# Práctica 2: Tipología y ciclo de vida de los datos

*Código: titanic_passengers_R*

*Autores: Luis Alberto Bayo Martín y Miguel A Bermejo Agueda*

*9 de June, 2019*



## 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.

Se entregará un solo archivo `github` con la solución:

- https://github.com/luisalbertobayo/titanic

## 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.

# 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.

# 4 Contenido

El objetivo de esta actividad será el tratamiento de el dataset:

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

Las diferentes tareas a realizar siguiendo las principales etapas de un proyecto analítico 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.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 3.2. Identificación y tratamiento de valores extremos.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 4.2. Comprobación de la normalidad y homogeneidad de la varianza. 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.

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.

# 5 Librerías

Librerías necesarias para el desarrollo de la actividad:

***'knitr'***, permite integrar código R en archivos de distintos formato.

***'lubridate'***, facilita las operaciones básicas con variables de tipo fecha y/o tiempo.

***'stringr'***, entre otras funciones, permite la manipulación de caracteres individuales dentro de las cadenas en los vectores de caracteres, incluyendo el trabajo con espacios en blanco.

***'plyr'***, conjunto de herramientas que resuelve un conjunto común de problemas. Permite dividir un problema en partes manejables, operar en cada pieza y luego volver a unir todas.

***'dplyr'***, permite trabajar con objetos del dataframe, tanto en la memoria como fuera de ella.

***'nortest'***, permite probar la hipótesis compuesta de normalidad mediante diferentes. pruebas.

***'ggplot2'***, permite generar gráficos para crear gráficos elegantes y complejos.

# 6 Resolución

## 6.1 Descripción del dataset

Se ha escogido el dataset *Titanic: Machine Learning from Disaster* (https://www.kaggle.com/c/titanic), recomendado en el enunciado de la práctica. El conjunto de datos para el análisis contiene 12 atributos (columnas) y 1309 muestras (filas).

- **PassengerId**, identificador del pasajero.
- **Survived**, variable booleana que indica si el pasajero sobrevivió al hundimiento (1) o no (0).
- **Pclass**, clase en la que viajaba el pasajero. Puede tomar los valores de: 1 = 1st; 2 = 2nd; 3 = 3rd.
- **Name**, nombre del pasajero.
- **Sex**, género del pasajero.
- **Age**, edad del pasajero.
- **SibSp**, número de familiares de segundo grado de parentesco (hermanos y hermanastros) más cónyuge abordo con los que viaja el pasajero.
- **Parch**, número de familiares de primer grado de parentesco (padres e hijos) abordo con los que viaja el pasajero.
- **Ticket**, número identificativo del ticket.
- **Fare**,tarifa del ticket.
- **Cabin**, número de camarote.
- **Embarked**, puerto en el que embarcó el pasajero: C = Cherbourg; Q = Queenstown; S = Southampton.

## 6.2 Importancia y objetivos de los análisis

Se quiere realizar un estudio demográfico para averiguar qué características comunes reúnen los pasajeros que no sobrevivieron en base a los datos conocidos y observar que atributos de estos pudieron influir más en el hecho.

Este dataset da opción al diseño de un modelo de *machine learning* al poder entrenar al modelo con el fin de predecir si un pasajero sobrevive en base a sus datos conocidos mediante modelos de regresión.

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

El dataset está divido en dos archivos:

- training set (**train.csv**)
- test set (**test.csv**)

El set de entrenamiento, *train.csv*, tiene la finalidad de usarse para construir el modelo de aprendizaje automático, pues contiene los 12 atributos definidos anteriormente, incluida la variable **Survived** para cada pasajero.

El fichero de pruebas, *test.csv*, debería usarse para la validación de la calidad del modelo de *machine learning*, pues contiene datos invisibles al no dar a conocer si el pasajero sobrevivió o no (exclusión del atributo **Survived**).

Comentar que en el *Kaggle* se incluye un tercer archivo, *gender_submission*, compuesto solo por dos atributos, **PassengerId** y una predicción donde se asume que sólo sobreviven las mujeres referente al dataset de pruebas *test.csv*. Por lo que no se hará uso de él.

## 6.4 Limpieza de los datos

Se hace uso de la función `read.csv()` para leer los archivos **train.csv** y **test.csv** a estudio.

```
# read data
datos_train <- read.csv(file="train.csv", header=TRUE, sep=",", strip.white=TRUE, encoding="UTF-8")
datos_test <- read.csv(file="test.csv", header=TRUE, sep=",", strip.white=TRUE, encoding="UTF-8")
```

Una vez leidos se unifican ambos dataset en uno, **master**, dimensionándolo con la estructura del dataset **train**.

```
datos_master <- bind_rows(datos_train, datos_test)
filas = dim(datos_train)

nomAtributos <- names(datos_master)
```

Para concer la asignación que R ha realizado de cada atributo se hace uso de `sapply()` .

```
#Tipos de variables
asignacionAtribR <- sapply(datos_master, class)

character <- c(4, 9, 11)
factor <- c(2, 3, 5, 10, 12)
integer <- c(1, 7, 8)
numeric <- c(6, 10)
claseAtributos <- vector(mode="character",length=ncol(datos_master))
claseAtributos[character] <- "character"
claseAtributos[factor] <- "factor"
claseAtributos[integer] <- "integer"
claseAtributos[numeric] <- "numeric"
difAtribConR <- nomAtributos[asignacionAtribR != claseAtributos]
```

Se han encontrado como asignaciones erróneas los atributos: Survived, Pclass, Embarked

De manera que la conversión necesaria a realizar es:

```
kable(data.frame(variables = difAtribConR, clase = c("factor", "factor", "factor")))
```

| variables | clase |
|---|---|
| Survived | factor |
| Pclass | factor |
| Embarked | factor |

Comentar que se ha considerado el atributo **age** como una variable cuantitativa continua, pues la edad que normalmente se da en años, se puede precisar más y dar en unidades más pequeñas al año como pueden ser los meses. Y sería el caso que se aplica aquí, pues hay varios registros que recogen decimales en su valor, definiendo al atributo **age** como una variable cuantitativa continua.

Visualización descriptiva de los datos.

```
summary(datos_master)
```

```
##   PassengerId   Survived   Pclass      Name              Sex
##  Min.   :   1   0   :549   1:323   Length:1309        female:466
##  1st Qu.: 328   1   :342   2:277   Class :character   male  :843
##  Median : 655   NA's:418   3:709   Mode  :character
##  Mean   : 655
##  3rd Qu.: 982
##  Max.   :1309
##
##       Age            SibSp            Parch          Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.000   Length:1309
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000   Class :character
##  Median :28.00   Median :0.0000   Median :0.000   Mode  :character
##  Mean   :29.88   Mean   :0.4989   Mean   :0.385
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :80.00   Max.   :8.0000   Max.   :9.000
##  NA's   :263
##       Fare            Cabin           Embarked
##  Min.   :  0.000   Length:1309         :  2
##  1st Qu.:  7.896   Class :character   C:270
##  Median : 14.454   Mode  :character   Q:123
##  Mean   : 33.295                      S:914
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
```

En base al planteamiento que se quiere llevar a cabo, hay ciertos atributos que no serían de interés para el estudio, pues no aportan datos significativos para el análisis. Como son **Name**, **Ticket** y **Cabin**. Por ello, se decide eliminarlos.

```
datos_master_Anls <- datos_master[, -(4)]
datos_master_Anls <- datos_master_Anls[, -(8)]
datos_master_Anls <- datos_master_Anls[, -(9)]

head(datos_master_Anls)
```

```
##   PassengerId Survived Pclass    Sex Age SibSp Parch    Fare Embarked
## 1           1        0      3   male  22     1     0  7.2500        S
## 2           2        1      1 female  38     1     0 71.2833        C
## 3           3        1      3 female  26     0     0  7.9250        S
## 4           4        1      1 female  35     1     0 53.1000        S
## 5           5        0      3   male  35     0     0  8.0500        S
## 6           6        0      3   male  NA     0     0  8.4583        Q
```

