M2.851 – TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS: PRA 2

Olga Garcés Ciemerozum / Carlos Acosta Quintas

Junio 2021

Contents

Introducción

El presente informe forma parte de la segunda práctica de la asignatura M2.851 - Tipología y ciclo de vida de los datos del Máster Universitario en Ciencia de Datos impartido por la Universitat Oberta de Catalunya.

En esta práctica se realizarán técnicas de limpieza de datos aplicadas a un juego de datos determinado y también se analizarán dichos datos para extraer información relevante y útil.

A su vez, se entregará, junto con la presente memoria, una serie de archivos con el código necesario para la realización de la limpieza y análisis con el que el usuario podrá realizar diferentes estudios analíticos a posteriori si lo desease.

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

Descripción del dataset

El dataset **Titanic** reune los datos sobre los pasajeros que viajaban a bordo del Titanic y registra para cada persona su supervivencia o no en el accidente. El Titanic transportaba a pasajeros con gran diversidad en sus niveles de renta y edad y a bordo se encontraban familias enteras.

La etiqueta (variable a predecir) es la variable dicotómica que indica si el viajero ha sobrevivido o no.

La ubicación en kaggle del dataset utilizado se muestra en el siguiente link:

https://www.kaggle.com/c/titanic/data

Los archivos disponibles son 3 y están en formato csv. Sus nombres son:

- train.csv
- test.csv
- gender_submission.csv: Ejemplo a seguir en la entrega de la competición Kaggle (no útil).

Según los registros, en el Titanic viajaban **2229 personas**, de las cuales 913 formaban parte de la tripulación del barco. El dataset que obtenemos de Kaggle tiene un total de **1309 registros**, por lo tanto, no todos los pasajeros que viajaban a bordo están incluidos en el dataset y podemos asumir que **el juego de datos es una muestra de toda la población a analizar**.

El dataset original está compuesto por dos ficheros: el fichero pensado para realizar el entrenamiento de un modelo (**train.csv**) y el fichero con los datos destinados a testear la calidad del modelo (**test.csv**). El fichero de entrenamiento contiene una columna más que el fichero de prueba. Esta columna corresponde a la columna de la clase "Survived".

El fichero de entrenamiento tiene **891** registros mientras que el fichero de test contiene **418** instancias.

Las variables de las que se compone el dataset son y sus unidades o magnitudes de las características son:

PassengerId:
Identificador del pasajero
Tipo: Entero indicando un identificador único de casa instancia.

Survived:
Indica si el pasajero ha sobrevivido la catástrofe
Tipo: Entero (categórica) 0 = No ha sobrevivido; 1 = Ha sobrevivido

Pclass:
Clase en la que viajaba el pasajero
Tipo: String (categórica) 1 = Primera clase; 2 = Segunda clase; 3 = Tercera Clase

Name:
Nombre del pasajero
Tipo: String

Sex:
Sexo del pasajero
Tipo: String (categórica) female = Mujer; male = hombre

Age:
Edad del pasajero
Tipo: Entero

SibSp:
Indica si el pasajero tenía hermanos o pareja a bordo
Tipo: Entero

Parch:
Indica si el pasajero tenía padres o hijos a bordo
Tipo: Entero

Ticket:
Número del billete
Tipo: String alfanumérico

Fare:
Precio del billete sin especificar si es un billete individual o grupal
Tipo: Número Real

Cabin:
Número de camarote
Tipo: String

Embarked:
Indica si el pasajero ha embarcado o no y donde
Tipo: String (categórica) C = Cherbourg; Q = Queenstown; S = Southampton

Los datos no han pasado por un proceso de preprocesado o limpieza, por lo que aún pueden existir inconsistencias y el formato no es necesariamente el más adecuado para un análisis directo.

Carga del dataset:

Cargamos el dataset y mostramos sus dimensiones, estructura y tipo de datos:

```
# Carga de los archivos que contienen los datos del train y test
test <- read.csv("dataset_titanic/test.csv")
train <- read.csv("dataset_titanic/train.csv")

train_rows <- dim(train)
test_rows <- dim(test)

train_rows
```

```
## [1] 891  12
```

```r
test_rows
```

```
## [1] 418  11
```

```r
# Estructura de los archivos train.csv y test.csv
options(width = 100)
str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "He
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```r
str(test)
```

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

Visualizamos las primeras líneas del conjunto de entrenamiento y de test.

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

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000		S
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875		Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625		S
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875		S
897	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250		S

¿Por qué es importante el dataset?

Este dataset es importante porque nos permite esclarecer qué factores pudieron influir en la supervivencia de viajeros del Titanic y obtener el conocimiento necesario para poder hacer predicciones con nuevas instancias.

Estos factores intuimos que pueden ser el estatus social, el sexo, la edad y también tener familiares cerca.

Asimismo, podemos ver si las pautas marcadas por la sociedad de "mujeres y niños primero" se cumplen cuando las personas se encuentran en situaciones de estres extremo.

De igual forma, y en el ámbito de la ciencia de datos, este dataset es importante porque es considerado un clásico y ha ayudado a muchos estudiantes a enfrentarse por primera vez a un problema de limpieza de datos, análisis estadísticos e incluso a técnicas de machine learning.

¿Qué problema pretende responder el dataset?

Este dataset pretende responder a cuáles son los diferentes factores que afectaron a la posibilidad de supervivencia de personas en el accidente del Titanic.

Integración y selección de los datos

Integración de los Datos

La integración es un proceso que forma parte de la fase de limpieza de datos y se entiende como la fusión de datos para crear una estructura única que tenga la información necesaria para el posterior análisis de datos.

Existe la integración horizontal, que básicamente se compone de la adición de nuevos atributos a partir de otras fuentes mediante sus relaciones usando claves primarias y la integración vertical, que se basaría en añadir más instancias al juego de datos (siempre manteniendo la integridad de los atributos).

En nuestro caso, tenemos dos archivos **train.csv** y **test.csv**, dónde la diferencia entre ambos es que el test no tiene las etiquetas de la variable "Survived".

Integración Vertical:

Con la finalidad de observar las distribuciones de las variables que serán base del estudio en la predicción de **"Survived"** integraremos verticalmente los dos archivos y así obtendremos un mayor número de datos **para ver sus medidas de tendencia central y dispersión.**

Para que la integración vertical sea satisfactoria, las variables y estructura de ambos archivos debe coincidir, por tanto, crearemos un dataframe **train_sin_etiqueta** que se integrará con las instancias de test.csv al cual llamaremos **df_total_sin_etiqueta**.

Observamos que la integración es satisfactoria puesto que las instancias ahora son **1309 (891 + 418)**. Generaremos a su vez un nuevo archivo csv auxiliar que mostrará este dataframe **"Titanic_global_sin_etiqueta.csv"**

```r
# Creación archivo train_sin_etiqueta.csv

etiquetas <- subset(train, select = Survived)
train_sin_etiqueta <- subset(train, select = -Survived)

# Integración archivos train.csv y test.csv
df_total_sin_etiqueta = rbind(train_sin_etiqueta, test)
dim(df_total_sin_etiqueta)
```

```
## [1] 1309   11
```

```r
# Guardamos el archivo con el nombre Titanic_global_sin_etiqueta.csv
write.csv(df_total_sin_etiqueta, "Titanic_global_sin_etiqueta.csv", row.names = FALSE)
```

Integración Horizontal:

Los archivos en la plataforma Kaggle no exponen ni fuentes externas ni csv adicionales que definan nuevas variables que se puedan integrar horizontalmente a nuestro juego de datos.

Selección de los Datos

Antes de seleccionar los datos, haremos una pequeña comprobación sobre posibles duplicidades de instancias:

Comprobación de líneas duplicadas:

Comprobamos si hay líneas duplicadas en el dataframe usando `duplicated`. No existen registros duplicados, pero sí detectamos dos pares de personas con el mismo nombre. **Para asegurarnos que se trata de personas diferentes**, buscamos los registros que tengan los nombres Connolly, Miss. Kate o Kelly, Mr. James.

Podría tratarse de la misma persona que ha comprado dos billetes, pero en estos registros vemos que las personas tienen edades diferentes y **no hay motivo para pensar que se trata de duplicados**.

```r
# Chequeo de líneas duplicadas
options(width = 100)
df_total_sin_etiqueta[duplicated(df_total_sin_etiqueta),]
```

```
##  [1] PassengerId Pclass      Name        Sex         Age         SibSp       Parch       Ticket
##  [9] Fare        Cabin       Embarked
## <0 rows> (or 0-length row.names)
```

```r
df_total_sin_etiqueta[duplicated(df_total_sin_etiqueta[c("Name","Sex")]),]
```

```
##     PassengerId Pclass                 Name    Sex  Age SibSp Parch Ticket   Fare Cabin Embarked
## 892         892      3     Kelly, Mr. James   male 34.5     0     0 330911 7.8292              Q
## 898         898      3 Connolly, Miss. Kate female 30.0     0     0 330972 7.6292              Q
```

```r
df_total_sin_etiqueta[df_total_sin_etiqueta$Name=="Kelly, Mr. James" |
                        df_total_sin_etiqueta$Name == "Connolly, Miss. Kate",]
```

```
##     PassengerId Pclass                 Name    Sex  Age SibSp Parch Ticket   Fare Cabin Embarked
## 290         290      3 Connolly, Miss. Kate female 22.0     0     0 370373 7.7500              Q
## 697         697      3     Kelly, Mr. James   male 44.0     0     0 363592 8.0500              S
## 892         892      3     Kelly, Mr. James   male 34.5     0     0 330911 7.8292              Q
```

```
## 898         898      3 Connolly, Miss. Kate female 30.0     0      0 330972 7.6292             Q
```

La selección se puede entender como un primer filtro de los datos, no solamente a través de poner límites a los valores de algunas instancias o elegir algún valor cualitativo específico, sino también a través de la inspección de las correlaciones entre los atributos y la posterior eliminación del dataset de aquellos que sean redundantes.

Debido a que el problema planteado es interpretar qué factores influyen en la supervivencia, a priori, no sabríamos si debemos descartar alguna variable o no (eliminación de la variable del estudio) o si deberíamos filtrar los datos, ya sean numérica o categóricamente.

No obstante, en esta sección eliminaremos la variable **"Name"** porque no es de mucha utilidad para nuestros análisis ya que el nombre no debería influir a priori en la supervivencia de los viajeros y también la variable **"PassengerId"**puesto que simplemente es un identificador.

Por lo tanto, además de esta primera selección relizada, **esta fase del proceso la dejaremos abierta en este punto y retomaremos una vez la exploración y análisis nos vaya indicando qué debemos seleccionar y/o filtrar**. A continuación, se hace una lista de las selecciones realizadas en este apartado y a posteriori.

```r
# Eliminamos variables Name
keep.cols <- c("Pclass", "Sex", "Age", "SibSp", "Parch", "Ticket", "Fare", "Cabin",
               "Embarked")
df_total_sin_etiqueta <- df_total_sin_etiqueta[keep.cols]
```

Variable Modificada	Tipo de selección	Apartado realizado	Motivo
Name	Eliminación	2.2	Variable no útil al ser independiente al estudio
PassengerId	Eliminación	2.2	Variable no útil al ser un simple identificador
Ticket	Eliminación	2.3	Usada para crear nueva variable y ya no es útil
Fase	Eliminación	2.3	Usada para crear nueva variable y ya no es útil
Cabin	Eliminación	3.1	Existencia masiva de valores nulos

Creación de nuevas variables

Se ha detectado que **hay números de billetes duplicados**. Esto indica que hay dos tipos de tickets:

- **Individuales**
- **Grupales**

Se observa que la variable **"Fare"** muestra el mismo precio para los tickets grupales, por tanto, para saber realmente el precio del ticket por viajero y también para poder usar correctamente la variable "Fare", **deberíamos saber de cuántas personas es el ticket grupal y después dividir la variable "Fare" for dicha cantidad**.

Crearemos una columna con el recuento de billetes (**Count.ticket**) con el mismo id para cada pasajero y otra con el precio unitario (**Unit.price**).

```r
# Creación variable con el contaje de los tickets con mismo nombre
df_total_sin_etiqueta$Count.ticket <- (df_total_sin_etiqueta%>%
```

```
                                          group_by(Ticket)%>%
                                          mutate(count=n()))$count
```

```
# Creación variable con el precio unitario de los tickets con mismo nombre
df_total_sin_etiqueta$Unit.price <- df_total_sin_etiqueta$Fare / df_total_sin_etiqueta$Count.ticket
```

Selección de datos inicial a posteriori

```
# Eliminamos variable Name
keep.cols <- c("Pclass", "Sex", "Age", "SibSp", "Parch", "Cabin",
               "Embarked", "Count.ticket", "Unit.price")

df_total_sin_etiqueta <- df_total_sin_etiqueta[keep.cols]
```

Ahora disponemos de una variable consistente con el precio del billete y que se puede aplicar a cada viajero, **pues la tarifa grupal se ha convertido en individual.**

Limpieza de datos

NOTA:

Hay que mencionar que se la limpieza de datos en este proyecto en particular **debe afectar tanto al archivo train.csv como al test.csv**, por tanto, limpiaremos los datos en base al dataframe global creado anteriormente (**df__total__sin__etiqueta**).

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

Elementos vacíos en el dataset

Comprobaremos si existen valores nulos o inexistentes en el juego de datos.

```
# Exploración de los datos - Type = 1

kbl(ExpData(data=df_total_sin_etiqueta,type=1),booktabs =T)%>%
  kable_styling(latex_options =c("striped", "hold_position"))
```

```
# Estructura de los datos - Type = 2

kbl(ExpData(data=df_total_sin_etiqueta,type=2),booktabs =T)%>%
  kable_styling(latex_options =c("striped","scale_down", "hold_position"))
```

Descriptions	Value
Sample size (nrow)	1309
No. of variables (ncol)	9
No. of numeric/interger variables	6
No. of factor variables	0
No. of text variables	3
No. of logical variables	0
No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	55.56% (5)
%. of variables having >0% and <50% missing cases	33.33% (3)
%. of variables having >=50% and <90% missing cases	11.11% (1)
%. of variables having >=90% missing cases	0% (0)

Index	Variable_Name	Variable_Type	Sample_n	Missing_Count	Per_of_Missing	No_of_distinct_values
1	Pclass	integer	1309	0	0.000	3
2	Sex	character	1309	0	0.000	2
3	Age	numeric	1046	263	0.201	98
4	SibSp	integer	1309	0	0.000	7
5	Parch	integer	1309	0	0.000	8
6	Cabin	character	295	1014	0.775	187
7	Embarked	character	1307	2	0.002	4
8	Count.ticket	integer	1309	0	0.000	9
9	Unit.price	numeric	1308	1	0.001	260

Una vez que sabemos que **tenemos valores nulos**, cuántos tenemos y sabemos las variables afectadas, se decide la estrategia para imputar dichos valores.

Variable Cabin:

Observamos que la variable "Cabin" tiene 1014 valores nulos de 1309, por tanto, **se decide eliminar dicha variable** por la imposibilidad de realizar una imputación generalizada.

```
# Eliminamos variable Cabin
keep.cols <- c("Pclass", "Sex", "Age", "SibSp", "Parch", "Embarked", "Count.ticket",
               "Unit.price")
df_total_sin_etiqueta <- df_total_sin_etiqueta[keep.cols]
```

Variable Age:

El número de registros de Age que son NA representan aproximadamente el **20% de los registros totales**.

Este dataset contiene variables categóricas y numéricas y para imputar los valores nulos de la variable Age podemos usar **el método `kNN`**. Aplicamos la función e imputamos los valores NA usando todos los demás campos del dataset y con un valor de k igual a 3. El algoritmo busca los registros de los 3 pasajeros más parecidos (cercanos según la distancia Gower) al que contiene un valor nulo y usa los datos de edades de estos pasajeros para imputar el valor faltante.

Una vez ejecutado el algoritmo para imputar los valores, volvemos a comprobar si existen valores NA y **podemos confirmar que todos los NA para la variable edad han sido imputados**.

```r
# Imputación de valores a los valores nulos de la variable age

options(width = 100)
df_total_sin_etiqueta <- kNN(df_total_sin_etiqueta, k=3)[1:10]

head(df_total_sin_etiqueta[is.na(df_total_sin_etiqueta$Age),])
```

```
##  [1] Pclass       Sex          Age          SibSp        Parch        Embarked     Count.ticket
##  [8] Unit.price   Pclass_imp   Sex_imp
## <0 rows> (or 0-length row.names)
```

```r
df_total_sin_etiqueta <- df_total_sin_etiqueta[keep.cols]
```

```r
# Comprobacion de la no existencia de registros nulos para la variable age después de la imputación.
sum(is.na(df_total_sin_etiqueta$Age))
```

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

Variable Embarked:

Se observa que la mayoría de las instancias pertenecen a la categoría S, por tanto, las instancias con valores nulos en esta variable, las imputaremos a S.

```r
# Exploración del resumen de los datos de la varibale Embarked
df_total_sin_etiqueta$Embarked <- as.factor(df_total_sin_etiqueta$Embarked)
summary(df_total_sin_etiqueta$Embarked)
```

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

```r
# Imputación clase mayoritaria a variable Embarked
options(width = 100)
df_total_sin_etiqueta$Embarked[df_total_sin_etiqueta$Embarked == ""] <- "S"
summary(df_total_sin_etiqueta$Embarked)
```

```
##       C   Q   S
##   0 270 123 916
```

Variable Unit.price:

Actuaremos de igual forma que con la variable Age e imputaremos a través del uso del kNN

```r
# Imputación de valores a los valores nulos de la variable age
df_total_sin_etiqueta <- kNN(df_total_sin_etiqueta, k=3)[1:10]
head(df_total_sin_etiqueta[is.na(df_total_sin_etiqueta$Unit.price),])
```

```
##  [1] Pclass       Sex          Age          SibSp        Parch        Embarked     Count.ticket
##  [8] Unit.price   Pclass_imp   Sex_imp
## <0 rows> (or 0-length row.names)
```

```
df_total_sin_etiqueta <- df_total_sin_etiqueta[keep.cols]
```

Gestión de los valores iguales a "cero" en el dataset:

Ahora comprobamos las variables que toman valores igual a cero sin que tenga sentido que tomen este tipo de valor.

La variable que representa las "etiquetas" (variable **Survived**) toma valores iguales a cero y consideramos que es correcto puesto que es parte de la variable dicotómica del dataset. Lo mismo ocurre con las variables **SibSp**, **Parch**, donde consideramos normal que existan valores iguales a cero, **significa que los pasajeros no tenían familia a bordo o viajaban solos**.

En cambio, los valores iguales a cero para la variable **Unit.price** son algo más extraños. Entre los pasajeros que tienen un Unit.price igual a cero hay personas que viajaban en primera, segunda y tercera clase.

La idea que un ticket sea gratuito, en principio, no sería posible, por tanto, volveremos a aplicar el método kNN para imputar estos valores.

