# Tipología y ciclo de vida de los datos

PRÁCTICA 2: Limpieza y análisis de datos

Autores: Raúl Vicente Ferrer y Carmen Lobato Cassinello

Enero 2022

## Contents

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
# https://cran.r-project.org/web/packages/ggpubr/index.html
if (!require('ggpubr')) install.packages('ggpubr'); library('ggpubr')
# https://cran.r-project.org/web/packages/caTools/index.html
if (!require('caTools')) install.packages('caTools'); library('caTools')
# https://cran.r-project.org/web/packages/caret/index.html
if (!require('caret')) install.packages('caret'); library('caret')
# https://cran.r-project.org/web/packages/rpart/index.html
if (!require('rpart')) install.packages('rpart'); library('rpart')
# https://cran.r-project.org/web/packages/rpart.plot/index.html
if (!require('rpart.plot')) install.packages('rpart.plot'); library('rpart.plot')
```

## 1. Descripción del dataset.

### ¿Qué conjunto de datos se ha utilizado?

Para la realización de esta práctica se ha utilizado el conjunto de entrenamiento de *Titanic: Machine Learning from Disaster* (kaggle), el cual se puede encontrar en la siguiente URL. Se ha renombrado como "train_titanic_og.csv".

El conjunto de datos recoge información sobre los pasajeros a bordo del Titanic cuando el buque chocó contra el iceberg:

- PassengerId: contiene el identificador numérico del pasajero.
- Survived [0, 1]: describe si el pasajero sobrevivió o no al hundimiento del buque.
- Pclass [1, 2, 3]: describe la clase en la que viajaba el pasajero.
- Name: describe el nombre completo del pasajero.
- Sex [male, female]: describe el género del pasajero.
- Age: describe la edad del pasajero en años.
- SibSp: describe el número de hermanos/as o esposos/as del pasajero a bordo del buque.
- ParCh: descibe el número de padres o hijos del pasajero a bordo del buque.
- Ticket: describe el número de billete del pasajero.
- Fare: describe el precio del billete que llevaba el pasajero.
- Cabin: describe la cabina en la que viajaba el pasajero.
- Embarked [C, Q, S]: describe el puerto de embarque del pasajero.

Cada uno de los registros se corresponde con un único viajero.

Los datos del dataset son de tipo carácter o numérico, quedando excluidos otros tipos de variables.

### ¿Por qué es importante y qué pregunta pretende responder?

El conjunto tratado resulta relevante porque nos ayuda a comprender mejor el funcionamiento de la sociedad en 1912, permitiéndonos analizar la evolución hasta la época actual y encontrar potenciales puntos de mejora.

Los datos analizados pretenden dar respuesta a varias preguntas; entre ellas:

1. ¿Cómo se relacionan la clase, el género y la edad con la tasa de supervivencia?
2. ¿Cuántas personas viajaban solas? ¿Y con familia? ¿Cómo se relaciona esto con la tasa de supervivencia?
3. ¿Tiene algo que ver el puerto de embarque con la tasa de supervivencia?

## 2. Integración y selección de los datos de interés a analizar.

Como primer paso, realizamos la carga del conjunto a estudiar:

```
titanic_data <- read.csv("./train_titanic_og.csv", header=T, sep=",", stringsAsFactors = FALSE, fileEnc
```

Calcularemos ahora las dimensiones del conjunto y analizaremos el tipo de variable correspondiente con cada columna:

```
str(titanic_data)
```

```
## '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)"
##  $ 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  NA "C85" NA "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Observamos que el fichero contiene 891 registros, correspondientes a las 12 variables descritas anteriormente (apartado 1):

- Categóricas: Survived, Pclass, Sex, Embarked
- Numéricas discretas: PassengerId, SibSp, ParCh
- Numéricas continuas: Age, Fare
- Texto: Name, Ticket, Cabin

Para el estudio a realizar, no todas las variables serán necesarias. En concreto, las variables que no aportan información global no resultan relevantes, por lo que se eliminan: * PassengerId * Name * Ticket * Cabin

```
titanic_data <- select(titanic_data, -PassengerId, -Name, -Ticket, -Cabin)
```

Cambiamos el tipo de variables a Factor donde sea necesario:

```
titanic_data$Survived <- as.factor(titanic_data$Survived)
titanic_data$Pclass <- as.factor(titanic_data$Pclass)
titanic_data$Sex <- as.factor(titanic_data$Sex)
titanic_data$Embarked <- as.factor(titanic_data$Embarked)
summary(titanic_data)
```

```
##  Survived Pclass      Sex           Age            SibSp           Parch
##  0:549    1:216   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  1:342    2:184   male  :577   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##           3:491                Median :28.00   Median :0.000   Median :0.0000
##                                Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                                3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                                Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                                NA's   :177
##       Fare        Embarked
##  Min.   :  0.00   C   :168
##  1st Qu.:  7.91   Q   : 77
##  Median : 14.45   S   :644
##  Mean   : 32.20   NA's:  2
##  3rd Qu.: 31.00
##  Max.   :512.33
```

##

## 3. Limpieza de los datos.

### 3.1 Elementos vacíos.

Estudiaremos ahora si nuestros datos contienen elementos vacíos:

```
colSums(is.na(titanic_data))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0      177        0        0        0        2
```

Observamos que tenemos elementos vacíos en las columnas "Age" y "Embarked":

- Age: Nos faltan 177 datos de los 891 registros (19,9%). Realizaremos una inferencia en base al resto de variables para asociar una edad.
- Embarked: Nos faltan 2 datos de los 891 registros (0,2%). Evaluaremos si tiene sentido asumir que pertenecen a un determinado grupo (en función del resto de características); si no es posible categorizar a estos pasajeros, eliminaremos los registros.

**Age** Como primer paso para realizar nuestra inferencia, seleccionamos las variables que resultan de interés a la hora de predecir la edad de los pasajeros: su género, la clase en la que viajaban, y la supervivencia al hundimiento. Agrupamos los datos por Sex, Pclass y Survived y calculamos la mediana para cada grupo (esta medida es más resistente a outliers que la media):

```
grouped_titanic_data_median <- titanic_data %>% group_by(Sex, Pclass, Survived) %>% summarise(median = m
```

```
## `summarise()` has grouped output by 'Sex', 'Pclass'. You can override using the `.groups` argument.
```

```
grouped_titanic_data_median
```

```
## # A tibble: 12 x 4
## # Groups:   Sex, Pclass [6]
##    Sex    Pclass Survived median
##    <fct>  <fct>  <fct>     <dbl>
##  1 female 1      0          25
##  2 female 1      1          35
##  3 female 2      0          32.5
##  4 female 2      1          28
##  5 female 3      0          22
##  6 female 3      1          19
##  7 male   1      0          45.5
##  8 male   1      1          36
##  9 male   2      0          30.5
## 10 male   2      1           3
## 11 male   3      0          25
## 12 male   3      1          25
```

Rellenaremos ahora los valores perdidos con la medida que acabamos de calcular:

```
for (sex in c("female", "male")){
  for (class in c("1", "2", "3")){
    for (survived in c("0", "1")){
      titanic_data$Age[titanic_data$Sex == sex & titanic_data$Pclass == class & titanic_data$Survived ==
    }
  }
}
```

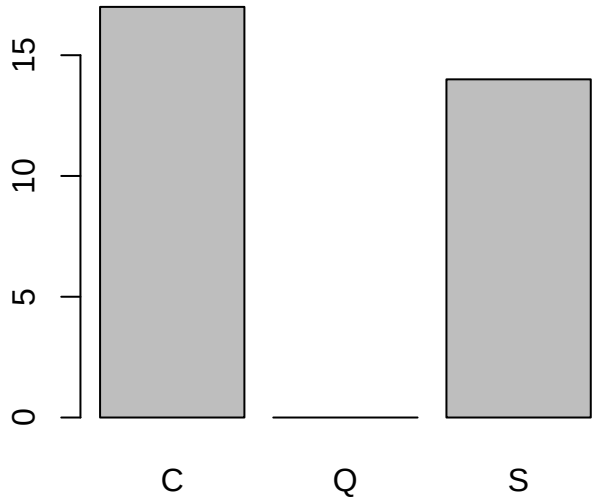
**Embarked** Observamos las características de los dos pasajeros cuyo puerto de embarque desconocemos:

```
titanic_data[is.na(titanic_data$Embarked),]
```

```
##     Survived Pclass    Sex Age SibSp Parch Fare Embarked
## 62         1      1 female  38     0     0   80     <NA>
## 830        1      1 female  62     0     0   80     <NA>
```

Vemos que ambas pasajeras eran mujeres de primera clase que viajaban solas y sobrevivieron al accidente. Observamos la distribución de puertos en pasajeras con las mismas características:

```
barplot(table(titanic_data$Embarked[titanic_data$Sex == "female" & titanic_data$Pclass == "1" & titanic_
```




Observamos que hay una proporción similar de pasajeras con estas características que embarcaron en Cherbourg y en Southampton. Como no podemos inferir en qué puerto embarcaron nuestras pasajeras, eliminaremos los dos registros:

```
titanic_data <- titanic_data[!is.na(titanic_data$Embarked),]
```

Comprobamos que no haya ningún valor vacío en el conjunto modificado:

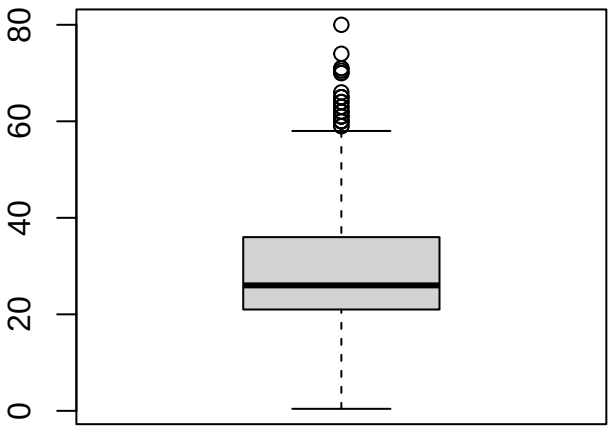
```
colSums(is.na(titanic_data))
```

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

### 3.2 Identificación y tratamiento de valores extremos.

Realizaremos un análisis de las variables numéricas continuas para detectar posibles valores extremos:

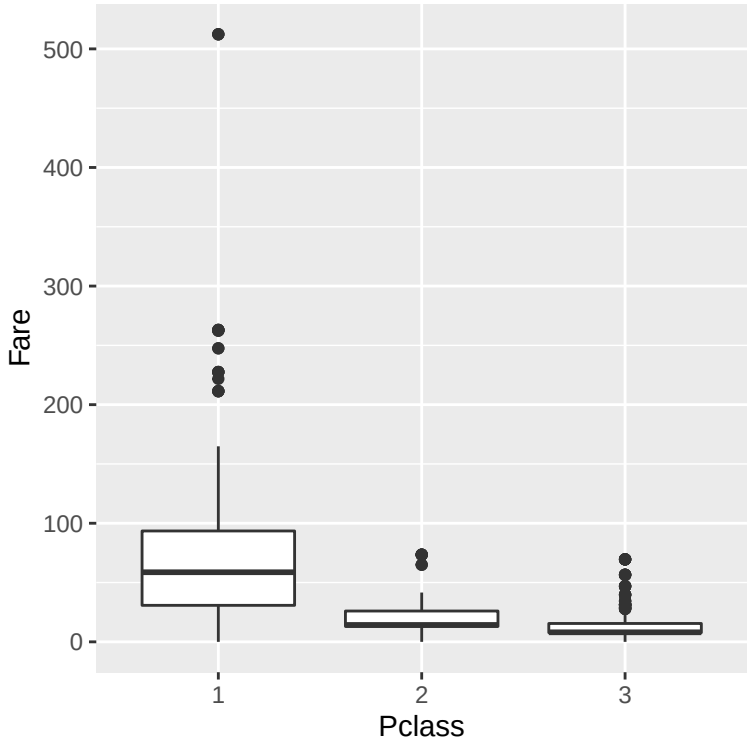
```
boxplot(titanic_data$Age)
```




**Age**

No se observa ningún valor atípico, todos los pasajeros tienen edades por debajo de 80 años.

```
ggplot(titanic_data, aes(x=Pclass, y=Fare)) + geom_boxplot()
```




**Fare**

Observamos diferentes outliers en función de la clase. Vamos a estudiarlas por separado:

- Primera clase

```
first_class <- titanic_data[titanic_data$Pclass == "1" & titanic_data$Fare > 200,]
first_class[order(first_class$Fare),]
```

```
##     Survived Pclass    Sex  Age SibSp Parch     Fare Embarked
## 690        1      1 female 15.0     0     1 211.3375        S
## 731        1      1 female 29.0     0     0 211.3375        S
## 780        1      1 female 43.0     0     1 211.3375        S
## 378        0      1   male 27.0     0     2 211.5000        C
## 528        0      1   male 45.5     0     0 221.7792        S
## 381        1      1 female 42.0     0     0 227.5250        C
## 558        0      1   male 45.5     0     0 227.5250        C
## 701        1      1 female 18.0     1     0 227.5250        C
## 717        1      1 female 38.0     0     0 227.5250        C
## 119        0      1   male 24.0     0     1 247.5208        C
## 300        1      1 female 50.0     0     1 247.5208        C
## 312        1      1 female 18.0     2     2 262.3750        C
## 743        1      1 female 21.0     2     2 262.3750        C
## 28         0      1   male 19.0     3     2 263.0000        S
## 89         1      1 female 23.0     3     2 263.0000        S
## 342        1      1 female 24.0     3     2 263.0000        S
## 439        0      1   male 64.0     1     4 263.0000        S
## 259        1      1 female 35.0     0     0 512.3292        C
## 680        1      1   male 36.0     0     1 512.3292        C
## 738        1      1   male 35.0     0     0 512.3292        C
```

- Segunda clase

```
second_class <- titanic_data[titanic_data$Pclass == "2" & titanic_data$Fare > 50,]
second_class[order(second_class$Fare),]
```

```
##     Survived Pclass    Sex Age SibSp Parch Fare Embarked
## 616        1      2 female  24     1     2 65.0        S
## 755        1      2 female  48     1     2 65.0        S
## 73         0      2   male  21     0     0 73.5        S
## 121        0      2   male  21     2     0 73.5        S
## 386        0      2   male  18     0     0 73.5        S
## 656        0      2   male  24     2     0 73.5        S
## 666        0      2   male  32     2     0 73.5        S
```

- Tercera clase

```
third_class <- titanic_data[titanic_data$Pclass == "3" & titanic_data$Fare > 25,]
third_class[order(third_class$Fare),]
```

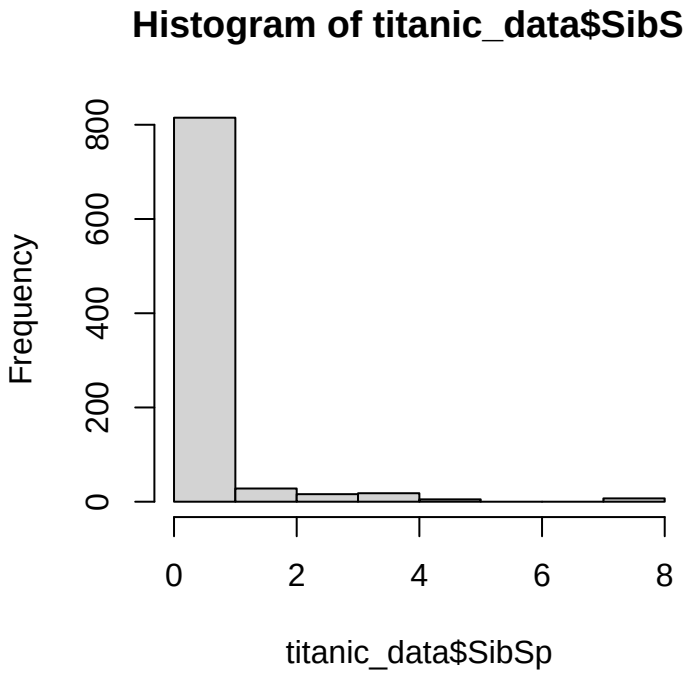
```
##     Survived Pclass    Sex Age SibSp Parch    Fare Embarked
## 177        0      3   male  25     3     1 25.4667        S
## 230        0      3 female  22     3     1 25.4667        S
## 410        0      3 female  22     3     1 25.4667        S
## 486        0      3 female  22     3     1 25.4667        S
## 64         0      3   male   4     3     2 27.9000        S
## 168        0      3 female  45     1     4 27.9000        S
## 361        0      3   male  40     1     4 27.9000        S
## 635        0      3 female   9     3     2 27.9000        S
## 643        0      3 female   2     3     2 27.9000        S
```

```
## 820        0      3   male  10     3     2 27.9000        S
## 17         0      3   male   2     4     1 29.1250        Q
## 172        0      3   male   4     4     1 29.1250        Q
## 279        0      3   male   7     4     1 29.1250        Q
## 788        0      3   male   8     4     1 29.1250        Q
## 886        0      3 female  39     0     5 29.1250        Q
## 14         0      3   male  39     1     5 31.2750        S
## 120        0      3 female   2     4     2 31.2750        S
## 542        0      3 female   9     4     2 31.2750        S
## 543        0      3 female  11     4     2 31.2750        S
## 611        0      3 female  39     1     5 31.2750        S
## 814        0      3 female   6     4     2 31.2750        S
## 851        0      3   male   4     4     2 31.2750        S
## 26         1      3 female  38     1     5 31.3875        S
## 183        0      3   male   9     4     2 31.3875        S
## 234        1      3 female   5     4     2 31.3875        S
## 262        1      3   male   3     4     2 31.3875        S
## 87         0      3   male  16     1     3 34.3750        S
## 148        0      3 female   9     2     2 34.3750        S
## 437        0      3 female  21     2     2 34.3750        S
## 737        0      3 female  48     1     3 34.3750        S
## 51         0      3   male   7     4     1 39.6875        S
## 165        0      3   male   1     4     1 39.6875        S
## 267        0      3   male  16     4     1 39.6875        S
## 639        0      3 female  41     0     5 39.6875        S
## 687        0      3   male  14     4     1 39.6875        S
## 825        0      3   male   2     4     1 39.6875        S
## 60         0      3   male  11     5     2 46.9000        S
## 72         0      3 female  16     5     2 46.9000        S
## 387        0      3   male   1     5     2 46.9000        S
## 481        0      3   male   9     5     2 46.9000        S
## 679        0      3 female  43     1     6 46.9000        S
## 684        0      3   male  14     5     2 46.9000        S
## 75         1      3   male  32     0     0 56.4958        S
## 170        0      3   male  28     0     0 56.4958        S
## 510        1      3   male  26     0     0 56.4958        S
## 644        1      3   male  25     0     0 56.4958        S
## 693        1      3   male  25     0     0 56.4958        S
## 827        0      3   male  25     0     0 56.4958        S
## 839        1      3   male  32     0     0 56.4958        S
## 160        0      3   male  25     8     2 69.5500        S
## 181        0      3 female  22     8     2 69.5500        S
## 202        0      3   male  25     8     2 69.5500        S
## 325        0      3   male  25     8     2 69.5500        S
## 793        0      3 female  22     8     2 69.5500        S
## 847        0      3   male  25     8     2 69.5500        S
## 864        0      3 female  22     8     2 69.5500        S
```

En todas las clases observamos que los rangos de precios se corresponden con el mismo puerto de embarque. Como consecuencia de esta consistencia en los precios, entendemos que son correctos.

