Universitat Oberta de Catalunya

# Tipología y ciclo de vida de los datos

# Práctica 2: Limpieza y análisis de datos

Autores: Xián Pérez Pérez y Daniel Santos Rivilla

Junio 2021



# 1 Introducción

## 1.1 Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

## 1.2 Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

# 1.3 Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

# 1.4 Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com (https://www.kaggle.com)). Algunos ejemplos de dataset con los que podéis trabajar son: * Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009 (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)) * Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic (https://www.kaggle.com/c/titanic))

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Integración y selección de los datos de interés a analizar.
3. Limpieza de los datos.
    1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?
    2. Identificación y tratamiento de valores extremos.
4. Análisis de los datos.
    1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).
    2. Comprobación de la normalidad y homogeneidad de la varianza.
    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.
5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

# 1.5 Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world (https://guides.github.com/activities/hello-world).


## 1.6 Criterios de evaluación

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2 y 6 valen 0,5 puntos.
- Los apartados 3, 5 y 7 valen 2 puntos.
- El apartado 4 vale 2,5 puntos.

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.

## 1.7 Formato y fecha de entrega

En referente a la entrega final, hay que entregar un único fichero que contenga el enlace Github, el cual no se podrá modificar posteriormente a la fecha de entrega, donde haya:

1. Una Wiki con los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, deberá aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales.
3. Una carpeta con el código generado para analizar los datos.
4. El fichero CSV con los datos originales.
5. El fichero CSV con los datos finales analizados.

Este documento de entrega final de la Práctica 2 se debe entregar en el espacio de Entrega y Registro de AC del aula antes de las 23:59 del día 8 de junio. No se aceptarán entregas fuera de plazo.

---

# 2 Desarrollo de la práctica

---

## 2.1 Introducción.

Para el desarrollo de la práctica 2 de **Tipología y ciclo de vida de los datos** se ha seleccionado el dataset de Titanic expuesto en kaggle. Este dataset consta de dos ficheros .csv (test y train) en los que se encuentra recogida la información a cerca de los pasajeros del célebre crucero. Esta separación inicial está fundamentada por su objetivo de aplicar un modelo de predicción para determinar los supervivientes del accidente, de modo que uno de los conjuntos se emplee para el entrenamiento del modelo y el otro para ponerlo a prueba. Para nuestro estudio emplearemos el conjunto de entrenamiento **train** puesto que es el que cuenta con la información relativa a la supervivencia de los pasajeros.

Se trata de un conjunto de datos que refieren al accidente sufrido por el transatlántico *Titanic* que terminó hundiéndose al colisionar con un iceberg. El conjunto de datos es un resumen de los pasajeros de la nave junto con el desenlace de la tragedia para cada uno, es decir si ha sobrevivido o no. Gracias a este conjunto de datos se puede llevar a cabo un estudio para determinar qué factores influyen en la supervivencia de los pasajeros de un navío de este tipo, de modo que se puedan fortalecer las medidas de seguridad para minimizar las perdidas o simplemente orientar a los compradores de billetes a la hora de seleccionar la configuración de su viaje.

# 2.2 Carga de datos.

Se procede a llevar a cabo la carga de datos de **train.csv** y almacenarlos en forma de dataframe.

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)
library(gridExtra)
library(ggcorrplot)
library(GGally)
library(ggpubr)
library(broom)
library(pROC)
```

```
# Cargamos el fichero de datos
train <- read.csv('train.csv', stringsAsFactors = FALSE)
# Verificamos la estructura del conjunto de datos
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)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
...
##  $ 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" ...
```

# 2.3 Descripción del dataset.

En este apartado se tratará de llevar a cabo una descripción detallada del conjunto de datos, con lo que se definirán los atributos del mismo, las proporciones y posibles valores de los mismos e información relevante para procesos posteriores. En primer lugar, los atributos del conjunto de datos:

- Name: un atributo de tipo texto con el nombre del pasajero.

- Sex: un atributo de tipo texto que especifica el género del pasajero. Puede tomar los valores: *male*, *female*.
- Age: atributo numérico que determina la edad del pasajero. La edad de los bebés por debajo de 12 meses se representa como una fracción del tipo 1/*meses de vida*.En caso de tratarse de una edad estimada se representa con la forma *xx.5*.
- Pclass: un atributo de tipo numérico que representa la clase a la que pertenece el pasajero. Existen las clases: *1*, *2* y *3* que coinciden con clase alta, clase media y clase baja respectivamente.
- Embarked: atributo de tipo texto que determina la pasarela de embarque del pasajero.Toma los valores: *S* (Southampton), *C* (Cherbourg) y *Q* (Queenstown).
- Ticket: atributo de tipo texto con el código del billete del pasajero.
- Fare: atributo numérico con el precio que ha pagado el pasajero. Entendemos aquellos valores de 0, como que el pasajero en cuestión es un empleado en la nave.
- Sibsp: atributo numérico que especifica la cantidad de hermanos y/o cónyuges a bordo del barco.
- Prch: atributo numérico que especifica la cantidad de padres y/o hijos a bordo del barco.
- Survived: atributo booleano que especifica si el pasajero ha sobrevivido al accidente o no. Se entiende *0* como defunción y *1* como supervivencia.
- Cabin: atributo de tipo texto que especifica el camarote en el que reside el pasajero.

Una vez definidos los atributos del conjunto, se procede a estudiar superficialmente los datos de modo que se puedan detectar ciertas tendencias o naturalezas.

```
# Comprobamos los distintos valores de los atributos que parecen ser categóricos
unique(train$Survived)
```

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

```
unique(train$Pclass)
```

```
## [1] 3 1 2
```

```
unique(train$Sex)
```

```
## [1] "male"   "female"
```

```
unique(train$Embarked)
```

```
## [1] "S" "C" "Q" ""
```

```
#Estadísticas básicas
summary(train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   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
##     Ticket              Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Tras la ejecución de un *summary* se puede observar la naturaleza de los datos y parte de su distribución. Aunque este tipo de resumen puede resultar de gran utilidad, es necesario tener en cuenta que en ciertos casos puede llevar a equívocos puesto que se podría obviar el significado de ciertos atributos, fundamentalmente en aquellos categóricos de tipo numérico.En cambio, en el caso de atributos no categóricos de tipo numérico se puede extraer información de mayor interés. En este caso se observa que el precio de los billetes es 32.2 de media y se aprecia que el valor máximo se aleja mucho de éste, por lo que podría ser un valor erroneo. En cuanto al número de familias viajando en la nave se podría decir que conforman una minoría y que lo más común son viajeros individuales o con un único familiar o pareja. La distribución de edad es relativamente baja por lo que se puede suponer que la mayoría de pasajeros son jóvenes (por debajo de 40 años).

```
#Previsualizacion de datos
head(train)
```

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

```
tail(train)
```

```
##     PassengerId Survived Pclass                                     Name    Sex
## 886         886        0      3     Rice, Mrs. William (Margaret Norton) female
## 887         887        0      2                     Montvila, Rev. Juozas   male
## 888         888        1      1              Graham, Miss. Margaret Edith female
## 889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female
## 890         890        1      1                     Behr, Mr. Karl Howell   male
## 891         891        0      3                       Dooley, Mr. Patrick   male
##     Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886  39     0     5     382652 29.125              Q
## 887  27     0     0     211536 13.000              S
## 888  19     0     0     112053 30.000   B42        S
## 889  NA     1     2 W./C. 6607 23.450              S
## 890  26     0     0     111369 30.000  C148        C
## 891  32     0     0     370376  7.750              Q
```

En ciertas ocasiones, la previsualización de los datos tal y como se han cargado puede ayudar a la identificación de ciertos detalles de los distintos atributos, tanto en cuanto a su significado como en cuanto a los valores que pueden tomar. En este caso se puede observar que los tipos de datos y sus valores coinciden con la definición de los atributos aunque salen a la luz otro factores como:

- Existencia de valores no numéricos para el atributo *Age*.
- Valores vacíos en el atributo no numérico *Cabin*

Aunque este método aporta cierta luz sobre los datos, no es suficiente, y por ello se deben llevar a cabo otras tareas para la correcta comprensión del conjunto de datos.

```
# Estadísticas de valores vacíos
colSums(is.na(train))
```

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

```
colSums(train=="")
```

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

Aparentemente, en el conjunto existe una alta cantidad de valores inválidos para el atributo *Age*, teniendo en cuenta que el conjunto cuenta con 891 registros. En cuanto a las cabinas se puede suponer que no todos los pasajeros contaban con una propia y por ello ese campo aparece vacío ya que la mayoría de pasajeros se encuentran con esta carencia. Sin embargo, en cuanto al atributo *Embarked* si que podemos concluir en que existen 2 registros que no fueron tomados y por lo tanto el conjunto de datos carece de dicha información.

# 2.4 Preparación del dataset.

Tras las observaciones llevadas a cabo en el apartado anterior se realizarán ciertas tareas y acciones sobre el conjunto de datos en vistas a facilitar y mejorar los resultados del futuro análisis.

## 2.4.1 Tratamiento de valores nulos o vacíos.

En cuanto a los valores no numéricos encontrados para el atributo *Age*, dado que su eliminación sería una pérdida significativa para los análisis posteriores, se lleva a cabo la sustitucion de estos valores inválidos por el valor medio del atributo.

```
# Tomamos la media para valores vacíos de la variable "Age"
train$Age[is.na(train$Age)] <- mean(train$Age,na.rm=T)
```

Tratándose del atributo *Embarked*, como se ha mostrado anteriormente solo existen 2 registros vacíos con lo que se puede eliminar perfectamente la información de estos pasajeros.

```
# Se eliminan las filas co
train <- train[train$Embarked!="",]
```

En cuanto al atributo *Cabin* dada la suposición antes mencionada se considera que es un valor válido que implica la carencia de un camarote por lo que se procede a asignar un valor más representativo como podría ser “Sin camarote”.

```
# Asignamos un valor significativo a los registros sin camarote
train$Cabin[train$Cabin == ""] <- "Sin camarote"
```

## 2.4.2 Discretización de variables.

Para el caso del atributo de edad será mas interesante tratarlo como si fuese una categoría por lo que se lleva a cabo una discretización que resultará en la existencia de 8 rangos de edad.

```
# se crea nueva variable discretizando Age
train$Age.disc <- cut(train$Age, breaks = c(0, 10, 20, 30, 40, 50, 60, 70, 100), labels
 = c("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-80"))
# Se previsualizan los datos
head(train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex      Age SibSp
## 1                            Braund, Mr. Owen Harris   male 22.00000     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00000     1
## 3                             Heikkinen, Miss. Laina female 26.00000     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00000     1
## 5                           Allen, Mr. William Henry   male 35.00000     0
## 6                                   Moran, Mr. James   male 29.69912     0
##   Parch            Ticket    Fare        Cabin Embarked Age.disc
## 1     0        A/5 21171  7.2500 Sin camarote        S    20-29
## 2     0         PC 17599 71.2833          C85        C    30-39
## 3     0 STON/O2. 3101282  7.9250 Sin camarote        S    20-29
## 4     0           113803 53.1000         C123        S    30-39
## 5     0           373450  8.0500 Sin camarote        S    30-39
## 6     0           330877  8.4583 Sin camarote        Q    20-29
```