Primero cambiaremos el valor de cero a NA y después actuaremos como en el apartado anterior.

```
df_total_sin_etiqueta$Unit.price[df_total_sin_etiqueta$Unit.price == "0"] <- NA

# Imputación de valores a los valores ceros de la variable Unit.price
df_total_sin_etiqueta <- kNN(df_total_sin_etiqueta, k=3)[1:10]

head(df_total_sin_etiqueta[is.na(df_total_sin_etiqueta$Unit.price),])
```

```
##  [1] Pclass       Sex          Age          SibSp        Parch        Embarked     Count.ticket
##  [8] Unit.price   Pclass_imp   Sex_imp
## <0 rows> (or 0-length row.names)
```

```
df_total_sin_etiqueta <- df_total_sin_etiqueta[keep.cols]

# Comprobacion de la no existencia de registros ceros para
# la variable Unit.price después de la imputación.
sum(is.na(df_total_sin_etiqueta$Unit.price))
```

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

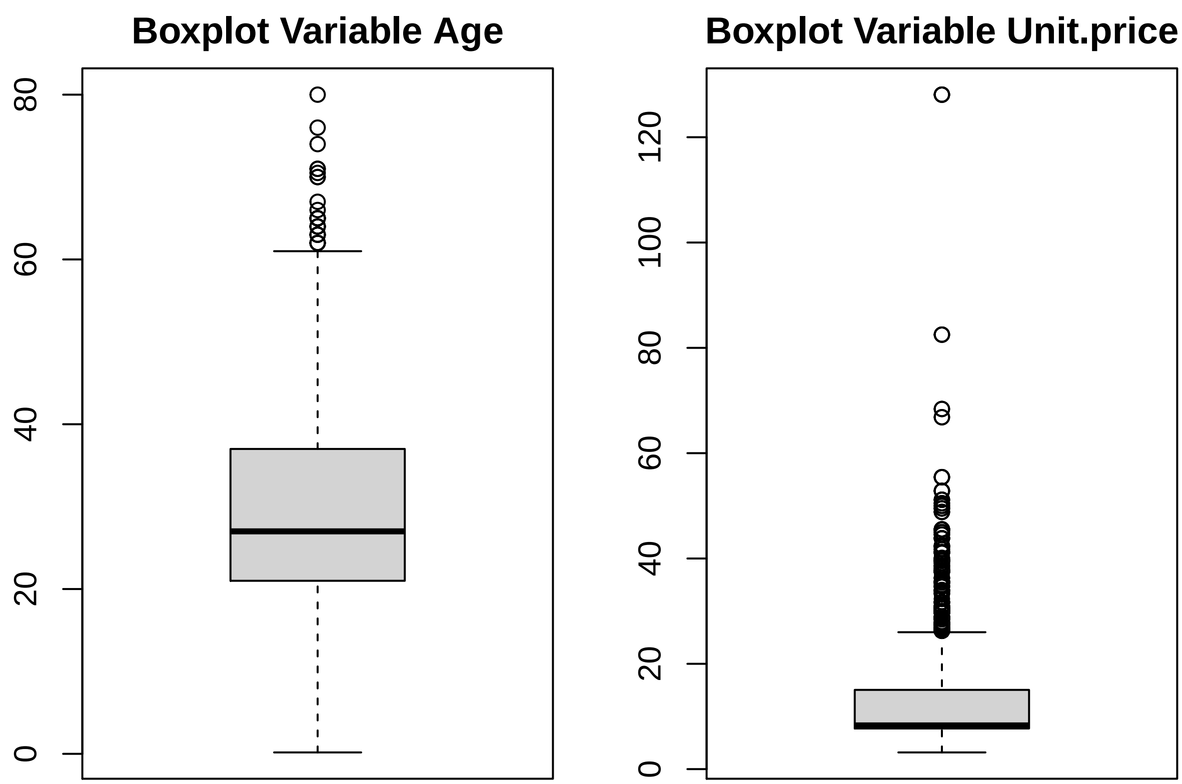
Valores extremos

Las comprobaciones de valores extremos las hemos hecho anteriormente con `sapply(df, summary)`.

Visualizamos boxplots para las variables numéricas que tenemos: **Age** y **Unit.price**.

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

# Boxplot variable Age
boxplot(df_total_sin_etiqueta$Age,
        main="Boxplot Variable Age",
        ylab="Edad (años)")

# Boxplot variable Unit.price
boxplot(df_total_sin_etiqueta$Unit.price,
```

```r
        main="Boxplot Variable Unit.price",
        ylab="Precio billete unitario")
```



Boxplot Stats Variable Age

```r
#Boxplot Stats Variable Age
boxplot.stats(df_total_sin_etiqueta$Age, coef = 1.5, do.conf = TRUE, do.out = TRUE)
```

```
## $stats
## [1]  0.17 21.00 27.00 37.00 61.00
##
## $n
## [1] 1309
##
## $conf
## [1] 26.30127 27.69873
##
## $out
##  [1] 66.0 65.0 71.0 65.0 70.5 62.0 63.0 65.0 64.0 65.0 63.0 71.0 64.0 62.0 62.0 80.0 65.0 70.0 70.0
## [20] 62.0 74.0 62.0 63.0 67.0 65.0 76.0 63.0 64.0 64.0 64.0
```

Boxplot Stats Variable Unit.price

```r
#Boxplot Stats Variable Fare
options(width = 100)
boxplot.stats(df_total_sin_etiqueta$Unit.price, coef = 1.5, do.conf = TRUE, do.out = TRUE)
```

```
## $stats
```

```
## [1]  3.1708  7.7208  8.3000 15.0500 26.0000
##
## $n
## [1] 1309
##
## $conf
## [1] 7.979931 8.620069
##
## $out
##    [1]  35.64165  26.55000  26.55000  35.50000  43.83333  27.72080  48.84027  41.08540  30.98960
##   [10]  35.50000  40.00000  41.73750  27.72080  43.83333  30.58750  34.65420  31.67915  38.64375
##   [19]  82.50693  38.64375  26.28330  26.55000  39.60000  33.30000  30.68960  27.50000  33.50000
##   [28]  30.69580  28.71250  50.00000  26.55000  27.72080  48.84027  31.00000  37.75833  38.14585
##   [37]  45.00000  41.73750  30.00000  26.27710  26.55000  39.60000  28.83333 128.08230  26.55000
##   [46]  51.15417  33.90832  29.70000  26.28333  45.53960  27.72080  30.50000  82.50693  27.72083
##   [55]  36.30000  28.46460  27.71943  37.48214  41.21668  26.90000  33.90832  28.98960  28.50000
##   [64]  51.15417  66.82500  33.30000  26.90000  35.50000  43.83333  35.00000  27.50000  37.62500
##   [73]  34.65000  27.72085  33.90832  41.08540  42.30000  45.50500  30.00000  37.75833  30.00000
##   [82]  26.55000  27.95000  30.00000  43.83333  27.28610  26.55000  30.50000  27.75000  44.55210
##   [91]  26.55000  26.55000  38.50000  26.55000  45.53960  45.00000  29.70000  30.50000  49.50420
## [100]  39.13335  28.83333  36.30000  26.55000  26.28750  29.70000  34.02080  28.98960  55.44480
## [109]  26.55000  35.47500  49.50000  35.50000  35.47500  27.72083  26.55000  39.60000  45.50500
## [118]  26.55000  26.38750  27.95000  27.72083  40.12500  26.55000  39.60000  39.13335  28.46460
## [127]  26.55000  30.50000  51.15417  26.27710  32.32080  30.00000  30.50000  34.65000  35.50000
## [136]  37.75833  66.82500 128.08230  52.83438  28.50000  26.55000  27.72083  45.50500  26.28750
## [145]  26.28750  49.50420  26.55000  45.50500  26.55000  52.83438 128.08230  30.00000  26.28333
## [154]  37.48214  35.50000  26.55000  28.83333  30.00000  39.60000  52.83438  28.50000  30.00000
## [163]  39.60000  30.69580  30.00000  31.68330  26.55000  26.55000  31.68330  40.00000  27.71943
## [172]  29.70000  44.55210  41.21668  26.55000  50.49580  26.27710  27.71943  30.00000  30.00000
## [181]  41.13335  30.58750  29.70000  31.68330  30.68960  37.48214  30.98960  30.50000  28.87500
## [190]  26.55000  26.27710  29.70000  38.14585  30.00000  43.83333  37.48214  37.48214  28.53750
## [199]  43.83333  27.72080  42.30000  42.30000  55.44480  26.28333  27.72085  31.67920  55.44480
## [208]  37.62085  28.87500  28.50000  37.48214  26.55000  26.55000  27.71943  55.44480  26.55000
## [217]  50.49580  27.72080  27.72085  27.71943  27.71943  26.55000  82.50693  26.90000  45.50500
## [226]  27.72080  42.50000  41.21668  42.30000  27.44580  26.55000  35.64165  37.62500  35.47500
## [235]  27.72080  26.90000  68.38960  37.62085  68.38960  41.13335  39.60000  27.28610  45.50000
## [244]  26.55000  33.90832  48.84027  26.55000  52.83438  39.60000  29.70000 128.08230  31.67915
## [253]  27.72085  29.70000  26.90000  27.28610  37.48214  50.00000  30.00000  39.60000  41.21668
## [262]  29.70000  27.72080  42.30000  30.00000  36.30000
```

El boxplot muestra que hay outliers en estas dos variables. La mayoría de los pasajeros eran jóvenes, aunque también encontramos pasajeros de más de 65 años. En cuanto a la variable **Unit.price**, se comprueba que a medida que el precio sube, la clase va bajando de 3 a 2 y de 2 a 1, con lo cual no hay razón porqué pensar que los precios no son reales. En la tabla siguiente podemos visualizar algunos de los pasajeros que pagaron un precio de billete alto. Notamos que todos son de primera clase.

```r
# Muestra de varios precios de billetes unitarios en el rango alto
options(width = 100)
kbl(tail(df_total_sin_etiqueta[df_total_sin_etiqueta$Unit.price>30,], 15),booktabs =T)%>%
  kable_styling(latex_options =c("striped","scale_down", "hold_position"))
```

Aunque se acepte que los precios son reales, es cierto que hay uno que es extremadamente alto y, aunque cierto, podría desvirtuar posibles futuras predicciones, por tanto se estima que se podría cambiar por **la media de Unit.price agrupado por la primera clase**, para tener un valor imputado **más real**.

	Pclass	Sex	Age	SibSp	Parch	Embarked	Count.ticket	Unit.price
1179	1	male	24	1	0	S	2	41.13335
1182	1	male	44	0	0	S	1	39.60000
1190	1	male	30	0	0	S	1	45.50000
1206	1	female	55	0	0	C	4	33.90832
1208	1	male	57	1	0	C	3	48.84027
1216	1	female	39	0	0	S	4	52.83438
1219	1	male	46	0	0	C	2	39.60000
1235	1	female	58	0	1	C	4	128.08230
1242	1	female	45	0	1	C	2	31.67915
1267	1	female	45	0	0	C	7	37.48214
1270	1	male	55	0	0	S	1	50.00000
1289	1	female	48	1	1	C	2	39.60000
1292	1	female	30	0	0	S	4	41.21668
1299	1	male	50	1	1	C	5	42.30000
1306	1	female	39	0	0	C	3	36.30000

```
# Imputación del valor máximo de Unit.price
df_price_Pclass <- aggregate(df_total_sin_etiqueta$Unit.price ~ df_total_sin_etiqueta$Pclass,
                             df_total_sin_etiqueta, mean)
mean_Unit.price <- df_price_Pclass[2][[1]][1]
```

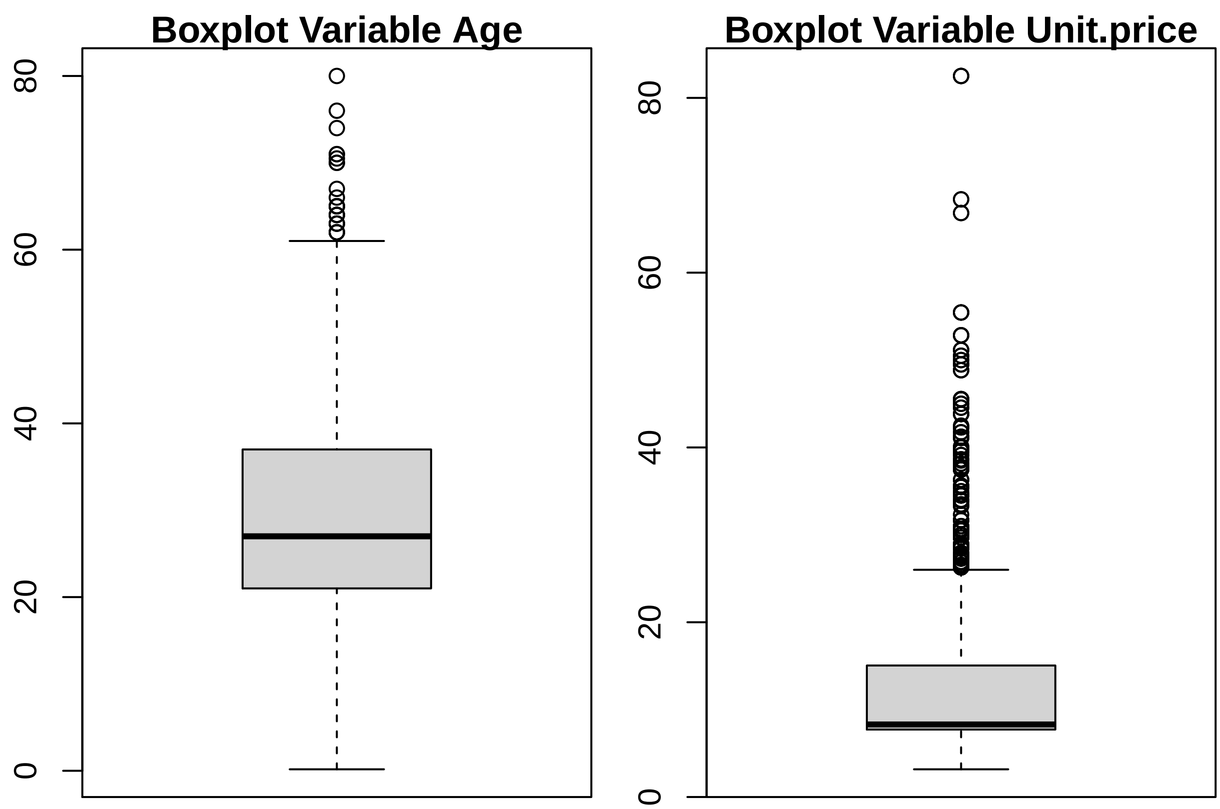
```
# Muestra de varios precios de billetes unitarios en el rango alto
max_Unit.price <- max(df_total_sin_etiqueta$Unit.price)
max_Unit.price
```

```
## [1] 128.0823
```

```
df_total_sin_etiqueta$Unit.price[df_total_sin_etiqueta$Unit.price == max_Unit.price] <- mean_Unit.price
mean_Unit.price
```

```
## [1] 34.50998
```

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

# Boxplot variable Age
boxplot(df_total_sin_etiqueta$Age,
        main="Boxplot Variable Age",
        ylab="Edad (años)")




# Boxplot variable Unit.price
boxplot(df_total_sin_etiqueta$Unit.price,
        main="Boxplot Variable Unit.price",
        ylab="Precio billete unitario")
```



Observamos que hemos eliminado el valor extremo (el máximo) de Unit.price.

Análisis de datos

Una vez limpiado el único archivo que contenía las líneas de los conjuntos train y test, **deberemos separar otra vez los conjuntos ya que únicamente tenemos datos de la etiqueta para el conjunto de entrenamiento**.

```
# Guardamos en disco el archivo global
write.csv(df_total_sin_etiqueta, "Titanic_global_sin_etiqueta.csv", row.names = FALSE)
```

```
# Creación del conjunto de train y test después de la limpieza del dataset
train <- df_total_sin_etiqueta[1:train_rows, ]
test <- df_total_sin_etiqueta[(train_rows + 1):(train_rows+test_rows), ]
```

Y añadimos las etiquetas al conjunto de train:

```r
# Adición de las etiquetas al conjunto de entrenamiento
train <- cbind(train, etiquetas )
```

Una vez llegados a este punto, **guardamos en disco los archivos train y test procesados y "limpios" preparados para su posterior análisis**.

```r
# Guardamos en disco los archivos procesados
write.csv(train, "train_processed.csv", row.names = FALSE)
write.csv(test, "test_processed.csv", row.names = FALSE)
```

Screening

Antes de crear las visualizaciones determinamos que las variables numéricas que son potencialmente importantes para el análisis son: **Age** y **Unit.price** que corresponden a la edad de los pasajeros y el precio unitario del billete.

Survived, Sex, Embarked son variables categóricas y Pclass, SibSp, Parch son variables categóricas ordinales (existen rangos en los valores de las variables).

Realizamos las transformaciones oportunas para guardar las variables con sus tipos correspondientes.

```r
# Categóricas
train$Survived <- as.factor(train$Survived)
train$Sex <- as.factor(train$Sex)
```

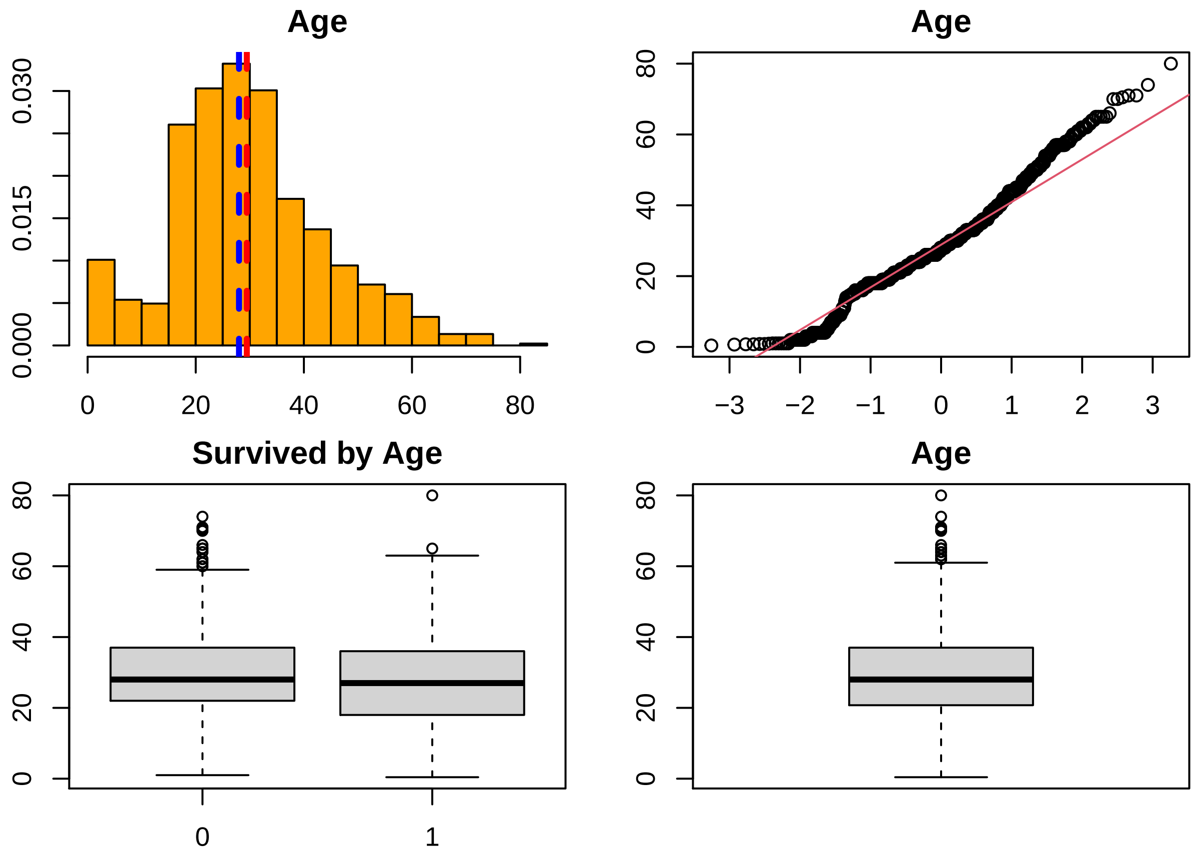
Creamos una función que nos facilite visualizar las distribuciones de las variables. Se mostrará su histograma con su media (en rojo) y su mediana (en azul), su gráfico Q-Q y su boxplot filtrado for la variable etiqueta (Survived).

```r
visualiz <- function(data, colm, facet.colm, title, facetPlot=FALSE){
oldpar = par(mfrow = c(2,2), mar=c(2,2,2,2))
truehist(data[[colm]], main = title, col = "orange")
abline(v = mean(data[[colm]]), col="red", lwd=3, lty=2);
abline(v = median(data[[colm]]), lwd=3, lty=2, col="blue");
qqnorm(data[[colm]], main = title);qqline(data[[colm]], col = 2 )
if (facetPlot==TRUE) {
  boxplot(data[[colm]]~data[[facet.colm]], main = paste("Survived by", title))
}

boxplot(data[[colm]], main = title)
}
```

Visualización de las variables numéricas Age y Unit.price

Los pasajeros que tienen edades entre los cuantiles 25 y 75 tienen entre 20 y 40 años. La edad mediana para los pasajeros que sobrevivieron y los que no es muy similar.