Adicionalmente, analizaremos el histograma de las variables numéricas discretas para detectar posibles valores extremos:

```
hist(titanic_data$SibSp)
```




**SibSp**

Observamos que la mayoría de pasajeros viajan solos o con un número de hermanos + pareja entre 1 y 4. Estudiaremos a los viajeros que viajan con más acompañantes.

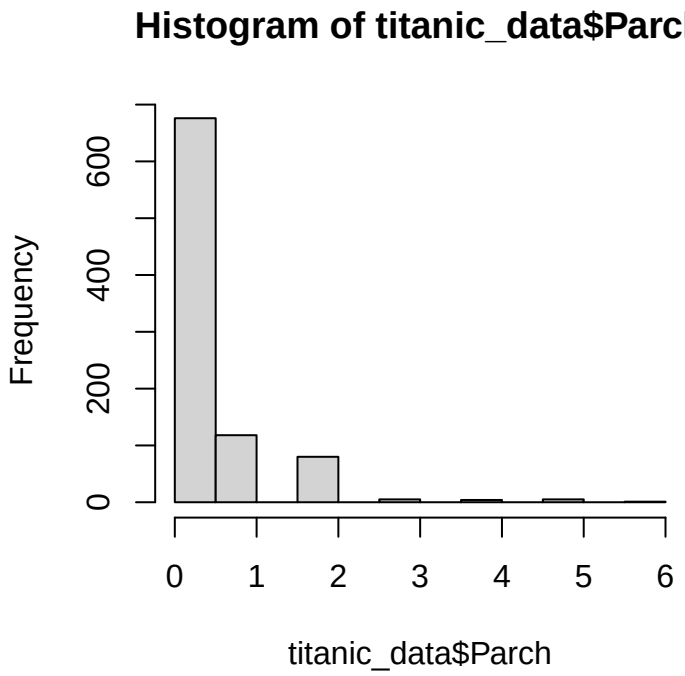
```
titanic_data[titanic_data$SibSp > "4",]
```

```
##     Survived Pclass    Sex Age SibSp Parch  Fare Embarked
## 60         0      3   male  11     5     2 46.90        S
## 72         0      3 female  16     5     2 46.90        S
## 160        0      3   male  25     8     2 69.55        S
## 181        0      3 female  22     8     2 69.55        S
## 202        0      3   male  25     8     2 69.55        S
## 325        0      3   male  25     8     2 69.55        S
## 387        0      3   male   1     5     2 46.90        S
## 481        0      3   male   9     5     2 46.90        S
## 684        0      3   male  14     5     2 46.90        S
## 793        0      3 female  22     8     2 69.55        S
## 847        0      3   male  25     8     2 69.55        S
## 864        0      3 female  22     8     2 69.55        S
```

Observamos que hay cinco pasajeros que viajan con cinco acompañantes, todos embarcando desde el mismo puerto y con una tarifa similar. Entendemos que los datos son correctos. Adicionalmente, observamos que hay siete pasajeros que viajan con ocho acompañantes; de nuevo, todos desde el mismo puerto y con una tarifa similar, por lo que asumimos que los datos son correctos.

Entendemos que los pasajeros restantes (uno para 5 acompañantes y dos para 8 acompañantes) están incluidos en el conjunto de test.

```
hist(titanic_data$Parch)
```




**Parch**

Observamos, de nuevo, que la mayoría de pasajeros viaja solos o con un número de padres + hijos entre uno y dos. Estudiaremos aquellos que viajan con más acompañantes.

```
titanic_data[titanic_data$Parch > "2",]
```

```
##     Survived Pclass    Sex Age SibSp Parch     Fare Embarked
## 14         0      3   male  39     1     5  31.2750        S
## 26         1      3 female  38     1     5  31.3875        S
## 87         0      3   male  16     1     3  34.3750        S
## 168        0      3 female  45     1     4  27.9000        S
## 361        0      3   male  40     1     4  27.9000        S
## 438        1      2 female  24     2     3  18.7500        S
## 439        0      1   male  64     1     4 263.0000        S
## 568        0      3 female  29     0     4  21.0750        S
## 611        0      3 female  39     1     5  31.2750        S
## 639        0      3 female  41     0     5  39.6875        S
## 679        0      3 female  43     1     6  46.9000        S
## 737        0      3 female  48     1     3  34.3750        S
## 775        1      2 female  54     1     3  23.0000        S
## 859        1      3 female  24     0     3  19.2583        C
## 886        0      3 female  39     0     5  29.1250        Q
```

Todas las edades son coherentes con tener por lo menos un hijo y/o viajar con padres (el varón de 16 estaría al límite, pero dada la época se entiende dentro del rango asumible). No se eliminará, por tanto, ningún registro.

Exportamos el archivo final a csv:

```
write.csv(titanic_data,"./train_titanic.csv", row.names = FALSE)
```

## 4. Análisis de los datos.

En este apartado dividiremos el conjunto de entrenamiento en varios grupos, los cuales nos ayudarán a responder las preguntas planteadas en el primer apartado.

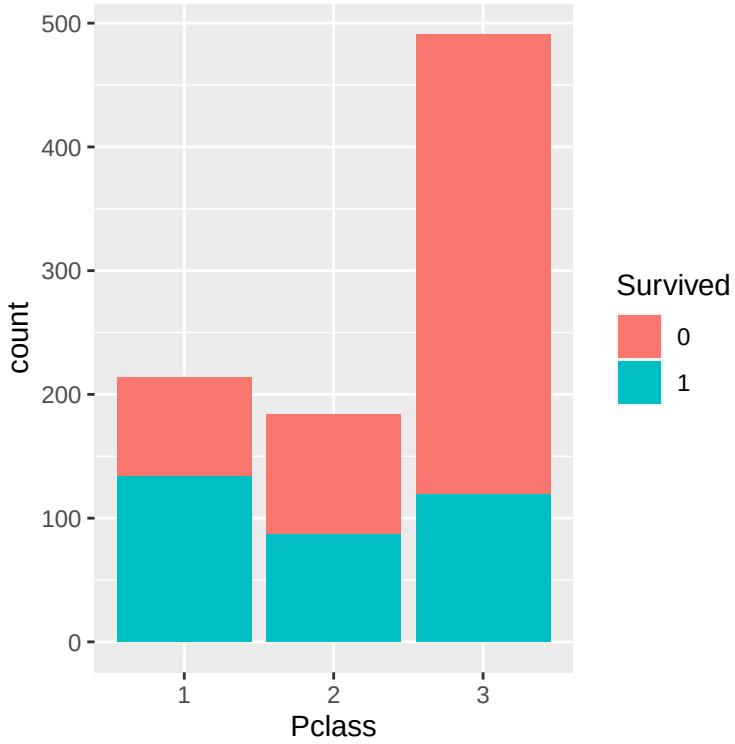
### 4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Los grupos que utilizaremos para nuestro análisis son los siguientes:

- Clase: analizaremos si la clase en la que viajaban los pasajeros influyó en su supervivencia.
- Género: estudiaremos si el género de los pasajeros afectó a la supervivencia.
- Rango de edad: estudiaremos si el rango de edad de los pasajeros tuvo algo que ver en su supervivencia. Dividiremos los datos en varios grupos en función del rango de edad.
- Acompañantes: analizaremos si las personas que viajaban solas tuvieron mayor o menor tasa de supervivencia que las que viajaban con acompañantes. Para esto crearemos una nueva variable que combine "SibSp" y "Parch".
- Puerto de embarque: estudiaremos si el puerto de embarque tuvo algún efecto en la supervivencia.

**Clase**

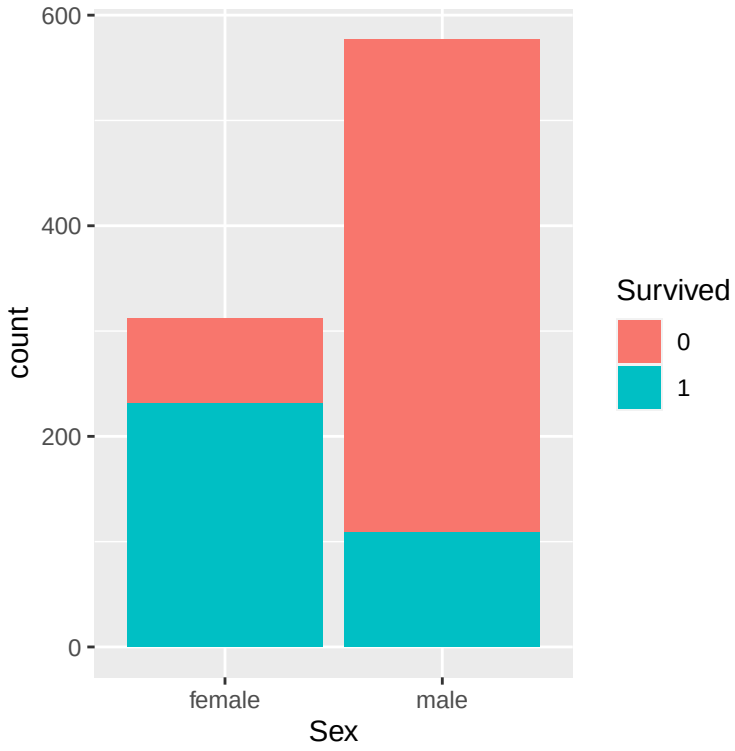
```
class <- ggplot(titanic_data, aes(x=Pclass, fill=Survived)) + geom_bar()
class
```




Observamos una supervivencia claramente superior en pasajeros que viajaban en primera clase con respecto a aquellos que viajaban en tercera clase. Parece que la clase en la que los pasajeros viajaban fue determinante para su supervivencia.