## 6.4.1 Ceros y elementos vacíos

Es muy importante conocer si existen valores nulos (campos vacíos) y la distribución de los valores que poseen las variables para asi poder realizar una correcta intepretación de ellos.

```
#Búsqueda de valores vacíos en atributos declarados como números, (NA), o como cadena de #caracteres, ("").
colSums(is.na(datos_master_Anls))
```

```
## PassengerId    Survived      Pclass         Sex         Age       SibSp
##           0         418           0           0         263           0
##       Parch        Fare    Embarked
##           0           1           0
```

```
colSums(datos_master_Anls == "")
```

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

Como era de esperar, se observan los 418 registros procedentes del archivo **test.csv** que no tienen valores para el atributo **Survived**. De esta manera, se decide eliminarlos, pues **Survived** es un atributo esencial en el que se centra el estudio de los datos, y haciéndolo, aún habría datos suficientes para obtener conclusiones sin sesgar los resultados.

```
datos_master_Anls <- datos_master_Anls[which(datos_master_Anls$Survived!="NA"),]
```

Para los valores vacíos del atributo **Age**, se decide completarlos con la media del propio atributo.

```
datos_master_Anls$Age[is.na(datos_master_Anls$Age)] <- round(mean(datos_master_Anls$Age, na.rm = T), d
igits = 2)
```

El valor vacío encontrado en el atributo **Fare** no se considera ya que corresponde a un registro procedente del archivo **test.csv**, de manera que se ha eliminado con el borrado de estos.

En el atributo **Embarked** hay dos registros vacíos para los que se les asigna el valor *C* (*Cherbourg*).

```
datos_master_Anls$Embarked[datos_master_Anls$Embarked == ""] = "C"
```

Una vez trabajodos los datos vacíos y no habiendo encontrado ningún registro no interpretable bajo definición de los atributos, el dataset a estudio presenta la siguente dimensión.

```
dim(datos_master_Anls)
```

```
## [1] 891   9
```

```
summary(datos_master_Anls)
```

```
##   PassengerId    Survived Pclass      Sex           Age
##  Min.   :  1.0   0:549    1:216   female:314   Min.   : 0.42
##  1st Qu.:223.5   1:342    2:184   male  :577   1st Qu.:22.00
##  Median :446.0            3:491                Median :29.70
##  Mean   :446.0                                 Mean   :29.70
##  3rd Qu.:668.5                                 3rd Qu.:35.00
##  Max.   :891.0                                 Max.   :80.00
##      SibSp           Parch             Fare        Embarked
##  Min.   :0.000   Min.   :0.0000   Min.   :  0.00    :  0
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:  7.91   C:170
##  Median :0.000   Median :0.0000   Median : 14.45   Q: 77
##  Mean   :0.523   Mean   :0.3816   Mean   : 32.20   S:644
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   Max.   :512.33
```

**master_clean** contiene 891 registros y 9 atributos, siendo estos: PassengerId, Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked.

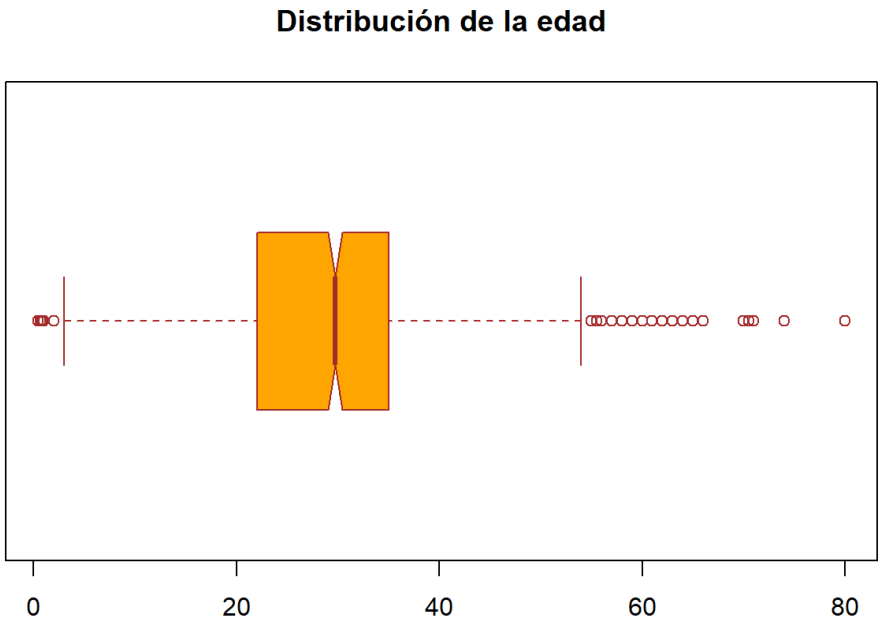
## 6.4.2 Valores extremos

A continuación se comprueba si existe la presencia en el dataset de valores extremos/atípicos(*outliers*), tanto máximos como mínimos en cada uno de sus atributos, de manera que se identifiquen aquellos posibles valores que parecen inconsistentes con el resto de registros. El uso de estos valores en cálculos probabilísticos puede dar lugar a errores o desviaciones en las estimaciones, por ejemplo en la media.

Uno de los métodos más visuales para poder identificar estos valores es mediante la representación gráfica de diagramas de cajas o *boxplot*, ya que permite observar todos estos valores (mínimos y máximos) y la distribución que toma el resto del conjunto.

Se procede con ello para el atributo **Age**.

```
boxplot(datos_master_Anls$Age, main = "Distribución de la edad", col = "orange", border = "brown", hor
izontal = TRUE, notch = TRUE)
```

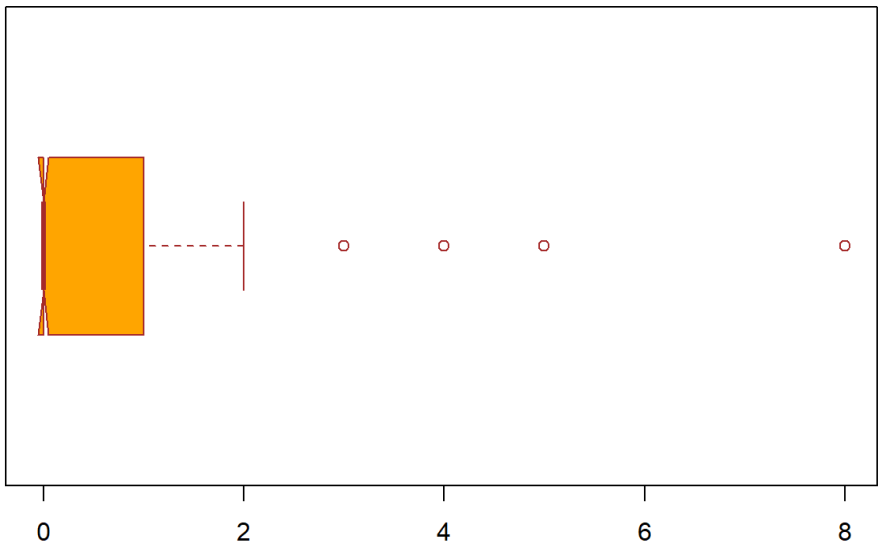



Se observa que existen *outliers* máximos (cercanos a 80) y mínimos (cercanos a 0) pero que están dentro del rango válido para el atributo que representan, ya que corresponderían a bebés y a pasajeros de edad avanzada. Por ello se mantienen en la muestra.