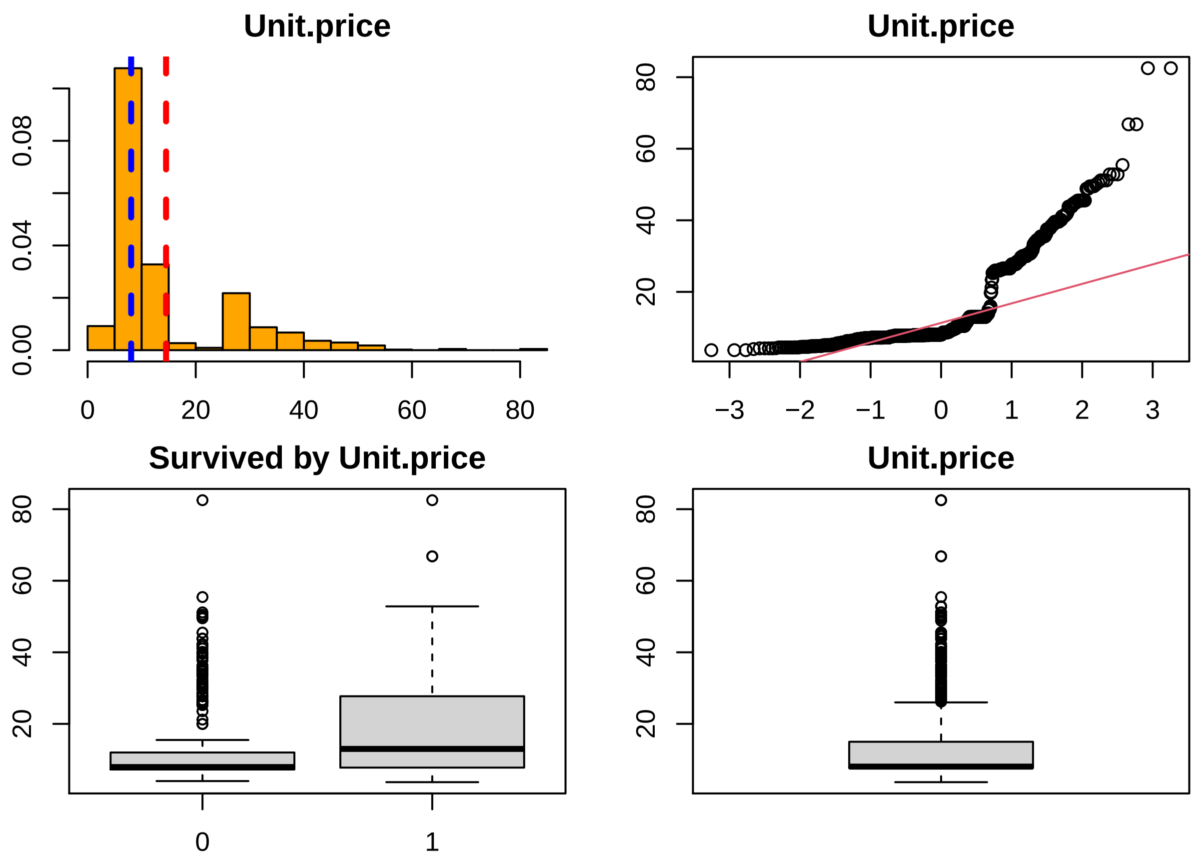
```r
visualiz(train, "Age", "Survived", "Age", facetPlot=TRUE)
```



Si observamos la variable **Unit.price**(precio unitario del billete), la mayoría de los pasajeros pagaron muy poco por el billete. Aquí en el boxplot encontramos outliers pero, como se ha comentado en el apartado anterior, consideramos solamente eliminar el máximo y no descartar el resto de estas entradas ya que hay pocos pasajeros que pagaron un importe alto por el billete y esta información puede ser muy interesante para predecir la posibilidad de supervivencia: ¿Viajar en una clase privilegiada aumenta la posibilidad de sobrevivir?

En el boxplot además podemos ver que las personas que sobreviven tienen una media más alta en el precio del billete.

```
visualiz(train, "Unit.price", "Survived", "Unit.price", facetPlot=TRUE)
```



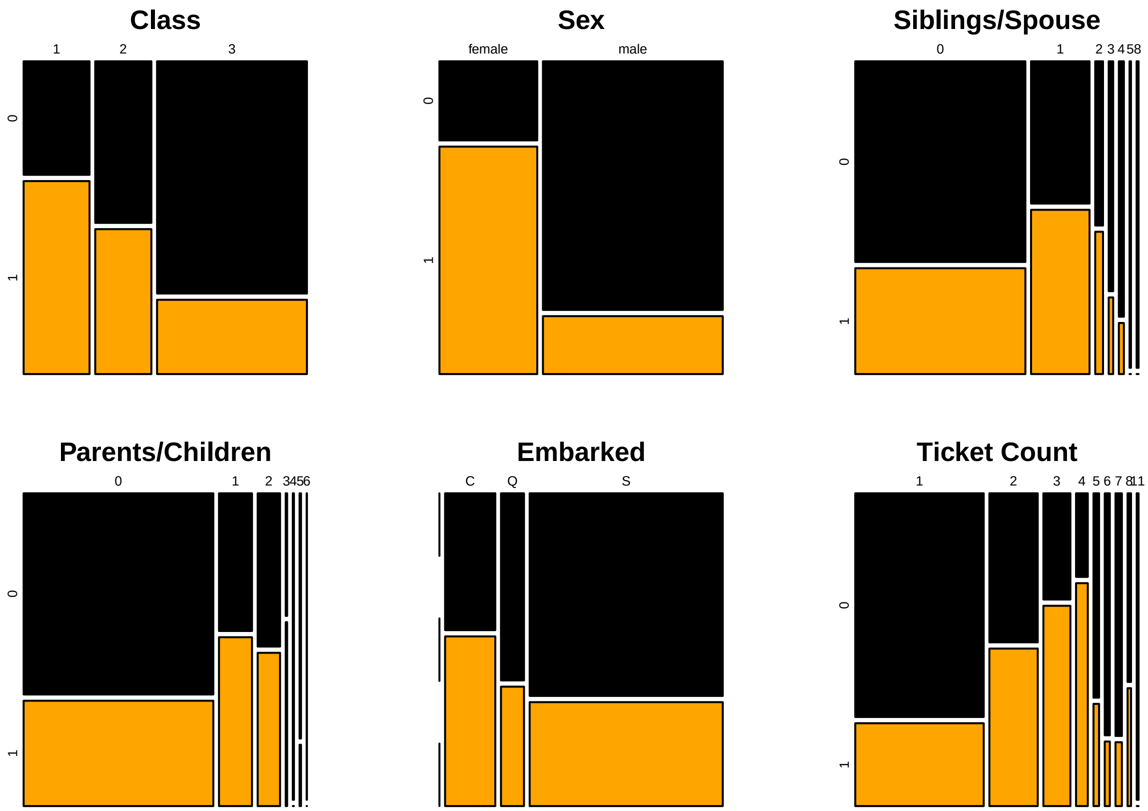
Visualización de las variables Class, Sex, Siblings/Spouse, Parents/Children

A continuación visualizamos los datos para las variables Class, Sex, Siblings/Spouse, Parents/Children:

Color Naranja: Indica que el pasajero ha sobrevivido a la catástrofe
Color Negro: Indica que el pasajero **NO** ha sobrevivido a la catástrofe

```
oldpar = par(mfrow = c(2,3), mar=c(2,2,2,2))

plot(table(train$Pclass, train$Survived),
     col = c("black", "orange"), main="Class")

plot(table(train$Sex, train$Survived),
     col = c("black", "orange"), main="Sex")

plot(table(train$SibSp, train$Survived),
     col = c("black", "orange"), main="Siblings/Spouse")

plot(table(train$Parch, train$Survived),
     col = c("black", "orange"), main="Parents/Children")

plot(table(train$Embarked, train$Survived),
     col = c("black", "orange"), main="Embarked")
```

```r
plot(table(train$Count.ticket, train$Survived),
     col = c("black", "orange"), main="Ticket Count")
```



Las personas que viajaban en tercera clase tienen la menor proporción de supervivencia comparando con las personas que viajaban en primera clase. Las mujeres que viajaban en el titanic sobrevivieron en su mayoría. Los hombres sobrevivieron en mucho menor proporción.

La mayoría de personas viajaba sin familiares (hermanos, pareja, padres o hijos) y parece ser que el porcentaje de supervivencia es algo más alto en personas que tenían familiares a bordo.

La mayoría de personas embarcaron en el punto S, pero el mayor porcentaje de supervivencia lo tienen las personas que embarcaron en C.

Los pasajeros que viajaban varias personas con el mismo billete también tienen una mayor supervivencia (hasta 4 pasajeros). 1 o más de 4 pasajeros con el mismo billete tienen la misma proporción de supervivencia.

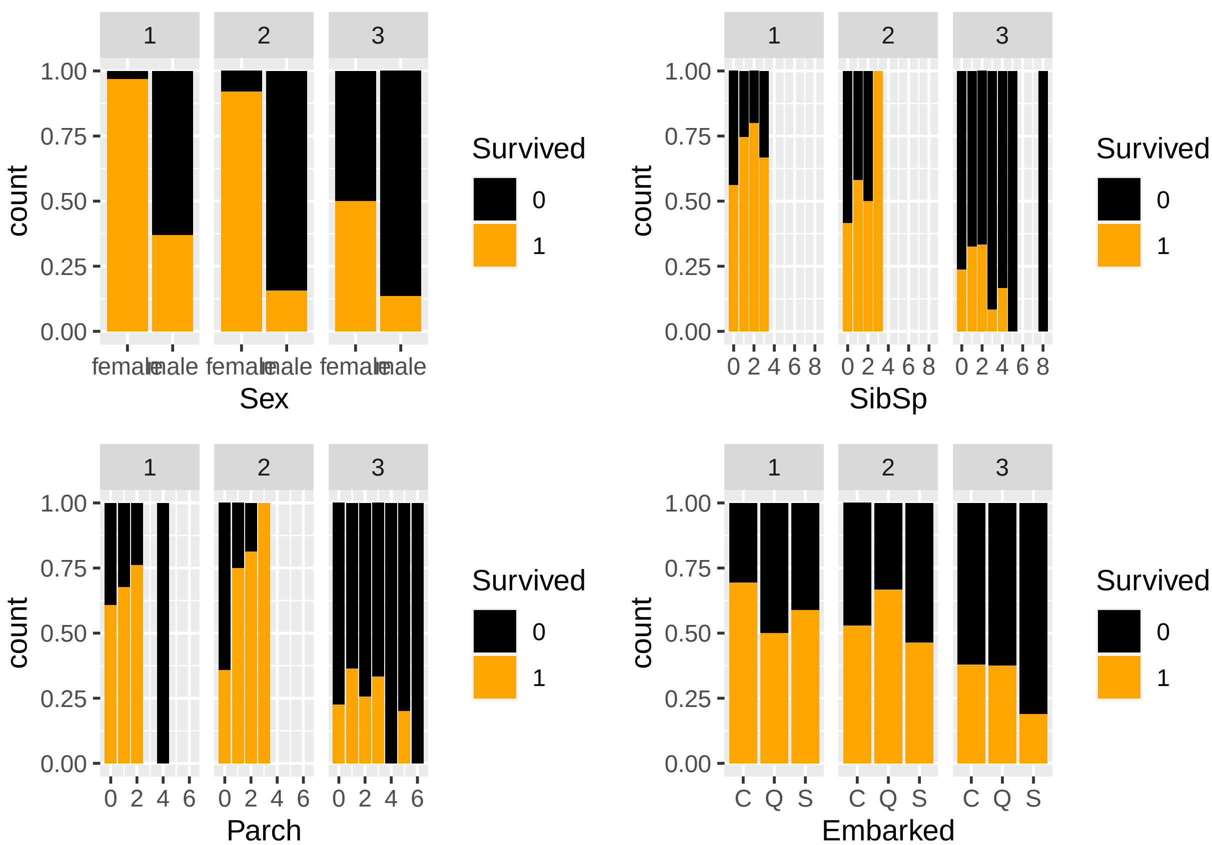
Análisis bivariable

Vamos a visualizar algunos plots que nos permiten ver la supervivencia de pasajeros combinando dos variables.

En función de la variable Clase

```r
grid.newpage()

sex.class <- ggplot(data=train, aes(x=Sex, fill=Survived))+
  geom_bar(position = 'fill')+
```

```
  facet_wrap(~Pclass)+
  scale_fill_manual(values=c("black", "orange"))

sibs.class <- ggplot(data=train, aes(x=SibSp, fill=Survived))+
  geom_bar(position = 'fill')+
  facet_wrap(~Pclass)+
  scale_fill_manual(values=c("black", "orange"))


parch.class <- ggplot(data=train, aes(x=Parch, fill=Survived))+
  geom_bar(position = 'fill')+
  facet_wrap(~Pclass)+
  scale_fill_manual(values=c("black", "orange"))

embark.class <- ggplot(data=train, aes(x=Embarked, fill=Survived))+
  geom_bar(position = 'fill')+
  facet_wrap(~Pclass)+
  scale_fill_manual(values=c("black", "orange"))

grid.arrange(sex.class, sibs.class, parch.class, embark.class,  ncol=2)
```



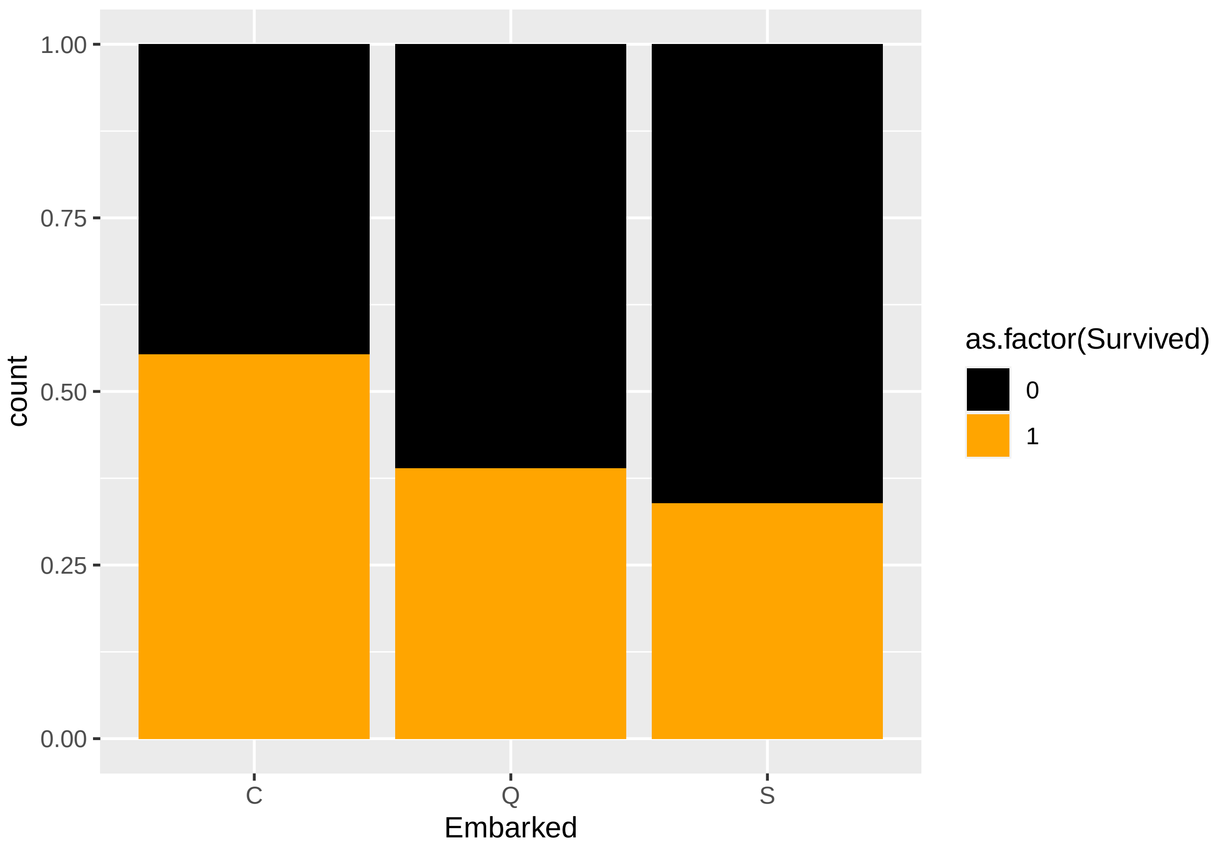
Las mujeres que viajaban en primera y segunda clase sobrevivieron casi todas. Los hombres sobrevivieron en mucho menor medida, incluso los hombres que viajaron en primera clase.

Las personas que viajaban con hermanos o pareja en primera clase sobrevivieron en mayor medida que las personas que viajaron solas. En primera y segunda clase no hay personas que viajasen con más de 3 hermanos. En tercera clase hay personas que viajaron con hasta 8 hermanos en este caso un menor número de hermanos (excepto cero) parece indicar una mayor supervivencia.

En función de la localización de Embarque, Clase y Sexo

También hemos visualizado la supervivencia en función del lugar donde los pasajeros embarcaron en el Titanic. Los pasajeros de tercera clase que embarcaron en el punto S han tenido menos proporción de supervivencia que los demás. Los pasajeros de segunda clase que embarcaron en Q sobrevivieron en mayor proporción que los pasajeros de segunda clase que embarcaron en otros puntos. Para los pasajeros de primera clase el punto de embarque con mayor porcentaje de superviviencia es C.

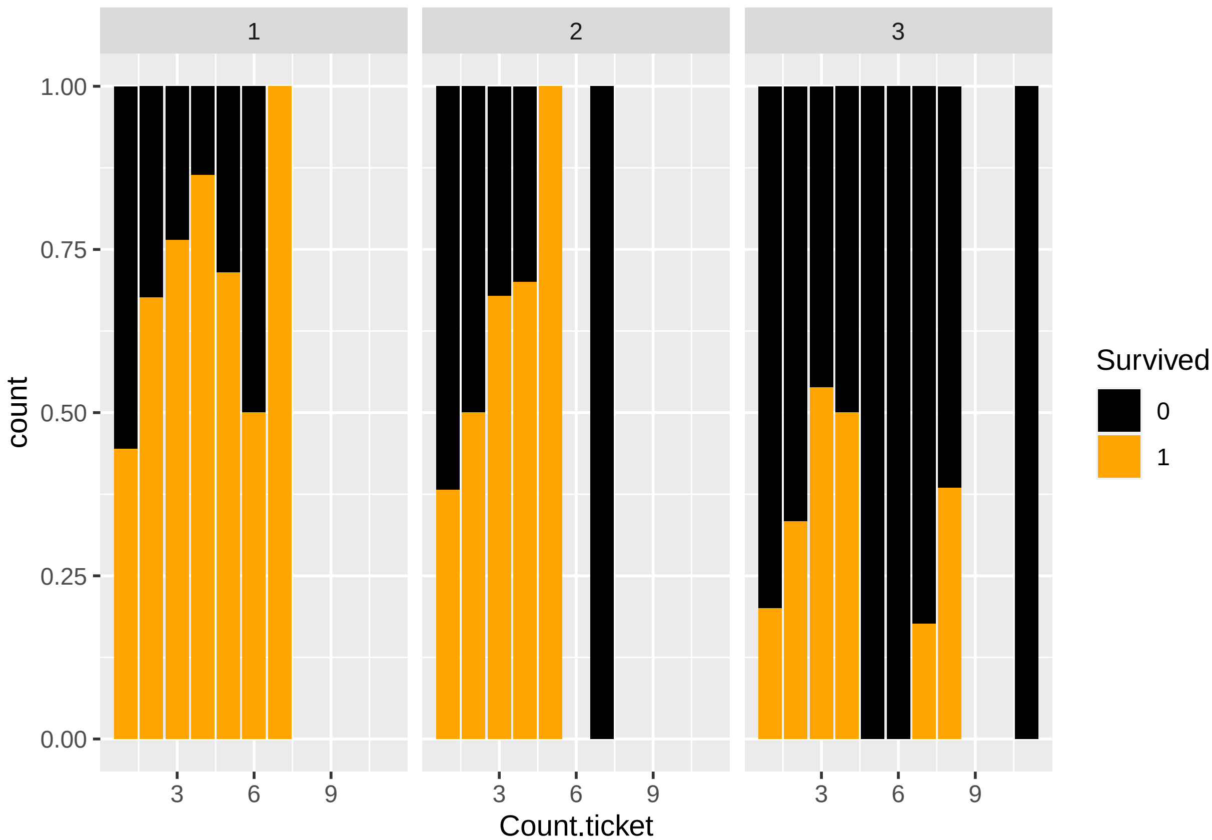
```
ggplot(data=train, aes(x=Embarked, fill=as.factor(Survived)))+geom_bar(position = 'fill')+
  scale_fill_manual(values=c("black", "orange", "red"))
```



Visualizamos la supervivencia usando las variables de la clase en la que viaja el pasajero y el número de personas que viajan con el mismo billete.

En el Titanic viajaron familias enteras de hasta 7 personas en primera clase e incluso de 11 personas en tercera clase. La supervivencia de personas que viajaban sobre el mismo billete en tercera clase es comparable con el mismo dato en segunda clase.

```
ggplot(data=train, aes(x=Count.ticket, fill=Survived))+
  geom_bar(position = 'fill')+
  facet_wrap(~Pclass)+
  scale_fill_manual(values=c("black", "orange", "red"))
```



Análisis de los datos.