**Género**

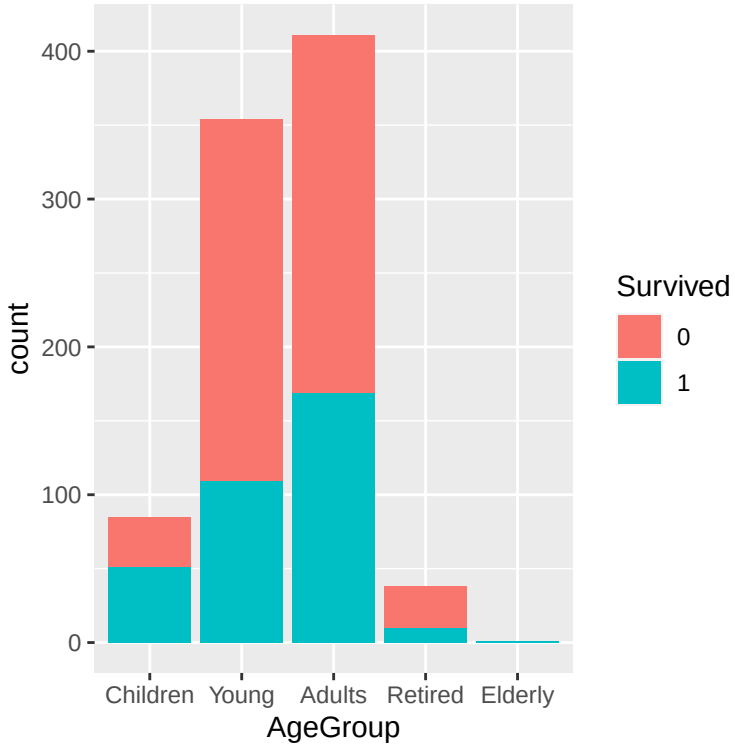
```
gender <- ggplot(titanic_data, aes(x=Sex, fill=Survived)) + geom_bar()
gender
```




Observamos que las mujeres tuvieron un porcentaje de supervivencia significativamente superior al de los hombres. Parece que esta variable fue determinante para la supervivencia de los pasajeros.

**Rango de edad**

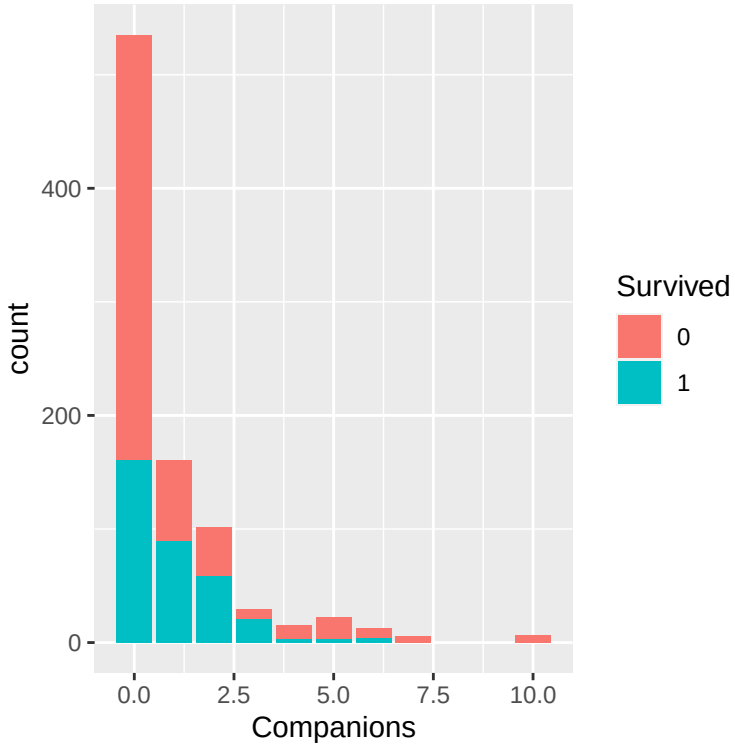
```
titanic_data$AgeGroup <- cut(titanic_data$Age, breaks = c(0,15,25,55,75,100), labels = c("Children","Yo
agegroup <- ggplot(titanic_data, aes(x=AgeGroup, fill=Survived)) + geom_bar()
agegroup
```




Observamos que el porcentaje de supervivencia disminuye a medida que aumenta la edad, salvo en los muy ancianos (poca muestra, todos sobrevivieron). Esto encaja con la frase "Mujeres y niños primero", conocida por aplicarse en este incidente. El rango de edad parece ser, por tanto, determinante para la supervivencia.

**Acompañantes**

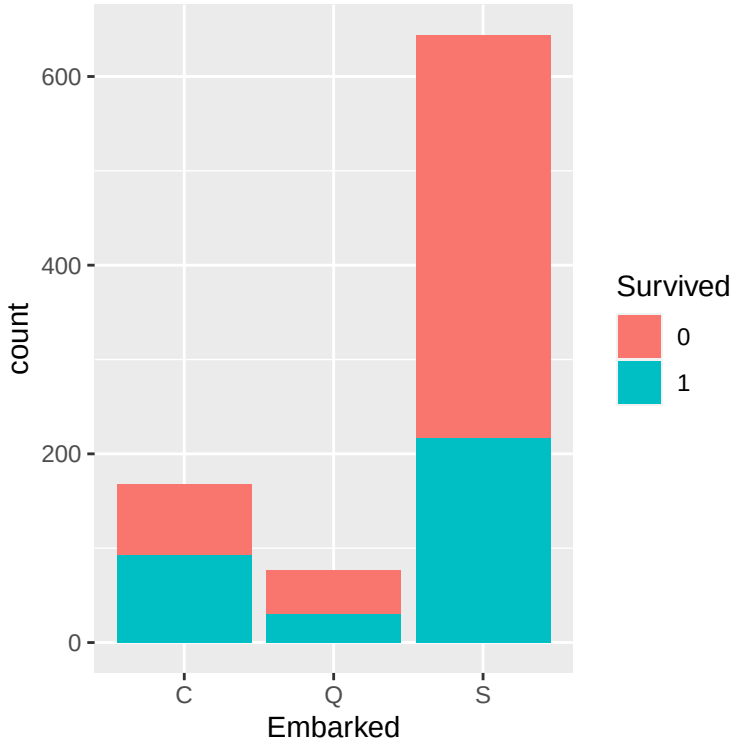
```
titanic_data$Companions <- titanic_data$SibSp + titanic_data$Parch
companions <- ggplot(titanic_data, aes(x=Companions, fill=Survived)) + geom_bar()
companions
```




No se observa una diferencia significativa en el porcentaje de supervivencia en función del número de acompañantes, por lo que parece que esta variable no fue relevante.

**Puerto de embarque**

```
port <- ggplot(titanic_data, aes(x=Embarked, fill=Survived)) + geom_bar()
port
```

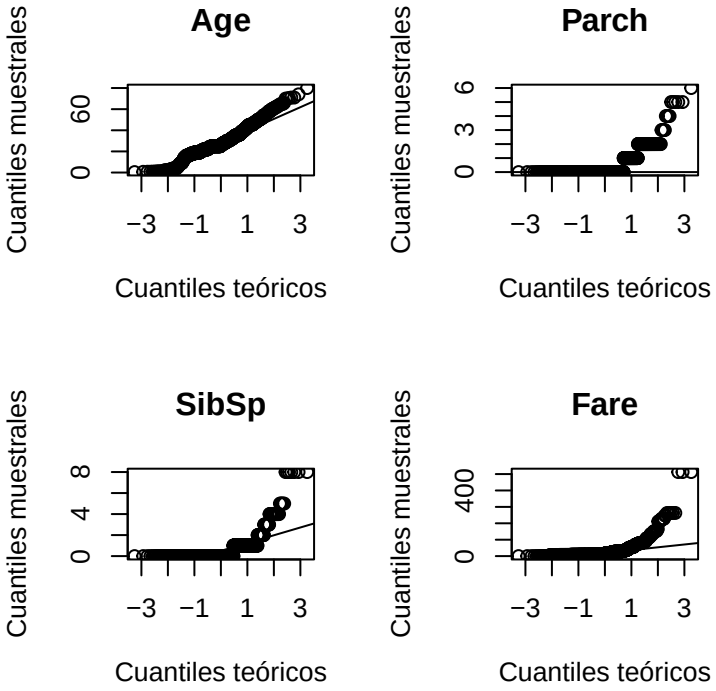



Observamos proporciones inferiores en la supervivencia de aquellos pasajeros embarcados en Southampton con respecto a los pasajeros embarcados en los otros dos puertos. Queda por estudiar si la diferencia se debe al efecto de alguna otra variable.

