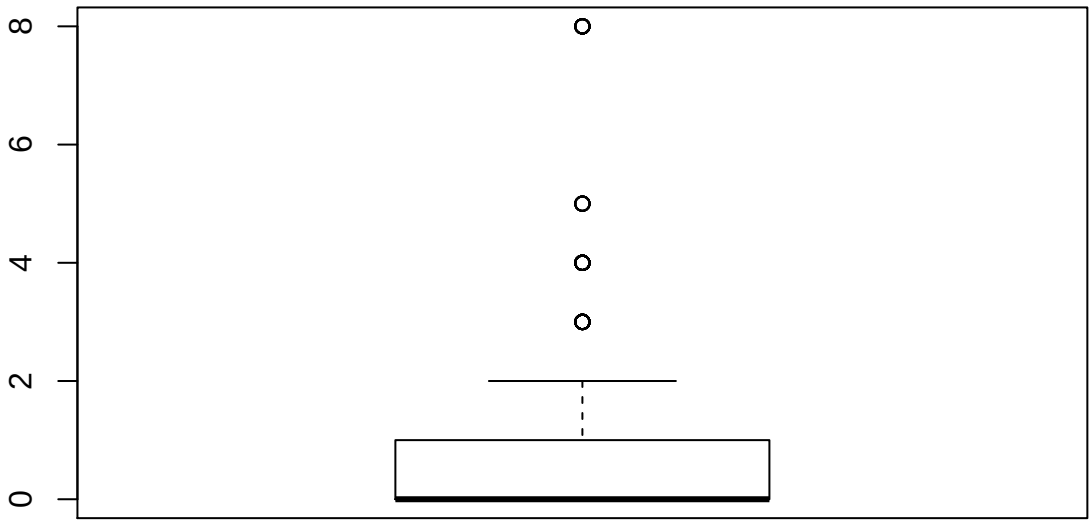

```
boxplot.stats(titanic.full$Age)$out
```

```
##  [1] 58 66 65  0 59 71 70  1 61  1 56  1 58 59 62 58 63 65  0 61 60  1  1
## [24] 64 65 56  0 63 58 71 64 62 62 60 61 57 80  0 56 58 70 60 60 70  0 57
## [47]  1  0  1 62  0 74 56 62 63 60 60 67 76 63  1 61 60 64 61  0 60 57 64
## [70]  0  1  0  1 64  0 57 58  0 59 57
```

Detecta varios valores como outliers, pero son valores válidos en la representación de la edad (valores entre cero y 82)

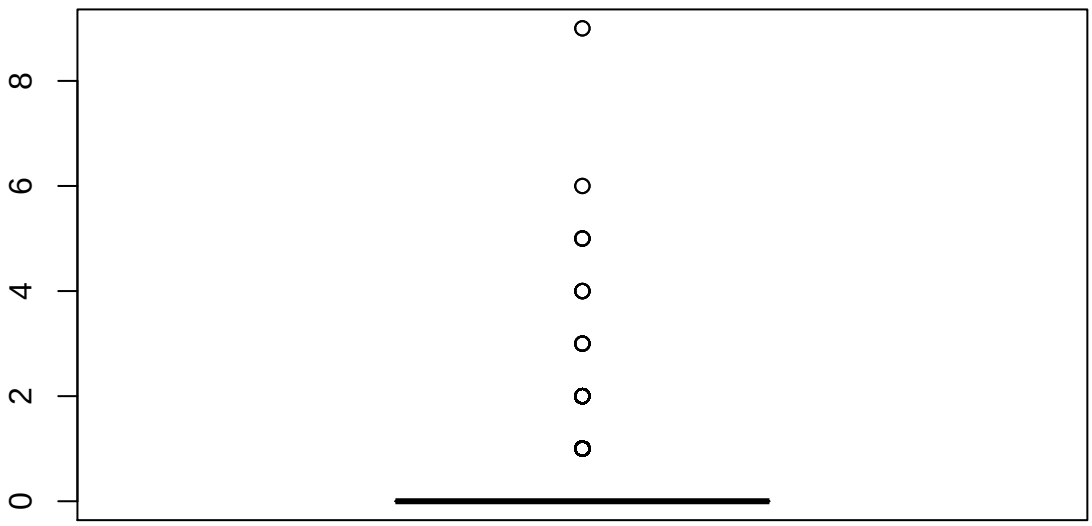
- **Sibsp:** Podemos ver si tiene extremos, pero ya vemos que va de cero a ocho, que son valores lógicos, Son datos correctos

```
boxplot(titanic.full$SibSp)
```



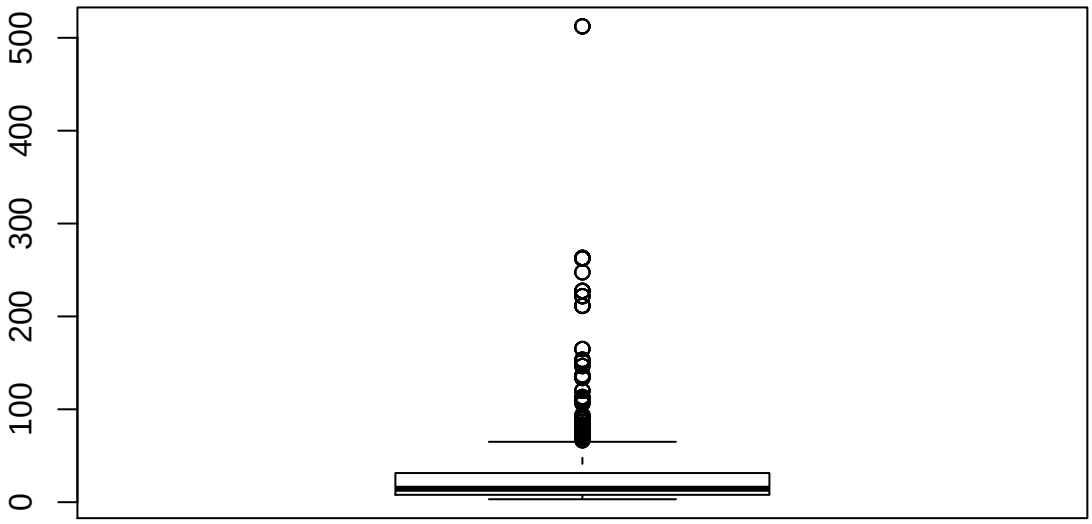
- **Parch:** Al igual que el caso anterior, está en valores lógicos por lo que si se detectan outliers, pero los consideramos como válidos

```
boxplot(titanic.full$Parch)
```