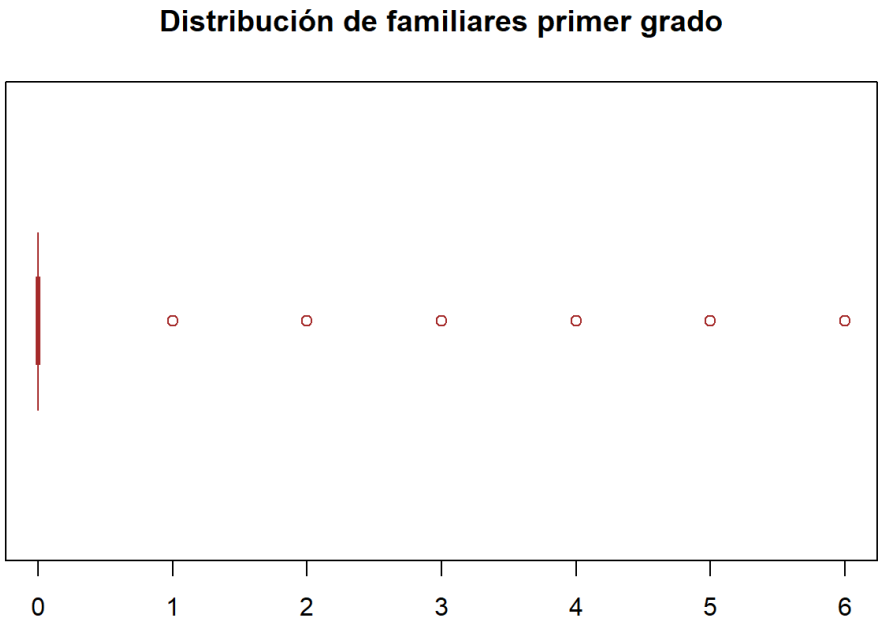
Extendiendo la misma representación para las variables de **SibSp**, **Parch** y **Fare**:

```
boxplot(datos_master_Anls$SibSp, main = "Distribución de familiares segundo grado", col = "orange", bo
rder = "brown", horizontal = TRUE, notch = TRUE)
```

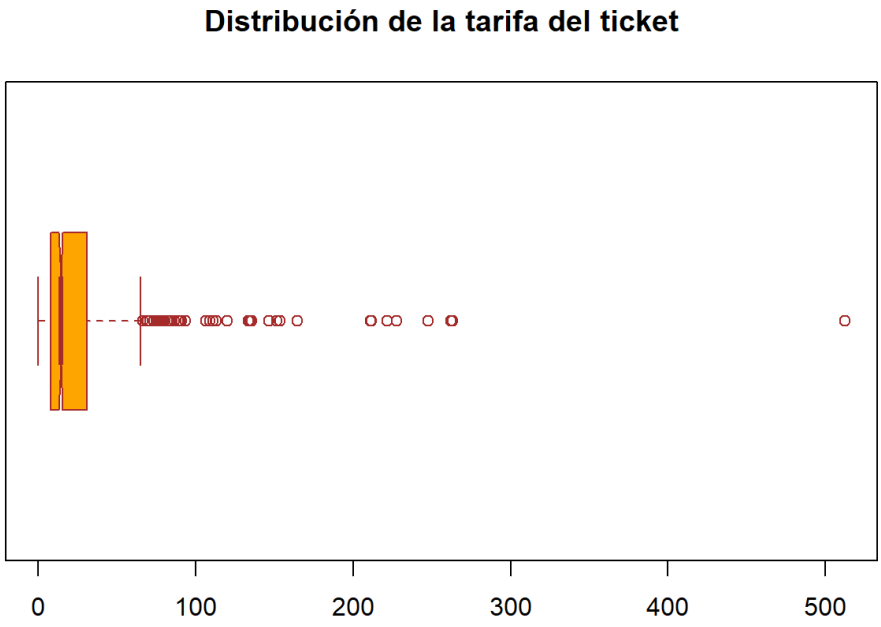
```
## Warning in bxp(list(stats = structure(c(0, 0, 0, 1, 2), .Dim = c(5L, 1L),
## class = structure("integer", .Names = "")), : some notches went outside
## hinges ('box'): maybe set notch=FALSE
```



```
boxplot(datos_master_Anls$Parch, main = "Distribución de familiares primer grado", col = "orange", bor
der = "brown", horizontal = TRUE, notch = TRUE)
```




```
boxplot(datos_master_Anls$Fare, main = "Distribución de la tarifa del ticket", col = "orange", border
= "brown", horizontal = TRUE, notch = TRUE)
```




A pesar de que en estos casos también se observan *outliers*, las variables toman valores dentro de los márgenes reales del atributo que definen. Los pasajeros viajan en su mayoría no acompañados de familiares de segundo grado de parentesco (hermanos y hermanastros) o cónyuges, definido por la variable *SibSp*, ni por familiares de primer grado de parentesco, definido por la variable *Parch*. Aún así se necesitaría un análisis más profundamente para ver la distribución de los atributos y darles una interpretación detallada más allá de un primer vistazo gráfico.

Aunque es importante destacar que con esta representación se ha conseguido el objetivo de observar que no exsiten valores fuera de un rango real en cuanto a valores que toman las variables según lo que representan.

Este análisis es aplicable al atributo **Fare** en el que existen *outliers* que también se consideran dentro de rango del de precios. De hecho, se

ha comprobado que el valor *outlier* más alejado de la media, que corresponde a un valor de billete de £512,13 (presumublemente en libras esterlinas), que se aproxima mucho a la información encontrada en diferentes portales web que cita como fuente los archivos de las Cortes de Distrito de los Estados Unidos y lo asocia a las personas relacionadas en el dataset con ese billete. Aún así, la información puede variar muy ligeramente de un portal web a otro por lo que se comenta solo como una curiosidad sin nombrar a las fuentes, pero que para el análisis da cierta credibilidad en ese *outlier*, y de ahí la decisión de mantenerlo.

Una vez que se ha acometido sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, se genera un nuevo fichero de salida denominado **master_clean.csv**.

```
new.datos_master_Anls <- "master_clean.csv"
write.csv(datos_master_Anls, file = new.datos_master_Anls, row.names = FALSE)
```

# 6.5 Análisis de los datos

Con objetivo de explicar las principales características de los mismos, y así tratar de responder a las preguntas planteadas en el marco del proyecto de datos.

## 6.5.1 Selección de los grupos de datos a analizar

Se generan los grupos de datos que pueden resultar interesantes para el análisis.

```
# Asociación en base a supervivencia
pasajeros.SIsobrevive <- datos_master_Anls[datos_master_Anls$Survived == "1",]
pasajeros.NOsobrevive <- datos_master_Anls[datos_master_Anls$Survived == "0",]

# Asociación en base al género
pasajeros.hombre <- datos_master_Anls[datos_master_Anls$Sex == "male",]
pasajeros.mujer <- datos_master_Anls[datos_master_Anls$Sex == "female",]

# Asociación en base a la clase de billete
pasajeros.hombre <- datos_master_Anls[datos_master_Anls$Pclass == "1",]
pasajeros.mujer <- datos_master_Anls[datos_master_Anls$Pclass == "2",]
pasajeros.mujer <- datos_master_Anls[datos_master_Anls$Pclass == "3",]

# Asociación en base al acompañamientos de familiares.
pasajeros.SIconSegun <- datos_master_Anls[datos_master_Anls$SibSp > "0",]
pasajeros.NOconSegun <- datos_master_Anls[datos_master_Anls$SibSp == "0",]
pasajeros.SIconPrim <- datos_master_Anls[datos_master_Anls$Parch > "0",]
pasajeros.NOconPrim <- datos_master_Anls[datos_master_Anls$Parch == "0",]
```

## 6.5.2 Normalidad y homogeneidad de la varianza

Como comprobación de que los registros de los atributos cuantitativos tienen origen en una población con distribución normal, se hace uso de la prueba de normalidad de **Anderson-Darling**. Es decir, se comprueba que para cada prueba se obtiene un p-valor superior al **nivel de significación**. Para esto, dado que no se conoce su \(\alpha\) (**nivel de significación**), y sabiendo que su fijación no es un probelma estrictamente matemático, se decide utilizar el valor estándar que se suele dar de \(\alpha\) = 0,05, lo que significa que, aunque la hipótesis nula sea cierta, los datos de cinco de cada cien muestras harán rechazarla. Es decir, se acepta que se puede rechazar la hipótesis nula de forma equivocada cinco de cada cien veces.