### 4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para comprobar la normalidad de las variables numéricas (Age, SibSp, Parch, y Fare) primero lo haremos de forma visual mediante gráficos Q-Q normal:

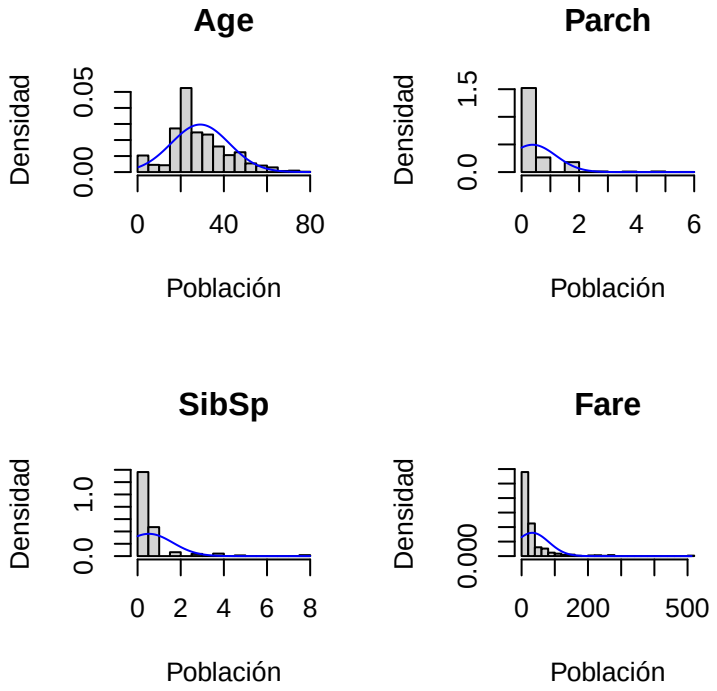
```
Conf2x2 = matrix(c(1:4), nrow=2, byrow=FALSE)
layout(Conf2x2)
qqnorm(titanic_data$Age, main = "Age", xlab = "Cuantiles teóricos", ylab = "Cuantiles muestrales")
qqline(titanic_data$Age)
qqnorm(titanic_data$SibSp, main = "SibSp", xlab = "Cuantiles teóricos", ylab = "Cuantiles muestrales")
qqline(titanic_data$SibSp)
qqnorm(titanic_data$Parch, main = "Parch", xlab = "Cuantiles teóricos", ylab = "Cuantiles muestrales")
qqline(titanic_data$Parch)
qqnorm(titanic_data$Fare, main = "Fare", xlab = "Cuantiles teóricos", ylab = "Cuantiles muestrales")
qqline(titanic_data$Fare)
```




La variable Age parece aproximarse ligeramente a una distribución normal, pero el resto de variables no.

Visualicemos con histogramas de densidad y una curva de distribución normal superpuesta para confirmar:

```
Conf2x2 = matrix(c(1:4), nrow=2, byrow=FALSE)
layout(Conf2x2)

hist(titanic_data$Age, probability = TRUE, main = "Age", xlab = "Población", ylab = "Densidad",  breaks=
x   <- seq(min(titanic_data$Age), max(titanic_data$Age), length = 1000)
y <- dnorm(x, mean(titanic_data$Age), sd(titanic_data$Age))
lines(x, y, col = "blue")

hist(titanic_data$SibSp, probability = TRUE, main = "SibSp", xlab = "Población", ylab = "Densidad",  br
x   <- seq(min(titanic_data$SibSp), max(titanic_data$SibSp), length = 1000)
y <- dnorm(x, mean(titanic_data$SibSp), sd(titanic_data$SibSp))
lines(x, y, col = "blue")

hist(titanic_data$Parch, probability = TRUE, main = "Parch", xlab = "Población", ylab = "Densidad",  br
x   <- seq(min(titanic_data$Parch), max(titanic_data$Parch), length = 1000)
y <- dnorm(x, mean(titanic_data$Parch), sd(titanic_data$Parch))
lines(x, y, col = "blue")

hist(titanic_data$Fare, probability = TRUE, main = "Fare", xlab = "Población", ylab = "Densidad",  breal
x   <- seq(min(titanic_data$Fare), max(titanic_data$Fare), length = 1000)
y <- dnorm(x, mean(titanic_data$Fare), sd(titanic_data$Fare))
lines(x, y, col = "blue")
```




Observando los histogramas se puede confirmar lo observado en los diagramas Q-Q, la variable Age se aproxima ligeramente a una normal, pero SibSp, Parch y Fare presentan un sesgo hacia la izquierda.

Para salir de dudas, aplicamos el Test Shapiro-Wilk:

- Age:

```
shapiro.test(titanic_data$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_data$Age
## W = 0.96732, p-value = 3.177e-13
```

La probabilidad, p-value, es menor que el valor de significación, 0.05, por lo que se acepta la hipótesis nula del test y se puede afirmar (al contrario de lo que parecían indicar las gráficas) que la variable Age no sigue una distribución normal.

- SibSp:

```
shapiro.test(titanic_data$SibSp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_data$SibSp
## W = 0.51353, p-value < 2.2e-16
```

- Parch:

```
shapiro.test(titanic_data$Parch)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  titanic_data$Parch
## W = 0.53345, p-value < 2.2e-16
```

- Fare:

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

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_data$Fare
## W = 0.5197, p-value < 2.2e-16
```

Con el mismo razonamiento que para la variable Age, confirmamos que SibSp, Parch y Fare tampoco siguen una distribución normal.

Para el estudio de la homocedasticidad usamos el test de Fligner-Killeen, que se trata de la alternativa no paramétrica, utilizada cuando los datos no cumplen con la condición de normalidad. Comprobamos si la varianza es significativamente distinta a la de Survived, con un nivel de significación del 5%.

- Age:

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 10.027, df = 1, p-value = 0.001543
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de homocedasticidad y se concluye que la variable Age presenta varianzas estadísticamente diferentes para los diferentes grupos de Survived.

- SibSp:

```
fligner.test(SibSp ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp by Survived
## Fligner-Killeen:med chi-squared = 1.3696, df = 1, p-value = 0.2419
```

Dado que p-value es superior al nivel de significancia ($> 0{,}05$), se acepta la hipótesis nula de homocedasticidad y se concluye que la variable SibSp no presenta varianzas estadísticamente diferentes para los diferentes grupos de Survived.

- Parch:

```
fligner.test(Parch ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Parch by Survived
## Fligner-Killeen:med chi-squared = 11.57, df = 1, p-value = 0.0006704
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de homocedasticidad y se concluye que la variable Parch presenta varianzas estadísticamente diferentes para los diferentes grupos de Survived.

- Fare:

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 94.7, df = 1, p-value < 2.2e-16
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de homocedasticidad y se concluye que la variable Fare presenta varianzas estadísticamente diferentes para los diferentes grupos de Survived.

### 4.3 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.

Teniendo en cuenta las preguntas que se han planteado inicialmente:

1. ¿Cómo se relacionan la clase, el género y la edad con la tasa de supervivencia?

Para dar respuesta a esta pregunta empleamos contrastes de hipótesis entre Pclass, Sex, AgeGroup y Survived:

#### Contraste de hipótesis entre Pclass y Survived

Emplearemos el test Chi-square, con las siguientes hipótesis:

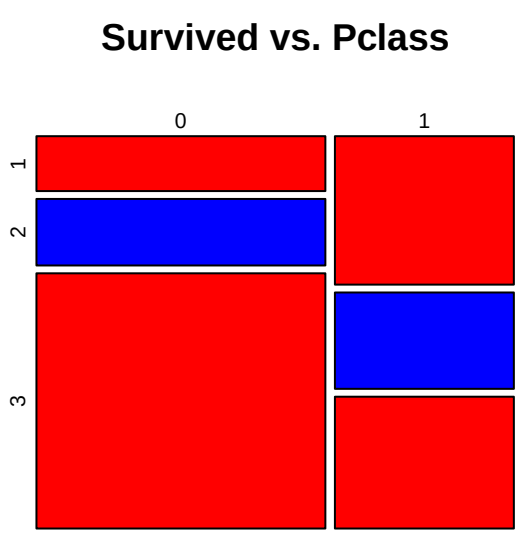
- Hipótesis nula: las variables son independientes.
- Hipótesis alternativa: las variables son dependientes.