- **Ticket:** Es un campo que no vamos a usar
- **Fare:** Tiene unos valores muy extremos, pero entendemos que son tarifas especiales y válidas.

```
boxplot(titanic.full$Fare)
```



- **Cabin:** Son textos, en su mayor parte vacíos. No los vamos a usar
- **Embarked:** Es un factor con tres elementos distintos.

Análisis de los datos

Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Como indicamos al principio, no vamos a usar todas las columnas. Quitamos el nombre, el código de billete y el número de camarote. Además, anteriormente fusionamos los datos de test y entranamiento para realizar mejor la limpieza. Ahora, después de limpiar los datos y elminar las columnas innecsarias, los volvemos a segregar.

```
titanic.transfomed<-titanic.full[,-c(4,9,11)]
titanic.transfomed$tittle<-as.factor(titanic.transfomed$tittle)
titanic.transformed.test<-titanic.transfomed[is.na(titanic.transfomed$Survived),]
titanic.transformed.train<-titanic.transfomed[!(is.na(titanic.transfomed$Survived)),]
```

Vamos a hacer una operación mas. Uno de los objetivos es ver la relación entre las variables y la supervivencia. Por eso vamos a dejar preparado otro conjunto de datos, sin el identificador de viajero y con las columnas de tipo numérico, para poder hacer la correlación.

```
titanic.corr<-titanic.transformed.train[,-c(1)]
titanic.corr$Survived<-as.numeric(titanic.corr$Survived)
titanic.corr$Pclass<-as.numeric(titanic.corr$Pclass)
titanic.corr$Sex<-as.numeric(titanic.corr$Sex)
titanic.corr$SibSp<-as.numeric(titanic.corr$SibSp)
titanic.corr$Parch<-as.numeric(titanic.corr$Parch)
titanic.corr$Embarked<-as.numeric(titanic.corr$Embarked)
titanic.corr$tittle<-as.numeric(titanic.corr$tittle)
```

Comprobación de la normalidad y homogeneidad de la varianza.

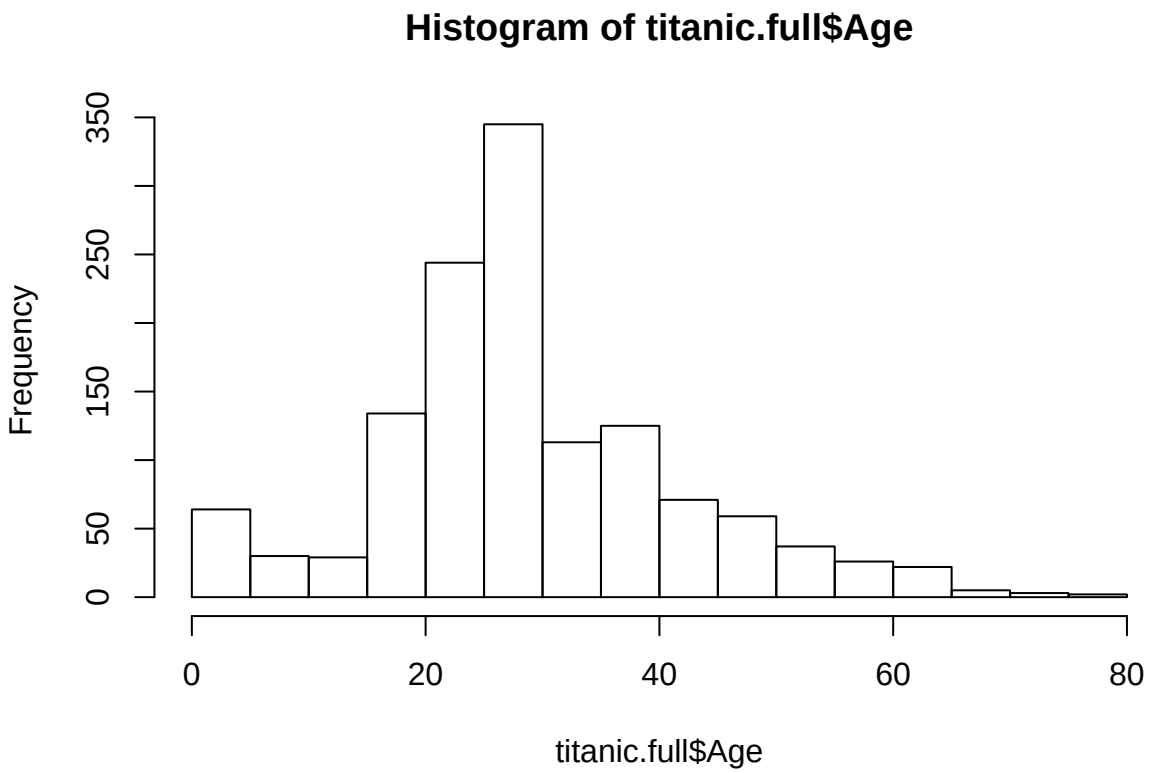
Vamos a hacer el test de normalidad para las variables cuantitativas.

```
shapiro.test(titanic.full$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic.full$Age
## W = 0.96567, p-value < 2.2e-16
```