## 2.4.3 Factorización de variables

Cambiamos el tipo de los atributos que lo necesitan a factor

```
# Transformación
train$Survived <- as.factor(train$Survived)
train$Pclass <- as.factor(train$Pclass)
train$Sex <- as.factor(train$Sex)
train$Embarked <- as.factor(train$Embarked)
```

## 2.4.4 Valores extremos.

Como hemos observado previamente es probable que existan valores atípicos en los extremos de la variable *Fair* (outliers). En caso de llevar a cabo un estudio que tenga en cuenta el precio del billete, estos valores podrían afectar a las conclusiones, por lo que deben ser tratados.

```
# Muestra de outliers
boxplot.stats(train$Fare)
```

```
## $stats
## [1]  0.0000  7.8958 14.4542 31.0000 65.0000
##
## $n
## [1] 889
##
## $conf
## [1] 13.22987 15.67853
##
## $out
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  83.4750  73.5000 263.0000
##   [9]  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500  69.5500
##  [17] 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000  79.2000
##  [25]  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500  91.0792
##  [33] 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750 164.8667
##  [41] 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000 263.0000
##  [49]  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000 120.0000
##  [57] 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792  90.0000
##  [65]  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250  71.0000
##  [73] 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000  78.2667
##  [81] 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500  73.5000
##  [89] 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250 211.3375
##  [97] 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583 211.3375
## [105]  79.2000  69.5500 120.0000  93.5000  83.1583  69.5500  89.1042 164.8667
## [113]  69.5500  83.1583
```

```
out <- boxplot.stats(train$Fare)$out
min(out)
```

```
## [1] 66.6
```

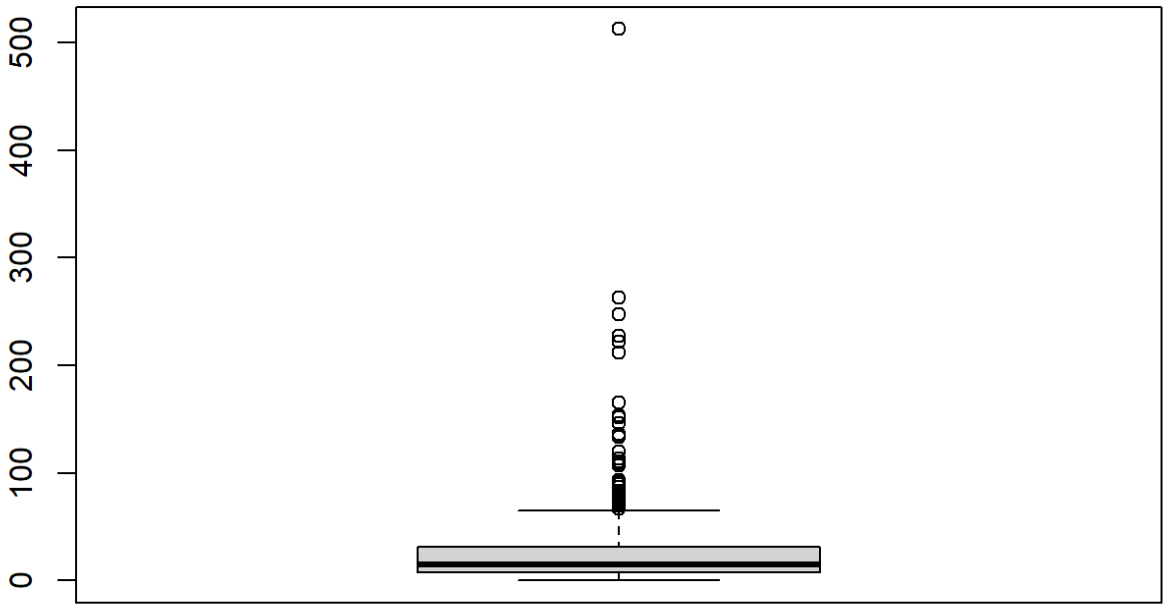
```
max(out)
```

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

```
length(out)
```

```
## [1] 114
```

```
boxplot(train$Fare)
```

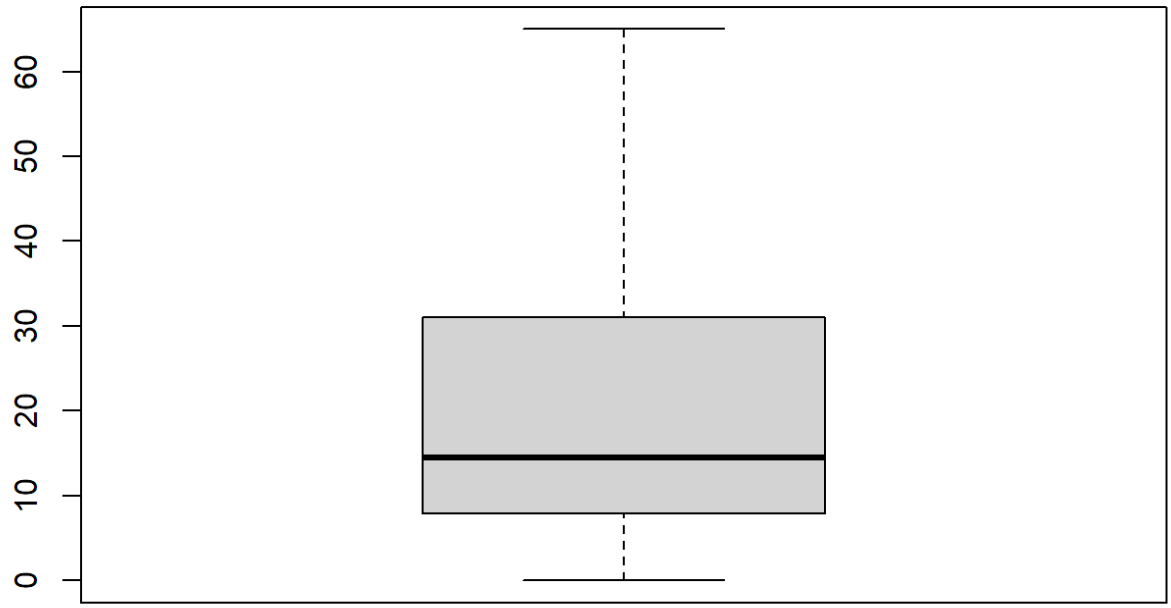



Como se puede observar tanto en el diagrama de caja, los valores que se encuentran entre 66.6 y 512.3292 son anómalos y se encuentran a demasiada distancia del resto de datos del conjunto. Aunque es cierto que el precio de los billetes en un crucero puede variar en función de los servicios contratados y la fecha de adquisición, en este caso se sustituirán todos estos valores por el valor medio del conjunto.

```
# Tomamos la media para valores vacíos de la variable "Fare"
train$Fare[train$Fare >= min(out)] <- mean(train$Fare,na.rm=T)
```

Una vez sustituidos los valores anómalos se vuelve a comprobar la distribución de este atributo en un diagrama de caja y se puede ver que dicjos valores han desaparecido

```
# Imprimimos boxplot
boxplot(train$Fare)
```




### 2.4.5 Eliminación de atributos prescindibles

En este conjunto de datos existen ciertos atributos que no representan ningún interés para los posibles estudios a realizar. Los datos relativos al nombre de los pasajeros y su identificador en el conjunto no aportan ningún valor, y por ello estas columnas serán desechar.

```
# Drop columns
train <- subset(train, select = -c(Name, PassengerId))
```

## 2.5 Análisis de los datos.

A continuación se llevará a cabo un estudio en profundidad del conjunto de datos, las distintas distribuciones de sus atributos y el efecto de unos atributos en los demás.

### 2.5.1 Selección de los grupos de datos que se quieren analizar

Para el análisis de datos se van a utilizar las siguientes variables, que agrupan los datos según su valor:

- Survived: el pasajero sobrevive o no.
- Pclas: clase en la que viaja el pasajero.
- Sex: género del pasajero.
- Embarked: puerto por el que embarca el pasajero

### 2.5.2 Comprobación de la normalidad y homogeneidad de la varianza

**Normalidad:**

Podemos utilizar test de significancia para comparar las distribuciones de las variables con la distribución normal. Para ello vamos a utilizar el test de Shapiro-Wilk. Es considerado uno de los test más potentes para el contraste de normalidad.

Las hipótesis son las siguientes:

$$\begin{cases} Hipótesis\ \ nula & H_0: \ distribución\ \ normal \\ Hipótesis\ \ alternativa & H_1: \ distribución\ \ no\ \ \ \ normal \end{cases}$$

Comenzamos con el análisis de la variable *Age*

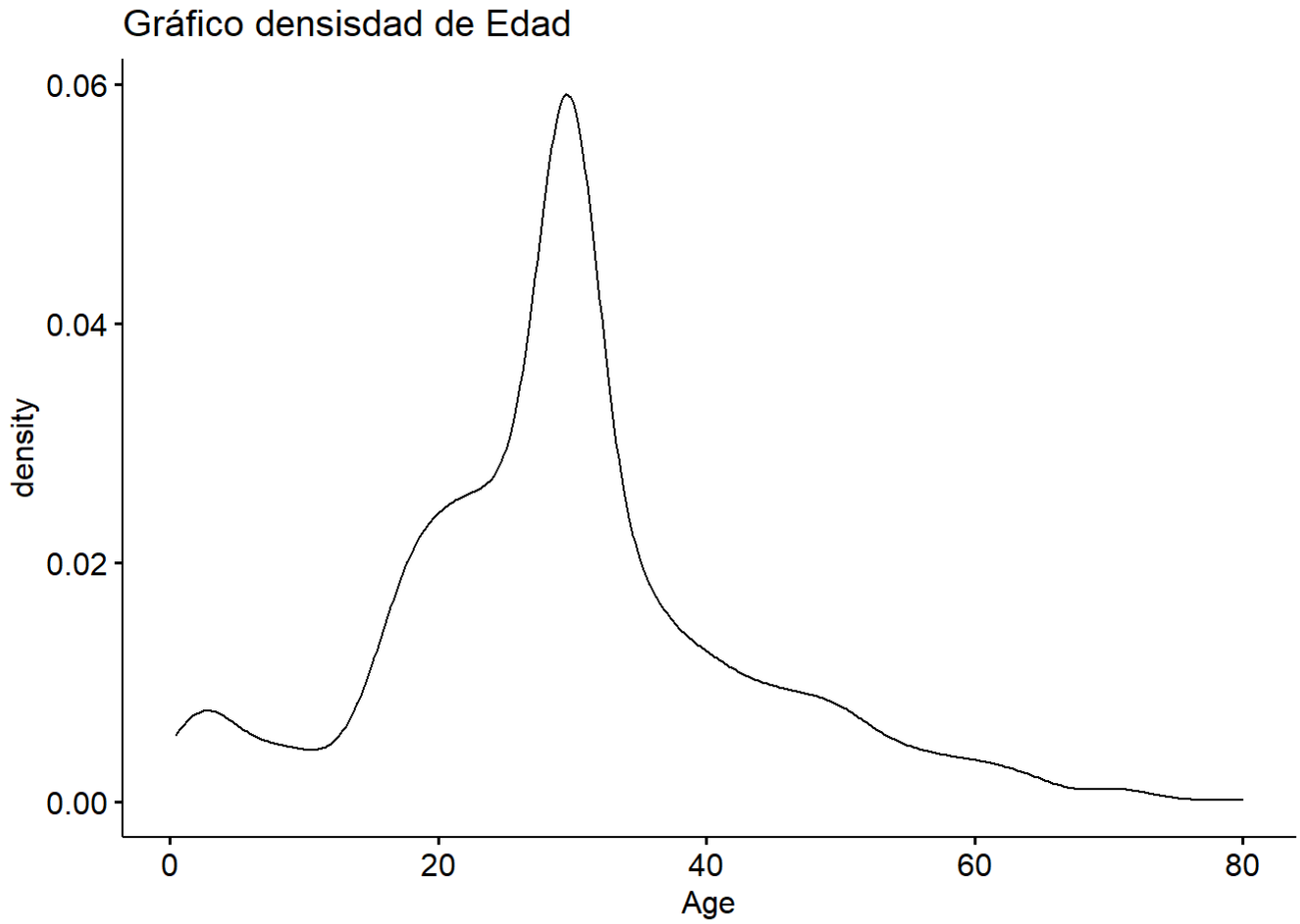
```
shapiro.test(train$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$Age
## W = 0.95882, p-value = 4.146e-15
```