Para comprobar la dependencia entre dos variables categóricas, se aplica el test chi-cuadrado. Pero antes es necesario crear la tabla de contingencia:

```
tabla1 <- table(titanic_data$Survived, titanic_data$Pclass)
tabla1
```

```
##
##       1   2   3
##   0  80  97 372
##   1 134  87 119
```

```
plot(tabla1, col = c("red", "blue"), main = "Survived vs. Pclass")
```




A continuación se aplica el test chi-square sobre la tabla de contingencia:

```
chisq.test(tabla1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla1
## X-squared = 100.98, df = 2, p-value < 2.2e-16
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de independencia y se concluye que la variable Survived depende de la variable Pclass, es decir, la supervivencia del pasajero depende de la clase del pasajero.

**Contraste de hipótesis entre AgeGroup y Survived**

Emplearemos el test Chi-square, con las siguientes hipótesis:

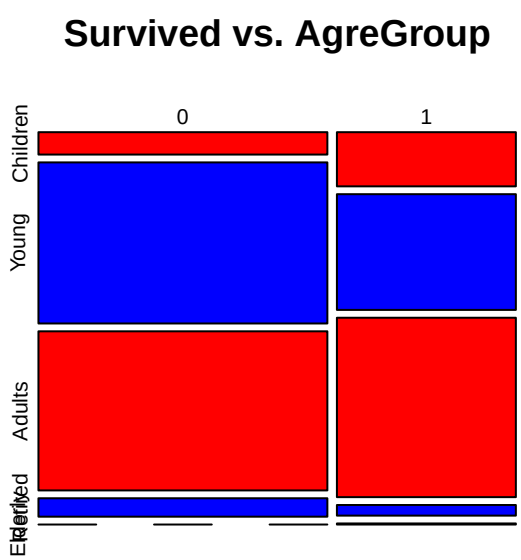
- Hipótesis nula: las variables son independientes.
- Hipótesis alternativa: las variables son dependientes.

Para comprobar la dependencia entre dos variables categóricas, se aplica el test chi-cuadrado. Pero antes es necesario crear la tabla de contingencia:

```
tabla2 <- table(titanic_data$Survived, titanic_data$AgeGroup)
tabla2
```

```
##
##     Children Young Adults Retired Elderly
##   0       34   245    242      28       0
##   1       51   109    169      10       1
```

```
plot(tabla2, col = c("red", "blue"), main = "Survived vs. AgreGroup")
```




A continuación se aplica el test chi-square sobre la tabla de contingencia:

```
chisq.test(tabla2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla2
## X-squared = 30.703, df = 4, p-value = 3.52e-06
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de independencia y se concluye que la variable Survived depende de la variable AgeGroup, es decir, la supervivencia del pasajero depende de la edad del pasajero.

**Contraste de hipótesis entre Sex y Survived**

Emplearemos el test Chi-square, con las siguientes hipótesis:

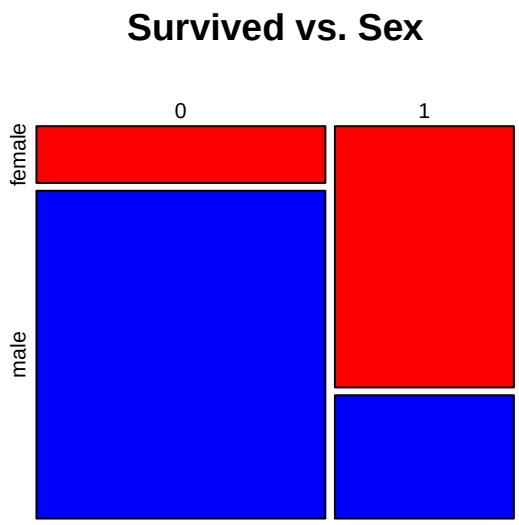
- Hipótesis nula: las variables son independientes.
- Hipótesis alternativa: las variables son dependientes.

Para comprobar la dependencia entre dos variables categóricas, se aplica el test chi-cuadrado. Pero antes es necesario crear la tabla de contingencia:

```
tabla3 <- table(titanic_data$Survived, titanic_data$Sex)
tabla3
```

```
##
##     female male
##   0     81  468
##   1    231  109
```

```
plot(tabla3, col = c("red", "blue"), main = "Survived vs. Sex")
```




A continuación se aplica el test chi-square sobre la tabla de contingencia:

```
chisq.test(tabla3)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla3
## X-squared = 258.43, df = 1, p-value < 2.2e-16
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de independencia y se concluye que la variable Survived depende de la variable Sex, es decir, la supervivencia del pasajero depende del sexo del pasajero.

2. ¿Cuántas personas viajaban solas? ¿Y con familia? ¿Cómo se relaciona esto con la tasa de supervivencia?

Calculamos primero cuántas personas viajaban solas y cuántas con familia:

```
length(titanic_data$Companions[titanic_data$Companions == 0]) # Viajan solos
```

```
## [1] 535
```

```
length(titanic_data$Companions[titanic_data$Companions > 0]) # Viajan acompañados
```

```
## [1] 354
```

Observamos que 535 pasajeros viajaban solos, mientras que 354 viajaban acompañados.

Para dar respuesta a la siguiente pregunta emplearemos un análisis regresión logística entre Companions y Survived:

```
rm_companions <- glm(Survived ~ Companions, family = binomial(link='logit'), data = titanic_data)
summary(rm_companions)
```

```
##
```

```
## Call:
## glm(formula = Survived ~ Companions, family = binomial(link = "logit"),
##     data = titanic_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0653  -0.9737  -0.9737   1.3856   1.3959
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.50021    0.07925  -6.312 2.75e-10 ***
## Companions   0.02306    0.04234   0.545    0.586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.8  on 888  degrees of freedom
## Residual deviance: 1182.5  on 887  degrees of freedom
## AIC: 1186.5
##
## Number of Fisher Scoring iterations: 4
```

Observamos que nuestra variable no es estadísticamente significativa (p-value superior a 0,05).

Ejecutamos el test de ANOVA para confirmar que la diferencia entre el modelo con y sin la variable no es significativa:

```
anova(rm_companions, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         888     1182.8
## Companions  1  0.29483       887     1182.5   0.5871
```

Observamos que la desviación de residuales es prácticamente la misma sin la variable (NULL, 1182.8) que con la variable (1182.5). Concluimos así que la variable Survived no depende de la variable Companions, es decir, la supervivencia del pasajero no depende del número de personas que lo acompañan.

3. ¿Tiene algo que ver el puerto de embarque con la tasa de supervivencia?

Para dar respuesta a esta pregunta emplearemos, de nuevo, un análisis regresión logística:

```
rm_embarked <- glm(Survived ~ Embarked, family = binomial(link='logit'), data = titanic_data)
summary(rm_embarked)
```

```
##
## Call:
## glm(formula = Survived ~ Embarked, family = binomial(link = "logit"),
##     data = titanic_data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2700  -0.9065  -0.9065   1.3730   1.4750
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2151     0.1552   1.386   0.1657
## EmbarkedQ    -0.6641     0.2805  -2.367   0.0179 *
## EmbarkedS    -0.8920     0.1762  -5.063 4.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.8  on 888  degrees of freedom
## Residual deviance: 1157.0  on 886  degrees of freedom
## AIC: 1163
##
## Number of Fisher Scoring iterations: 4
```

Observamos que embarcar en los puertos de Queenston o Southampton tiene una relación significativamente negativa con la supervivencia (p-value inferior a 0,05). Esto puede deberse a que un porcentaje mayor de pasajeros de tercera clase embarcara en estos puertos.

Ejecutamos el test de ANOVA para confirmar que la diferencia entre el modelo con y sin la variable es significativa:

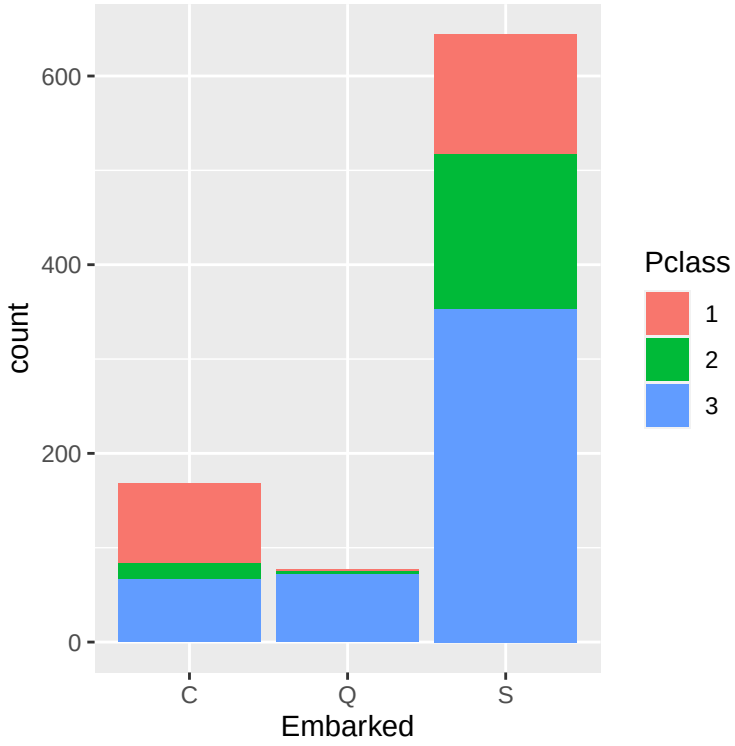
```
anova(rm_embarked, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       888     1182.8
## Embarked  2   25.865       886     1157.0 2.418e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La desviación de residuales varía considerablemente al añadir la variable Embarked con respecto a la observada sin la variable (1157 frente a 1182.8). Concluimos así que la variable Survived depende de la variable Embarked, es decir, la supervivencia del pasajero depende del puerto en el que embarcó.

Analizamos la relación entre el puerto de embarque y la clase para ver si había un porcentaje mayor de pasajeros de tercera clase en Queenston o Southampton con respecto a Cherbourg:

```
ggplot(titanic_data, aes(x=Embarked, fill=Pclass)) + geom_bar()
```




```
port_class <- table(titanic_data$Embarked, titanic_data$Pclass)
port_class
```

```
##
##        1   2   3
##   C  85  17  66
##   Q   2   3  72
##   S 127 164 353
```

Observamos que el porcentaje de viajeros de primera clase (mayor supervivencia) es de un 50,6% en Cherbourg, mientras que solo representan el 2,6% en Queenston y el 19,7% en Southampton. Con respecto a la tercera clase (menor supervivencia), observamos que representan un 39,3% en Cherbourg, un 93,5% en Queenston y un 54,8% en Southampton.