El test nos indica que los datos de edad no están normalizados ya que su p-valor es inferior a 0.05. Indicar que al tener mas de 30 registros si que podemos aproximarla a una distribución normal. Este es su histograma:

```
hist(titanic.full$Age)
```

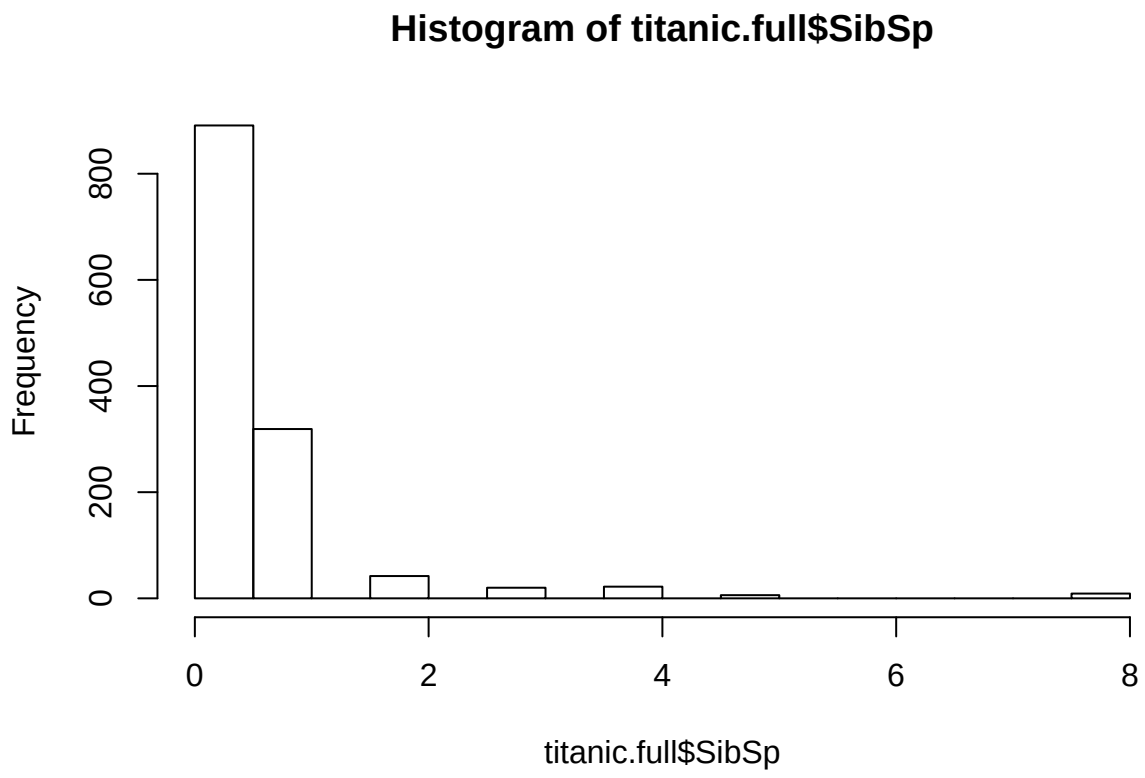


Para las variables de unidad familiar, obtenemos la misma conclusión.

```r
shapiro.test(titanic.full$SibSp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic.full$SibSp
## W = 0.51108, p-value < 2.2e-16
```

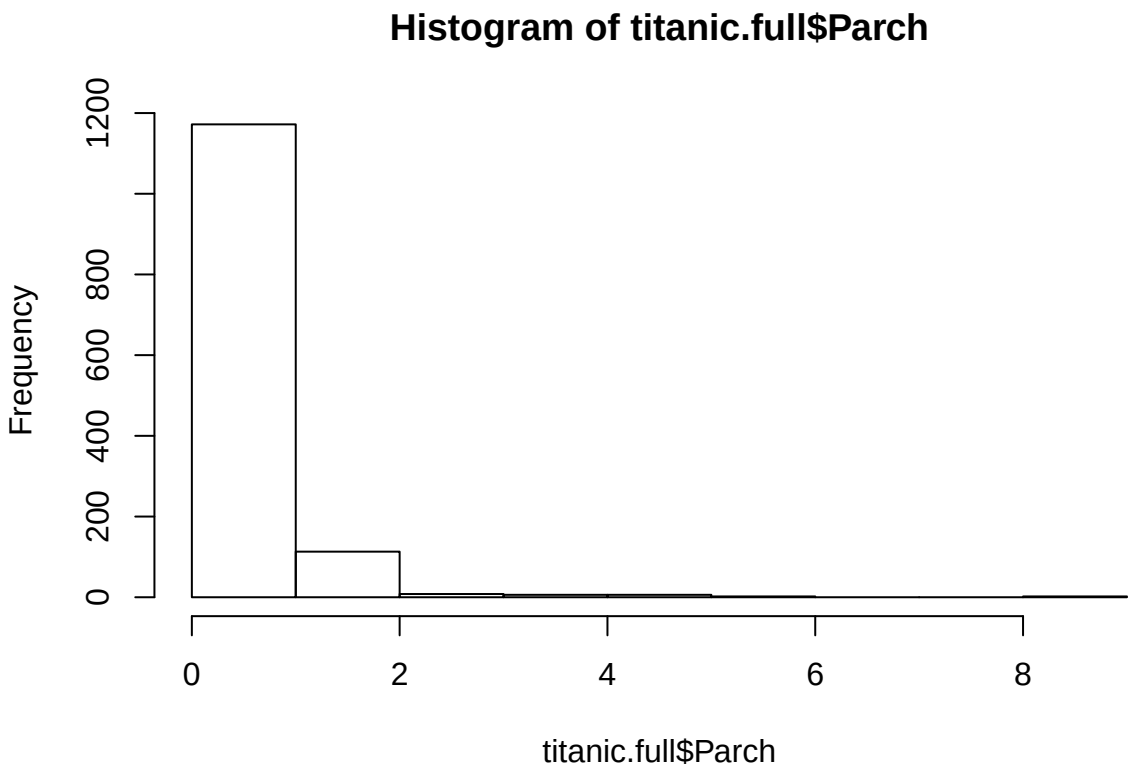
```r
hist(titanic.full$SibSp)
```




```r
shapiro.test(titanic.full$Parch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic.full$Parch
## W = 0.49797, p-value < 2.2e-16
```

```r
hist(titanic.full$Parch)
```



Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc

Correlación

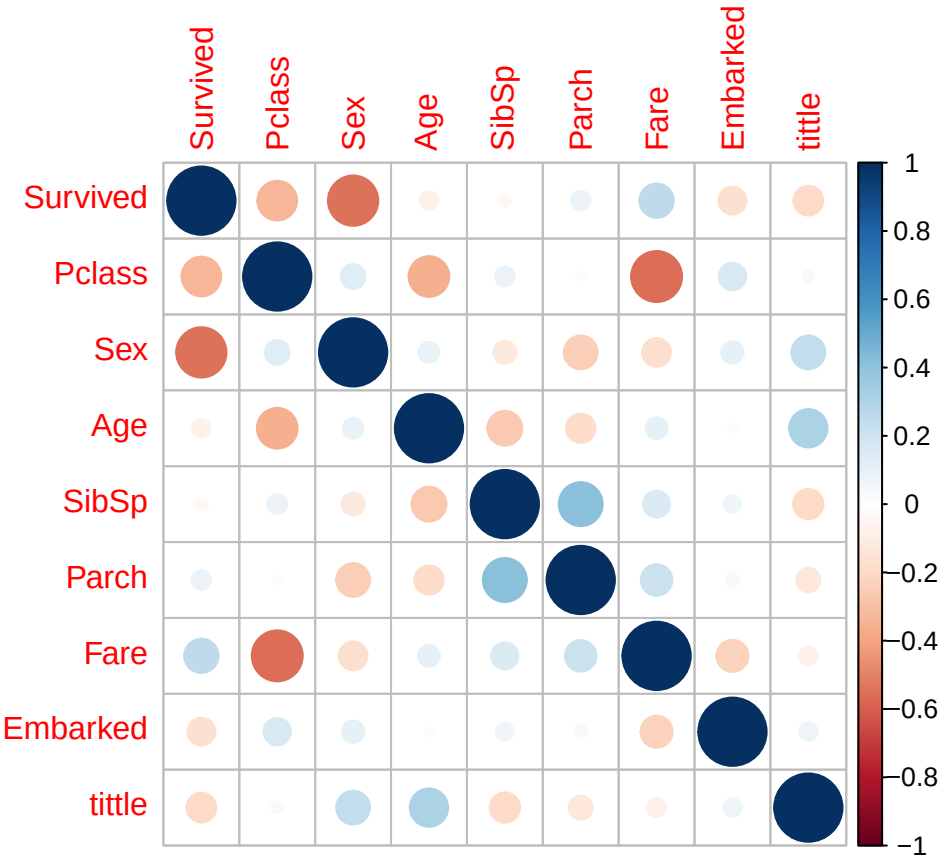
Vamos a estudiar la correlación de las variables, para ver cual puede afectar mas a la supervivencia, además de detectar altas correlaciones entre las variables.

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.2
```

```
## corrplot 0.84 loaded
```

```
titanic.correlacion<-cor(titanic.corr)
corrplot(titanic.correlacion,method="circle")
```




```
titanic.correlacion
```

```
##              Survived      Pclass         Sex         Age       SibSp
## Survived  1.00000000 -0.33848104 -0.54335138 -0.07146567 -0.03532250
## Pclass   -0.33848104  1.00000000  0.13190049 -0.35634144  0.08308136
## Sex      -0.54335138  0.13190049  1.00000000  0.09230282 -0.11463081
## Age      -0.07146567 -0.35634144  0.09230282  1.00000000 -0.26171004
## SibSp    -0.03532250  0.08308136 -0.11463081 -0.26171004  1.00000000
## Parch     0.08162941  0.01844267 -0.24548896 -0.18471417  0.41483770
## Fare      0.25318641 -0.55723289 -0.17798473  0.10266109  0.15696087
## Embarked -0.16767531  0.16209780  0.10826220 -0.01857073  0.06823029
## tittle   -0.19048835  0.03251975  0.24508508  0.31548383 -0.19688362
##                Parch        Fare    Embarked      tittle
## Survived  0.08162941  0.25318641 -0.16767531 -0.19048835
## Pclass    0.01844267 -0.55723289  0.16209780  0.03251975
## Sex      -0.24548896 -0.17798473  0.10826220  0.24508508
## Age      -0.18471417  0.10266109 -0.01857073  0.31548383
## SibSp     0.41483770  0.15696087  0.06823029 -0.19688362
## Parch     1.00000000  0.21368319  0.03979839 -0.12440730
## Fare      0.21368319  1.00000000 -0.22146722 -0.07740623
## Embarked  0.03979839 -0.22146722  1.00000000  0.07333313
## tittle   -0.12440730 -0.07740623  0.07333313  1.00000000
```

La clase de billete afecta mucho a la supervivencia, pero podemos ver como la que mas afecta es el género. Las que menos afectan son la edad y las condiciones familiares. Se aprecia además una fuerte relación (logicamente) entre la clase de tarifa y su precio.

Regresión

Además de la correlación, vamos a crear un modelo de regresión logística que nos indique la importancia de las variables en la dependiente Survived