De cumplirse esto, se considera que el atributo a estudio contiene una distribución normal.

```
alpha = 0.05

nomAtributos = names(datos_master_Anls)

if (ncol(datos_master_Anls) >0){
  cat("Atributos que no están definidos bajo una distribución normal:\n")
}
```

```
## Atributos que no están definidos bajo una distribución normal:
```

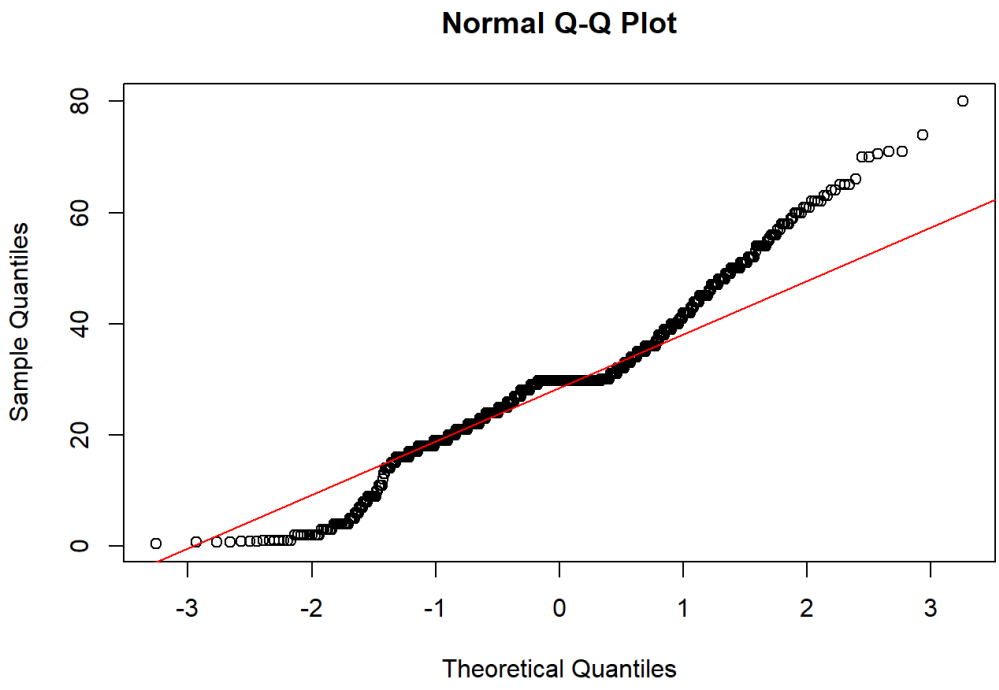
```
for (i in 1:ncol(datos_master_Anls)) {

  if (is.integer(datos_master_Anls[,i]) | is.numeric(datos_master_Anls[,i])) {
    p_value = ad.test(datos_master_Anls[,i])$p.value

    if (p_value < alpha) {
      cat(nomAtributos[i])
      if (i < ncol(datos_master_Anls) - 1) {
        cat(", ")
      }
    }
  }
}
```

```
## PassengerId, Age, SibSp, Parch, Fare
```

Para corroborar esta asunción de normalidad en los atributos comentados: **Age**, **SibSp**, **Parch** y **Fare**, se hace uso del gráfico Q-Q. Se obvia **PassengerID**, porque es tan solo el identificador del pasajero y no aportaría información al análisis de pruebas estadísticas. Para una muestra de tamaño *n*, se dibujan *n* puntos con los (n+1)-cuantiles de la distribución normal, en el eje horizontal el estadístico de k-ésimo orden, (para k = 1, …, n) de la muestra en el eje vertical. Si la distribución de la variable es la misma que la distribución de comparación se obtendrá, aproximadamente, una línea recta, especialmente cerca de su centro, cargando de normalidad al conjunto de valores.

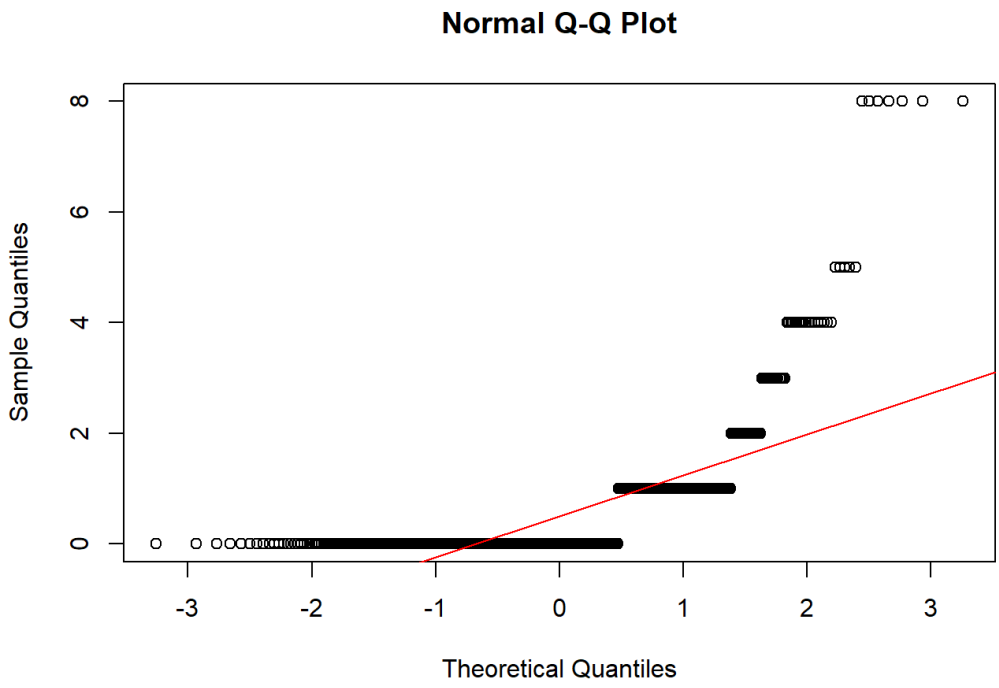
- Normal frente a representación Q-Q de **Age**:

```
qqnorm(datos_master_Anls$Age);qqline(datos_master_Anls$Age, col = 'red')
```