Alpha es menor que el nivel de significancia (0.05), por lo que debemos aceptar la hipótesis alternativa que implica que la variable no sigue una distribución normal.

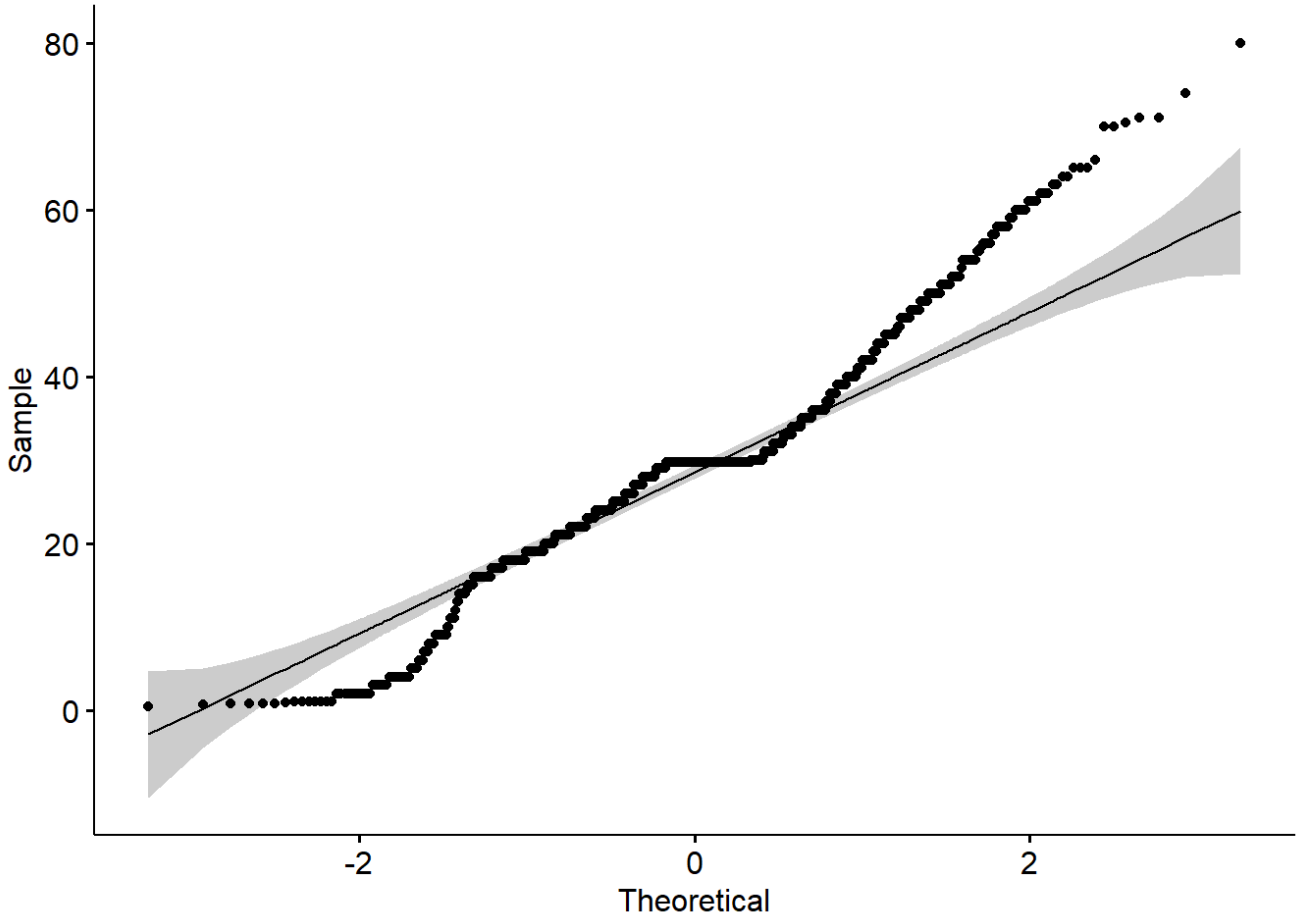
Mostramos el gráfico de densidad para comprobarlo visualmente:

```
# gráfico de denisdad
ggdensity(train$Age,
          main = "Gráfico densisdad de Edad",
          xlab = "Age")
```



Y el diagrama q-q:

```
# q-q plot
ggqqplot(train$Age)
```




A continuación, realizamos el test para la variable *Fare*:

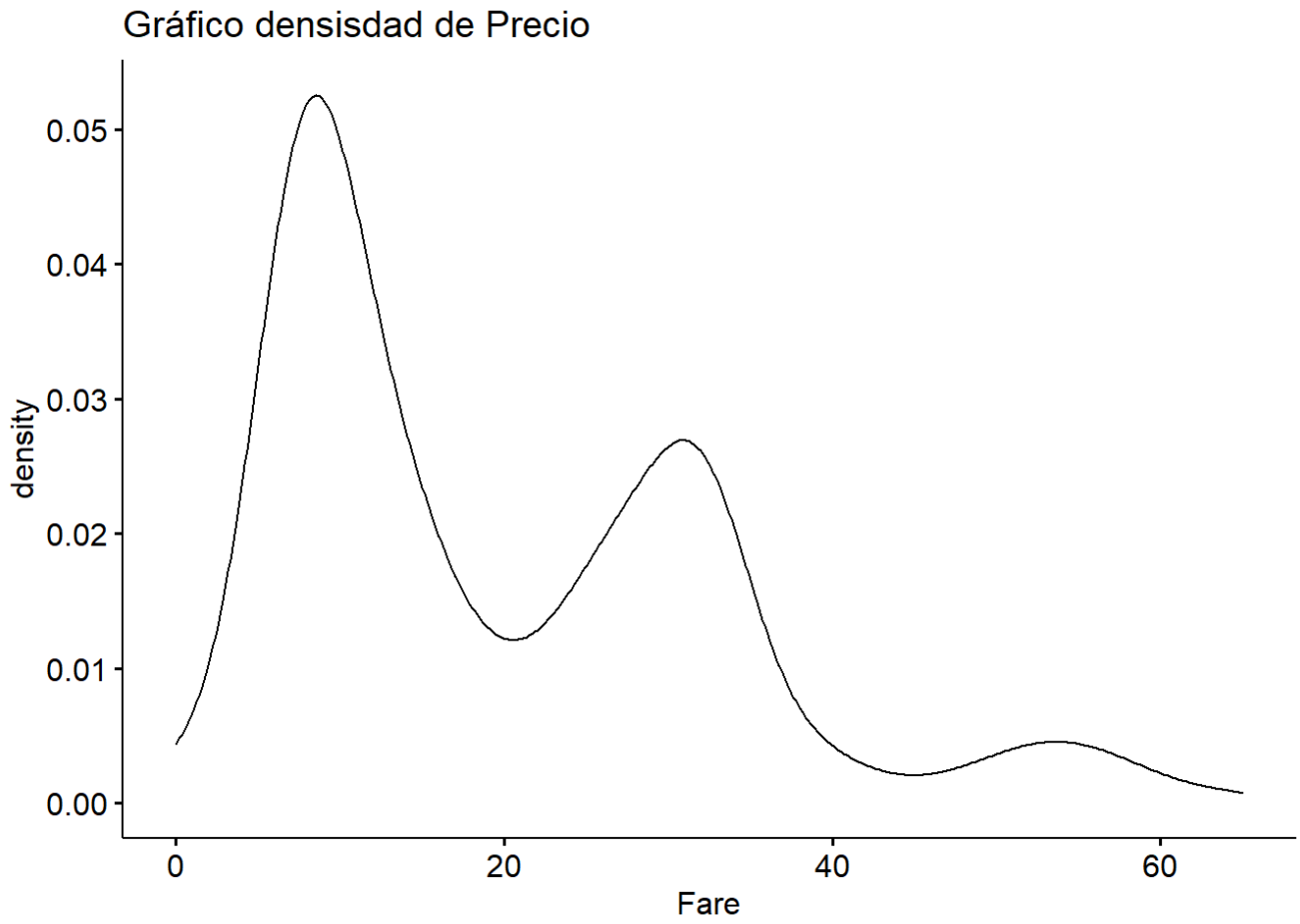
```
shapiro.test(train$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$Fare
## W = 0.86743, p-value < 2.2e-16
```

Alpha es menor que el nivel de significancia (0.05), por lo que debemos aceptar la hipótesis alternativa que implica que la variable no sigue una distribución normal.