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

La pregunta que planteamos inicialmente es si entre las variables que tenemos en el dataset existen algunas que influyen en mayor medida en la supervivencia de los pasajeros del Titanic. En los análisis anteriores hemos comprobado que:

-Visualmente que las personas de primera clase han sobrevivido en mayor medida
-Las mujeres y los niños tienen una mayor proporción de supervivencia
-Las personas que viajan acompañadas, ya sea por familia o amigos, tiene una proporcion de supervivencia algo mejor

Vamos a usar estos datos para analizar el dataset en mayor profundidad. Para ello necesitamos datasets especializados.

Subconjunto Supervivientes Vs. No Supervivientes

Este subconjunto se utilizará practicamente en la totalidad de los análisis por ser la variable dependiente y principal a la pregunta realizada.

```
train.survived <- train[train$Survived==1,]
train.not.survived <- train[train$Survived==0,]
```

Subconjunto Hombres Vs. Mujeres

Este subconjunto nos ayudará a poder tener diferenciados los datos por la variable Sexo (hombres y mujeres) y poder realizar contrastes de hipótesis, regresiones logísticas, árboles de decisión y un modelo Random Forest.

```
# Dataframe por sexo del pasajero
train.male <- train[train$Sex=="male", ]
train.female <- train[train$Sex=="female",]
```

Subconjunto Jóvenes (<18 años) Vs. Adultos

Este subconjunto nos ayudará a poder tener diferenciados los datos por la variable Age (menores y adultos) y poder realizar contrastes de hipótesis, regresiones logísticas, árboles de decisión y un modelo Random Forest.

```
# Dataframe por edad del pasajero
train.young <- train[train$Age<18,]
train.older <- train[train$Age>=18,]
```

Subconjunto Billete Único Vs. Billete Grupal

Este subconjunto nos ayudará a poder tener diferenciados los datos por la variable Count.ticket (únicos o grupales) y poder realizar contrastes de hipótesis (incluyendo un Xi cuadrado), regresiones logísticas, árboles de decisión y un modelo Random Forest.

```
# Tabla de contingencia Survived - Tipo Billete

train$ticket_tipo[train$Count.ticket==1] = "Único"
train$ticket_tipo[train$Count.ticket!=1] = "Grupal"

train$ticket_tipo <- as.factor(train$ticket_tipo)
```

Usaremos estos datos para realizar el contraste de hipótesis sobre la proporción de supervivientes entre los pasajeros que viajaban solos frente a los pasajeros que viajaban acompañados.

```
# Dataframe por pasajeros que viajaban solos o acompañados
train.single <- train[train$Count.ticket=="1",]
train.many <- train[train$Count.ticket>1,]
```

Preparación de datos para aplicar algoritmos (Regresión logística, árboles de decisión, random forest).:

Los datos de test que tenemos no disponen de etiqueta de clase.

Partimos el conjunto train en subconjuntos de train y validación. De esta manera dispondremos de datos para comprobar el funcionamiento de los modelos que construyamos.

```
# Variables independientes
X <- train[c("Pclass", "Sex", "Age", "SibSp", "Parch",
             "Count.ticket", "Unit.price", "Embarked")]

# Etiquetas de clase
y <- train[9]
```

```
set.seed(3)
indexes = sample(1:nrow(train), size=floor((2/3)*nrow(train)))
trainX<-X[indexes,]
trainy<-y[indexes,]
validX<-X[-indexes,]
validy<-y[-indexes,]
```

Para algunos modelos necesitaremos la variable en forma de texto, por lo tanto guardamos un dato alternativo sobre la supervivencia de los pasajeros: **"Survived" para el valor 1 y "Died" para el valor 0** de la variable dependiente.

```
train$survived <- ifelse(train$Survived==1, "Survived", "Died")
```

Preparamos los datos para el algoritmo de Random Forest:

```
# Para random forest
y.1 <- train["survived"]
trainy.1 <- y.1[indexes,]
validy.1 <- y.1[-indexes,]
```

Para la regresión logística es mejor tener todos los datos juntos en un dataframe:

```
regr.data <- data.frame(cbind(trainX, trainy))
```

Comprobación de la normalidad y homogeneidad de la varianza.

Función para visualizar los datos por pares.

```
 visualiz1 <- function(D1, D2, name1, name2, title){
   oldpar = par(mfrow = c(2,2), mar=c(2,2,2,2))
  truehist(D1, main = paste(name1," ", title), col="orange")
  abline(v = mean(D1), col="red", lwd=3, lty=2);
  abline(v = median(D1), lwd=3, lty=2, col="blue");
  qqnorm(D1, main = paste(name1, " ", title));qqline(D1, col = 2 )
  truehist(D2, main = paste(name2," ", title), col="grey")
```

```
  abline(v = mean(D2), col="red", lwd=3, lty=2);
  abline(v = median(D2), lwd=3, lty=2, col="blue");
  qqnorm(D2, main = paste(name2," ", title)); qqline(D2, col = 2 )

}
```

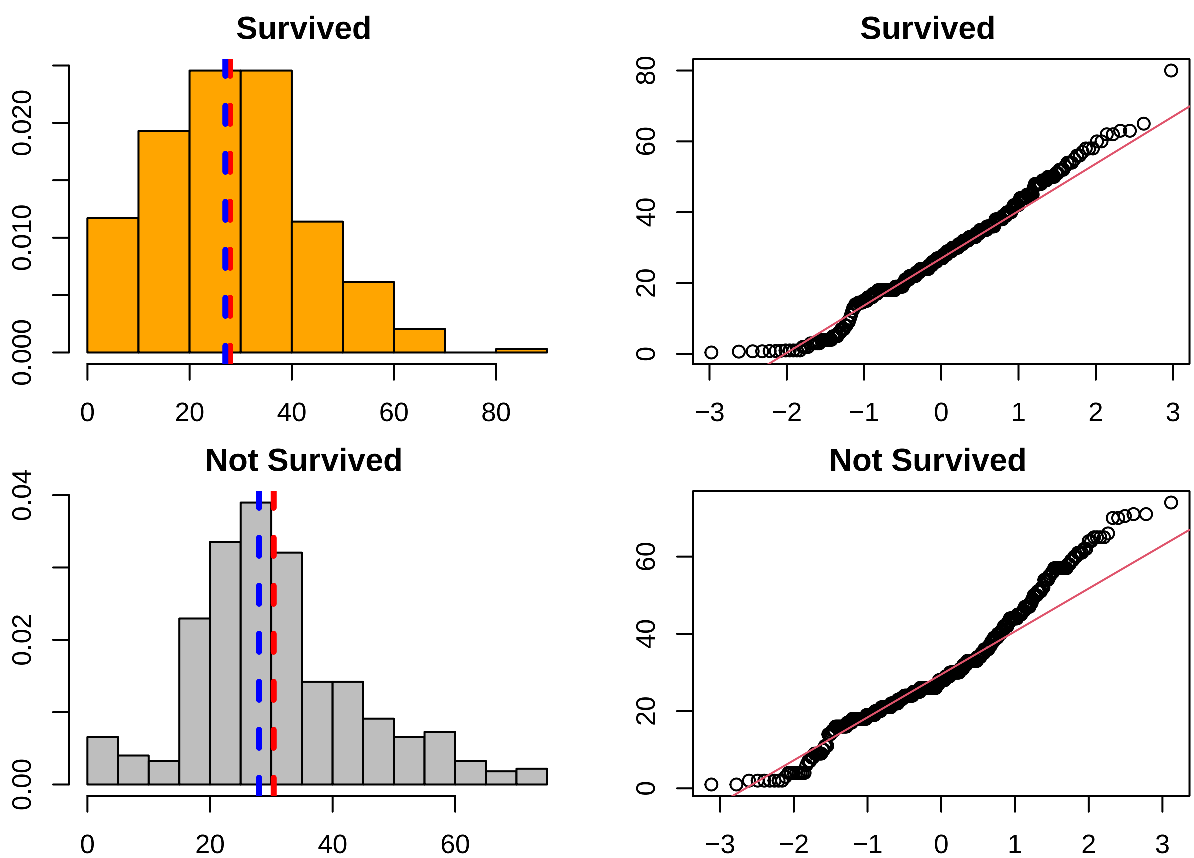
Variable age
Visualizamos los datos de la edad para personas que sobrevivieron en el hundimiento frente a personas que no sobrevivieron.

La media (en rojo) y la mediana (en azul) de los pasajeros que sobrevivieron están muy cerca. Las personas que no sobrevivieron tienen una edad mediana más baja que la edad media. Los valores de estos dos indicadores de tendencia central son más altos para personas que no sobrevivieron: **las personas que murieron tenían una media de edad algo más alta que los que sobrevivieron**.

Según el qqplot las dos distribuciones son cercanas a la normal.

Según el teorema del límite central podemos asumir la distribución normal de la media de estas dos muestras ya que sus tamaños son mayores que 30 observaciones.

```
visualiz1(train.survived$Age, train.not.survived$Age,
          "Survived", "Not Survived", "")
```



Comprobaremos de la igualdad de las varianzas entre los conjuntos de datos de pasajeros que sobrevivieron y no. Para ello usamos la función `var.test`.

El contraste de varianzas se realiza mediante un contraste de hipótesis, donde aceptar H0 significaría que las varianzas son iguales.

El valor **p-value** que obtenemos es de **0.15**. Este valor indica que rechazando la hipótesis nula de igualdad de varianzas probablemente estaríamos comentiendo un error, por tanto:

-Las varianzas en la edad de los pasajeros que sobrevivieron y los que no sobrevivieron son iguales y podemos hacer esta afirmación con un nivel de confianza del 95%.

```
var.test(train.survived$Age, train.not.survived$Age)
```

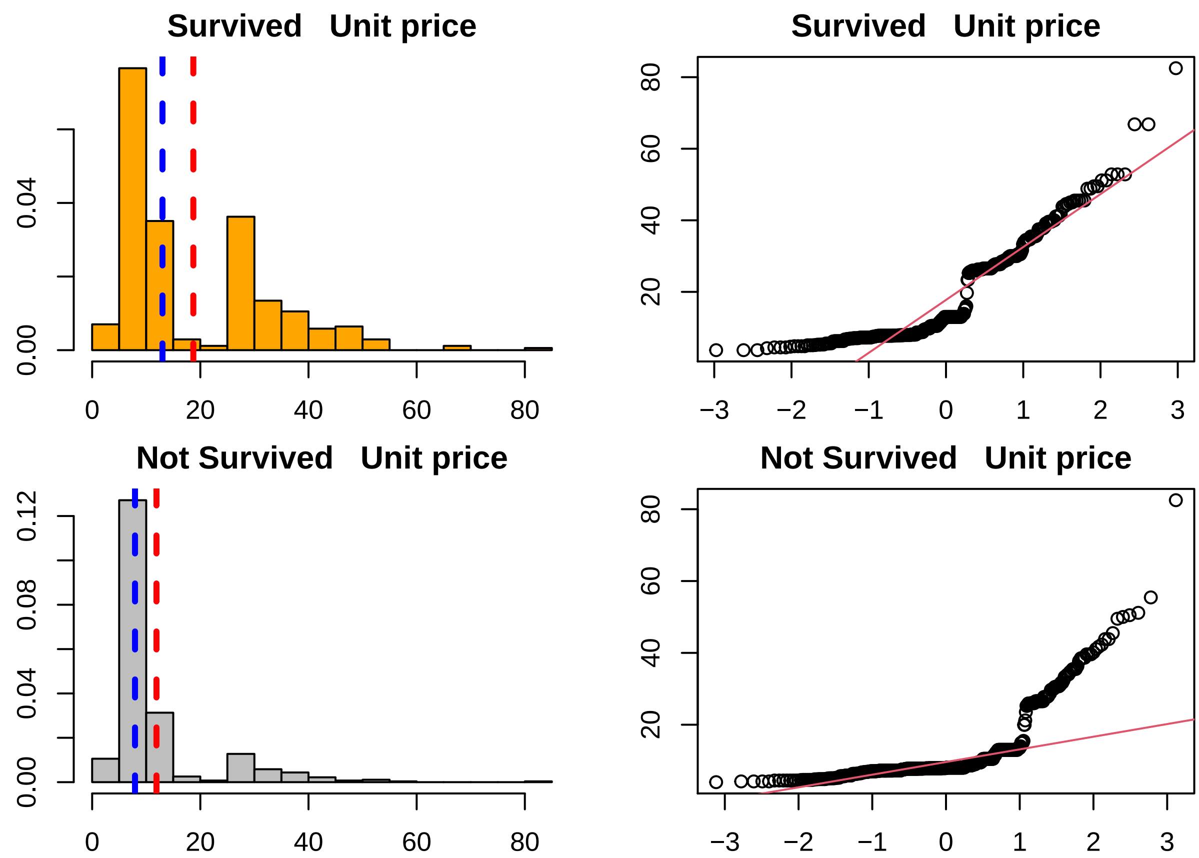
```
##
##  F test to compare two variances
##
## data:  train.survived$Age and train.not.survived$Age
## F = 1.1489, num df = 341, denom df = 548, p-value = 0.1508
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9507576 1.3943200
## sample estimates:
## ratio of variances
##           1.148887
```

Variable Unit price: precio del billete unitario
Visualizamos el precio del billete unitario para los pasajeros que sobrevivieron frente a los que no. Podemos ver que la media (en rojo) y la mediana (en azul) del precio de billete para los pasajeros que sobrevivieron están bastante separadas. El precio que corresponde a la mediana es alrededor de **13** y la media es más cercana a **20**. Existe esta diferencia porque la distribución tiene una cola larga a la derecha, tenemos pasajeros que pagaron un precio muy alto por sus billetes.

La mayoría de los pasajeros que no sobrevivieron pagaron un precio bajo por el billete, hay pocas personas que pagaron precios más altos y no sobrevivieron.

Ninguna de las dos distribuciones es normal, aunque para el fin de hacer un contraste de hipótesos sobre la media podemos asumir que la media poblacional se distribuye normalmente (**Teorema del límite central**).

```
visualiz1(train.survived$Unit.price, train.not.survived$Unit.price,
          "Survived", "Not Survived", "Unit price")
```



Realizamos el test de igualdad de las varianzas para el precio del billete para pasajeros que sobrevivieron frente a los que no sobrevivieron. El valor **p-value** nos indica que podemos rechazar H0, es decir, **las varianzas entre estos dos grupos de pasajeros son diferentes**.

En los histogramas podemos ver que los pasajeros que sobrevivieron de media pagaron más por sus billetes. Tanto en el grupo de supervivientes como de no supervivientes la media es mayor que la mediana, dado que tenemos outliers de precio de billete unitario muy alto.

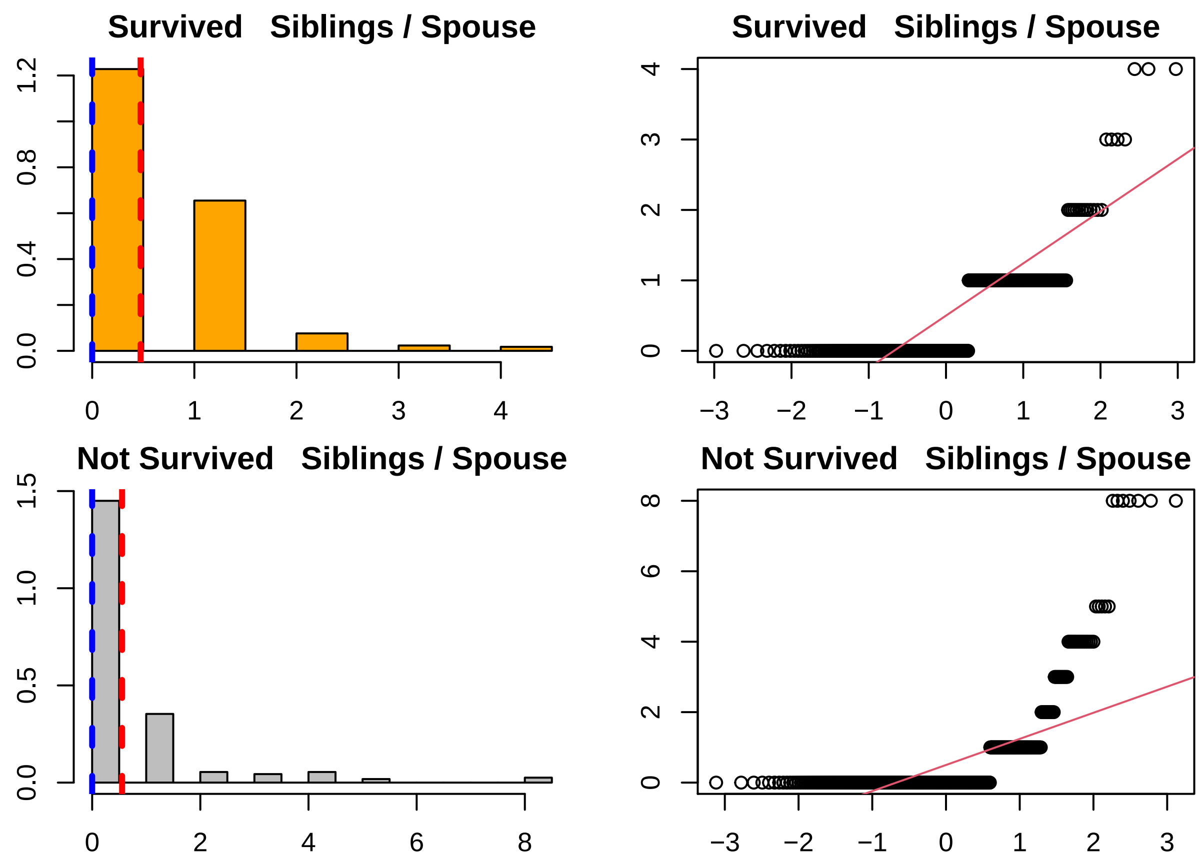
```
var.test(train.survived$Unit.price, train.not.survived$Unit.price)
```

```
##
##  F test to compare two variances
##
## data:  train.survived$Unit.price and train.not.survived$Unit.price
## F = 2.0651, num df = 341, denom df = 548, p-value = 3.73e-14
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.708968 2.506262
## sample estimates:
## ratio of variances
##           2.065101
```

Variable SibSp: pareja o hermanos

Aunque la variable no sigue una distribución normal, dado el tamaño grande de la muestra podemos asumir que la media muestral sí sigue una distribución normal (**Teorema del límite central**)

```r
visualiz1(train.survived$SibSp, train.not.survived$SibSp,
          "Survived", "Not Survived", "Siblings / Spouse")
```



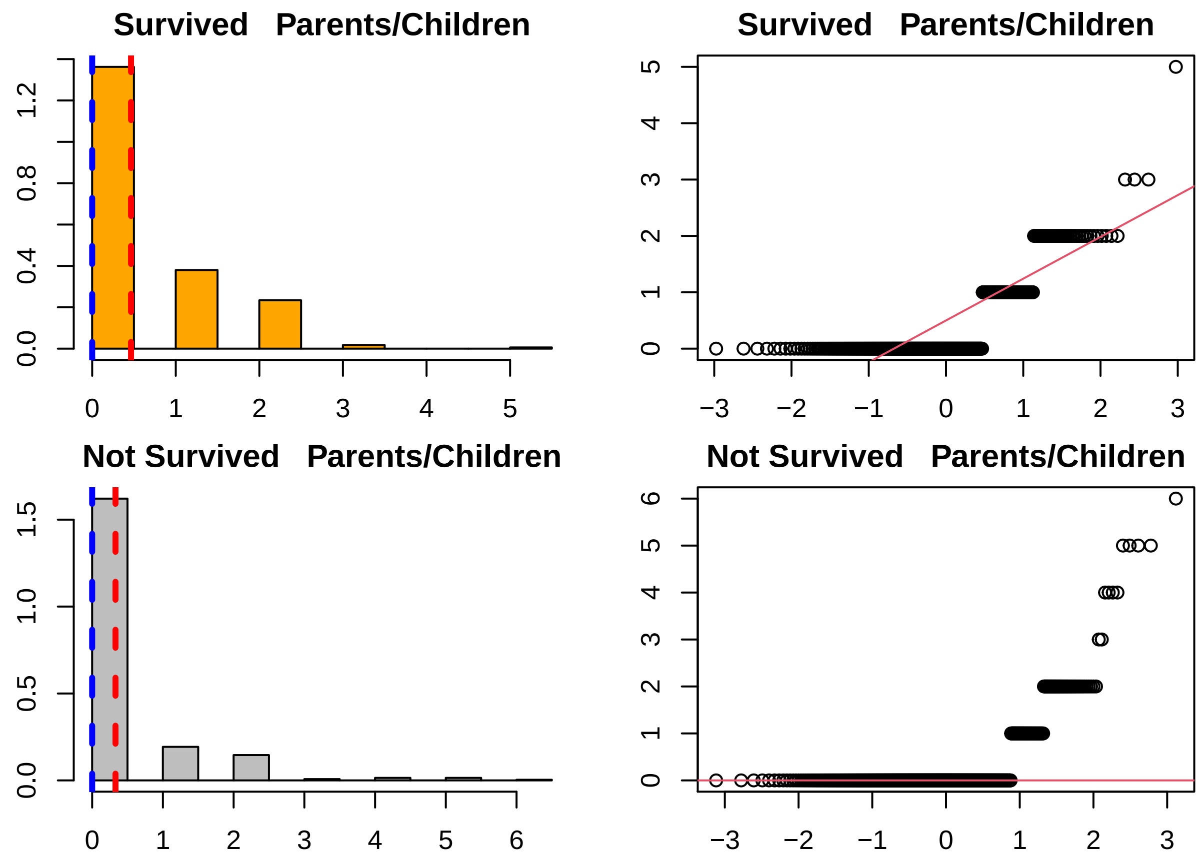
Realizamos el test de igualdad de las varianzas para SibSp para pasajeros que sobrevivieron frente a los que no sobrevivieron. El valor **p-value** nos indica que podemos rechazar H0, es decir, **las varianzas entre estos dos grupos de pasajeros son diferentes**.

```r
var.test(train.survived$SibSp, train.not.survived$SibSp)
```

```
##
##  F test to compare two variances
##
## data:  train.survived$SibSp and train.not.survived$SibSp
## F = 0.30256, num df = 341, denom df = 548, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2503810 0.3671926
## sample estimates:
## ratio of variances
##          0.3025581
```