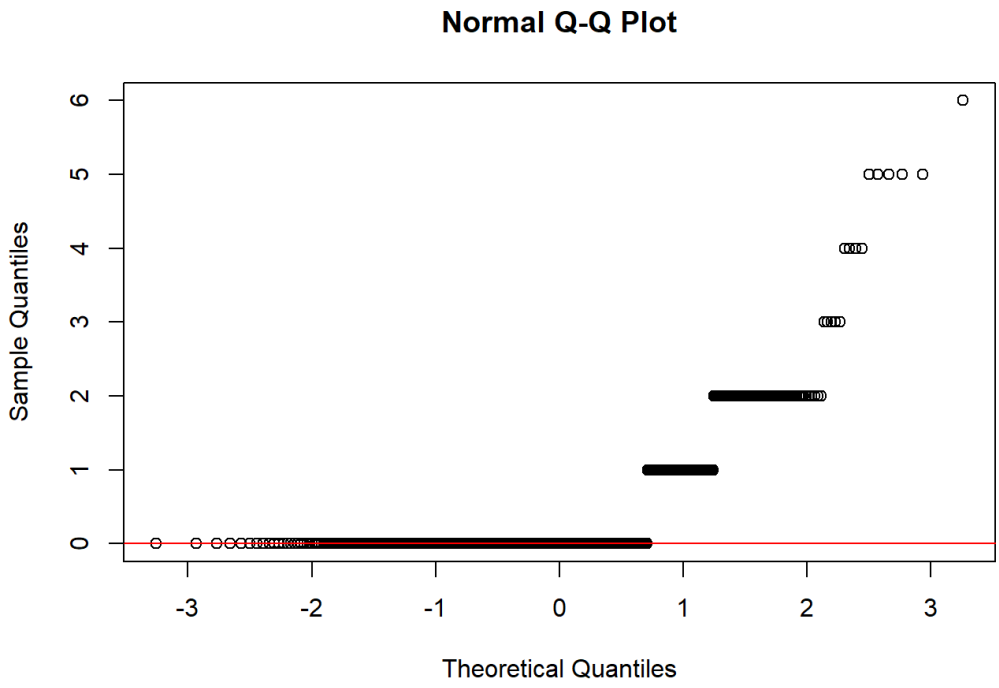

- Normal frente a representación Q-Q de **SibSp**:

```
qqnorm(datos_master_Anls$SibSp);qqline(datos_master_Anls$SibSp, col = 'red')
```



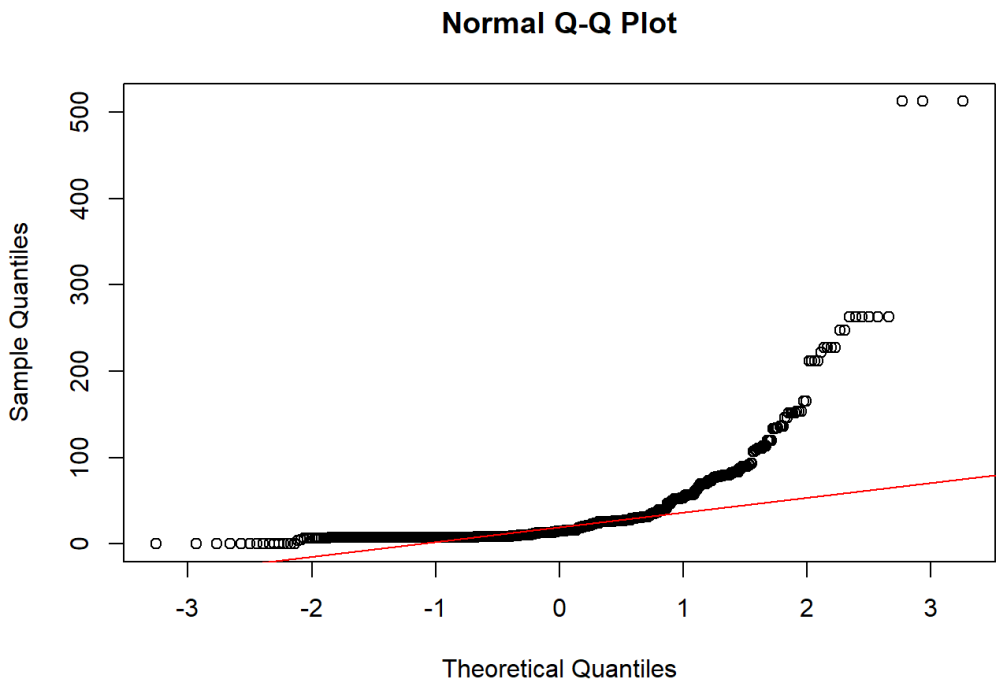

- Normal frente a representación Q-Q de **Parch**:

```
qqnorm(datos_master_Anls$Parch);qqline(datos_master_Anls$Parch, col = 'red')
```




- Normal frente a representación Q-Q de **Fare**:

```
qqnorm(datos_master_Anls$Fare);qqline(datos_master_Anls$Fare, col = 'red')
```




Como se observan desviaciones sustanciales de la linealidad, se rechazan sus hipótesis nulas \(h_0\) de similitud y se aceptarían sus hipótesis alternativas \(h_1\), que afirman que la muestra de datos de **Age**, **SibSp**, **Parch** y **Fare** no siguen una distribución normal.

Respecto a la homogeneidad de la varianza se hace uso del test de **Fligner-Killeen**. Este trata de la alternativa no paramétrica, utilizada cuando los datos no cumplen con la condición de normalidad. De esta manera se decide comprobar la homogeneidad referente a los grupos definidos de pasajeros que viajan solos o acompañados (por parientes de primer o segundo grado). Como se sabe, para el test, la hipótesis nula, \(H_0\), asume igualdad de varianzas en los diferentes grupos de datos.

Comentar que, como el atributo **Survived** es una variable cualitativa nominal, se decide duplicarla en un nuevo atributo, **SurvivedCD**, con el fin de redefinirla como una variable cuantitativa discreta para poder resolver la homogeneidad de la varianza mediante el test de **Fligner-Killeen**, y así no generar inconsistencias en análisis posteriores al cambiar su identidad.

```
datos_master_Anls$SurvivedCD <- datos_master_Anls$Survived
datos_master_Anls$SurvivedCD <- as.integer(datos_master_Anls$SurvivedCD)

fligner.test(SurvivedCD ~ SibSp, data = datos_master_Anls)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SurvivedCD by SibSp
## Fligner-Killeen:med chi-squared = 21.832, df = 6, p-value =
## 0.001298
```

```
fligner.test(SurvivedCD ~ Parch, data = datos_master_Anls)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SurvivedCD by Parch
## Fligner-Killeen:med chi-squared = 17.231, df = 6, p-value =
## 0.00847
```

Dado que en ambos resultados el valor de p-value es menor que \(\alpha\) (**nivel de significación**), se rechaza la hipótesis nula de homocedasticidad conlcuyendo que el atributo **Survived** presenta varianzas estadísticamente diferentes para los grupos de **SibSp** y **Parch**.

## 6.6 Pruebas estadísticas

Aplicación de diversas pruebas estadísticas con el objetivo de responder a preguntas que relacionen los distintos atributos del juego de datos.

Este tipo de análisis tiene por objetivo modelar los datos a través de la distribución conocida. Sabiendo que el conjunto de datos estudiado representa una fracción de la totalidad de la población *titanic*, su objetivo es inferir cómo es esa población, asumiendo un grado de error en las estimaciones por el hecho de disponer de una muestra reducida de los datos.

Los siguientes apartados describen algunos ejemplos de análisis de este tipo, como son la comparación de grupos mediante los contrastes de hipótesis, las regresiones o las correlaciones.

## 6.6.1 Correlación

Se realiza un análisis de *correlación* entre los distintos atributos, con el fin de aprender cuál de ellos influye más sobre la supervivencia de un pasajero.

Como el coeficiente de correlación es una medida asociada entre dos variables, se estudia en las variables cuantitativas: **Age**, **SibSp**, **Parch** y **Fare**. Se obvia **PassengerID**, porque es tan solo el identificador del pasajero y no aportaría información al análisis de pruebas estadísticas.