```
titanic.regresion<-glm(Survived~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked,data=titanic.transformed.train
summary(titanic.regresion)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, family = "binomial", data = titanic.transformed.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6402  -0.6028  -0.4123   0.6227   2.4703
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.235081   0.479100   8.840  < 2e-16 ***
## Pclass2     -0.983858   0.299242  -3.288  0.00101 **
## Pclass3     -2.231304   0.300526  -7.425 1.13e-13 ***
## Sexmale     -2.720214   0.201523 -13.498  < 2e-16 ***
## Age         -0.041165   0.007859  -5.238 1.63e-07 ***
## SibSp       -0.337357   0.109182  -3.090  0.00200 **
## Parch       -0.090051   0.119324  -0.755  0.45044
## Fare         0.001762   0.002433   0.724  0.46896
## EmbarkedQ   -0.153177   0.386797  -0.396  0.69209
## EmbarkedS   -0.448196   0.238660  -1.878  0.06039 .
## ---
## 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:  781.69  on 881  degrees of freedom
## AIC: 801.69
##
## Number of Fisher Scoring iterations: 5
```

Volvemos a ver como la clase (concretamente la tercera clase) y el sexo masculino son variables muy significativas para la supervivencia y vemos como lo son también la edad del pasajero y el número de hermanos o cónyuges.

Predicción

Vamos a aplicar un algoritmo **Random Forest** para generar una predicción de la supervivencia de los viajeros del grupo test. Vamos a incluir en la fórmula la clase, el sexo, el embarque y la tarifa

```
library('randomForest')
```

```
## Warning: package 'randomForest' was built under R version 3.5.2
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
set.seed(123)
titanic.random.forest <- randomForest(factor(Survived) ~ Pclass + Sex  + Fare + Embarked +tittle, data =
```

```r
# prediction
titanic.random.predct = predict(titanic.random.forest)
titanic.random.fitted = rep(NA,891)
for(i in 1:891){
  titanic.random.fitted[i] = as.integer(titanic.random.predct[[i]]) - 1
}
# Résultat
table(titanic.random.fitted)
```

```
## titanic.random.fitted
##   0   1
## 624 267
```

```r
print(titanic.random.forest)
```

```
##
## Call:
##  randomForest(formula = factor(Survived) ~ Pclass + Sex + Fare +      Embarked + tittle, data = titan
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 17.17%
## Confusion matrix:
##     0   1 class.error
## 0 510  39  0.07103825
## 1 114 228  0.33333333
```

```r
titanic.prediction <- predict(titanic.random.forest, titanic.transformed.test)

# Solution 2 columns (prediction)
titanic.solution <- data.frame(Survived = titanic.prediction, PassengerID = titanic.transformed.test$Pas
```

```r
titanic.test.result<-titanic.test
titanic.test.result$Survived<-titanic.solution$Survived
```

Representación de los resultados a partir de tablas y gráficas

---

Vamos a comparar el resultado de nuestro test con el resultado real.

```r
titanic.gender_submission=read.csv("dataset\\gender_submission.csv")
titanic.test<-titanic.test.result[,c(1,12)]
table(titanic.gender_submission$Survived,titanic.test$Survived)
```

```
##
##       0   1
##   0 248  18
##   1  30 122
```

Hemos tenido 30 + 18 predicciones incorrectas, de 418 registros (un 11.5%), por lo que vemos que el modelo

es bastante efectivo.

Para terminar de evaluar el modelo, vamos a representar en varios gráficos las tasas de supervivencia respecto a diversos factores, representando primero del resultado de nuestra evaluación y a su derecha de los de entranamiento, para así pode comparar visualmente las tasas de supervivencia reales con las que acabamos de calcular.

- Supervivencia respecto al género.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

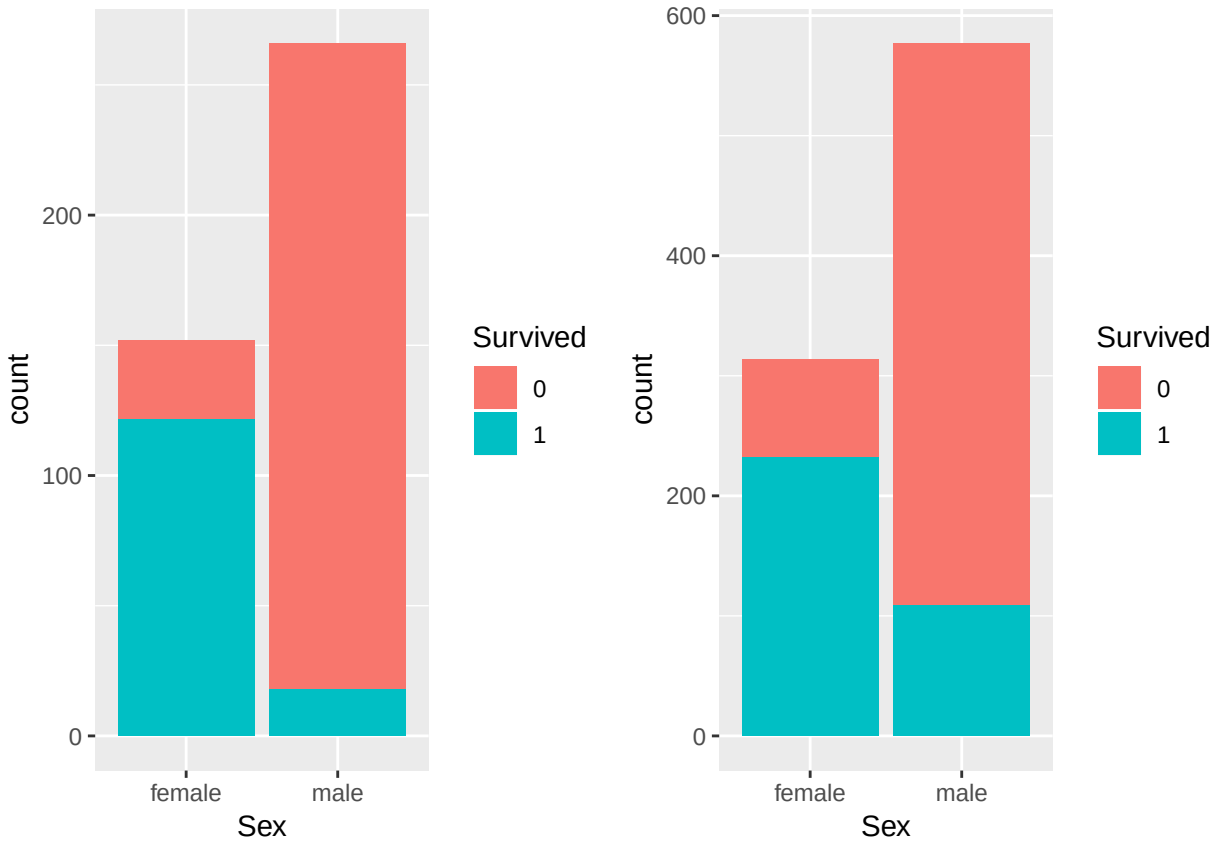
```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.2
```