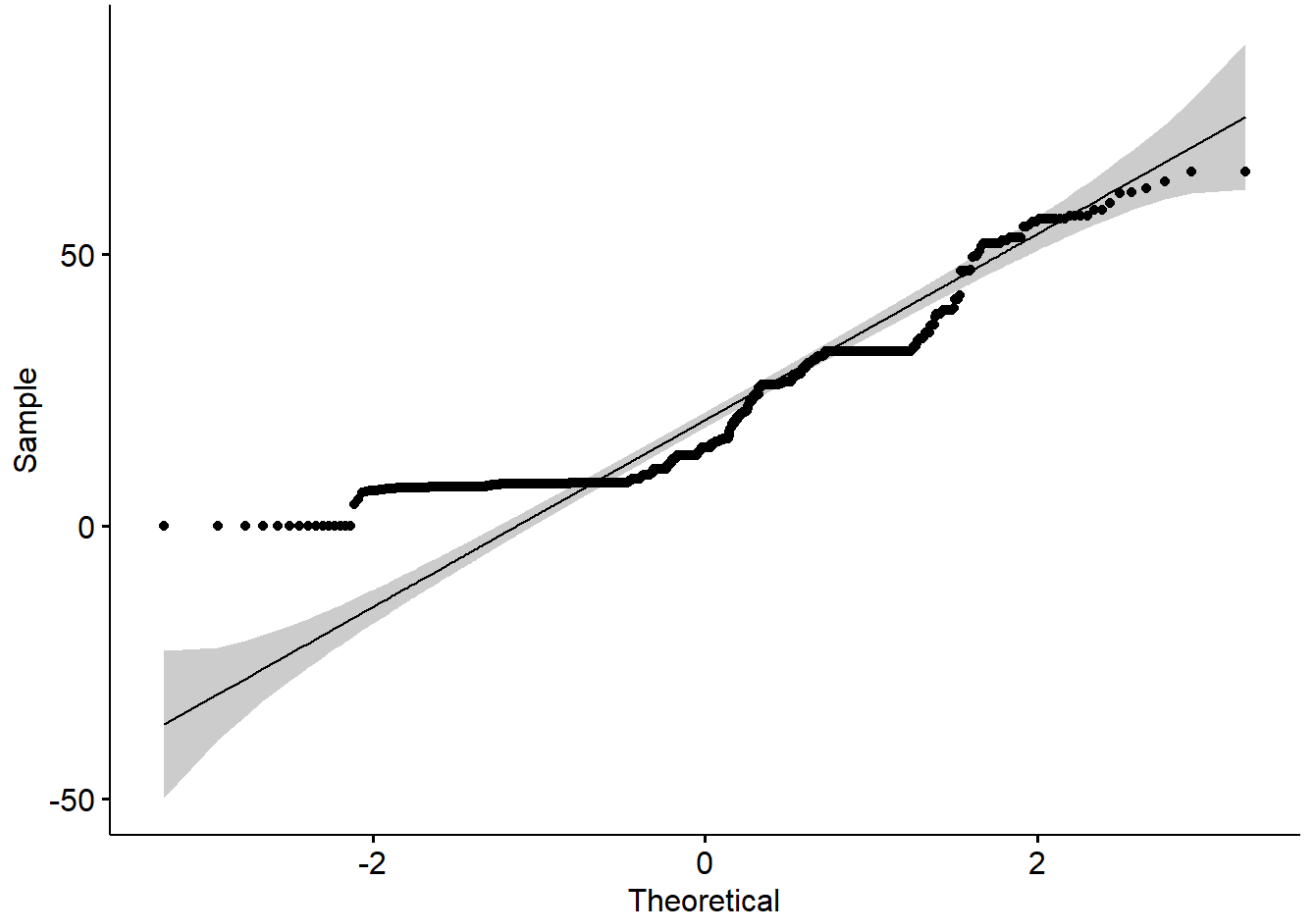
Mostramos el gráfico de densidad para comprobarlo visualmente:

```
# gráfico de denisdad
ggdensity(train$Fare,
          main = "Gráfico densisdad de Precio",
          xlab = "Fare")
```




Y el diagrama q-q

```
# q-q plot
ggqqplot(train$Fare)
```




Las variables *Age* y *Fare* no siguen una distribución normal.

**Homogeneidad de la varianza**

Test de Fligner-Killeen

Se trata de un test no paramétrico que compara las varianzas basándose en la mediana. Es también una alternativa cuando no se cumple la condición de normalidad en las muestras, como es nuestro caso. Considera como hipótesis nula que la varianza es igual entre los grupos y como hipótesis alternativa que no lo es.

```
fligner.test(Fare ~ Age, data = train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Age
## Fligner-Killeen:med chi-squared = 89.419, df = 88, p-value = 0.4378
```

Puesto que obtenemos un p-valor inferior a 0,05, aceptamos la hipótesis alternativa, las varianzas de ambas muestras no son homogéneas.

Ahora vamos a comprobar el resultado gráficamente.

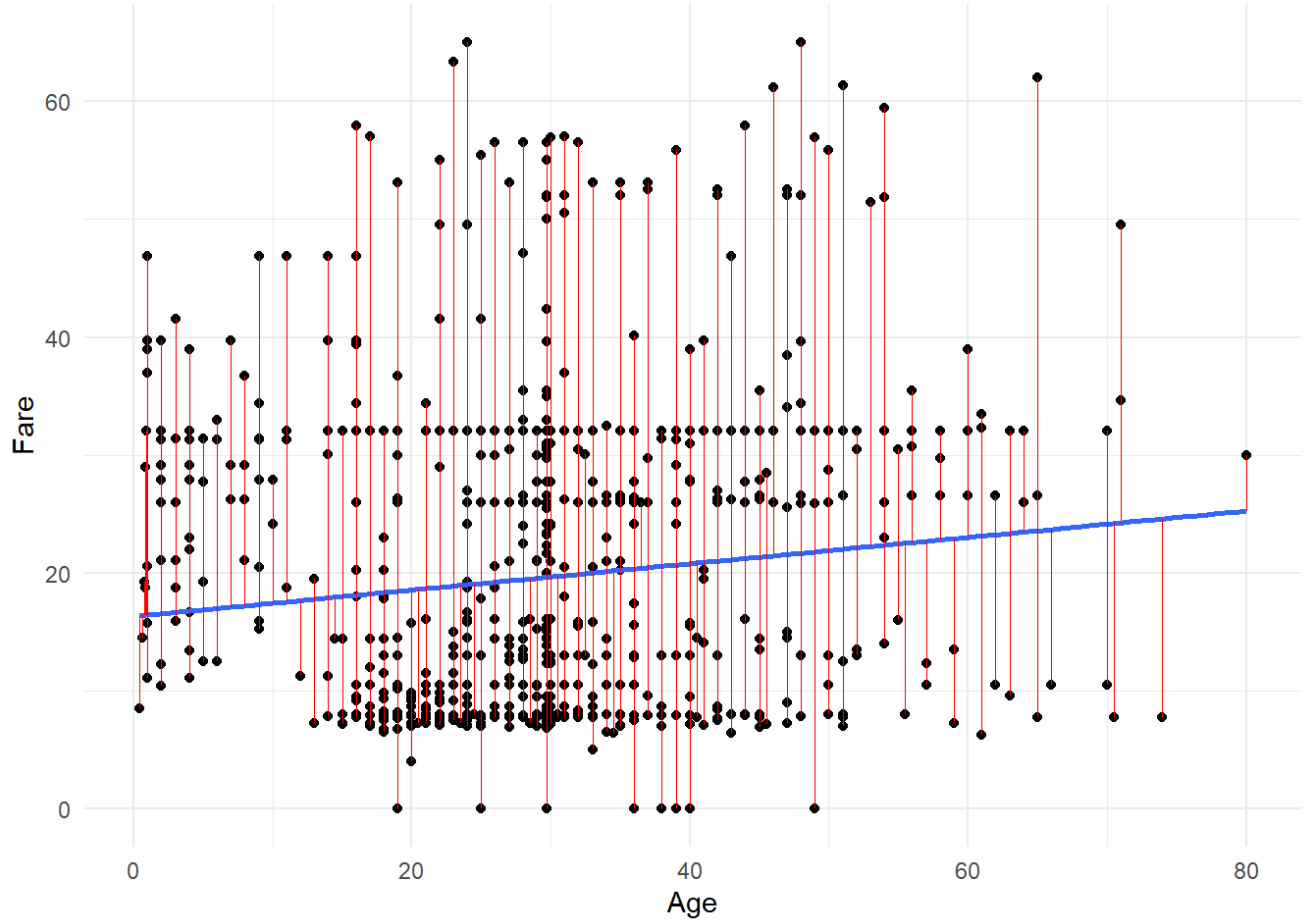
```
# modelo regresión lineal
model <- lm(Fare ~ Age, data = train)
model
```

```
##
## Call:
## lm(formula = Fare ~ Age, data = train)
##
## Coefficients:
## (Intercept)          Age
##     16.3422       0.1116
```

```
# obtenemos las métricas
model.diag.metrics <- augment(model)
head(model.diag.metrics)
```

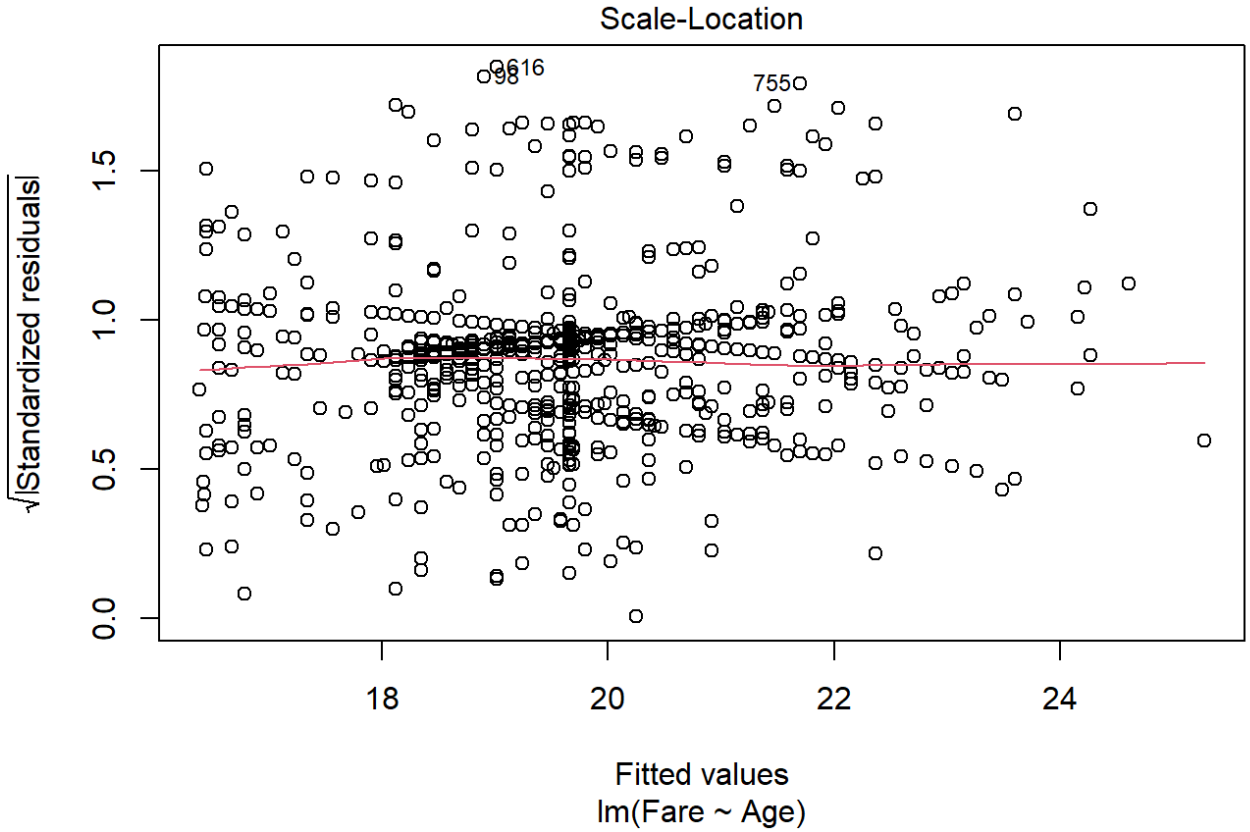
```
## # A tibble: 6 x 9
##   .rownames  Fare   Age .fitted .resid    .hat .sigma  .cooksd .std.resid
##   <chr>     <dbl> <dbl>   <dbl>  <dbl>   <dbl>  <dbl>    <dbl>      <dbl>
## 1 1          7.25  22      18.8  -11.5 0.00152   13.5 0.000559     -0.858
## 2 2         32.1   38      20.6   11.5 0.00159   13.5 0.000583      0.855
## 3 3          7.92  26      19.2  -11.3 0.00121   13.5 0.000429     -0.840
## 4 4         53.1   35      20.2   32.9 0.00132   13.4 0.00392       2.44
## 5 5          8.05  35      20.2  -12.2 0.00132   13.5 0.000541     -0.906
## 6 6          8.46  29.7    19.7  -11.2 0.00112   13.5 0.000389     -0.832
```

```
# mostramos el gráfico
ggplot(model.diag.metrics, aes(Age, Fare)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = Age, yend = .fitted), color = "red", size = 0.3) +
  theme_minimal()
```