Como se ha analizado antes, todas ellas no siguen una distribución normal, por lo que se aplica el *coeficiente de correlación de Spearman*, que aparece como una alternativa no paramétrica que mide el grado de dependencia entre dos variables. (Se continua trabajando con el atributo *survived* como variable cuantitativa discreta, **SurvivedCD**, para poder calcular el coeficiente de correlación de Spearman al operar este con vectores numéricos)

Relación entre la edad del pasajero y su posibilidad de supervivencia.

```
spearman_test_Age_Survived = cor.test(datos_master_Anls$Age, datos_master_Anls$SurvivedCD, method = "s
pearman")
spearman_test_Age_Survived = spearman_test_Age_Survived$estimate
```

Relación entre viajar con o sin familiares (el pasajero) y su posibilidad de supervivencia (la del pasajero a estudio).

```
spearman_test_SibSp_Survived = cor.test(datos_master_Anls$SibSp, datos_master_Anls$SurvivedCD, method
= "spearman")
spearman_test_SibSp_Survived = spearman_test_SibSp_Survived$estimate

spearman_test_Parch_Survived = cor.test(datos_master_Anls$Parch, datos_master_Anls$SurvivedCD, method
= "spearman")
spearman_test_Parch_Survived = spearman_test_Parch_Survived$estimate
```

Relación entre el valor del billete del pasajero y su posibilidad de supervivencia.

```
spearman_test_Fare_Survived = cor.test(datos_master_Anls$Fare, datos_master_Anls$SurvivedCD, method =
"spearman")
spearman_test_Fare_Survived = spearman_test_Fare_Survived$estimate
```

```
cat(" Coef. correlación Age =", spearman_test_Age_Survived, "\n",
     "Coef. correlación SibSp =", spearman_test_SibSp_Survived, "\n",
     "Coef. correlación Parch =", spearman_test_Parch_Survived, "\n",
     "Coef. correlación Fare =", spearman_test_Fare_Survived, "\n")
```

```
##  Coef. correlación Age = -0.03910946
##  Coef. correlación SibSp = 0.08887948
##  Coef. correlación Parch = 0.1382656
##  Coef. correlación Fare = 0.3237361
```

Para identificar cuáles son las variables más correlacionadas con el atributo **Survived**, se observa la proximidad del coeficiente de correlación de Spearman con los valores “-1” o “1”, pues estos extremos indican una correlación perfecta. Por el contrario, “0”, marca la ausencia de correlación. Comentar también que el signo es negativo cuando los atributos a estudio son indirectamente proporcionales, y el signo es positivo cuando los atributos a estudio son directamente proporcionales.

De esta manera, se aprecia una total ausencia de correlación para los atributos **Age**, **SibSp** y **Parch**, pues el valor de su coeficiente es muy cercano a 0. Y, aunque para el coeficiente de correlación para el atributo **Fare** es algo más elevado, **0,32**, sigue siendo un valor bajo que marca ausencia de correlación.

Se decide observar también si puede haber relación entre el precio del billete y la edad del pasajero. A priori, parece lógico pensar que la edad puede tener una relación directa con el precio del billete.

```
spearman_test_Fare_Age = cor.test(datos_master_Anls$Fare, datos_master_Anls$Age, method = "spearman")
spearman_test_Fare_Age = spearman_test_Fare_Age$estimate
cat(" Coef. correlación Fare_Age =", spearman_test_Fare_Age, "\n")
```

```
##  Coef. correlación Fare_Age = 0.1188471
```

Como se puede ver, el valor de correlación es muy pequeño, lo que indica una ausencia de correlación asumible.

## 6.6.2 Contraste de hipótesis

Como segunda prueba estadística se decide observar si el precio del billete es superior o inferior dependiendo del sexo del pasajero (hombre o mujer), dado que como se ha visto en el análisis de la correlación, los atributos cuantitativos presentan una ausencia de relación con la supervivencia de los pasajeros.

Para ello, se lleva acabo un contraste de hipótesis sobre las muestras a estudio: **Sex** y **Fare**.

```
pasajeros.mujer.precio <- datos_master_Anls[datos_master_Anls$Sex == "female",]$Fare
pasajeros.hombre.precio <- datos_master_Anls[datos_master_Anls$Sex == "male",]$Fare
```

Como hay ausencia de normalidad, hay que aplicar tests no paramétricos. El test de *suma de rangos de Wilcoxon*, también llamado *Mann-Whitney U-test* es el equivalente no paramétrico al *test t* para dos muestras independientes.

Como se parte del supuesto establecido en el que los datos no presentan una distribución normal, las hipótesis planteadas se hacen en base a las medianas \(Me\).

- Hipótesis nula: la mediana del precio de los billetes para mujeres (\(Me_M\)) es igual a la mediana del precio de los billetes para hombres (\(Me_H\)):

\(H_0: Me_M = Me_H\) \(H_0: Me_M - Me_H = 0\)

- Hipótesis alternativa: la mediana del precio de los billetes para mujeres (\(Me_M\)) es distinta de la mediana del precio de los billetes para hombres (\(Me_H\)):

\(H_1: Me_M \neq Me_H\)

Siendo un contraste bilateral para la hipótesis alternativa.

Se aplica ahora el **test de la U de Mann-Whitney**, sabiendo que la fijación de \(\alpha\) (**nivel de significación**) no es un probelma estrictamente matemático y se decide utilizar el valor estándar que se suele dar de \(\alpha\) = 0,05.

```
u.mannwhitney <- wilcox.test(pasajeros.mujer.precio, pasajeros.hombre.precio, alternative = "two.sided
", paired = FALSE, exact = FALSE, conf.level = 0.95)
u.mannwhitney
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pasajeros.mujer.precio and pasajeros.hombre.precio
## W = 119000, p-value = 9.612e-15
## alternative hypothesis: true location shift is not equal to 0
```

Se obtiene un **pvalor** mucho menor que el **nivel de significación** \(\alpha\) 0.05. Por tanto, se rechaza la hipótesis nula \(H_0\), pudiendo concluir que el precio del billete es significativamente diferente dependiendo de si el pasajero es una mujero o un hombre.

En base a esta conclusión, se ha analizado y se ha observado que:

```
prcioMed_billM <- mean(pasajeros.mujer.precio)
prcioMed_billH <- mean(pasajeros.hombre.precio)

cat(" Valor medio del prcio del billete para mujeres =", prcioMed_billM, "\n",
    "Valor medio del prcio del billete para hombres =", prcioMed_billH, "\n")
```

```
##  Valor medio del prcio del billete para mujeres = 44.47982
##  Valor medio del prcio del billete para hombres = 25.52389
```

El *valor medio del prcio del billete para mujeres* es casi del doble que el *valor medio del prcio del billete para hombres*. Con esto no se está concluyendo que el precio de venta a mujeres fuera diferente que a hombres, sino que el valor del precio medio de un billete adquirido por la muestra mujeres es casi del doble del precio medio del billete adquirido por la muestra hombres. Además, con estos resultados sería interesante ampliar la conclusión obtenida argumentándola desde el marco socio-económico de la época, dependiendo de la tendencia global de sus regresores.

## 6.6.3 Modelo de regresión

Como objtevio de la actividad, se quiere obtener predicciones de si una pasajero de unas determinadas características sobreviría o no.

Para ello, se calcula un modelo de regresión logística utilizando tan sólo regresores cualitativos, pues como se ha visto en los anteriores análisis de datos, los atributos cuantitativos, **Age**, **SibSp**, **Parch** y **Fare**, ofrecen una fuerte ausencia de correalción con el atributo dependiente a estudio **Survived**.