Parece haber una relación de dependencia entre el puerto de embarque y la clase de los pasajeros. Para confirmarla, aplicamos el test Chi-square sobre la tabla de contingencia:

```
chisq.test(port_class)
```

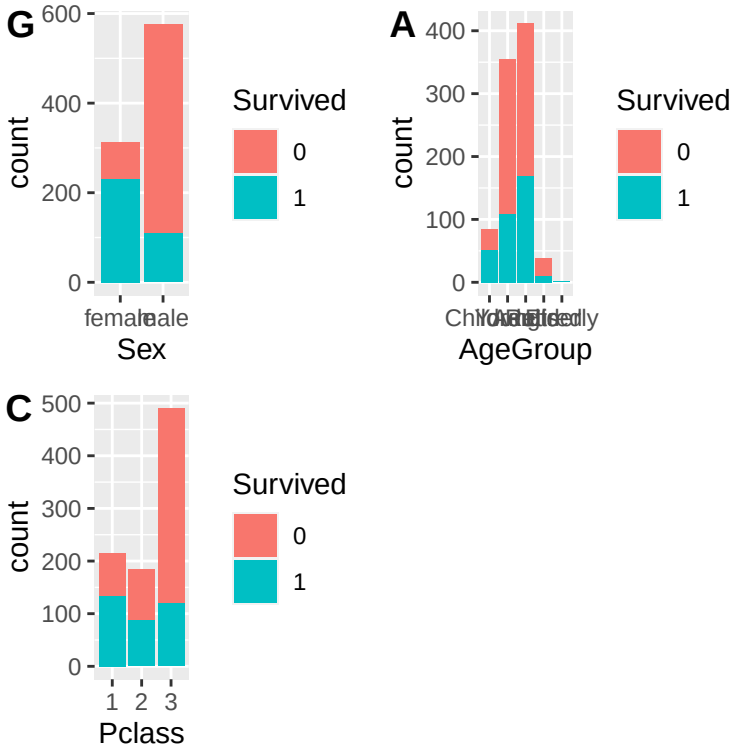
```
##
##  Pearson's Chi-squared test
##
## data:  port_class
## X-squared = 123.75, df = 4, p-value < 2.2e-16
```

Dado que p-value es inferior al nivel de significancia ($< 0{,}05$), se rechaza la hipótesis nula de independencia y se concluye que la variable Embarked depende de la variable Pclass, es decir, el puerto de embarque depende de la clase del pasajero.

## 5. Representación de los resultados a partir de tablas y gráficas.

A lo largo del estudio anterior, hemos encontrado una relación significativa entre la supervivencia de los viajeros y su género, edad y clase:

```
ggarrange(gender, agegroup, class, labels = c("G", "A", "C"), ncol = 2, nrow = 2)
```



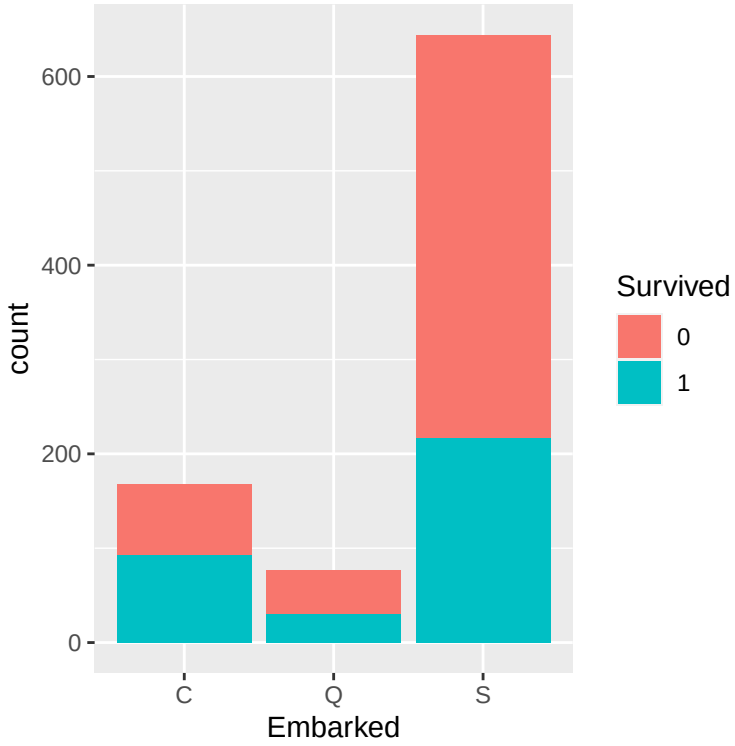

```
Category <- c("Pclass", "Agegroup","Sex")
Chisq <- c(chisq.test(tabla1)$statistic, chisq.test(tabla2)$statistic, chisq.test(tabla3)$statistic)
Pvalue <- c(chisq.test(tabla1)$p.value, chisq.test(tabla2)$p.value, chisq.test(tabla3)$p.value)

data.frame(Category, Chisq, Pvalue)
```

```
##   Category    Chisq       Pvalue
## 1   Pclass 100.9804 1.181362e-22
## 2 Agegroup  30.7028 3.519872e-06
## 3      Sex 258.4266 3.779910e-58
```

Observamos que las mujeres tuvieron una tasa de supervivencia mayor a los hombres, los jóvenes mayor a los adultos y los pasajeros de primera clase mayor a los que viajaban en segunda o tercera clase.

Adicionalmente, hemos concluido que el puerto de embarque de los pasajeros guarda relación con la clase en la que viajaban, lo que justifica la relación encontrada entre el puerto de embarque y la supervivencia:

```
port
```

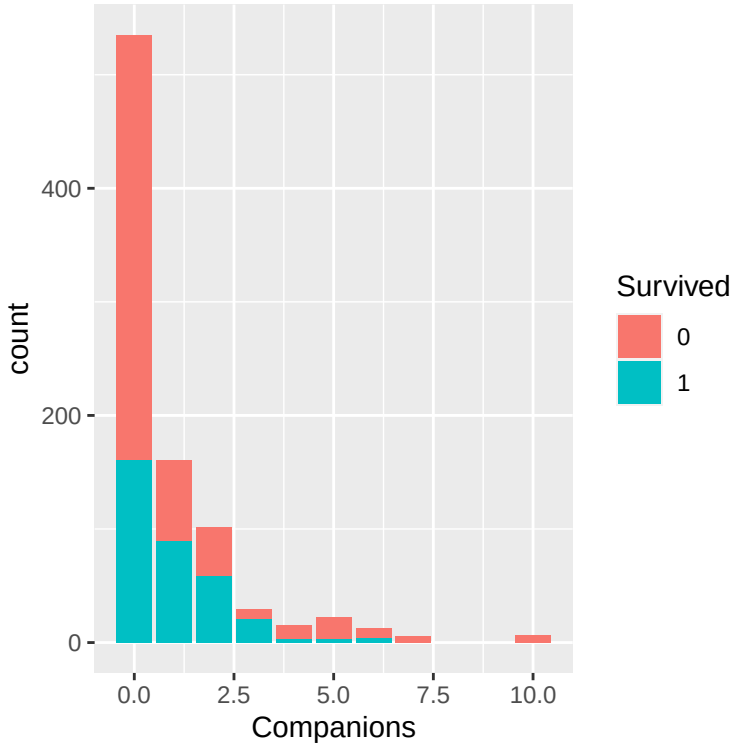



```
chisq.test(port_class)
```

```
##
##  Pearson's Chi-squared test
##
## data:  port_class
## X-squared = 123.75, df = 4, p-value < 2.2e-16
```

Finalmente, hemos concluido que el número de acompañantes no guarda relación con las probabilidades de supervivencia de los pasajeros a nivel individual:

```
companions
```




```
anova(rm_companions, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          888     1182.8
## Companions  1  0.29483       887     1182.5   0.5871
```

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

Para ser capaces de predecir la probabilidad de sobrevivir de un pasajero vamos a elaborar diferentes modelos predicción y a valorar su calidad.

Antes de nada creamos dividimos el dataset en train y test

```
set.seed(123)
split = sample.split(titanic_data$Survived, SplitRatio = 0.80)
titanic_train = subset(titanic_data, split == TRUE)
titanic_test = subset(titanic_data, split == FALSE)

str(titanic_train)
```

```
## 'data.frame':    711 obs. of  10 variables:
```

```
##  $ Survived  : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 2 2 2 1 ...
##  $ Pclass    : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 3 3 1 3 ...
##  $ Sex       : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 1 1 1 2 ...
##  $ Age       : num  22 38 26 35 35 25 27 4 58 20 ...
##  $ SibSp     : int  1 1 0 1 0 0 0 1 0 0 ...
##  $ Parch     : int  0 0 0 0 0 0 2 1 0 0 ...
##  $ Fare      : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked  : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 3 ...
##  $ AgeGroup  : Factor w/ 5 levels "Children","Young",..: 2 3 3 3 3 2 3 1 4 2 ...
##  $ Companions: int  1 1 0 1 0 0 2 2 0 0 ...
```

```
str(titanic_test)
```

```
## 'data.frame':    178 obs. of  10 variables:
##  $ Survived  : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 1 1 2 ...
##  $ Pclass    : Factor w/ 3 levels "1","2","3": 1 3 2 3 2 1 1 3 3 3 ...
##  $ Sex       : Factor w/ 2 levels "female","male": 2 2 1 1 2 2 2 2 1 1 ...
##  $ Age       : num  54 2 14 14 35 28 40 21 18 14 ...
##  $ SibSp     : int  0 3 1 0 0 0 0 0 2 1 ...
##  $ Parch     : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Fare      : num  51.86 21.07 30.07 7.85 26 ...
##  $ Embarked  : Factor w/ 3 levels "C","Q","S": 3 3 1 3 3 3 1 3 3 1 ...
##  $ AgeGroup  : Factor w/ 5 levels "Children","Young",..: 3 1 1 1 3 3 3 2 2 1 ...
##  $ Companions: int  0 4 1 0 0 0 0 0 2 1 ...
```