```
##
## Attaching package: 'gridExtra'
```

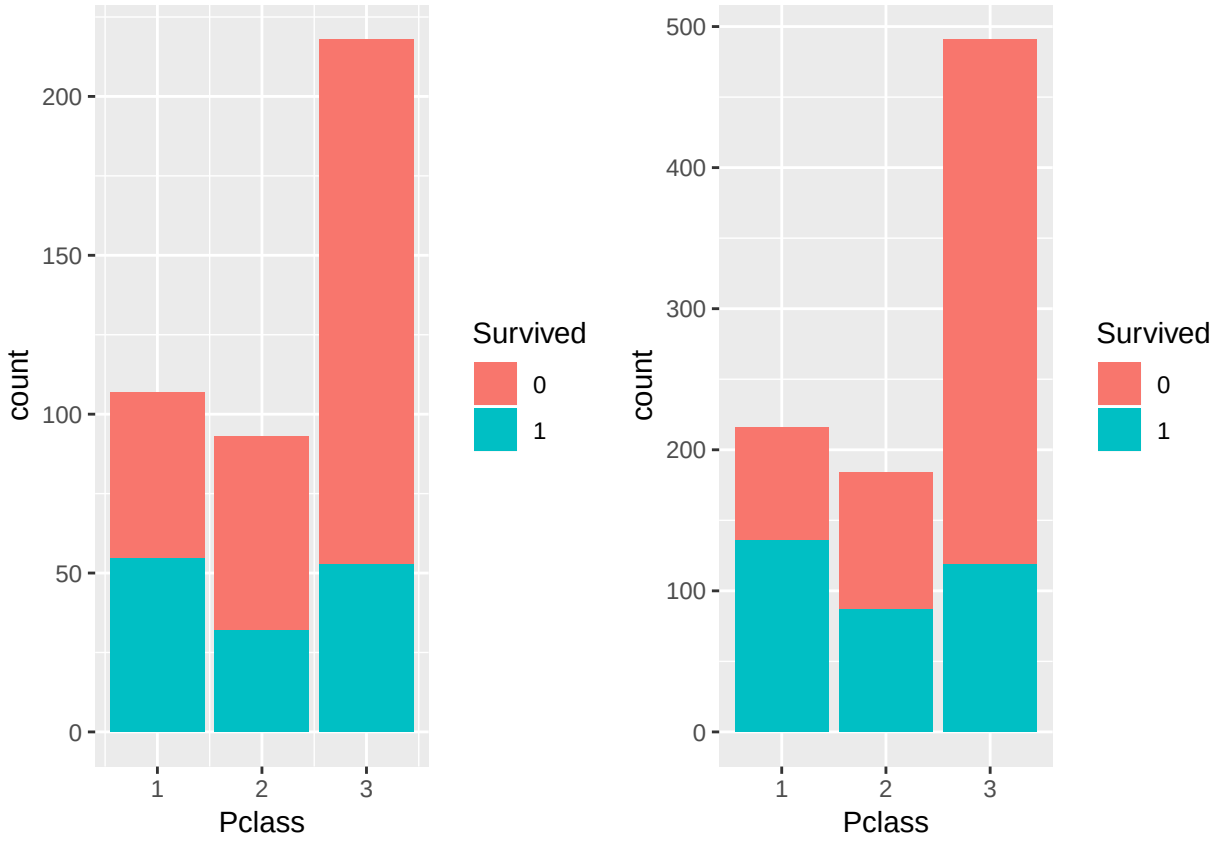
```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
g1<-ggplot(data=titanic.test.result, aes(x=Sex, fill=Survived))+geom_bar()
g2<-ggplot(data=titanic.transformed.train, aes(x=Sex, fill=Survived))+geom_bar()
grid.arrange(g1,g2,ncol=2)
```



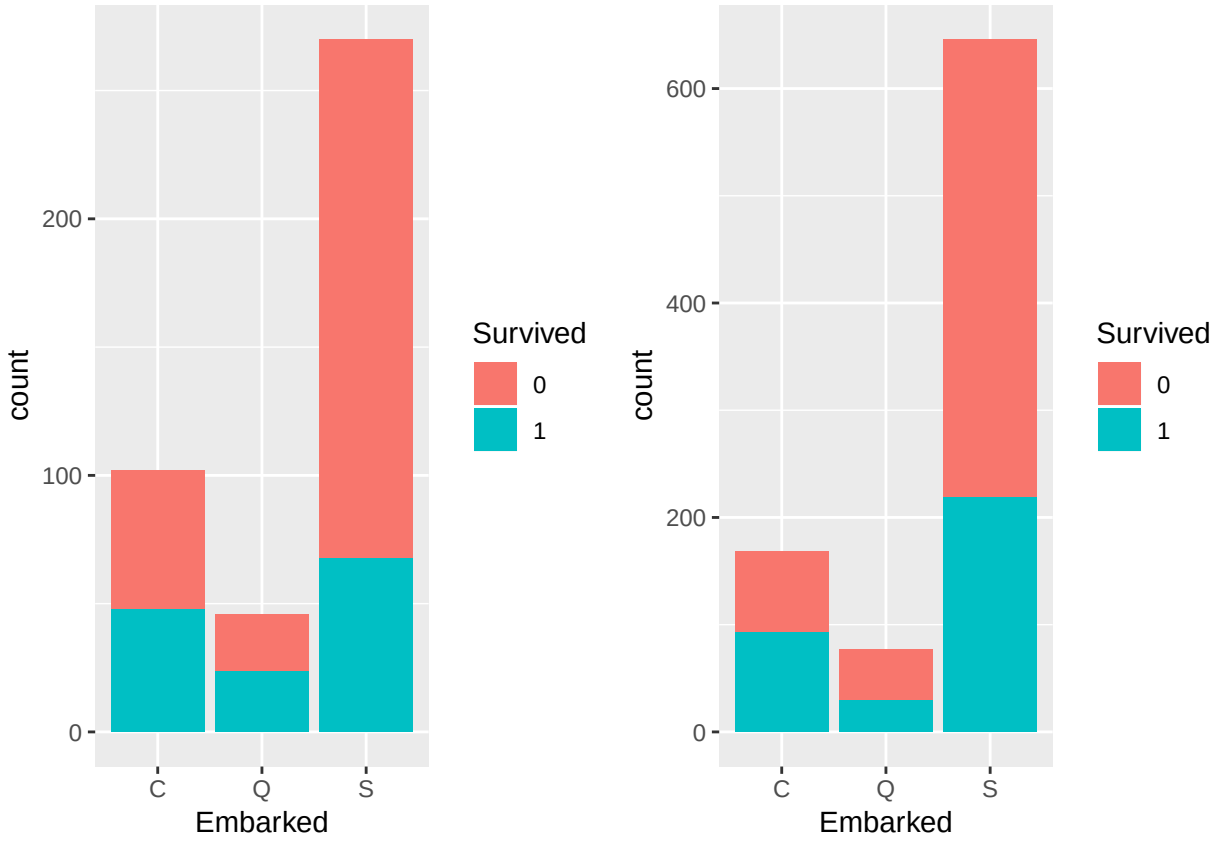
- Supervivencia respecto a clase de tarifa.