Variable Parch: padres o hijos
Asumimos la normalidad de distribución de la media muestral aplicando el **Teorema del límite central**.

```
visualiz1(train.survived$Parch, train.not.survived$Parch,
          "Survived", "Not Survived", "Parents/Children")
```



El test de igualdad de las varianzas indica que **no podemos rechazar la hipótesis nula**. Entre los grupos que sobrevivieron y los que no sobrevivieron las varianzas de la variable Parch **son iguales**.

```
var.test(train.survived$Parch, train.not.survived$Parch)
```

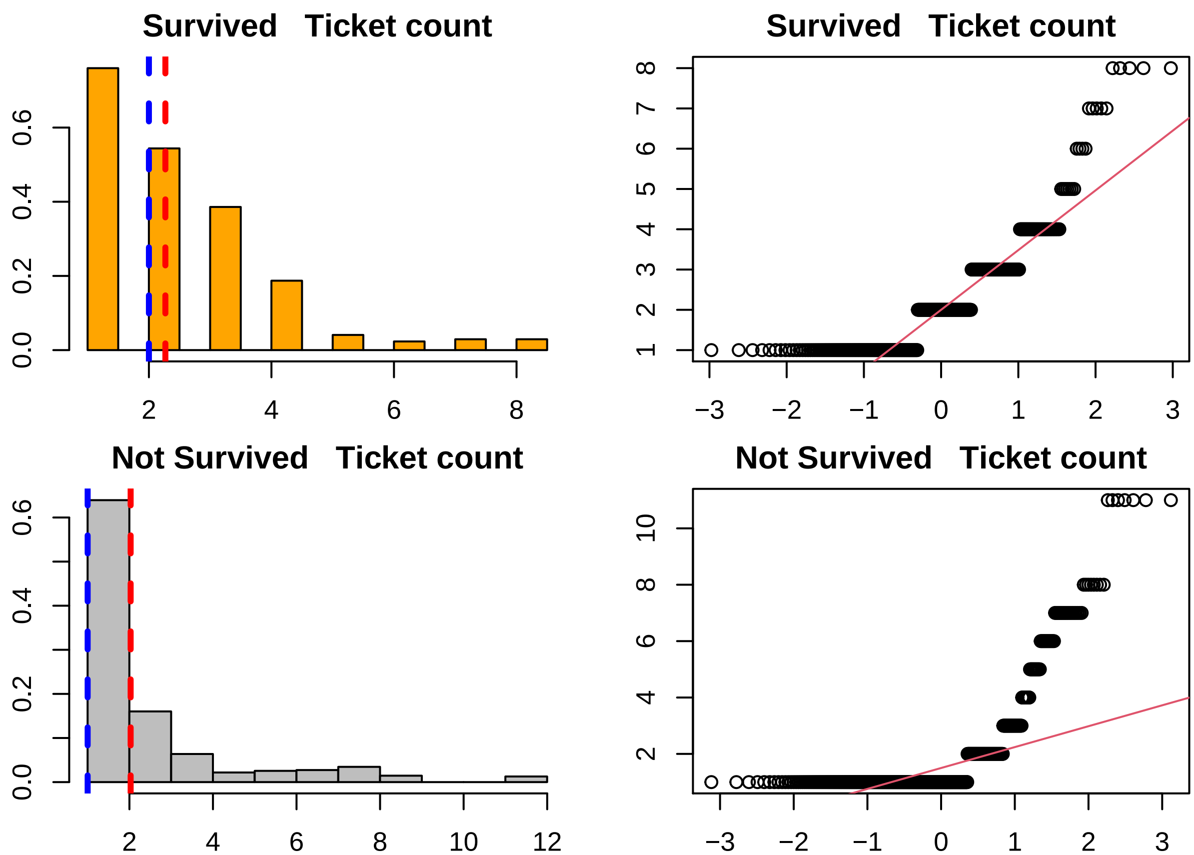
```
##
##  F test to compare two variances
##
## data:  train.survived$Parch and train.not.survived$Parch
## F = 0.87889, num df = 341, denom df = 548, p-value = 0.1908
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7273246 1.0666475
## sample estimates:
## ratio of variances
##          0.8788923
```

Variable Count.ticket: número de personas que viajaban con un mismo billete

Visualizamos un histograma que muestra la frecuencia de recuento de billetes: pasajeros que viajaban solos hasta pasajeros que viajaban con muchos acompañantes.

Otra vez, aplicamos el **Teorema del límite central** y asumimos la normalidad de la distribución de las medias muestrales.

```
visualiz1(train.survived$Count.ticket, train.not.survived$Count.ticket, "Survived", "Not Survived", "Ticke
```



En este caso tenemos un valor **p-value** muy bajo y por lo tanto **podemos aceptar H1 que indica que las varianzas de la variable Count.ticket son diferentes** en el grupo de supervivientes y no supervivientes.

```
var.test(train.survived$Count.ticket, train.not.survived$Count.ticket)
```

```
##
##  F test to compare two variances
##
## data:  train.survived$Count.ticket and train.not.survived$Count.ticket
## F = 0.55183, num df = 341, denom df = 548, p-value = 3.459e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4566685 0.6697206
## sample estimates:
## ratio of variances
##          0.5518339
```

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. Aplicar al menos tres métodos de análisis diferentes.

En esta práctica nos hemos planteado como objetivo encontrar cuales de las variables tuvieron una mayor influencia sobre la supervivencia de los pasajeros del Titanic. En este apartado aplicaremos varios contrastes de hipótesis, regresiones logísticas y también modelos de aprendizaje no supervisado (árboles de decisión y random forest).

Las tareas que nos hemos planteado se muestran a continuación:

1. Buscaremos correlaciones entre las variables del dataset.

2. Realizaremos contrastes de hipótesis para responder a las siguientes preguntas:

- **¿La proporción de hombres supervivientes es igual a la de mujeres supervivientes?**
- **¿La proporción de menores supervivientes es igual a la proporción de mayores de edad que sobrevivieron?**
- **¿La proporción de supervivientes entre los pasajeros que viajaban solos es igual a la proporción de supervivientes entre pasajeros que viajaban con más de una persona?**
- **¿Las personas que sobrevivieron eran más jóvenes que las que no sobrevivieron?**

3. Realizaremos un test Xi cuadrado para comprobar si:

- **Existe relación entre el tipo de billete (único o grupal) y la supervivencia de los distintos pasajeros**
- **Existe relación entre la clase en la que viajaban los pasajeros y su supervivencia.**

4. Planteamos las siguientes regresiones logísticas:

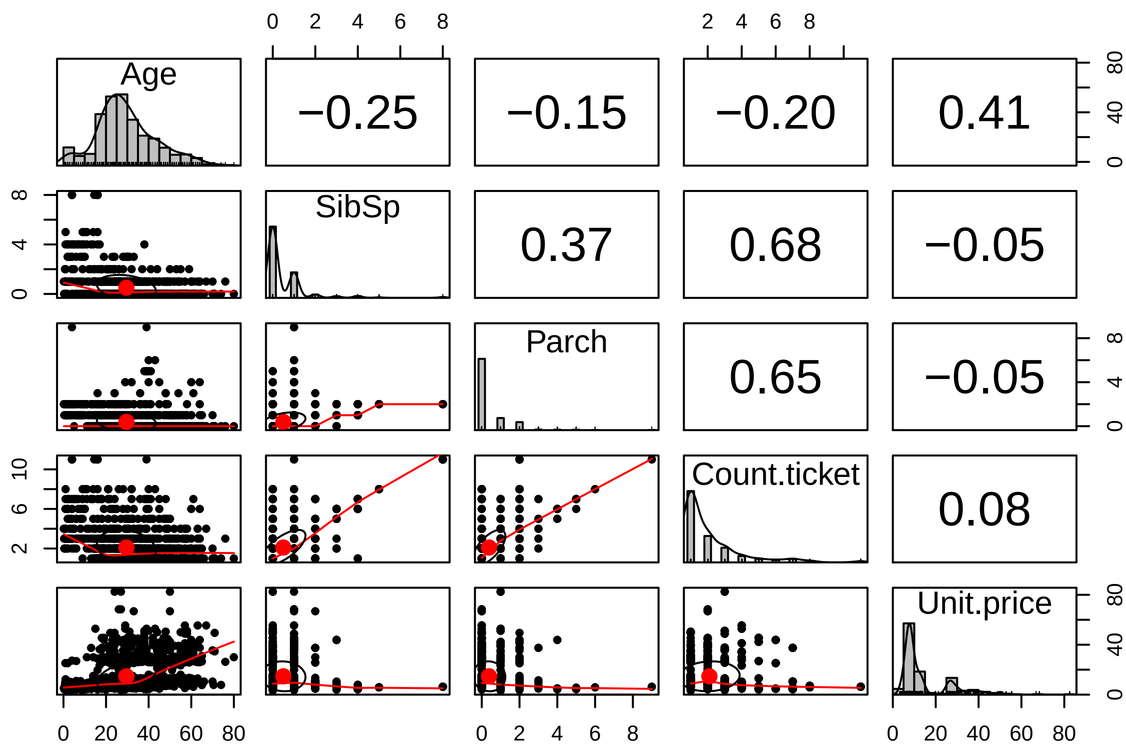
- **Variable dependiente Survived explicada por la variable Sex**
- **Variable dependiente Survived explicada por las variables Sex y Pclass**
- **Variable dependiente Survived explicada por todas las variables independientes disponibles.**

5. Aplicaremos un modelo de **árbol de decisión CART** con y sin validación cruzada y boosting.

6. Aplicaremos un algoritmo de **Random Forest**.

1. –> Correlaciones

```
# Relaciones cruzadas de las variables cuantitativas (correlaciones) para todos los datos.
col_var_cuantitativa_sin_ID <- c("Age", "SibSp", "Parch", "Count.ticket", "Unit.price")

df_var_cuantitativa <- df_total_sin_etiqueta[, col_var_cuantitativa_sin_ID]

pairs.panels(df_var_cuantitativa[,col_var_cuantitativa_sin_ID],
             method = "pearson",
             hist.col = "grey",
             density = TRUE,
             ellipses = TRUE
             )
```



No se observa a priori ninguna correlación fuerte entre ninguna de las variables numéricas, aunque se puede notar cierta relación entre el número de tickets (**Count.ticket**) grupales y las variables **SibS** y **Parch**, suceso lógico pues estas variables indican que los viajeros iban acompañados y por tanto es lógico pensar que compartían mismo billete múltiple.

El precio unitario del billete (**Unit.price**) versus la edad (**Age**), indica que hay viajeros de todas las edades en todas las clases, si asumimos, como hemos demostrado en apartados anteriores que el precio unitario es un indicador de la clase en la que viajaban los viajeros.

2. –> Contrastes de hipótesis

Contraste de hipótesis sobre el sexo y nivel de supervivencia

¿La proporción de supervivencia en hombres es inferior a la de las mujeres?

Acorde con la pregunta planteada, realizaremos un contraste de hipótesis con las siguientes hipótesis de partida:

Hipótesis nula - H0: La proporción de hombres que sobreviven es igual o mayor a la proporción de mujeres que sobreviven
Hipótesis alternativa - H1: La proporción de hombres que sobreviven es menor que la proporción de mujeres que sobreviven

H0: p.m >= p.f
H1: p.m < p.f

Este contraste de hipótesis representa un test de la proporción de la población por la cola izquierda, por tanto, el valor zcrit.R representará, bajo la hipótesis nula, el límite inferior de la supuesta verdadera proporción de la población.
Este contraste rechazará la HO si el estadístico calculado es menor o igual al valor crítico (con signo negativo) al nivel de significación estimado.

Aceptar la hipótesis nula implicará aceptar que los hombres sobrevivieron al menos en la misma proporción que las mujeres. En cambio, aceptar la hipótesis alternativa indicaría que los hombres sobrevivieron en menor proporcion que las mujeres.

El tamaño de la muestra es los suficientemente grande para poder asumir que la proporción muestral para la población de hombres sigue una distribución normal con media p.m y desviación estándar `sqrt(p.m(1-p.m) / n.m)`. De forma análoga se determina la media y desviación estándar para la proporción muestral de las mujeres.

Debido al planteamiento de la hipótesis nula y su alternativa y por cómo se han descrito las hipótesis H0 y H1, se observa que la hipótesis alternativa es unilateral, puesto que se plantea como un límite a un solo valor dado.

```
# Fijamos un nivel de significación

alfa <- 0.05
```

Determinamos el estadístico de contraste, en este caso, la muestra es grande y proviene de una distribución de Bernoulli de parámetro p, con lo cual, según el Teorema del Límite Central podremos utilizar el estadístico de contraste mostrado anteriormente.

```
train.survived_M <- train.survived[train.survived$Sex=="male",]
train.survived_F <- train.survived[train.survived$Sex=="female",]


n.m <- length(train.male$Sex)
n.f <- length(train.female$Sex)

# Proporción de supervivientes entre hombres
p.m <- length(train.survived_M$Sex) / n.m

# Proporción de supervivientes entre mujeres
p.f <- length(train.survived_F$Sex) / n.f

# Parámetro p
p<-(n.m*p.m + n.f*p.f) / (n.m+n.f)

# Estadístico observado
zobs <- (p.m-p.f)/( sqrt(p*(1-p)*(1/n.m+1/n.f)) )
```

```r
# Valor crítico
zcrit.L <- qnorm(alfa,  lower.tail=TRUE)

# valor p
pvalue<- pnorm(zobs, lower.tail=TRUE )

# Resultado
kbl(data.frame(zobs,  zcrit.L, pvalue, p.m, p.f, n.m, n.f),booktabs =T)%>%
  kable_styling(latex_options =c("striped", "hold_position"))
```

zobs	zcrit.L	pvalue	p.m	p.f	n.m	n.f
-16.21883	-1.644854	0	0.1889081	0.7420382	577	314

Tal y como se ha planteado H0 y H1, siendo H1 p.m < p.f, rechazamos H0 si el valor del estadístico zobs es menor que el valor crítico (en signo negativo). El resultado del contraste es **-16.22** < **-1.64** y esto nos permite rechazar H0 a favor de H1 con un nivel de confianza (1-alfa) igual al 95%: la hipótesis alternativa afirma que **la proporción de supervivencia en hombes es menor que la proporción de supervivencia en mujeres**.

De igual forma, al ser valor p casi nulo y por tanto menor que nuestro nivel de significación (0.05), el valor p es significativo y podemos rechazar, confirmando el contraste anterior, la hipótesis nula.

Siguiendo la misma metodología, responderemos a las siguientes preguntas:

Contraste de hipótesis sobre la edad y nivel de supervivencia

¿La proporción de supervivencia en menores es inferior a la de los mayores?

Hipótesis nula - H0: La proporción de menores que sobreviven es igual o mayor a la proporción de adultos que sobreviven
Hipótesis alternativa - H1: La proporción de menores que sobreviven es menor que la proporción de adultos que sobreviven

H0: p.y >= p.o
H1: p.y < p.o

```r
train.survived_Y <- train.survived[train.survived$Age<18,]
train.survived_O <- train.survived[train.survived$Age>=18,]


n.y <- length(train.young$Age)
n.o <- length(train.older$Age)

# Proporción de supervivientes entre hombres
p.y <- length(train.survived_Y$Age) / n.y

# Proporción de supervivientes entre mujeres
p.o <- length(train.survived_O$Age) / n.o
```

```r
# Parámetro p
p<-(n.y*p.y + n.o*p.o) / (n.y+n.o)

# Estadístico observado
zobs <- (p.y-p.o)/( sqrt(p*(1-p)*(1/n.y+1/n.o)) )

# Valor crítico
zcrit.L <- qnorm(alfa,  lower.tail=TRUE)

# valor p
pvalue<- pnorm(zobs, lower.tail=TRUE )

# Resultado
kbl(data.frame(zobs,  zcrit.L, pvalue, p.y, p.o, n.y, n.o),booktabs =T)%>%
  kable_styling(latex_options =c("striped", "hold_position"))
```

zobs	zcrit.L	pvalue	p.y	p.o	n.y	n.o
3.749007	-1.644854	0.9999112	0.530303	0.3583663	132	759

Obtenemos que el valor crítico (**-1.64**) es menor que el estadístico observado (**3.75**), por lo tanto se sitúa en el rango de aceptación de H0. Por lo tanto, con un nivel de confianza del 95% debemos aceptar la hipótesis nula que indica que **la proporción de personas menores de 18 años que sobrevivieron en el accidente es al menos igual (o mayor) que la proporción de mayores de 18 años que sobrevivieron**.

De igual forma, al ser valor p un valor muy cercano a 1 y mayor que alfa, si rechazamos la hipótesis nula, posiblemente estaremos comentiendo un error.

Contraste de hipótesis sobre el el tipo de billete (único o grupal) y nivel de supervivencia

¿La proporción de supervivencia en viajeros con billete único es inferior a la de los viajeros con billete en conjunto?

Hipótesis nula - H0: La proporción de supervivientes que viajaban solos es igual o mayor a la proporción de supervivientes que viajaban acompañados.
Hipótesis alternativa - H1: La proporción de supervivientes que viajaban solos es menor que la proporción de supervivientes que viajaban acompañados.

H0: p.s >= p.my
H1: p.s < p.my

```r
# Determinación del estadístico de contraste

train.survived_S <- train.survived[train.survived$Count.ticket==1,]
train.survived_My <- train.survived[train.survived$Count.ticket>1,]

n.s <- length(train.single$Count.ticket)
n.my <- length(train.many$Count.ticket)

# Proporción de supervivientes entre hombres
p.s <- length(train.survived_S$Count.ticket) / n.s

# Proporción de supervivientes entre mujeres
p.my <- length(train.survived_My$Count.ticket) / n.my
```

```r
# Parámetro p
p<-(n.s*p.s + n.my*p.my) / (n.s+n.my)

# Estadístico observado
zobs <- (p.s-p.my)/( sqrt(p*(1-p)*(1/n.s+1/n.my)) )

# Valor crítico
zcrit.L <- qnorm(alfa,  lower.tail=TRUE)

# valor p
pvalue<- pnorm(zobs, lower.tail=TRUE )

# Resultado
kbl(data.frame(zobs,  zcrit.L, pvalue, p.s, p.my, n.s, n.my),booktabs =T)%>%
  kable_styling(latex_options =c("striped", "hold_position"))
```

zobs	zcrit.L	pvalue	p.s	p.my	n.s	n.my
-7.550129	-1.644854	0	0.2702703	0.5170732	481	410

Para aceptar H0 el estadístico observado tendría que ser mayor que el valor crítico, dado el contraste por la izquierda. El valor del estadístico observado (**-7.55**) es menor que el valor crítico (**-1.64**), el estadístico observado se encuentra fuera del rango de aceptación de H0, por lo tanto, debemos rechazar la hipótesis nula a favor de la alternativa: **con un nivel de confianza del 95% podemos afirmar que la proporción de supervivientes entre los pasajeros que viajaban solos es menor que la proporción de supervivientes entre los pasajeros que viajaban acompañados**.

El valor bajo de p-value confima la conclusión que hemos extraido anteriormente, podemos aceptar H1.

3. –> Hipótesis sobre independencia de variables (Xi Cuadrado)
Relación entre supervivencia y billete único o grupal
Utilizamos el test Chi cuadrado para comprobar si los pasajeros de primera clase sobrevivieron en mayor medida que los pasajeros de segunda clase.