Este tipo de modelo de regresión (*regresión logística*), trata de predecir el resultado de una variable dicotómica dependiente, **Survived**, en función de una serie de variables predictoras, que para este caso se hace uso de los atributos cualitativos: **Pclass**, **Sex** y **Embarked**.

De esta manera, se decide evaluar la probabilidad en la que el pasajero sobrevive o no, y si alguno de sus regresores, **Pclass**, **Sex** y **Embarked**, tienen una influencia significativa, estableciendo un p-valor del contraste individual inferior al 5%.

Se utiliza como categoría de referencia de la variable **Pclass** *3*, de la variable **Sex** *female* y de la variable **Embarked** *S*, y se generan cuatro modelos de regresión logística con estos regresores dos a dos y uno con los tres, para escoger entre ellos el mejor al evaluar su bondad mediante la medida *AIC*. Para ello, dado que esta medida tiene en cuenta tanto la bondad del ajuste como la complejidad del modelo, al comparar los cuatro modelos candidatos, se seleccionará aquel que resulte con el menor AIC.

```
datos_master_Anls$PclassR = relevel(datos_master_Anls$Pclass, ref = '3')
datos_master_Anls$SexR = relevel(datos_master_Anls$Sex, ref = 'female')
datos_master_Anls$EmbarkedR = relevel(datos_master_Anls$Embarked, ref = 'S')

modRegLog1 = glm(Survived ~ PclassR + SexR, family=binomial, data = datos_master_Anls)
summary(modRegLog1)
```

```
##
## Call:
## glm(formula = Survived ~ PclassR + SexR, family = binomial, data = datos_master_Anls)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1877  -0.7312  -0.4476   0.6465   2.1681
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3916     0.1509   2.596  0.00944 **
## PclassR1      1.9055     0.2141   8.898  < 2e-16 ***
## PclassR2      1.0675     0.2205   4.842 1.28e-06 ***
## SexRmale     -2.6419     0.1841 -14.351  < 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: 1186.66  on 890  degrees of freedom
## Residual deviance:  826.89  on 887  degrees of freedom
## AIC: 834.89
##
## Number of Fisher Scoring iterations: 4
```

```
modRegLog2 = glm(Survived ~ PclassR + EmbarkedR, family=binomial, data = datos_master_Anls)
summary(modRegLog2)
```

```
##
## Call:
## glm(formula = Survived ~ PclassR + EmbarkedR, family = binomial,
##     data = datos_master_Anls)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6704  -0.8957  -0.6684   1.0671   1.7935
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3850     0.1243 -11.140  < 2e-16 ***
## PclassR1      1.6500     0.1869   8.829  < 2e-16 ***
## PclassR2      1.2006     0.1899   6.324 2.55e-10 ***
## EmbarkedRC    0.6789     0.1943   3.494 0.000475 ***
## EmbarkedRQ    0.8455     0.2648   3.193 0.001407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1063.7  on 886  degrees of freedom
## AIC: 1073.7
##
## Number of Fisher Scoring iterations: 4
```

```
modRegLog3 = glm(Survived ~ SexR + EmbarkedR, family=binomial, data = datos_master_Anls)
summary(modRegLog3)
```

```
##
## Call:
## glm(formula = Survived ~ SexR + EmbarkedR, family = binomial,
##     data = datos_master_Anls)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9726  -0.5994  -0.5994   0.8246   1.9800
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.9041     0.1417   6.380 1.77e-10 ***
## SexRmale     -2.5298     0.1712 -14.773  < 2e-16 ***
## EmbarkedRC    0.8873     0.2084   4.257 2.07e-05 ***
## EmbarkedRQ   -0.1826     0.2978  -0.613     0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  897.96  on 887  degrees of freedom
## AIC: 905.96
##
## Number of Fisher Scoring iterations: 4
```

```
modRegLog4 = glm(Survived ~ PclassR + SexR + EmbarkedR, family=binomial, data = datos_master_Anls)
summary(modRegLog4)
```

```
##
## Call:
## glm(formula = Survived ~ PclassR + SexR + EmbarkedR, family = binomial,
##     data = datos_master_Anls)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3312  -0.7145  -0.4157   0.6709   2.2324
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2065     0.1699   1.216   0.2241
## PclassR1      1.8425     0.2246   8.204 2.33e-16 ***
## PclassR2      1.1703     0.2274   5.147 2.64e-07 ***
## SexRmale     -2.6118     0.1855 -14.084  < 2e-16 ***
## EmbarkedRC    0.5999     0.2276   2.636   0.0084 **
## EmbarkedRQ    0.4503     0.3157   1.427   0.1537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  818.82  on 885  degrees of freedom
## AIC: 830.82
##
## Number of Fisher Scoring iterations: 4
```

El valor de AIC en el primero modelo es 834.8884, en el segundo modelo AIC es: 1073.6501, en el tercer modelo AIC es: 905.9579 y en el cuarto modelo AIC es: 830.8159, respectivamente.

De esta manera y sabiendo que el criterio para decidir el mejor modelo es el valor más pequeño de AIC obtenido, el mejor modelo es el modelo regresor que añade **Pclass**, **Sex** y **Embarked** al modelo de regresión logística definido, es decir, el *modelo 4*. Auqnue es importante comentar que el primer modelo con los regresores **Pclass** y **Sex** se sitúa muy próximo del *modelo 4* con menor AIC.

```
b <- which(summary(modRegLog4)$coefficients[-1,4] < 0.05)
b <- b + 1
```

El test parcial sobre los coeficientes de PclassR1, PclassR2, SexRmale, EmbarkedRC, ha sido significativo al ser la estimacion de sus coeficientes 1.84245, 1.1703, -2.61182, 0.59991.

Con estos datos, se conoce que la probabilidad de sobrevivir siendo mujer y viajando en primera o segunda clase era muy alta.

Por último se estudia la calidad del ajuste que se ha realizado en base a los regresores seleccionados a través de la matriz de confusión en el mejor modelo, *modelo 4*, suponiendo un umbral de discriminación del 75%.

```
datos_master_Anls$prob_Survived= predict(modRegLog4, datos_master_Anls, type="response")

datos_master_Anls$prob_Survived <- ifelse(datos_master_Anls$prob_Survived > 0.75,1,0)
table(datos_master_Anls$prob_Survived, datos_master_Anls$prob_Survived)
```

```
##
##       0   1
##   0 721   0
##   1   0 170
```