Podemos comprobar la homocedasticidad examinando el gáfico *scale-location plot*

```
plot(model, 3)
```



Este gráfico nos muestra si los valores residuales están distribuidos igualmente a lo largo del rango del predictor. La línea contiene una curva al principio, aunque después es horizontal, pero los puntos no se distribuyen homogénamente a ambos lados de la línea, podemos asumir que no existe homocedasticidad. Llegamos a la misma conclusión que el test de Fligner-Killeen.

## 2.5.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

Ahora, pasamos a comparar los supervivientes con los grupos de creados por las distitnas variables. Para ello vamos a utilizar el modelo de regresión logística, teniendo como variable dependiente Survived.

**Supervivientes - primera clase**

```
# nueva variable para saber si es primera clase o no
train$primera <- ifelse(train$Pclass == 1,1,0)
```

```
# frecuencias absolutas
primera.tab <- table(train$Survived, train$primera)
primera.tab_M <- addmargins(primera.tab, FUN = list(Total = sum), quiet = TRUE)
primera.tab_M
```

```
##
##            0   1 Total
##   0     469  80   549
##   1     206 134   340
##   Total 675 214   889
```

Creamos el primer modelo con la nueva variable que nos dice si el pasajero viaja en primera o no:

```
# modelo regresión logística
logit_model_1 <- glm(formula=Survived~primera, data=train, family=binomial)
summary(logit_model_1)
```

```
##
## Call:
## glm(formula = Survived ~ primera, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4028  -0.8534  -0.8534   0.9676   1.5407
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.82273    0.08359  -9.843  < 2e-16 ***
## primera      1.33854    0.16416   8.154 3.53e-16 ***
## ---
## 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: 1113.4  on 887  degrees of freedom
## AIC: 1117.4
##
## Number of Fisher Scoring iterations: 4
```

```
# coeficientes
exp(coefficients(logit_model_1))
```

```
## (Intercept)     primera
##   0.4392324   3.8134709
```

Observamos que ir en primera clase es un factor significativo para la supervivencia, además, como el coeficiente es positivo, indica que que si se viaja en primera, existen más posibilidades de sobrevivir.

Viajar en primera, con una OR de 3.81 implica que las odds (razón de probabilidad) de sobrevivir es 3.81 veces mayor que no viajar en primera.

**Supervivientes - primera clase + género**

```
# nueva variable para saber si es mujer o no
train$genero <- ifelse(train$Sex == "female",1,0)
```

```
# modelo regresión logística
logit_model_2 <- glm(formula=Survived~primera+genero, data=train, family=binomial)
summary(logit_model_2)
```

```
##
## Call:
## glm(formula = Survived ~ primera + genero, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1803  -0.5244  -0.5244   0.8939   2.0258
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9145     0.1332 -14.378  < 2e-16 ***
## primera       1.5682     0.1984   7.905 2.69e-15 ***
## genero        2.6256     0.1796  14.622  < 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: 1182.8  on 888  degrees of freedom
## Residual deviance:  850.2  on 886  degrees of freedom
## AIC: 856.2
##
## Number of Fisher Scoring iterations: 4
```

```
# coeficientes
exp(coefficients(logit_model_2))
```

```
## (Intercept)     primera      genero
##   0.1474184   4.7979453  13.8127756
```

La nueva variable de género, disminuye considerablemente el AIC, por lo que el modelo es significativamente mejor.

Ahora, viajar en primera implica que las odds de sebrevivir es 4.80 veces mayor que no viajar e primera, y ser mujer implica que las odds de sobrevivir es 13.81 veces mayor de sobrevivir que si eres hombre.

**Supervivientes - primera clase + genero + edad**

Añadimos al modelo la edad de los pasajeros:

```
logit_model_3 <- glm(formula=Survived~primera+genero+Age, data=train, family=binomial)
summary(logit_model_3)
```

```
##
## Call:
## glm(formula = Survived ~ primera + genero + Age, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5779  -0.5756  -0.4909   0.8017   2.4342
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.14449    0.22925  -4.992 5.96e-07 ***
## primera      1.88873    0.21988   8.590  < 2e-16 ***
## genero       2.59863    0.18163  14.307  < 2e-16 ***
## Age         -0.02847    0.00724  -3.932 8.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  834.04  on 885  degrees of freedom
## AIC: 842.04
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coefficients(logit_model_3))
```

```
## (Intercept)     primera      genero         Age
##   0.3183873   6.6109502  13.4453575   0.9719330
```

En este modelo, ambas variables son significativas, además el AIC es menor, por lo que mejora la calidad. Observando los coeficientes tenemos lo siguiente:

- Viajar en primera implica 6.61 veces más de sobrevivir.
- Ser mujer implica 13.44 veces más de sobrevivir
- La edad tiene coeficiente negativo, por lo que a medida que aumente la edad, la probabilidad de sobrevivir es menor. En concreto por cada año que aumente, las odds de sobrevivir es 0.97 veces menor.

**Supervivientes - primera clase + genero + edad + pruerto**

Como último modelo, añadimos el puerto de embarque.

```
puerto_Rel=relevel(train$Embarked, ref = 'S')
logit_model_4 <- glm(formula=Survived~primera+genero+Age+factor(puerto_Rel), data=train,
family=binomial)
summary(logit_model_4)
```

```
##
## Call:
## glm(formula = Survived ~ primera + genero + Age + factor(puerto_Rel),
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5025  -0.5787  -0.4892   0.7688   2.4498
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.231052   0.235785  -5.221 1.78e-07 ***
## primera              1.784256   0.228979   7.792 6.58e-15 ***
## genero               2.588909   0.183074  14.141  < 2e-16 ***
## Age                 -0.027722   0.007276  -3.810 0.000139 ***
## factor(puerto_Rel)C  0.441949   0.229065   1.929 0.053686 .
## factor(puerto_Rel)Q  0.125466   0.305250   0.411 0.681053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  830.31  on 883  degrees of freedom
## AIC: 842.31
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coefficients(logit_model_4))
```

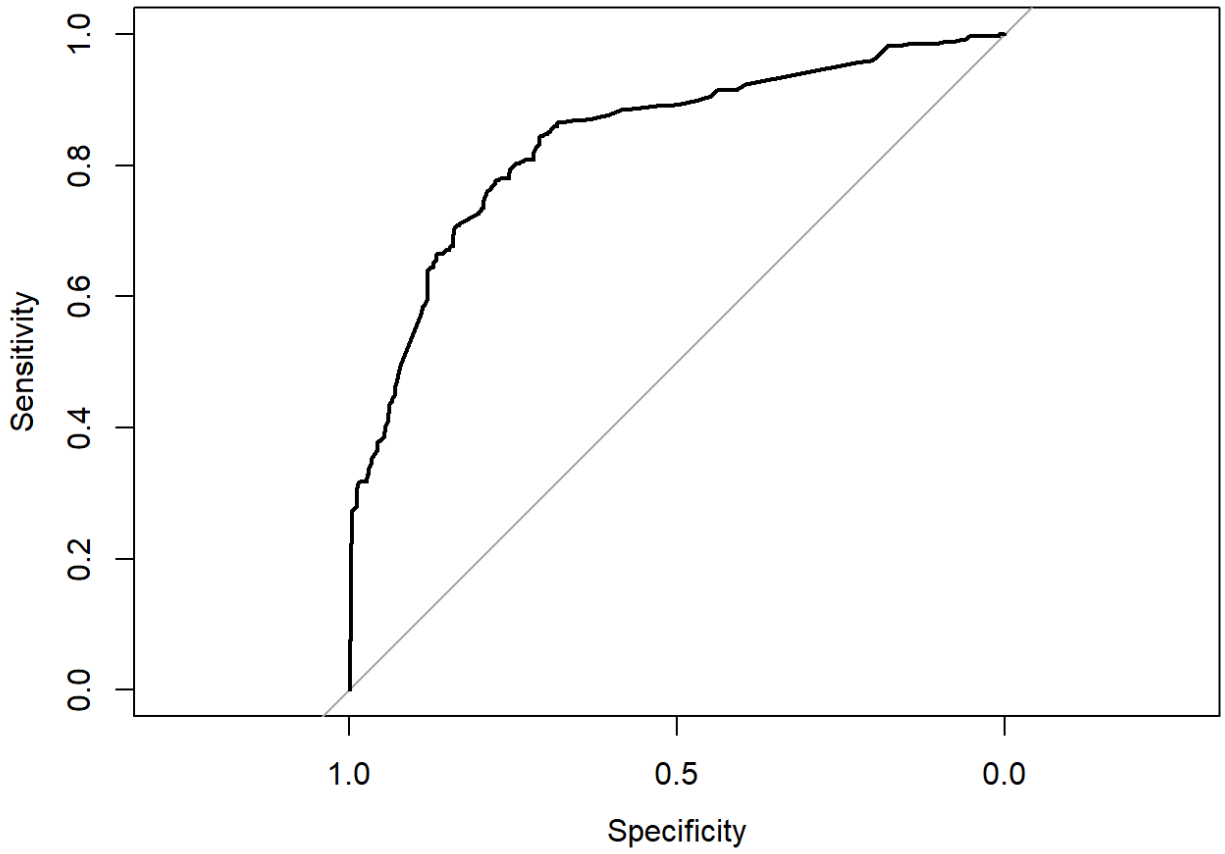
```
##         (Intercept)             primera              genero                 Age
##           0.2919852           5.9551454          13.3152314           0.9726586
## factor(puerto_Rel)C factor(puerto_Rel)Q
##           1.5557365           1.1336763
```

Observamos que el AIC no disminuye, y además, ninguno de los dos factores puerto, son significativos respecto al puerto de de referencia “S”. Por ello, desechamos este modelo.