```
g1<-ggplot(data=titanic.test.result, aes(x=Pclass, fill=Survived))+geom_bar()
g2<-ggplot(data=titanic.transformed.train, aes(x=Pclass, fill=Survived))+geom_bar()
grid.arrange(g1,g2,ncol=2)
```



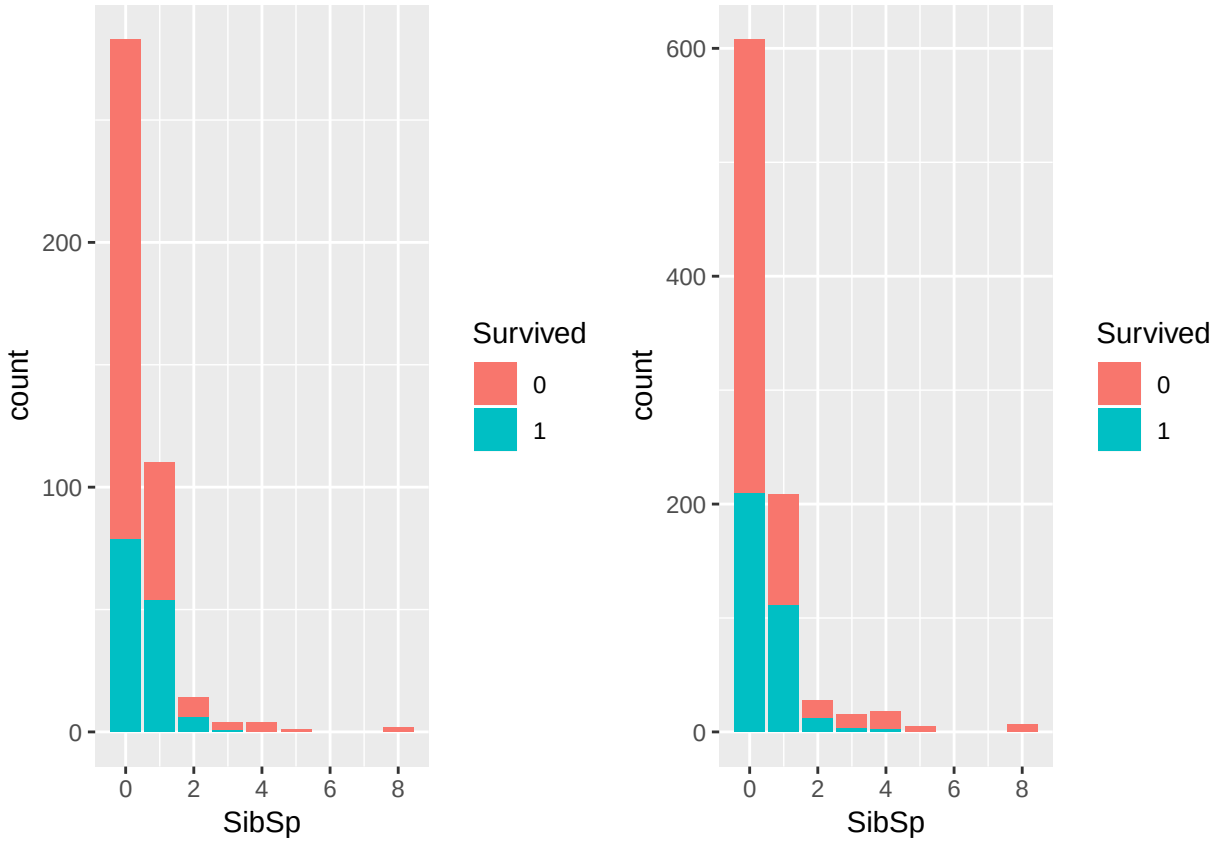
- Supervivencia respecto al origen de embarque.