El test se aplica en R con un nivel de confianza por defecto del 95%.

Crearemos dos tablas de contingencia entre las variables que indican supervencia y la que indica el tipo de billete. Una primera en valores absolutos, y una segunda mostrando las proporciones.

```r
# Tabla de contingencia Survived - Tipo Billete

table(train$Survived, train$ticket_tipo, dnn = c("Supervivencia", "Tipo Billete"))
```

```
##              Tipo Billete
## Supervivencia Grupal Único
##             0    198   351
##             1    212   130
```

```r
# Tabla de contingencia en proporciones Survived - Tipo Billete
prop.table(table(train$Survived, train$ticket_tipo, dnn = c("Supervivencia (%)",
                                                            "Tipo Billete (%)")))
```

```
##                  Tipo Billete (%)
## Supervivencia (%)    Grupal     Único
##                 0 0.2222222 0.3939394
##                 1 0.2379349 0.1459035
```

```r
length(train$ticket_tipo)
```

```
## [1] 891
```

```r
length(train$Survived)
```

```
## [1] 891
```

Dos variables categóricas que forman parte de una tabla de contingencia pueden ser sujetas a un test de independencia. Este test puede ser representado for el ChiSquare Test mediante un contraste de hipótesis.

El contraste se realizará para observar si las dos variables son independientes o no, por tanto, podemos plantear las hipótesis como se indica a continuación:

Hipótesis nula - H0: Las variables Survived y Tipo Billete son independientes.
Hipótesis alternativa - H1: Las variables Survived y Tipo Billete están relacionadas.

```r
# Tabla de contingencia Survived y Tipo Billete
tabla_cont <- table(train$Survived, train$ticket_tipo, dnn = c("Supervivencia",
                                                                "Tipo Billete"))
tabla_cont
```

```
##              Tipo Billete
## Supervivencia Grupal Único
##             0    198   351
##             1    212   130
```

```r
# Comprobación con la función chisq.test()
chisq.test(tabla_cont)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla_cont
## X-squared = 55.966, df = 1, p-value = 7.375e-14
```

El cálculo de los grados de libertad para una distribución Chi Square se calcula como df = (c - 1)(r - 1) donde c es el número de columnas y r el número de filas. En nuestro caso c = r = 2, por lo tanto df = (2 - 1)(2 - 1) = 1.

```r
# Grados de libertad
gdl <- 1

# Nivel de confianza
ndc <- 0.95


# Valor Crítico
val_cri <- qchisq(ndc, gdl)
val_cri
```

```
## [1] 3.841459
```

Debido a que el estadístico Chi Square es mayor al valor crítico calculado de la distribución Chi Square (con un grado de libertad y con un nivel de significación de 0.05), podemos rechazar la hipótesis nula de que no hay relación entre Survived y Tipo de Billete, es decir, rechazamos la hipótesis al 95 % de nivel de confianza de que tales variables sean independientes.

Como el valor de p-value es menor que el nivel de significancia, **podemos rechazar la hipótesis nula de que las variables Survived y Tipo de Billete son independientes**, cuadrando este resultado con los cálculos anteriores.

Por tanto, se puede afirmar al 95% de nivel de confianza que las variables Survived y Tipo de Billete **están relacionadas**.

Relación entre supervivencia y la clase en la que viajaban los pasajeros

Planteamos las hipótesis nula y alternativa:

Hipótesis nula - H0 : Las variables Survived y Pclass son independientes.
Hipótesis alternativa - H1: Las variables Survived y Pclass están relacionadas.

Igual que hicimos en el ejemplo anterior, calculamos la tabla de contingencia.

```
# Tabla de contingencia
tabla_cont1 <- table(train$survived, train$Pclass, dnn = c("Supervivencia", "Clase"))
tabla_cont1
```

```
##              Clase
## Supervivencia   1   2   3
##      Died      80  97 372
##      Survived 136  87 119
```

```
prop.table(table(train$survived, train$Pclass, dnn = c("Supervivencia (%)", "Clase (%)" )))
```

```
##                  Clase (%)
## Supervivencia (%)          1          2          3
##          Died     0.08978676 0.10886644 0.41750842
##          Survived 0.15263749 0.09764310 0.13355780
```

```
chisq.test(tabla_cont1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_cont1
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

```
# Grados de libertad
gdl1 <- 1

# Nivel de confianza
ndc1 <- 0.95

# Valor crítico
val_cri1 <- qchisq(ndc1, gdl1)
val_cri1
```

```
## [1] 3.841459
```

Los resultados que hemos obtenido son:

* El valor p-value es mucho menor que el nivel de significancia.
* El valor crítico es menor que el estadístico observado.

Con estos datos podemos concluir **que podemos rechazar H0 a favor de H1 con un nivel de confianza del 95%**, dados el valor p-value y el valor crítico.

4. –> Regresión Logística

Primera Regresión Logística

Aplicamos una regresión logística para predecir la probabilidad de supervivencia usando como variable predictora la variable **sexo**.

Según los datos que nos proporciona la función `summary` la variable sexo **es significativa para predecir la supervivencia**. El coeficiente negativo asociado a la variable dummy **Sexmale** indica que ser de sexo masculino reduce la probabilidad de sobrevivir.

```
model0=glm(formula=trainy~Sex, data=regr.data, family=binomial(link=logit))

summary(model0)
```

```
##
## Call:
## glm(formula = trainy ~ Sex, family = binomial(link = logit),
##     data = regr.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6035  -0.6561  -0.6561   0.8045   1.8121
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.9619     0.1518   6.338 2.33e-10 ***
## Sexmale      -2.3885     0.2001 -11.939  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 792.97  on 593  degrees of freedom
## Residual deviance: 626.45  on 592  degrees of freedom
## AIC: 630.45
##
## Number of Fisher Scoring iterations: 4
```

Calculamos los Odds Ratio. Los odds ratio indican que ser hombre es factor de protección para la clase 1 (sobrevivir). En un evento como el hundimiento del titanic, ser hombre reduce la probabilidad de sobrevivir con respecto a ser mujer.

El Odds Ratio se calcula como (Psobrevivir / (1 - Psobrevivir)) / (Pno sobrevivir / (1 - Pno sobevivir)). Cuando optenemos un valor menor que uno, significa que el denominador es mayor, por lo tanto, la probabilidad de no sobrevivir es mayor.

```
exp(coefficients(model0))
```

```
## (Intercept)     Sexmale
##  2.61666667  0.09177003
```

Usamos los datos de validación para realizar una predicción de supervivencia.

Tenemos que 76 instancias de supervivencia y 164 instancias de no supervivencia han sido predichas correctamente.
36 instancias de supervivencia real han sido predichas como no supervivencia
21 instancias de no supervivencia fueron predichas incorrectamente como supervivencia.

El modelo de regresión que utiliza únicamente la variable Sexo para predecir la supervivencia tiene una **precisión del 80.8%**.

```
newdata0 <- validX[c("Sex")]
probabilities <- model0 %>% predict(newdata0, type = "response")

predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

# Matriz de confusión
conf.1 <- table(validy, predicted.classes)
conf.1
```

```
##       predicted.classes
## validy   0   1
##      0 164  21
##      1  36  76
```

```
# Precisión
sum(diag(conf.1))/sum(colSums(conf.1))
```

```
## [1] 0.8080808
```

Segunda Regresión Logística

Aplicamos una regresión logística para comprobar si las variables **Sexo** y **Clase** son significativas para la supervivencia en el accidente.

Ambas variables son significativas y así lo indican los asteriscos junto a los valores Pr(>|z|).

Ser hombre, viajar en clase 2 o clase 3 reduce la posibilidad de supervivencia con respecto a ser mujer y viajar en primera clase.. Esto viene indicado por el signo negativo que acompaña el valor del coeficiente para Sexmale, Pclass2 y Pclass3.

El coeficiente para la variable Sexmale tiene el mayor valor absoluto del coeficiente y significa que un aumento de una unidad del número de hombres reduce en 2.5 unidades la posibilidad de supervivencia general.

```
# Modelo de regresión
model1=glm(formula=trainy~Sex+as.factor(Pclass), data=regr.data,
           family=binomial(link=logit))
```

```r
summary(model1)
```

```
##
## Call:
## glm(formula = trainy ~ Sex + as.factor(Pclass), family = binomial(link = logit),
##     data = regr.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1160  -0.7517  -0.4639   0.6717   2.1366
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.1260     0.2563   8.296  < 2e-16 ***
## Sexmale              -2.4934     0.2185 -11.413  < 2e-16 ***
## as.factor(Pclass)2  -0.7518     0.2989  -2.515   0.0119 *
## as.factor(Pclass)3  -1.8074     0.2539  -7.118  1.1e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 792.97  on 593  degrees of freedom
## Residual deviance: 568.28  on 590  degrees of freedom
## AIC: 576.28
##
## Number of Fisher Scoring iterations: 4
```

Calculamos los odds ratio. Igual que en el caso anterior vemos que **la probabilidad ser hombre y sobrevivir es mucho menor que la de ser mujer y sobrevivir**.

Por clases el OR indica que **estar en segunda o tercera clase es factor de protección** respecto a la primera clase. Para los pasajeros de segunda o tercera clase la probabilidad de sobrevivir disminuye con respecto a los pasajeros de primera clase.

```r
exp(coefficients(model1))
```

```
##        (Intercept)            Sexmale as.factor(Pclass)2 as.factor(Pclass)3
##         8.38086029         0.08263044         0.47150020         0.16408095
```

Realizamos la predicción sobre el conjunto de validación que obtuvimos y con el resultado creamos una matriz de confusión y calculamos la precisión. El resultado es igual que en la regresión anterior.

```r
newdata1 <- validX[c("Sex","Pclass")]
probabilities1 <- model1 %>% predict(newdata=newdata1, type = "response")
predicted.classes1 <- ifelse(probabilities1 > 0.5, 1, 0)
# Matriz de confusión
conf.2 <- table(validy, predicted.classes1)
conf.2
```

```
##       predicted.classes1
## validy   0   1
##      0 164  21
```

```
##      1  36  76
```

```
# Precisión
sum(diag(conf.2)) / sum(colSums(conf.2))
```

```
## [1] 0.8080808
```

Tercera Regresión Logística

Por último generamos usando como variables predictoras la clase, el sexo, el número de personas que viajaban con el mismo billete y el puerto de embarque.

```
# Modelo de regresión
model2=glm(formula=trainy~factor(Pclass)+Sex+Age+factor(Count.ticket)+Embarked,data=regr.data, family=binon

summary(model2)
```

```
##
## Call:
## glm(formula = trainy ~ factor(Pclass) + Sex + Age + factor(Count.ticket) +
##     Embarked, family = binomial(link = logit), data = regr.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2657  -0.6203  -0.3874   0.6005   2.8263
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.958791   0.567747   6.973 3.11e-12 ***
## factor(Pclass)2        -1.018130   0.349450  -2.914 0.003574 **
## factor(Pclass)3        -2.382016   0.351810  -6.771 1.28e-11 ***
## Sexmale                -2.531891   0.247112 -10.246  < 2e-16 ***
## Age                    -0.038464   0.009966  -3.860 0.000114 ***
## factor(Count.ticket)2   0.211339   0.288010   0.734 0.463077
## factor(Count.ticket)3   0.608194   0.368461   1.651 0.098814 .
## factor(Count.ticket)4   0.848109   0.552332   1.536 0.124659
## factor(Count.ticket)5  -1.401969   0.628952  -2.229 0.025810 *
## factor(Count.ticket)6  -3.445269   1.272196  -2.708 0.006766 **
## factor(Count.ticket)7  -0.743931   0.695531  -1.070 0.284806
## factor(Count.ticket)8   2.045373   0.795210   2.572 0.010108 *
## factor(Count.ticket)11 -15.020208 576.063041  -0.026 0.979198
## EmbarkedQ              -0.105479   0.488249  -0.216 0.828960
## EmbarkedS              -0.520022   0.301341  -1.726 0.084403 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 792.97  on 593  degrees of freedom
## Residual deviance: 506.80  on 579  degrees of freedom
## AIC: 536.8
##
## Number of Fisher Scoring iterations: 14
```

Las variables **significativas son: clase, sexo, edad**. Algunos de los valores del recuento de billetes, como **6 billetes, 5 billetes también son significativas** para esta regresión y sus coeficientes indican que este tipo de billetes influyeron negativamente en la probabilidad de supervivencia.

Por otra parte, en los casos de personas que viajaban con un billete grupal de 8 personas, el coeficiente es positivo y la variable es significativa, entre estos grupos la porobabilidad de supervivencia aumentaba.

Como ya hemos visto en regresiones anteriores, la clase en la que viajaban los pasajeros es relevante para la regresión, viajar en segunda o tercera clases reduce la probabilidad de sobrevivir con respecto a viajar en primera. Lo mismo ocurre para pasajeros de sexo masculino, en el accidente ser hombre redujo la probabilidad de supervivencia. La edad de los pasajeros es otro factor que ha influido.

Más adelante crearemos un árbol de decisión y podremos visualizar los resultados obtenidos desde otra perspectiva.

El resultado del modelo indica que **cuando de dos a cuatro personas viajaban juntas, la probabilidad de sobrevivir aumentaba** (signo positivo del coeficiente).

Calculamos los ORS. Para las personas que viajaban con otra persona, otras dos o tres personas el "riesgo" de sobrevivir es mucho mayor.

Factor de protección:
* Clase 2
* Clase 3
* Sexo Masculino
* Edad
* Viaje entre 5, 6, 7
* Embarque en Q o S

Las personas que tenían estas características (una o varias) muy probablemente no sobrevivieron

Factor de riesgo:
-Viaje entre 2, 3, 4, 8 personas

Las personas con estas características (una o ambas) tenían más probabilidad de sobrevivir

```
exp(coefficients(model2))
```

```
##           (Intercept)       factor(Pclass)2       factor(Pclass)3               Sexmale
##          5.239394e+01          3.612700e-01          9.236422e-02          7.950853e-02
##                   Age factor(Count.ticket)2 factor(Count.ticket)3 factor(Count.ticket)4
##          9.622665e-01          1.235330e+00          1.837111e+00          2.335228e+00
##  factor(Count.ticket)5 factor(Count.ticket)6 factor(Count.ticket)7 factor(Count.ticket)8
##          2.461119e-01          3.189618e-02          4.752423e-01          7.732041e+00
## factor(Count.ticket)11             EmbarkedQ             EmbarkedS
##          2.997828e-07          8.998931e-01          5.945076e-01
```

Usamos el conjunto de validación para comprobar la precisión del modelo.

Con respecto al modelo anterior, la precisión ha mejorado un poco, ahora tenemos 81.1% frente al 80.8% anterior.

```
newdata2 <- validX[c("Pclass", "Sex", "Age", "Count.ticket", "Embarked")]
probabilities2 <- model2 %>% predict(newdata2, type = "response")

predicted.classes2 <- ifelse(probabilities2 > 0.5, 1, 0)
# Matriz de confusión
```

```r
conf.3 <- table(validy, predicted.classes2)
conf.3
```

```
##       predicted.classes2
## validy   0   1
##      0 162  23
##      1  33  79
```

```r
# Precisión

sum(diag(conf.3)) / sum(colSums(conf.3))
```

```
## [1] 0.8114478
```

5. –> Árboles de decisión CART

En primer lugar creamos un árbol de decisión para predecir la supervivencia de los pasajeros del Titanic tomando en cuenta las variables explicativas del conjunto de entrenamiento y la etiqueta de clase (Supervivencia o no).

Usamos los datos del conjunto train para entrenar el modelo y visualizamos el árbol y los datos del modelo.

```r
treeFit <- rpart(trainy~.,data=trainX,method ='class')
print(treeFit)
```

```
## n= 594
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 594 230 0 (0.61279461 0.38720539)
##    2) Sex=male 377  73 0 (0.80636605 0.19363395)
##      4) Age>=12.5 348  57 0 (0.83620690 0.16379310)
##        8) Unit.price< 26.33542 282  33 0 (0.88297872 0.11702128) *
##        9) Unit.price>=26.33542 66  24 0 (0.63636364 0.36363636)
##         18) Age>=36.5 42  10 0 (0.76190476 0.23809524) *
##         19) Age< 36.5 24  10 1 (0.41666667 0.58333333)
##           38) Unit.price>=32.48958 11   3 0 (0.72727273 0.27272727) *
##           39) Unit.price< 32.48958 13   2 1 (0.15384615 0.84615385) *
##      5) Age< 12.5 29  13 1 (0.44827586 0.55172414)
##       10) SibSp>=2 13   1 0 (0.92307692 0.07692308) *
##       11) SibSp< 2 16   1 1 (0.06250000 0.93750000) *
##    3) Sex=female 217  60 1 (0.27649770 0.72350230)
##      6) Pclass>=2.5 103  50 0 (0.51456311 0.48543689)
##       12) Count.ticket>=4.5 19   2 0 (0.89473684 0.10526316) *
##       13) Count.ticket< 4.5 84  36 1 (0.42857143 0.57142857)
##         26) Unit.price>=8.0396 18   5 0 (0.72222222 0.27777778) *
##         27) Unit.price< 8.0396 66  23 1 (0.34848485 0.65151515)
##           54) Age>=27.5 12   3 0 (0.75000000 0.25000000) *
##           55) Age< 27.5 54  14 1 (0.25925926 0.74074074) *
##      7) Pclass< 2.5 114   7 1 (0.06140351 0.93859649) *
```

El árbol ha realizado la primera particiones de los datos en base a la variable Sex. Los pasajeros de sexo masculino, mayores de 13 años tienen como clase por defecto la no supervivencia. De ellos sobreviven los que pagaron entre 26 y

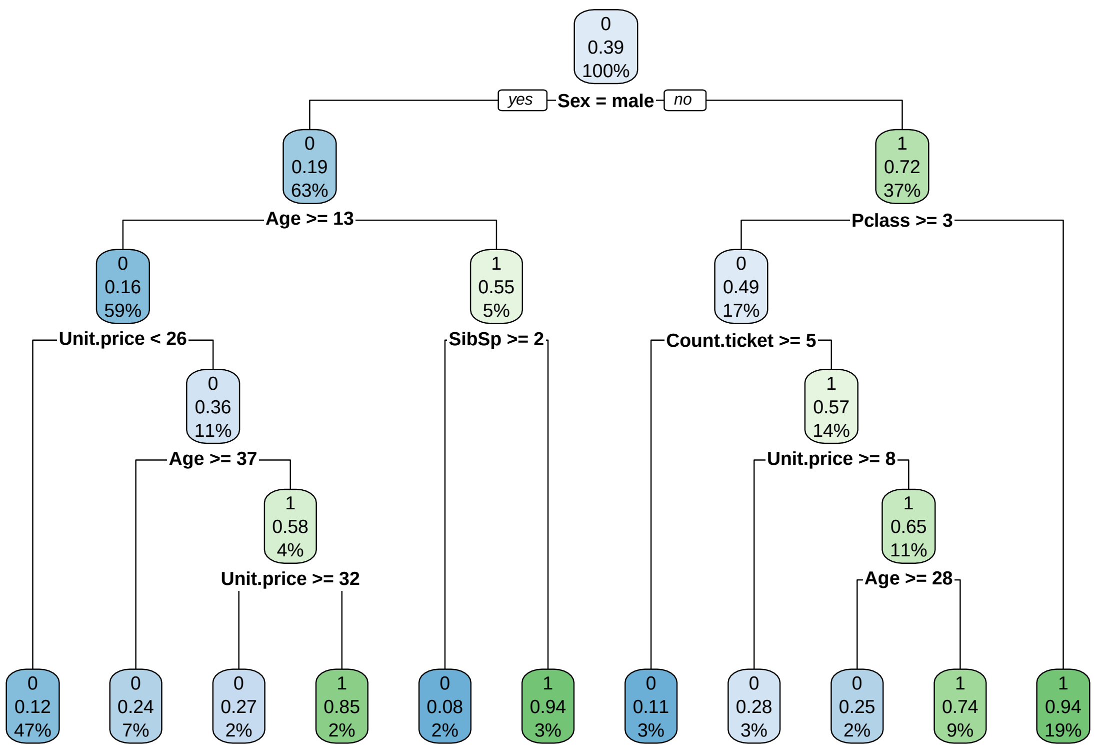
32 dólares por el billete y son menores de 37 años. El precio de entre 26 y 32 es bastante alto, se trata de pasajeros de primera o segunda clase.

La mayoría de menores de 13 años sobreviven, algo más si viajaban con 1 hermano o menos.

Por otra parte, los pasajeros de sexo Femenino en que viajaban en tercera clase sobrevivieron si viajaban con menos de 5 personas (count ticket), si pagaron menos de 8 dólares y si tenían edad menor de 28 años. Todos los demás se clasifican como no supervivientes. Las mujeres en otras clases distintas de la tercera sobrevivieron en su mayoría.

En el árbol podemos ver que se ha considerado variables bastante diferentes para clasificar a los hombres y a las mujeres. Para los hombres la edad ha sido un factor importante para la supervivencia, principalmente sobrevivieron los hombres menores de 14 años. En el caso de las mujeres el factor determinante ha sido la clase y las personas con las que viajaban.