**Curva ROC**

Una vez encontrado el mejor modelo, vamos a dibujar la curva ROC y calcular el área bajo la curva para obtener la bondad del modelo.

```
prob_low=predict(logit_model_3, train, type="response")
r=roc(train$Survived, prob_low, data=train)
plot(r)
```




```
auc(r)
```

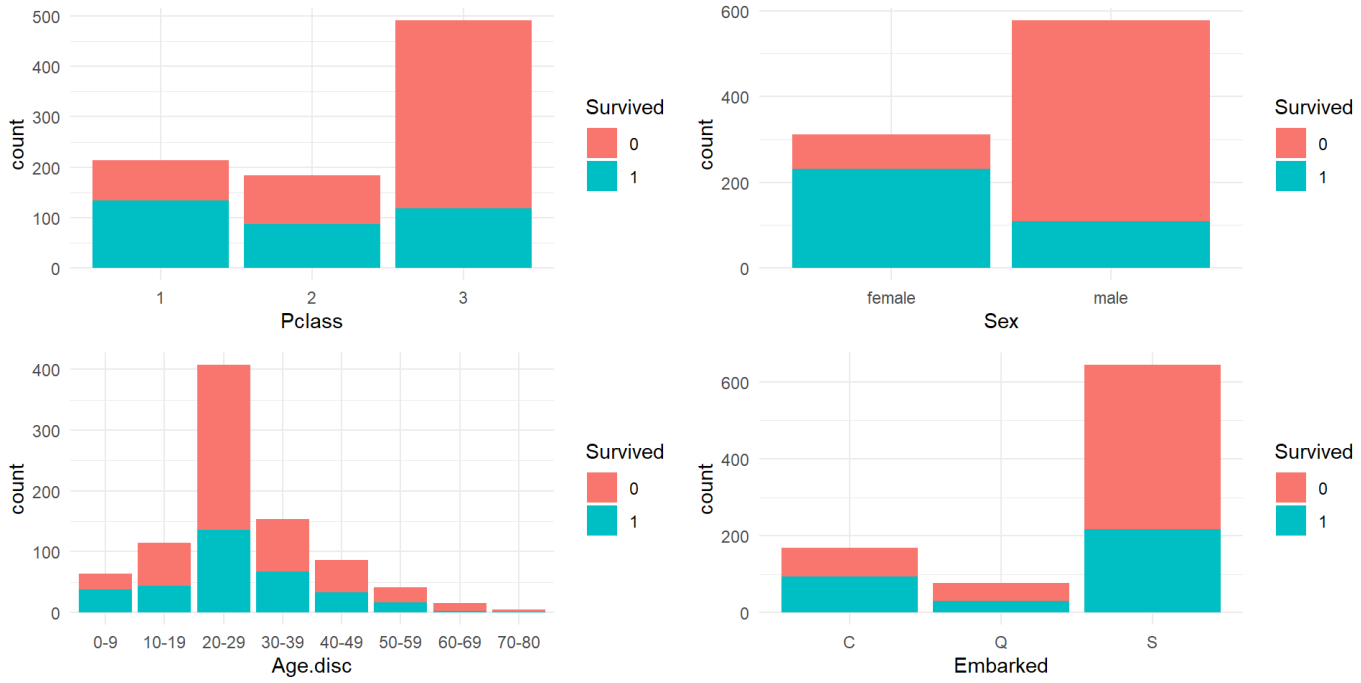
```
## Area under the curve: 0.8371
```

Un área bajo la curva de 0.84 implica que el modelo ayuda a predecir la supervivencia.

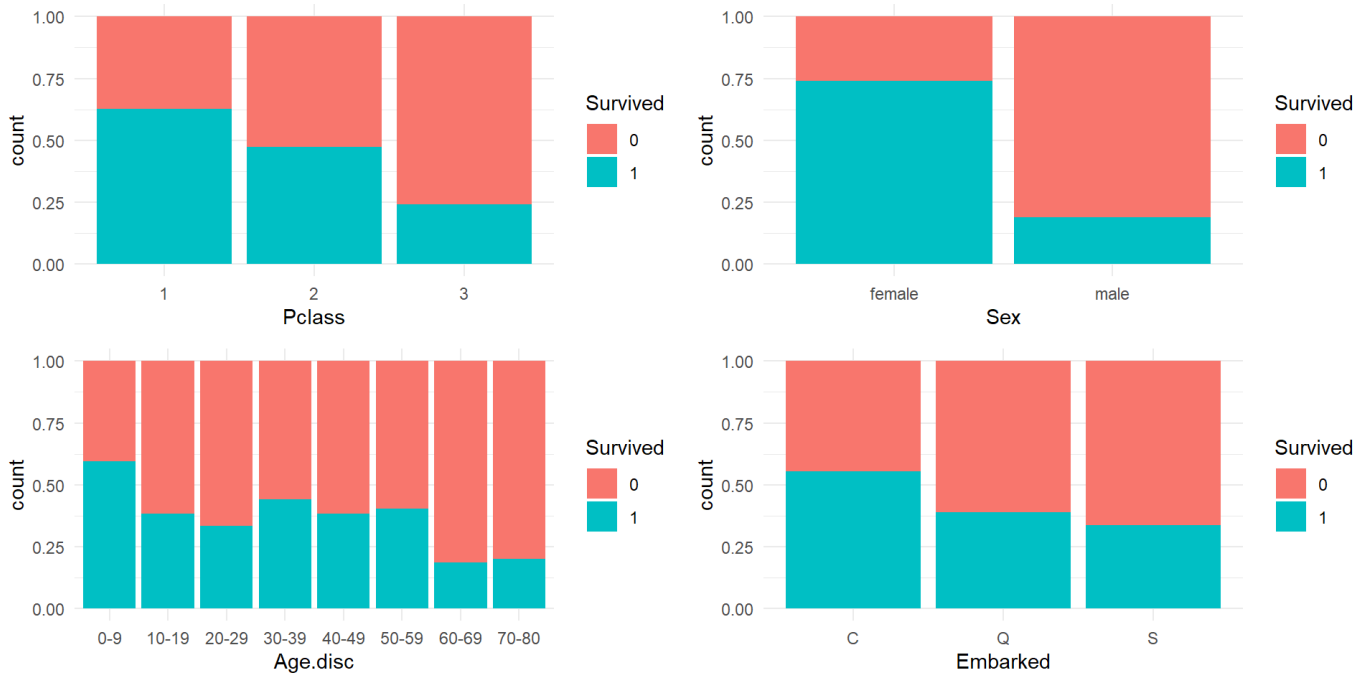
# 2.6 Representación de los resultados

Nos proponemos analizar las relaciones entre las diferentes variables del conjunto de datos para ver si se relacionan y como. Empezamos con las variables categóricas y su relación con Survived

```
# Visualizamos la relación entre las variables:
g1 <- ggplot(data=train, aes(x=Pclass, fill = Survived)) +
  geom_bar(position = "stack") +
  theme_minimal()

g2 <- ggplot(data=train, aes(x=Sex, fill = Survived)) +
  geom_bar(position = "stack") +
  theme_minimal()


g3 <- ggplot(data=train, aes(x=Age.disc, fill = Survived)) +
  geom_bar(position = "stack") +
  theme_minimal()

g4 <- ggplot(data=train, aes(x=Embarked, fill = Survived)) +
  geom_bar(position = "stack") +
  theme_minimal()

grid.arrange(g1, g2, g3, g4, nrow = 2, ncol = 2)
```



```
# Visualizamos la relación entre las variables, pero como porcentaje:
g1 <- ggplot(data=train, aes(x=Pclass, fill = Survived)) +
  geom_bar(position = "fill") +
  theme_minimal()

g2 <- ggplot(data=train, aes(x=Sex, fill = Survived)) +
  geom_bar(position = "fill") +
  theme_minimal()


g3 <- ggplot(data=train, aes(x=Age.disc, fill = Survived)) +
  geom_bar(position = "fill") +
  theme_minimal()

g4 <- ggplot(data=train, aes(x=Embarked, fill = Survived)) +
  geom_bar(position = "fill") +
  theme_minimal()

grid.arrange(g1, g2, g3, g4, nrow = 2, ncol = 2)
```




Con estos 4 pares de visaualizaciones podemos llegar a los siguientes razonamientos:

- A pesar de que el número de pasajeros de la clase 3 es muy superior, en porcentaje es más bajo que los de la clase 1 y 2.
- Ocurre algo parecido con el género, a pesar de haber menos mujeres, el porcentaje de supervivencia es muy superior, parece que se le dio preferencia a las mujeres a la hora de embarcar en los botes salvavidas.
- Del mismo, llegamos a la conclusión de preferencia para los niños menores de 10 años, a partir de esta edad, el porcentaje se estabiliza hasta llegar a la decada de los 60 y 70 años, donde el porcentaje de supervivientes disminuye de manera significativa. Esto implicaría que se priorizó a las personas más jóvenes frente a las de edad avanzada.
- Respecto a la variable Embarked, solo podemos observar que el número de pasajeros que embarcó en Southampton es mucho mayor que el resto, sin embargo, el porcentaje de supervivientes es menor. Se podría trabajar el puerto C (Cherburgo) para intentar explicar la diferencia en los datos. Quizás, porcentualmente embarcaron más mujeres, niños o gente de primera clase

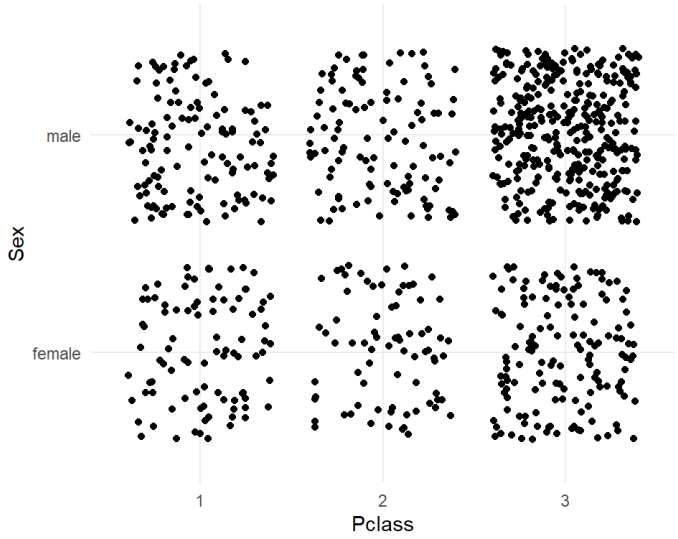
Obtenemos ahora una matriz de porcentajes de frecuencia. Vemos, por ejemplo que la probabilidad de sobrevivir si se embarcó en “C” es de un 55.36%