```
g1<-ggplot(data=titanic.test.result, aes(x=Embarked, fill=Survived))+geom_bar()
g2<-ggplot(data=titanic.transformed.train, aes(x=Embarked, fill=Survived))+geom_bar()
grid.arrange(g1,g2,ncol=2)
```



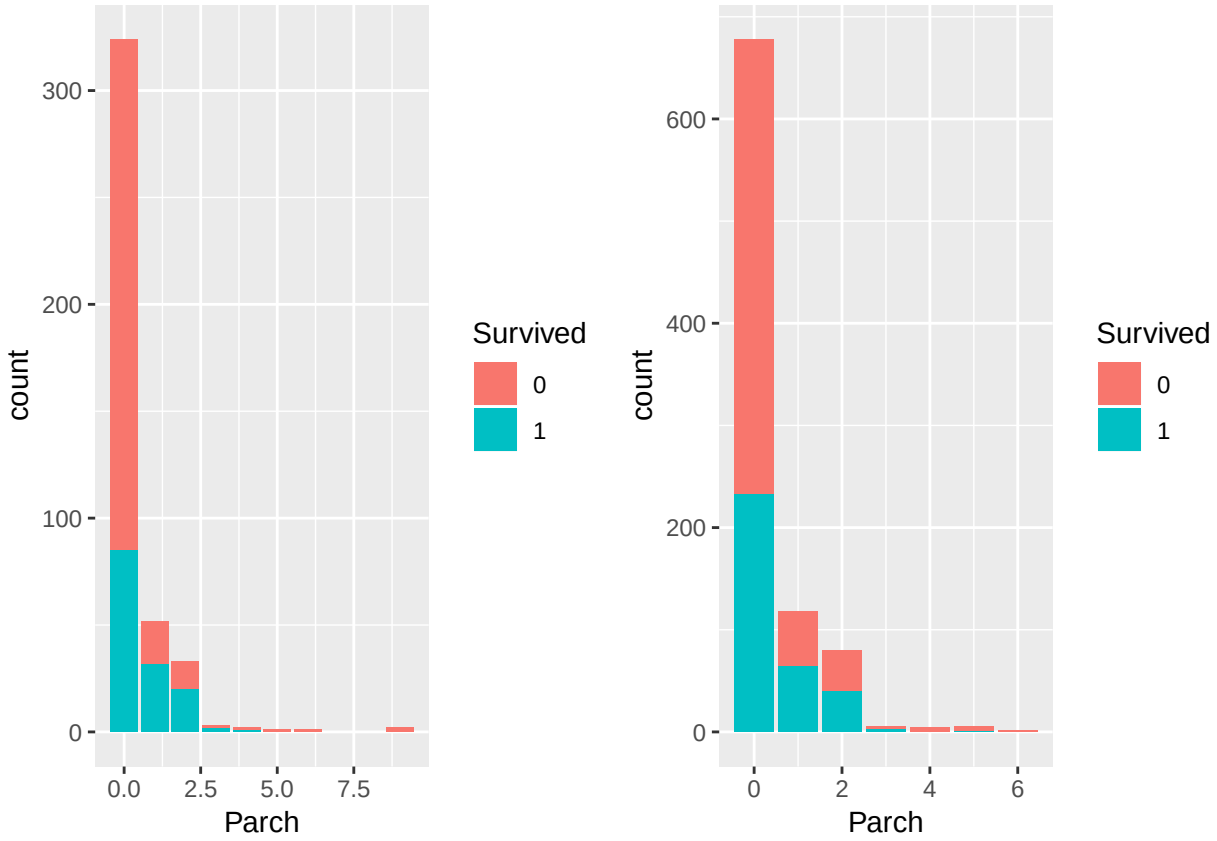
- Supervivencia respecto al número de conyuges y hermanos.

```
g1<-ggplot(data=titanic.test.result, aes(x=SibSp, fill=Survived))+geom_bar()
g2<-ggplot(data=titanic.transformed.train, aes(x=SibSp, fill=Survived))+geom_bar()
grid.arrange(g1,g2,ncol=2)
```



- Supervivencia respecto al número de ascendientes y descendientes.

```
g1<-ggplot(data=titanic.test.result, aes(x=Parch, fill=Survived))+geom_bar()
g2<-ggplot(data=titanic.transformed.train, aes(x=Parch, fill=Survived))+geom_bar()
grid.arrange(g1,g2,ncol=2)
```



Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Viendo los datos, vemos que el mito de que salvaron mas pasajeros de tarifas altas es cierto, pero se saca a la luz un nuevo dato, que es que el género es la mayor condición para sobrevivir al Titanic. Aunque nos planteamos una pregunta, ¿no será que la tercera clase está compuesta mayoritariamente por hombres, o por lo menos en mayor medida que en el resto de las clases?

```
table(titanic.full$Sex, titanic.full$Pclass)
```

```
##
##             1   2   3
##   female  144 106 216
##   male    179 171 493
```

El otro objetivo era ver si podíamos construir un modelo con estos datos. Como vimos tuvimos 48 fallos sobre 418 predicciones, lo que significa que hemos acertado con el 88,5 de los individuos del test.