```r
# Árbol resultante.
rpart.plot(treeFit)
```



Podemos ver la importancia de las variables para el árbol que hemos creado. **Las variables Sex, Unit.price, Age, Pclass, Count.ticket son las más importantes para decidir la supervivencia de los pasajeros**.

```r
treeFit$variable.importance
```

```
##          Sex   Unit.price          Age       Pclass Count.ticket        SibSp        Parch 
##    77.335802    48.179790    36.420760    30.815149    29.348424    20.743898    14.242406 
##     Embarked 
##     4.642216
```

Predecimos usando los datos de validación.

```
prediction <- ifelse(predict(treeFit,newdata=validX)[,1] > predict(treeFit,newdata=validX)[,2], 0, 1)
```

Con la matriz de confusión y el valor de Accuracy podemos ver que la predicción ha mejorado con respecto a la regresión logística. Conseguimos una mejor clasificación de los no supervivientes y una precisión del 85.9%.

```
# Matriz de confusión
cm.tree <- table(prediction, validy)
cm.tree
```

```
##           validy
## prediction   0   1
##          0 176  33
##          1   9  79
```

```
score.tree <- sum(diag(cm.tree))/sum(colSums(cm.tree))
score.tree
```

```
## [1] 0.8585859
```

Árbol de decisión usando cross validation

Podemos entrenar un árbol de decisión usando cross validation y la búsqueda en rejilla.

La validación cruzada consiste en dividir los datos en partes (parámetro number) y ejecutar el modelo varias veces usando como test set una de las particiones de los datos. El mejor modelo es aquel que tiene la mejor métrica y en nuestro caso la métrica a considerar es "Accuracy".

En el parámetro control que se le puede pasar a rpart podemos especificar una serie de condiciones, en nuestro caso especificamos **xval igual a 10** (número de iteraciones de la validación cruzada), **minbucket igual a 5** (número mínimo de observaciones en un nodo final), **maxdepth igual a 10** (profundidad máxima del árbol).

Imprimimos los datos del árbol y lo visualizamos.

En este caso el sexo del pasajero sigue siendo la primera variable que se usa para realizar la partición.

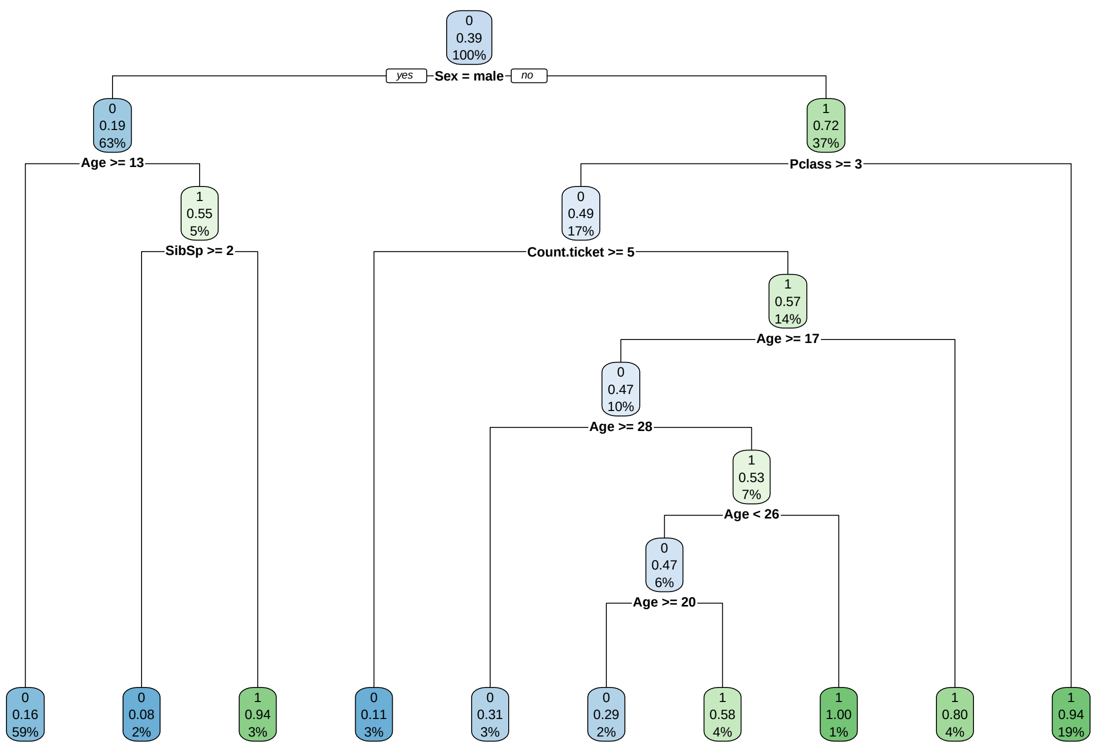
Los **hombres menores de 13 años sobreviven** si viajan con 1 hermano o ninguno. Para las mujeres sigue siendo determinante la clase: **las mujeres de clases diferentes a tercera clase sobreviven**. Las mujeres de tercera clase sobreviven si viajan con menos de 4 otras personas y si tienen una edad entre 17 y 28 años. Las mujeres de 17 años o menos sobreviven.

```
cvrpart  <- rpart(as.factor(trainy) ~  Age + Sex + Pclass + Embarked + Count.ticket + Parch + SibSp,
               data =data.frame(cbind(trainX,trainy)),
               method ='class',
               control = rpart.control(xval = 10, minbucket = 5, maxdepth=10, metric="Accuracy"))
print(cvrpart)
```

```
## n= 594
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 594 230 0 (0.61279461 0.38720539)
##     2) Sex=male 377  73 0 (0.80636605 0.19363395)
```

```
##       4) Age>=12.5 348  57 0 (0.83620690 0.16379310) *
##       5) Age< 12.5 29  13 1 (0.44827586 0.55172414)
##        10) SibSp>=2 13   1 0 (0.92307692 0.07692308) *
##        11) SibSp< 2 16   1 1 (0.06250000 0.93750000) *
##     3) Sex=female 217  60 1 (0.27649770 0.72350230)
##       6) Pclass>=2.5 103  50 0 (0.51456311 0.48543689)
##        12) Count.ticket>=4.5 19   2 0 (0.89473684 0.10526316) *
##        13) Count.ticket< 4.5 84  36 1 (0.42857143 0.57142857)
##          26) Age>=16.5 59  28 0 (0.52542373 0.47457627)
##            52) Age>=27.5 16   5 0 (0.68750000 0.31250000) *
##            53) Age< 27.5 43  20 1 (0.46511628 0.53488372)
##             106) Age< 25.5 38  18 0 (0.52631579 0.47368421)
##               212) Age>=19.5 14   4 0 (0.71428571 0.28571429) *
##               213) Age< 19.5 24  10 1 (0.41666667 0.58333333) *
##             107) Age>=25.5 5   0 1 (0.00000000 1.00000000) *
##          27) Age< 16.5 25   5 1 (0.20000000 0.80000000) *
##       7) Pclass< 2.5 114   7 1 (0.06140351 0.93859649) *
```

```
rpart.plot(cvrpart)
```



Visualizamos los errores cometidos por el árbol en cada nodo y buscamos el parámetro de complejidad donde rel error + xstd es menor que xerror, siendo xerror el cross validation error.

```
cvrpart$cptable
```

```
##           CP nsplit rel error    xerror       xstd
## 1 0.42173913      0 1.0000000 1.0000000 0.05161709
## 2 0.03260870      1 0.5782609 0.5782609 0.04417283
```

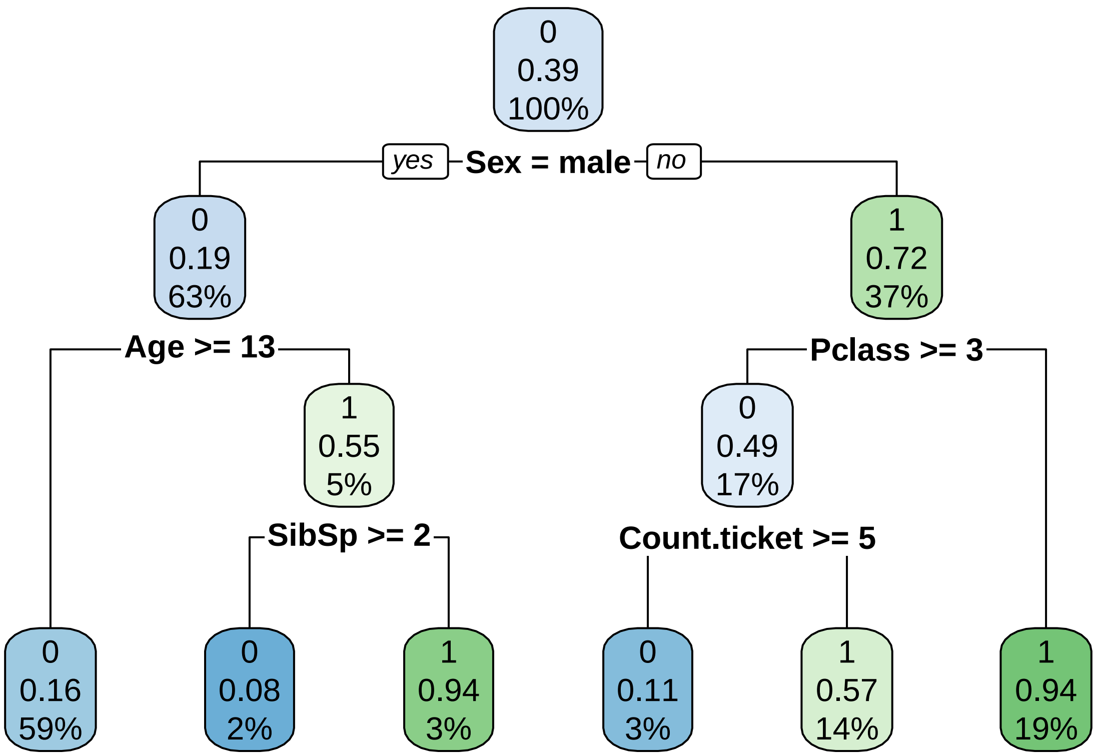
```
## 3 0.03043478      3 0.5130435 0.5434783 0.04319370
## 4 0.01304348      5 0.4521739 0.5260870 0.04267782
## 5 0.01000000      9 0.4000000 0.5434783 0.04319370
```

Podamos el árbol con el parámetro de complejidad **cp=0.03** y visualizamos el plot.

Para los hombres el árbol se comporta igual que el primer árbol que hemos creado. Para las mujeres simplifica las particiones. Las mujeres de tercera que no sobreviven son las que viajan con 4 o más otras personas. El resto sobrevive.

Probamos varios valores de cp para la poda y nos quedamos con cp=0.03. El resultado de usar el conjunto de validación en Accuracy es el mismo después de haber usado varios parámetros de complejidad. Nos quedamos con el árbol más simple.

```
prunedrpart <- prune(cvrpart, cp=0.03)
rpart.plot(prunedrpart)
```



Realizamos la predicción y vemos la matriz de confución. Obtenemos una precisión del 85.2%. Este valor es algo más bajo que en árbol anterior.

```
predictions <- ifelse(predict(prunedrpart,
                              newdata=validX)[,1]>0.50, "0", "1")
# Matriz de confusión
cm.cv.tree <- table(predictions, validy)
cm.cv.tree
```

```
##            validy
## predictions   0   1
##           0 171  30
```

```
##           1  14  82
```

```r
# Score
score.cv.tree <- sum(diag(cm.cv.tree))/sum(colSums(cm.cv.tree))
score.cv.tree
```

```
## [1] 0.8518519
```

Visualizamos la importancia de las variables para el árbol podado que acabamos de crear. **La variable de mayor importancia es el sexo de los pasajeros**, seguida de la clase, número de personas que viajaban con un billete.

```r
varImp(prunedrpart)
```

```
##                Overall
## Age          23.248145
## Count.ticket 50.039716
## Embarked     24.703340
## Parch         8.609076
## Pclass       59.847755
## Sex          77.335802
## SibSp        18.290021
```

6. –> Random Forest

Por último, aplicamos una validación cruzada de 5 folds para el Random Forest.

```r
cnt  <- trainControl(method="repeatedcv", number=5, repeats=3, savePredictions = TRUE)
model.rf <- train(as.factor(trainy)~.,
                  data=as.data.frame(cbind(trainX, trainy)),
                  method="rf",
                  metric="Accuracy",
                  trControl=cnt)
```

Calculamos la matriz de confusión y la precisión del modelo. La precisión es de 82%.

```r
# Matriz de confusión
cm.forest <- table(predict(model.rf,newdata=validX), validy)
cm.forest
```

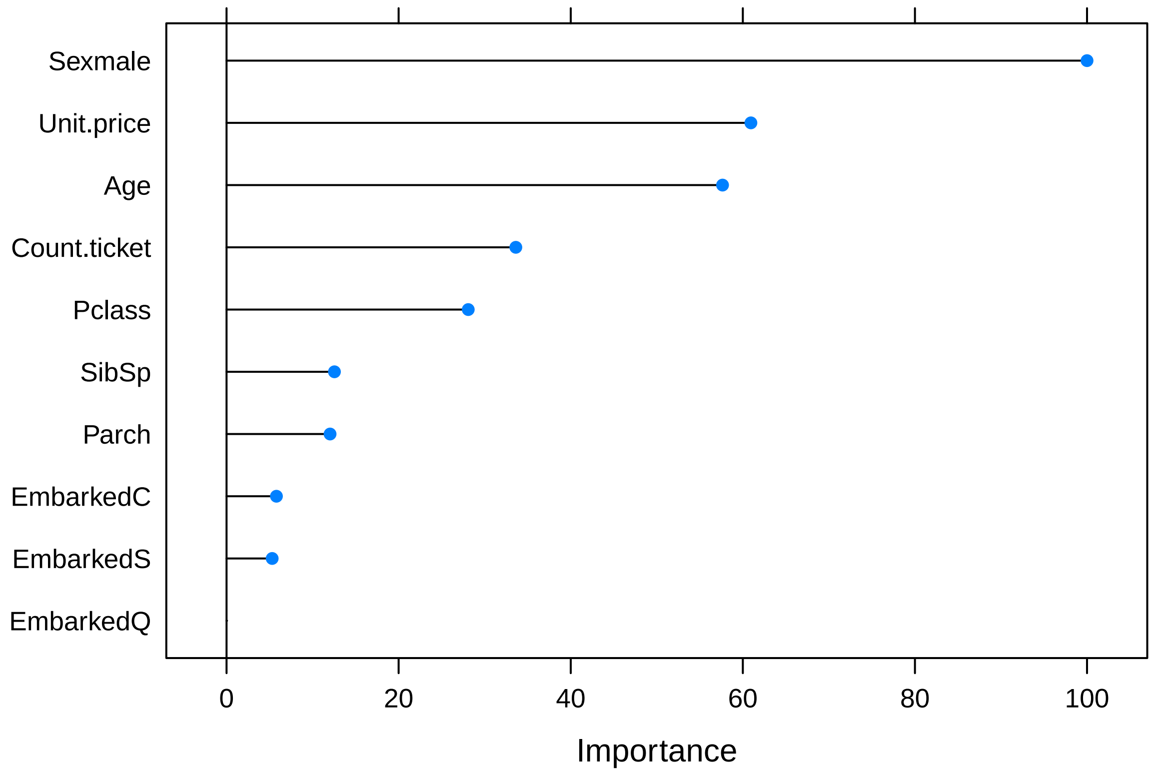
```
##    validy
##       0   1
##   0 170  37
##   1  15  75
```

```r
# Score
score.forest <- sum(diag(cm.forest))/sum(colSums(cm.forest))
score.forest
```

```
## [1] 0.8249158
```

Visualizamos la importancia de las variables. Entre todos los estimadores las variables más importantes han sido el sexo (ser hombre), el precio pagado por billete, la edad y el número de pasajeros que viajaban juntos.