- Regresión logística: variable dependiente Survived y variable explicativa Pclass

```
logit_model_1 <- glm( formula = Survived ~ Pclass, data = titanic_train, family = binomial)
summary(logit_model_1)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass, family = binomial, data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4074  -0.7540  -0.7540   0.9636   1.6713
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5261     0.1564   3.363 0.000772 ***
## Pclass2      -0.7432     0.2276  -3.265 0.001096 **
## Pclass3      -1.6385     0.1957  -8.370  < 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: 946.06  on 710  degrees of freedom
## Residual deviance: 868.50  on 708  degrees of freedom
## AIC: 874.5
##
## Number of Fisher Scoring iterations: 4
```

- Regresión logística: variable dependiente Survived y variables explicativas Pclass y Sex

```r
logit_model_2 <- glm( formula = Survived ~ Pclass + Sex, data = titanic_train, family = binomial)
summary(logit_model_2)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex, family = binomial, data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1745  -0.7031  -0.4724   0.6898   2.1205
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.2654     0.2420   9.362  < 2e-16 ***
## Pclass2      -0.9508     0.2717  -3.500 0.000466 ***
## Pclass3      -1.8161     0.2347  -7.737 1.01e-14 ***
## Sexmale      -2.5861     0.2037 -12.697  < 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: 946.06  on 710  degrees of freedom
## Residual deviance: 671.22  on 707  degrees of freedom
## AIC: 679.22
##
## Number of Fisher Scoring iterations: 4
```

- Regresión logística: variable dependiente Survived y variables explicativas Pclass, Sex y Age

```r
logit_model_3 <- glm( formula = Survived ~ Pclass + Sex + Age, data = titanic_train, family = binomial)
summary(logit_model_3)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = binomial,
##     data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7457  -0.6543  -0.4445   0.6816   2.4697
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.829263   0.419880   9.120  < 2e-16 ***
## Pclass2     -1.340450   0.291409  -4.600 4.23e-06 ***
## Pclass3     -2.451409   0.280195  -8.749  < 2e-16 ***
## Sexmale     -2.506926   0.207984 -12.053  < 2e-16 ***
## Age         -0.041605   0.008485  -4.903 9.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 946.06  on 710  degrees of freedom
```

```
## Residual deviance: 645.21  on 706  degrees of freedom
## AIC: 655.21
##
## Number of Fisher Scoring iterations: 5
```

- Regresión logística: variable dependiente Survived y variables explicativas Pclass, Sex, Age y Embarked

```
logit_model_4 <- glm(formula=Survived ~ Pclass + Sex + Age + Embarked, data = titanic_train, family = bi
summary(logit_model_4)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Embarked, family = binomial,
##     data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6209  -0.6417  -0.4320   0.6605   2.5137
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.132840   0.447995   9.225  < 2e-16 ***
## Pclass2     -1.131826   0.301578  -3.753 0.000175 ***
## Pclass3     -2.335849   0.289471  -8.069 7.07e-16 ***
## Sexmale     -2.480698   0.209786 -11.825  < 2e-16 ***
## Age         -0.039562   0.008483  -4.664 3.10e-06 ***
## EmbarkedQ   -0.147258   0.431186  -0.342 0.732714
## EmbarkedS   -0.651853   0.261531  -2.492 0.012687 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 946.06  on 710  degrees of freedom
## Residual deviance: 637.98  on 704  degrees of freedom
## AIC: 651.98
##
## Number of Fisher Scoring iterations: 5
```

Observando el AIC de cada modelo, se ve que cada variable que se añade mejora el modelo, aunque hay que destacar que la mejora provocada por Embarked es casi imperceptible, por lo que nos quedamos con el modelo sin esta variable.

Los resultados obtenidos permiten responder al problema planteado: la supervivencia de los pasajeros dependía de su género, edad y clase.

Estudiemos la bondad de nuestro modelo:

```
titanic_test$Survived_predict_regression = round(predict.glm(logit_model_3, titanic_test, type = 'respon

confusionMatrix(as.factor(titanic_test$Survived), as.factor(titanic_test$Survived_predict_regression))
```

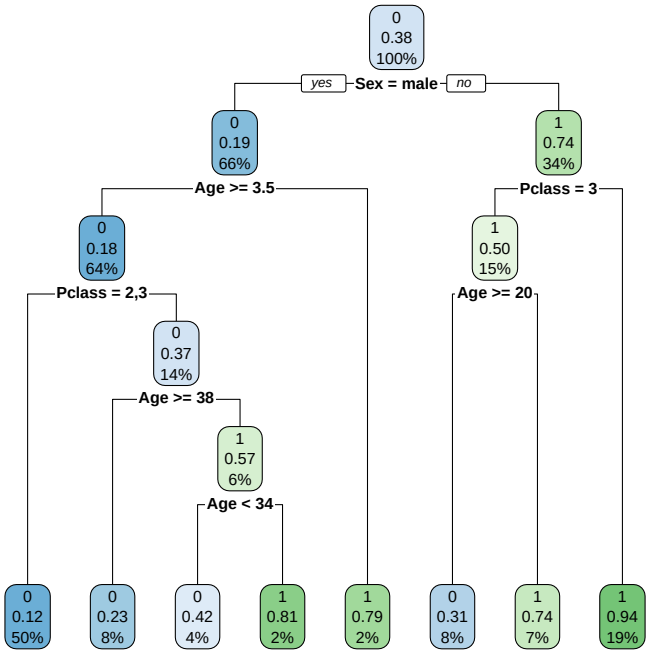
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 91 19
```

```
##          1 16 52
##
##                Accuracy : 0.8034
##                  95% CI : (0.7373, 0.8591)
##     No Information Rate : 0.6011
##     P-Value [Acc > NIR] : 6.444e-09
##
##                   Kappa : 0.587
##
##  Mcnemar's Test P-Value : 0.7353
##
##             Sensitivity : 0.8505
##             Specificity : 0.7324
##          Pos Pred Value : 0.8273
##          Neg Pred Value : 0.7647
##              Prevalence : 0.6011
##          Detection Rate : 0.5112
##    Detection Prevalence : 0.6180
##       Balanced Accuracy : 0.7914
##
##        'Positive' Class : 0
##
```

Obtenemos una precisión en la predicción del 80.34%

- Vamos a entrenar un árbol de decisión y lo compararemos con el modelo de regresión logística multivariable anterior:

```
tree_model <- rpart(Survived ~ Pclass + Sex + Age, data = titanic_train, method = "class")
rpart.plot(tree_model)
```



Evaluamos su precisión:

```
predict_tree <- predict(tree_model, newdata = titanic_test[-1], type="class")
titanic_test$Survived_predict_tree = predict(tree_model, newdata = titanic_test[-1], type="class")

confusionMatrix(titanic_test$Survived, as.factor(titanic_test$Survived_predict_tree))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 98 12
##          1 18 50
##
##                Accuracy : 0.8315
##                  95% CI : (0.7682, 0.8833)
##     No Information Rate : 0.6517
##     P-Value [Acc > NIR] : 7.94e-08
##
##                   Kappa : 0.6369
##
##  Mcnemar's Test P-Value : 0.3613
##
##             Sensitivity : 0.8448
##             Specificity : 0.8065
##          Pos Pred Value : 0.8909
##          Neg Pred Value : 0.7353
##              Prevalence : 0.6517
##          Detection Rate : 0.5506
##    Detection Prevalence : 0.6180
##       Balanced Accuracy : 0.8256
##
##        'Positive' Class : 0
##
```

Obtenemos una precisión en la predicción del 83.15%, por lo que es más preciso el árbol de decisión que la regresión logística.

## 7. Aplicación del mejor modelo al dataset de test para concurso.

```
titanic_data_test <- read.csv("./test_titanic.csv", header=T, sep=",", stringsAsFactors = FALSE, fileEn
titanic_data_test$Pclass <- as.factor(titanic_data_test$Pclass)
titanic_data_test$Sex <- as.factor(titanic_data_test$Sex)

predict_tree <- predict(tree_model, newdata = titanic_data_test[-1], type="class")
titanic_data_test$Survived = predict(tree_model, newdata = titanic_data_test[-1], type="class")
titanic_data_test <- titanic_data_test[,-(2:11),drop=FALSE]

write.csv(titanic_data_test, "test_titanic_survived.csv")
```

## 8. Código.

El código utilizado para el análisis de la información se puede encontrar en la siguiente URL.

## Agradecimientos.

Los datos utilizados han sido recolectados del repositorio Kaggle.

Por este motivo, ellos son los propietarios de los datos utilizados en esta práctica.

## Licencia.



## Tabla contribuciones

| **Contribuciones** | **Firma** |
|---|---|
| *Investigación Previa* | Carmen Lobato Cassinello, Raúl Vicente Ferrer |
| *Redacción de las respuestas* | Carmen Lobato Cassinello, Raúl Vicente Ferrer |
| *Desarrollo del código* | Carmen Lobato Cassinello, Raúl Vicente Ferrer |