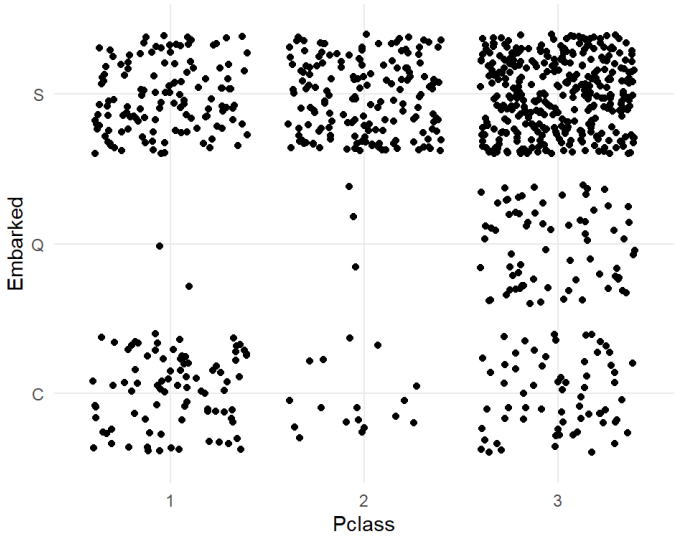
```
t<-table(train$Embarked,train$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

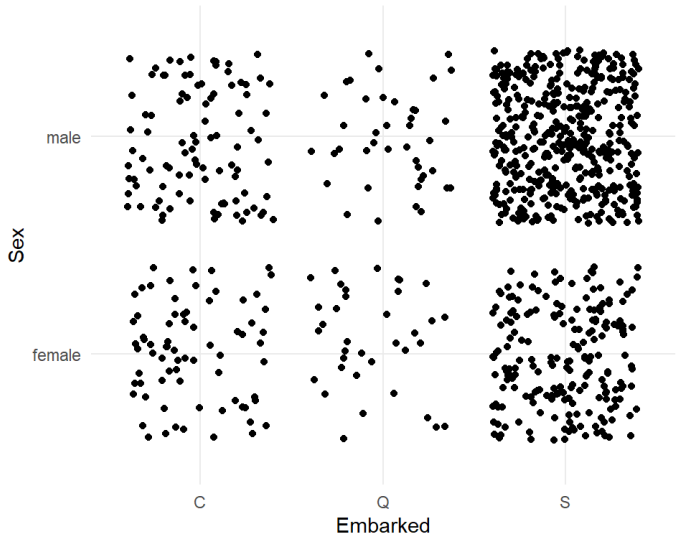
```
##
##            0        1
##   C 44.64286 55.35714
##   Q 61.03896 38.96104
##   S 66.30435 33.69565
```

A continuación, se muestran visualizaciones que relacionan las variables categóricas entre sí, sin tener en cuenta la superviviencia. De esta manera buscaremos correlaciones de una manera muy visual.

```
# Observamos las posibles relaciones entre las variables categóricas
g1 <- ggplot(data=train, aes(x=Pclass, y=Sex)) +
  geom_jitter() +
  theme_minimal()

g2 <- ggplot(data=train, aes(x=Pclass, y=Embarked)) +
  geom_jitter() +
  theme_minimal()

g3 <- ggplot(data=train, aes(x=Embarked, y=Sex)) +
  geom_jitter() +
  theme_minimal()

grid.arrange(g1, g2, g3, nrow = 2, ncol = 2)
```




Con estos tres gráficos se llega a las siguientes conclusiones:

- La distribución de mujeres en las diferentes clases es mucho más homogénea que la de hombres, que se concentran en la tercera clase.
- En la gráfica que relaciona el puerto de embarque con la clase observamos que proporcionalmente, los embarque de primera clase en el puerto C son mayores, por ello, junto con la gráfica porcentaje de supervivencia respecto a la clase, no se sobrevive más por embarcar en C, sino por ser de primera clase. Mostramos la tabla de probabilidades que nos ayuda a verlo numéricamente.

```
t<-table(train$Embarked,train$Pclass)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

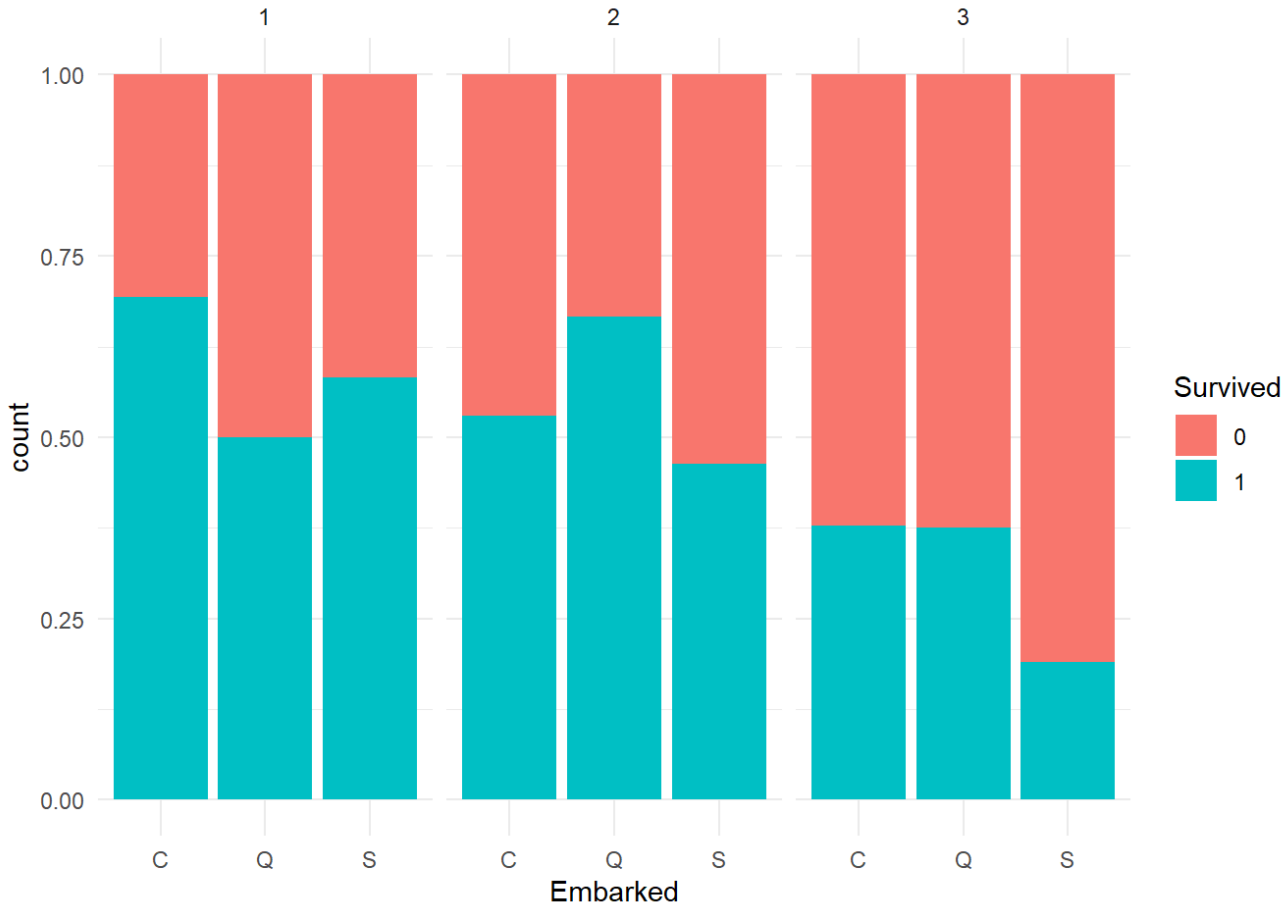
```
##
##            1         2         3
##   C 50.595238 10.119048 39.285714
##   Q  2.597403  3.896104 93.506494
##   S 19.720497 25.465839 54.813665
```

- El último gráfico relaciona el género con el puerto de embarque. Observamos que la mayoría de registros se encuentra en masculino-S, esto, junto con el resto de gráficos, nos ayuda a comprender que el puerto S

es donde embarcaron más personas de 3ª clase, que en su mayoría son hombres.

Veamos ahora como en un mismo gráfico de frecuencias podemos trabajar con 3 variables: Embarked, Survived y Pclass.

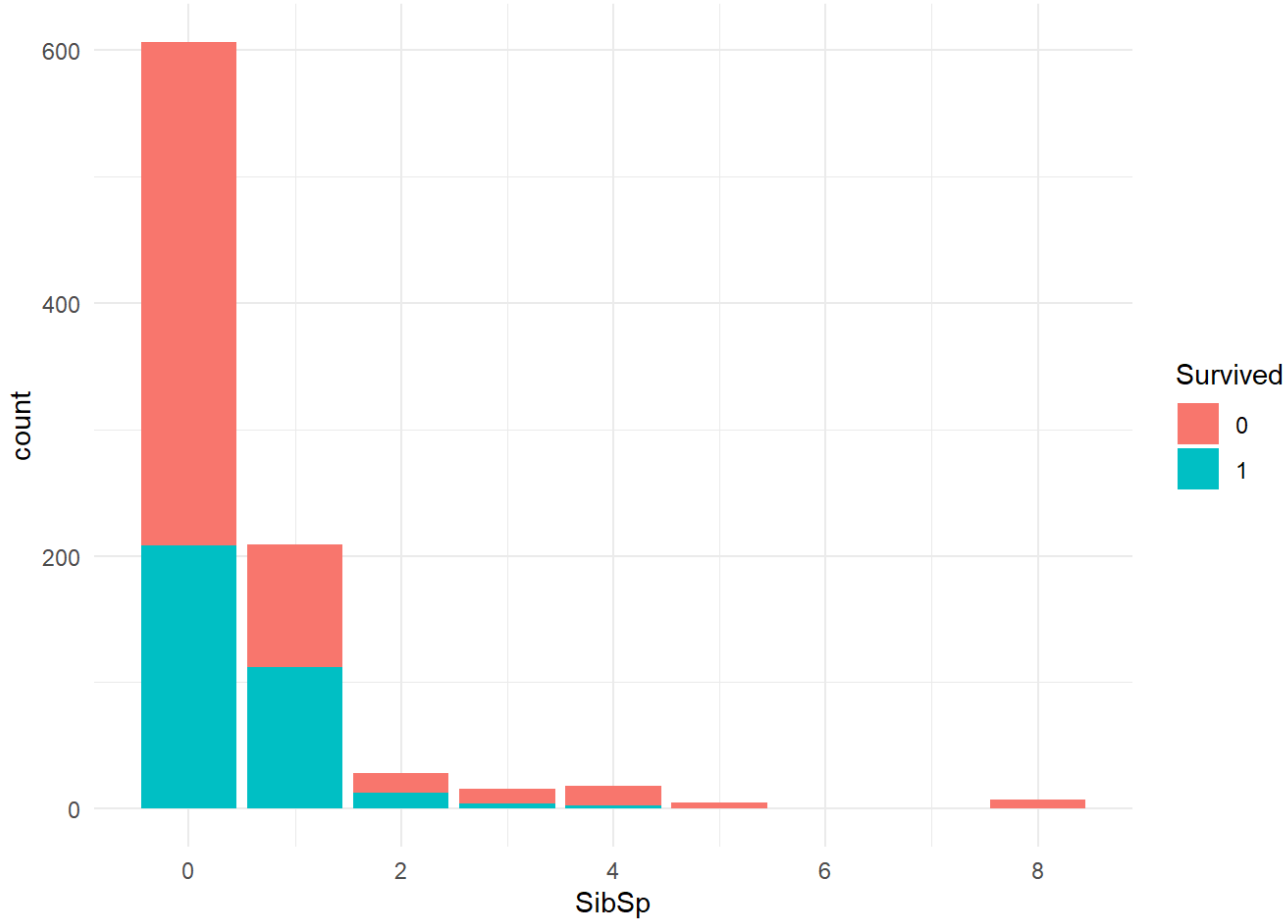
```
# Ahora, podemos dividir el gráfico de Embarked por Pclass:
ggplot(data = train,aes(x=Embarked,fill=Survived)) +
  geom_bar(position="fill") +
  facet_wrap(~Pclass) +
  theme_minimal()
```