```
plot(varImp(model.rf))
```



Representación de resultados

Una vez obtenido el juego de datos en su totalidad (train + test), y sabiendo que el propio juego de datos es una muestra solamente del juego de datos completo, es lógico realizar/calcular estadísticos en base a este juego de datos combinado train/test **para obtener un mayor conocimiento de las distribuciones y comportamientos de sus variables**.

Tabla resumen de las variables cualitativas (datos completos: train + test)

Realizaremos una tabla resumen con las frecuencias relativas y las frecuencias absolutas de las variables cualitativas. Creamos un dataframe auxiliar para generar nuestra tabla.

Calculamos, para todos los campos, la frecuencia relativa y absoluta a través del contaje dividido por el número total de filas del dataframe.

Creamos una tabla mediante la función kable().

Table 2: -TABLA RESUMEN DE LAS VARIABLES CUALITATIVAS- ‘ ‘ (Total observaciones: 1309. Suma de frecuencias relativas sin redondeo = 1.000)

Variable Cualitativa	Frecuencia Absoluta	Frecuencia Relativa
Clases Embarque		
1	323	0.247
2	277	0.212
3	709	0.542
Sexo		
female	466	0.356
male	843	0.644
Embarque		
C	270	0.206
Q	123	0.094
S	916	0.700

Tabla resumen de las variables cuantitativas (datos completos: train + test)

Realizaremos una tabla resumen con los **estadísticos principales de tendencia central y dispersión**, con **medidas robustas y no robustas**. Para ello, utilizaremos tres funciones que nos aportarán diferentes estadísticos a utilizar:

- describe()
- winsor.mean() (aplicaremos unos límites del 5 %)
- stat.desc()

Estas tres funciones nos darán diversos estadísticos que uniremos y ordenaremos en una única tabla para mostrar un completo resumen estadístico de las variables cuantitativas.

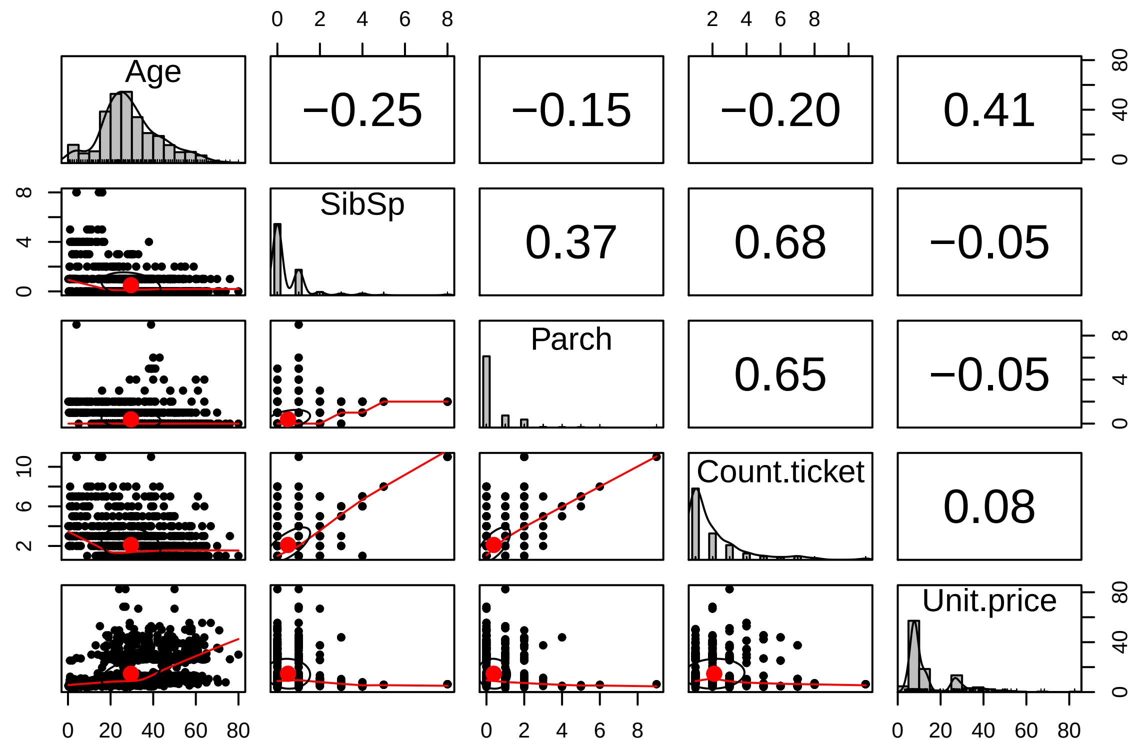
Finalmente, creamos una tabla mediante la función kable().

Resultado de la comprobación de la normalidad y homogeneidad de la varianza

Variable	Variable Dependiente	P-Value	Resultado
Age	Survived	0.15	Las varianzas en la edad de los pasajeros en función de la supervivencia son iguales
Unit.price	Survived	3.73e-14	Las varianzas en el precio unitario de los pasajeros en función de la supervivencia son diferentes
SibSp	Survived	2.2e-16	Las varianzas en SibSp de los pasajeros en función de la supervivencia son diferentes
Parch	Survived	0.1908	Las varianzas en Parch de los pasajeros en función de la supervivencian son iguales

Variable	Variable Dependiente	P-Value	Resultado
Count.ticket	Survived	3.459e-09	Las varianzas en la edad de los pasajeros en función de la supervivencia son diferentes

Gráfico de las correlaciones



Resultado de los Contrastes de Hipótesis

Variable	Variable	Tipo Contraste	P-value	Resultado
Sex(hombres/mujeres)	Survived	Proporción Supervivencia	2.174142e-14	H1:prop.h<prop.m
Age(menores/adultos)	Survived	Proporción Supervivencia	1	H0:prop.m > prop.a
Tipo Billete (único/conjunto)	Survived	Proporción Supervivencia	2.174142e-14	H1:prop.u<propc

Table 3: -TABLA RESUMEN DE LAS VARIABLES CUANTITATIVAS- ‘ ‘ (Total observaciones: 1309.)

	Age	SibSp	Parch	Count.ticket	Unit.price
Tendencia Central (medidas NO robustas)					
Mean	29.56	0.50	0.39	2.102	14.69
SE_Mean	0.38	0.03	0.02	0.049	0.33
CI_Mean_0.95	0.75	0.06	0.05	0.100	0.65
Tendencia Central (medidas robustas)					
Median	27.00	0.00	0.00	1.000	8.30
Trimmed_Median	28.93	0.27	0.18	1.689	12.40
Winsor_Mean_0.5	29.47	0.39	0.34	2.008	14.24
Dispersion (medidas NO robustas)					
Var	192.67	1.09	0.75	3.170	143.86
Coef_Var	0.47	2.09	2.25	0.850	0.82
St_Dev	13.88	1.04	0.87	1.780	11.99
Dispersion (medidas robustas)					
MAD	11.86	0.00	0.00	0.000	3.26
IQR	16.00	1.00	0.00	2.000	7.33
Información Adicional					
Number	1309.00	1309.00	1309.00	1309.000	1309.00
NA	0.00	0.00	0.00	0.000	0.00
Min	0.17	0.00	0.00	1.000	3.17
Max	80.00	8.00	9.00	11.000	82.51
Range	79.83	8.00	9.00	10.000	79.34

Resultado del estudio de independencia del Contraste Xi Cuadrado

Variable	Variable	Tipo Contraste	P-value	Resultado
Tipo Billete (único/conjunto)	Survived	Xi Cuadrado	7.375e-14	Las variables Tipo Billete y Survived están relacionadas
PClass (1/2/3)	Survived	Xi Cuadrado	2.2e-16	Las variables Survived y Pclass están relacionadas

Resultado de las regresiones logísticas

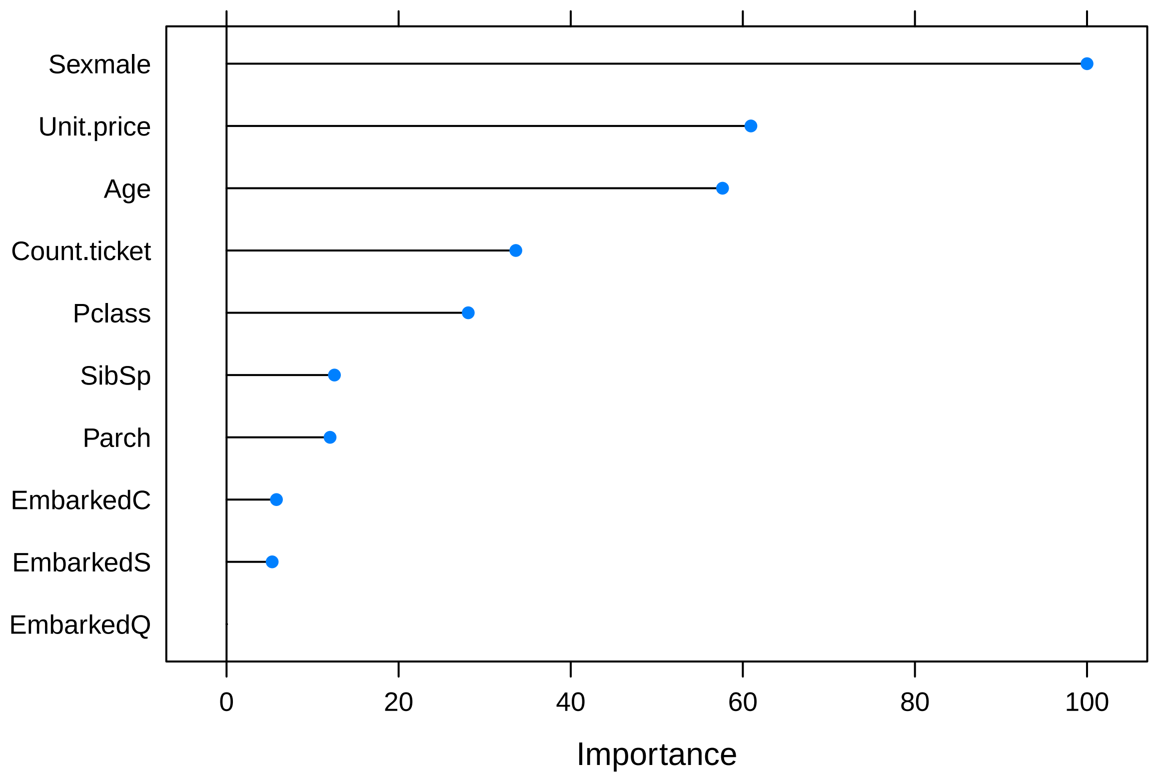
Variables Significativas	P-Value	Resultado
Pclass 2	0.003574	Reducción de probabilidad de sobrevivir
Pclass 3	1.28e-11	Reducción de probabilidad de sobrevivir
Sexmale	< 2e-16	Reducción de probabilidad de sobrevivir
Age	0.000114	Reducción de probabilidad de sobrevivir
Count.ticket 5	0.025810	Reducción de probabilidad de sobrevivir
Count.ticket 6	0.006766	Reducción de probabilidad de sobrevivir
Count.ticket 8	0.010108	Aumento de probabilidad de sobrevivir

```
##           0 171  30
##           1  14  82
```

```
## [1] 0.85
```

Resultado Gráfico Random Forest

Importancia de las variables en Random Forest:



Resultados de la validación del random forest

```
## [1] "Matriz de confusión y score para el bosque"
```

```
##    validy
##       0   1
##   0 170  37
##   1  15  75
```

```
## [1] 0.82
```

Resolución del problema

Al inicio de las práctica nos planteamos descubrir los posibles factores que pudieron influir en la supervivencia o no supervivencia de los pasajeros del Titanic.

Para poder responder a esta pregunta, hemos realizado una serie de pasos:

- Carga de los datos
- Integración de los datos
- Creación de nuevas variables y eliminación de variables innecesarias
- Imputación de valores nulos y tratamiento de outliers.

Hemos usamos los datos procesados para realizar una exploración visual de los datos, aplicar técnicas de estadística inferencial y modelos de aprendizaje automático.

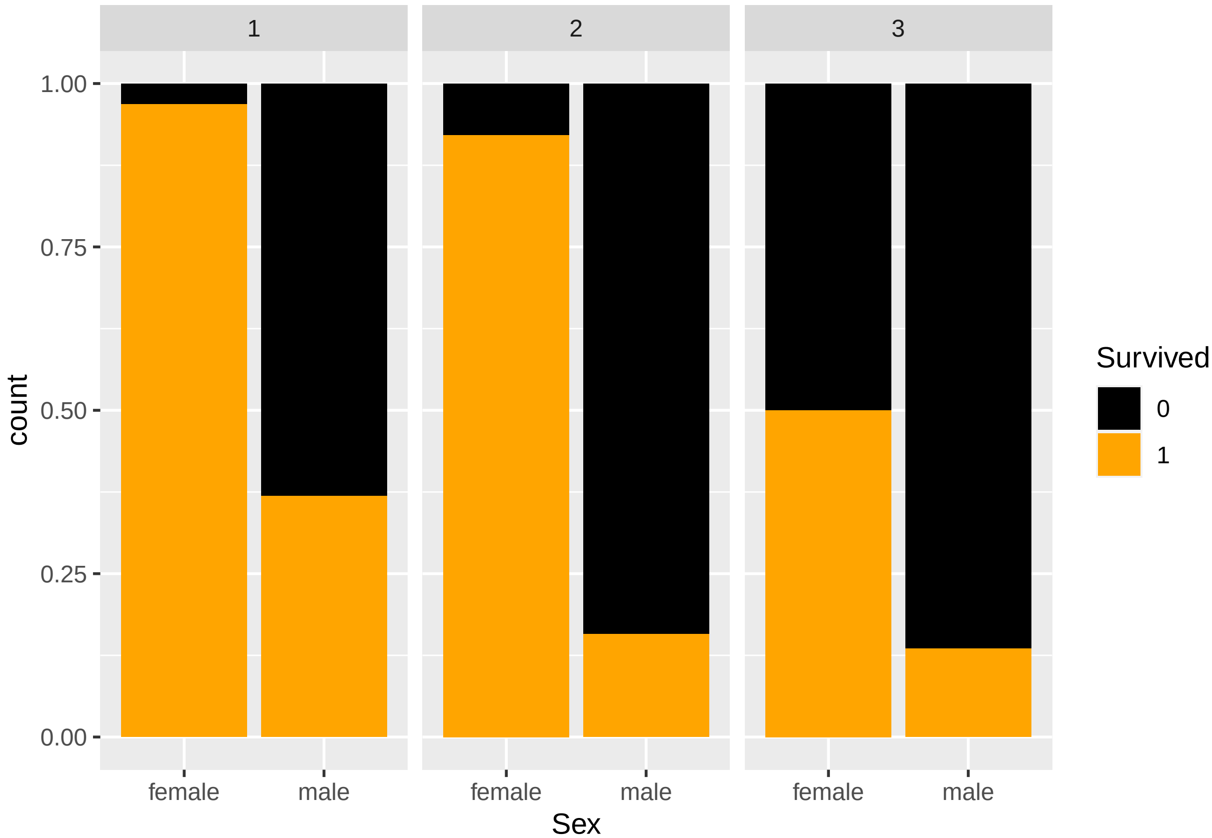
Las variables con las que trabajamos en nuestro dataset procesado, además de la variable dependiente *Survived** han sido:

- Pclass
- Sex
- Age
- SibSp
- Parch
- Embarked
- Count.ticket
- Unit.price, y
- Ticket_tipo

  siendo las últimas 3 variables calculadas.

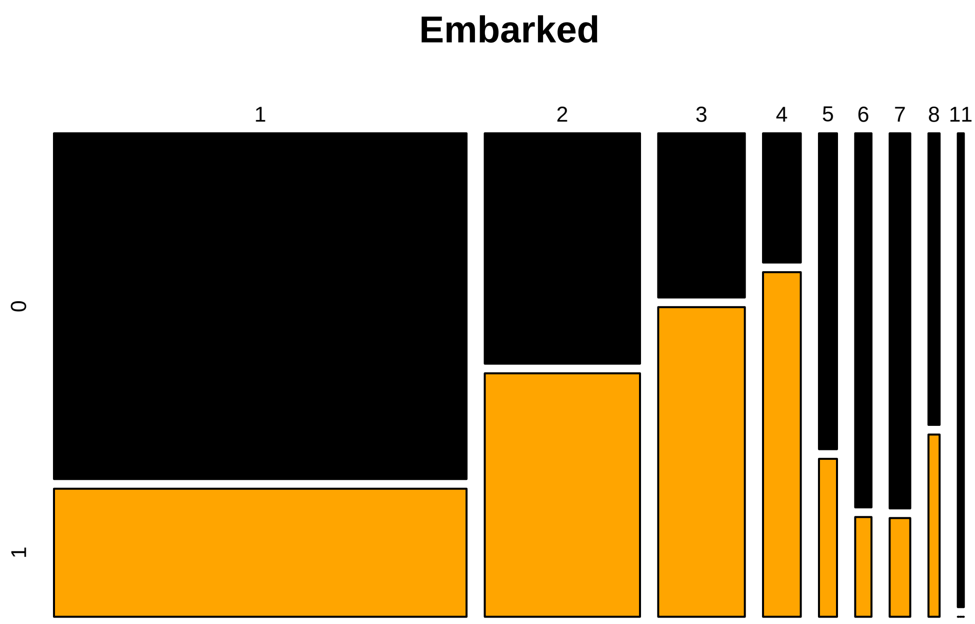
Conclusiones extraídas del análisis visual:

- Las personas que sobrevivieron eran algo más jóvenes (mediana), y pagaron más por sus billetes.
- La proporción de personas de primera clase que sobrevivió al accidente es mayor que la proporción de supervivientes en tercera y incluso en segunda clase.
- De los supervivientes en el accidente la mayoría de personas eran mujeres.
- En tercera clase la mitad de las mujeres no sobrevivieron.
- La proporción de mujeres supervivientes en primera y segunda clases es muy alta.
- En cuanto a la supervivencia de los hombres, los hombres de primera clase sobrevivieron en mayor proporción que los hombres de otras clases, sin embargo, la proporción de hombres supervivientes de tercera clase es menor que la de mujeres supervivientes de tercera clase.



Otro factor que nos ha parecido interesante es la compañía de los pasajeros en el crucero. La mayoría de los pasajeros viajaron solos, sin embargo, **había muchos billetes que tenían el mismo número para pasajeros distintos**.

Este número está relacionado con las variables Parch y SibSp, pero no indica el parentesco. Según las visualizaciones que realizamos, **las personas que viajaban con acompañantes sobrevivieron en mayor proporción que las personas que viajaban solas**.



Conclusiones del análisis inferencial:

En el análisis inferencial hemos realizado varios contrastes de hipótesis para confirmar o desmentir las conclusiones que obtuvimos durante el análisis visual de los datos.

Con el contraste sobre el sexo y el nivel de supervivencia hemos podido afirmar con un nivel de confianza del 95% que **la proporción de hombres que sobrevivieron al accidente era menor que la proporción de mujeres que sobrevivieron**. Asimismo **la proporción de menores de edad supervivientes es mayor que la de mayores de edad supervivientes**.

Conclusiones de regresión logística:

En las regresiones que hemos aplicado las principales conclusiones extraidas han sido que las variables **Sex**, **Pclass**, **Age** son significativa para predecir la supervivencia de los pasajeros del Titanic.

A partir de los coeficientes negativos en Pclass 2, Pclass3 y Sexmale deducimos que **las personas en segunda y tercera clase y los hombres tenían menos posibilidades de sobrevivir que las mujeres**.

La variable dummy **Count.ticket 6** también es significativa para el modelo final y su coeficiente indica que **para grupos de 6 viajeros las posibilidades de sobrevivir se reducían**.

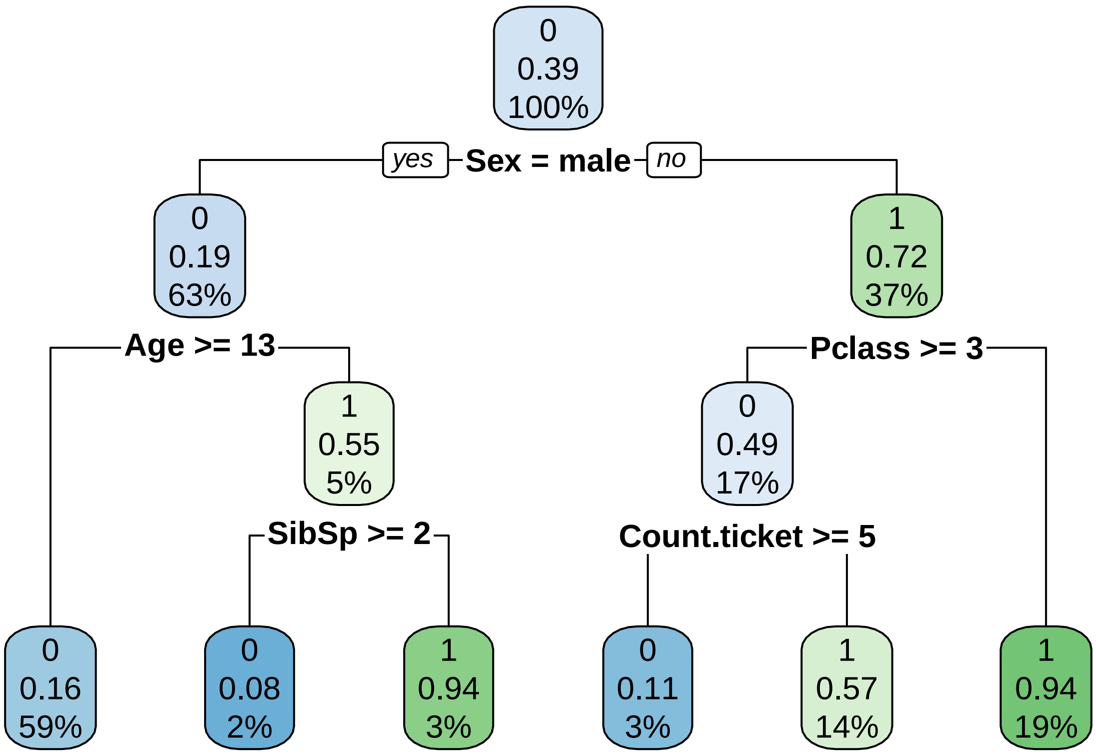
Conclusiones árbol de decisión:

Tras aplicar el árbol de decisión obtuvimos una mejor visión de los posibles motivos de supervivencia.

Para los hombres la supervivencia por defecto es 0 (no supervivencia), a no ser que sean menores de 13 años y viajen con un hermano o sin hermanos.

En el caso de las mujeres el factor determinante ha sido **la clase en la que viajaban**. La mayoría de mujeres que viajaban en primera y segunda clase sobrevivieron. De las mujeres que viajaban en tercera clase sobrevivieron las menores de 28 años, que viajaban con 4 o menos acompañantes.

En definitiva, para todos los modelos que creamos la variable Sexo es la más importante y otras variables de importancia han sido la edad, la clase, el número de personas que viajaban juntas.



CONCLUSIONES FINALES.

Tras realizar una inspección visual, aplicar métodos de estadística inferencial, regresiones y árboles de decisión podemos ver que entre todos los métodos hay consenso en cuanto a los factores que influyeron en la supervivencia o no de los pasajeros del Titanic.

- El sexo de las personas que viajaban tuvo mucha influencia en su supervivencia:
- En caso de los hombres, la supervivencia era mayor para niños.

- En el caso de las mujeres, no fue tan importante la edad como la clase en la que viajaban. Las mujeres de primera y segunda clase sobrevivieron en su mayoría.
- Las personas que viajaban con uno o más acompañantes también sobrevivieron en mayor proporción.

En definitiva, **el juego de datos nos ayuda a responder el problema/pregunta principal planteada en cuanto a los factores influyentes en la supervivencia del accidente del Titanic**

Los análisis mostrados reflejan qué y cómo pudieron influir los diferentes factores en la supervivencia de los viajeros.

Código y archivos CSV resultantes del proceso de limpieza

Código
El código empleado en la realización de este proyecto se ha mantenido visible para su potencial estudio y uso si fuera el caso excepto en el apartado de Representación de Resultados (ver archivo Rmd para ver la totalidad del código).

Archivos CSV procesados

Se han elaborado 3 archivos finales:

- train_processed.csv: Juego de datos de entrenamiento procesado con etiquetas
- test_processed.csv: Juego de datos de test procesado
- Titanic_global_sin_etiqueta.csv: Combinación de los dos archivos anteriores sin etiquetas.

Repositorio
Tanto el código como los resultados (output) y los archivos resultantes del proceso de limpieza están disponibles en el siguiente repositorio:

https://github.com/Carlos-Acosta/Limpieza-y-analisis-de-datos

Tabla de Contribuciones a la Práctica

Contribuciones	Firma
Investigación Previa	O. G. / C. A.
Redacción de las respuestas	O. G. / C. A.
Desarrollo código	O. G. / C. A.