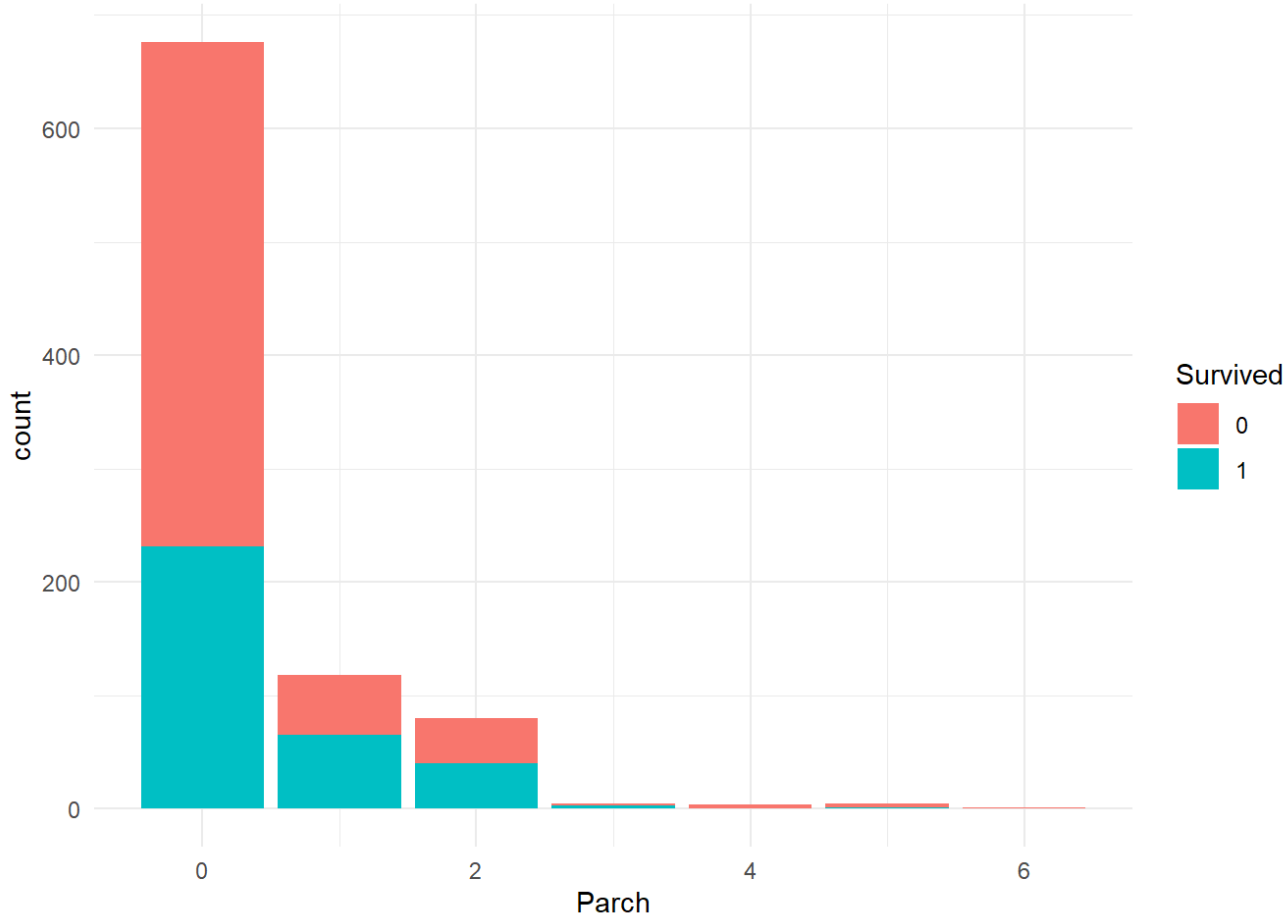

Esta visualzación nos ayuda a comprender que la probabilidad de supervivencia está más relacionada con la clase que con el lugar de embarque.

Comparemos ahora dos gráficos de frecuencias: Survived-SibSp y Survived-Parch

```
# Survivial como función de SibSp y Parch
ggplot(data = train,aes(x=SibSp,fill=Survived)) +
  geom_bar() +
  theme_minimal()
```



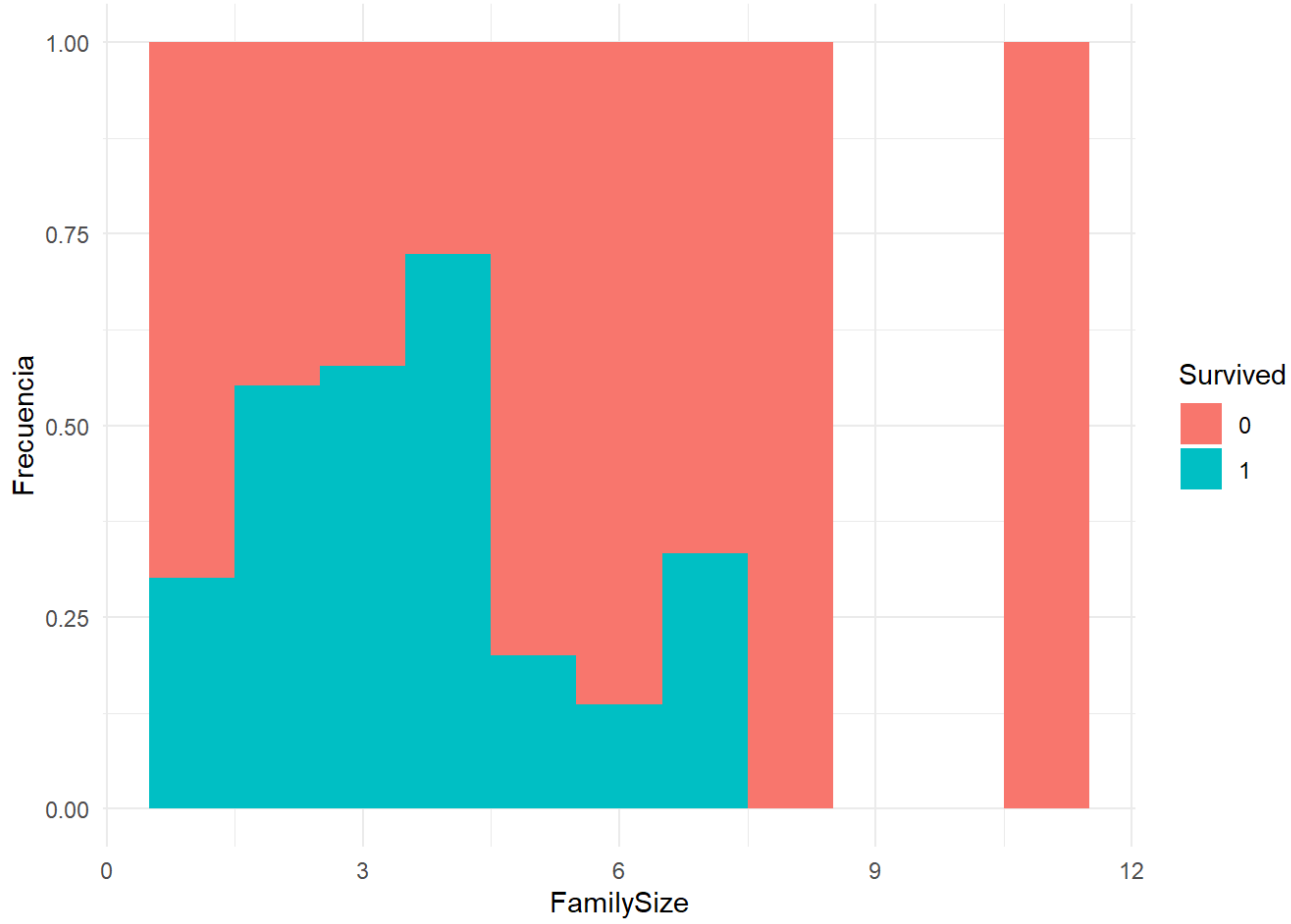

```
ggplot(data = train,aes(x=Parch,fill=Survived)) +
  geom_bar() +
  theme_minimal()
```



Observamos como las forma de estos dos gráficos es similar. Este hecho nos puede indicar presencia de correlaciones altas y que la mayoría de personsas viajaron sin familia.

Veamos un ejemplo de construcción de una variable nueva: Tamaño de familia

```
# Construimos un atributo nuevo: family size.
train$FamilySize <- train$SibSp + train$Parch +1;
ggplot(data = train[!is.na(train$FamilySize),],aes(x=FamilySize,fill=Survived)) +
  geom_histogram(binwidth =1,position="fill") +
  ylab("Frecuencia") +
  theme_minimal()
```




# 2.7 Conclusiones.

Como se ha podido observar gracias al análisis existen ciertos factores que pueden potenciar la supervivencia de los pasajeros, de modo que:

- Las mujeres y los niños tienen una tasa de supervivencia mucho más alta, lo cual puede corroborar que en un accidente de este tipo son los primeros en tener acceso a las balsas salvavidas.
- La clase en la que viaja el individuo influye de modo que, cuanto más alta sea mayores serán las probabilidades de sobrevivir.
- La zona de embarque no parece influir directamente en la supervivencia aunque podría dar lugar a dudas puesto que el acceso al barco de las distintas clases de pasajeros no se encuentra distribuido de forma uniforme.

La información proporcionada en el conjunto de datos es de gran utilidad a la hora de determinar los factores de supervivencia en un accidente de este tipo aunque probablemente podrían existir otros que no se recogen, como podrían ser: eventos activos durante el accidente, posición de las salidas de emergencia con respecto a la localización de los pasajeros...

```
# csv con datos tratados
write.csv(train, file = "train_tratados.csv")
```

# 2.8 Tabla de contribuciones

| Contribuciones | Firma |
|---|---|
| Investigación previa | XPP, DSR |
| Redacción de las respuestas | XPP, DSR |
| Desarrollo de código | XPP, DSR |