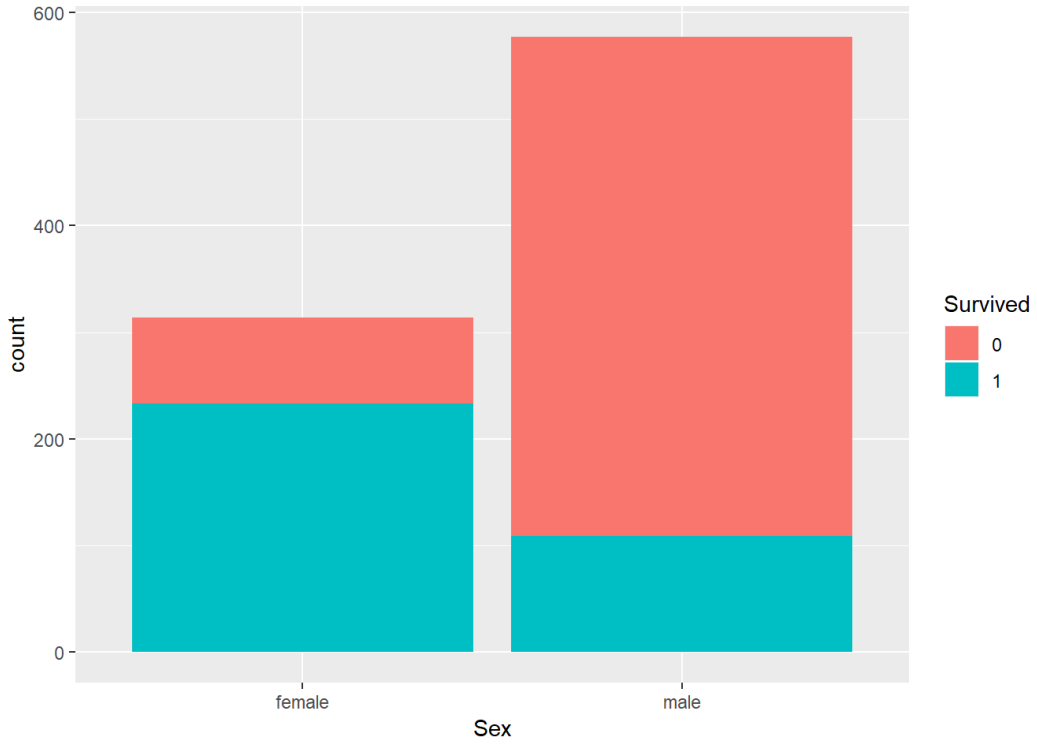
Dada el modelo de regresión logístico generado, no se aprecia ningún falso negativo y ningún falso positivo tampoco.

## 6.7 Representación de los resultados

Análisis de los atributos del juego de datos. Es decir, una vez trabajados los atributos, se analizan las relaciones entre algunos de los atributos más significativos del juego de datos.

Visualización de la relación entre supervivencia y género de los pasajeros. Se vuelve a redimensionar el atributo **Survived** a tipo *factor* para una correcta representación.

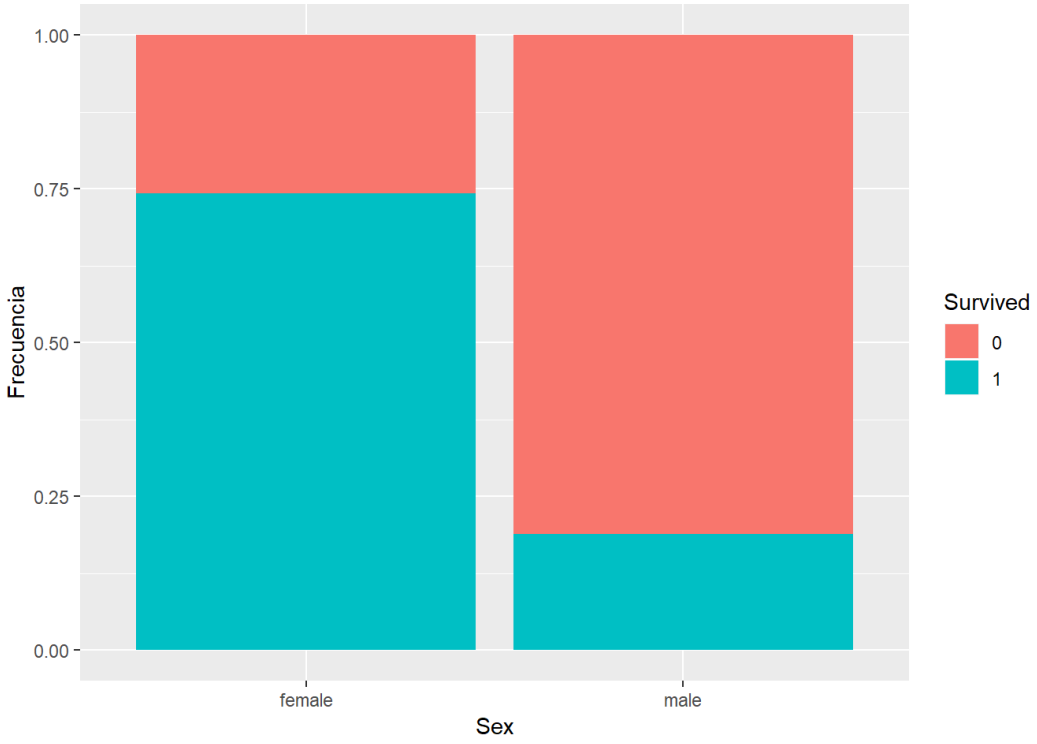
```
filas = dim(datos_master_Anls)
ggplot(data = datos_master_Anls[1:filas,], aes(x = Sex, fill = Survived)) + geom_bar()
```




Se observa que el porcentaje de hombres que no sobrevivieron al desastre del *titanic* es bastante más elevado que el de mujeres.

Como para este estudio se ha cogido una muestra del total de la población al desechar ciertos registros en la limpieza de datos, se decide llevar el estudio a un análisis en frecuncias normalizando los datos a 1 y así trabajar con porcentajes.

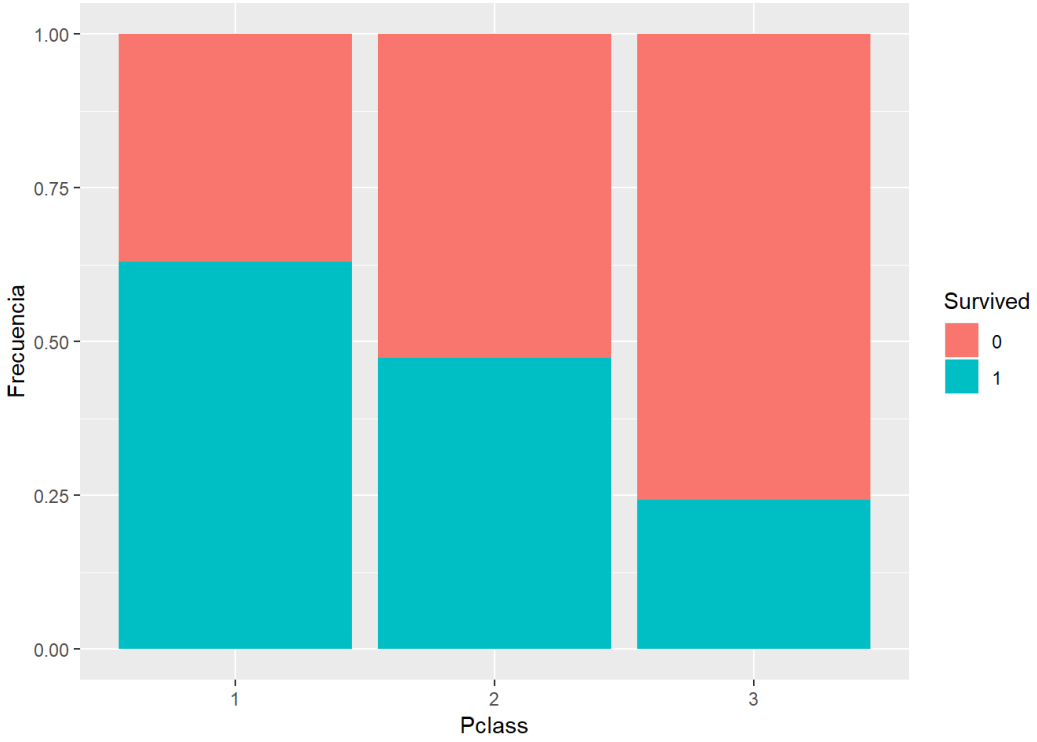
```
ggplot(data = datos_master_Anls[1:filas,], aes(x = Sex, fill = Survived)) + geom_bar(position = "fill"
) + ylab("Frecuencia")
```




El porcentajes de hombres que sobrevivieron no alcanza el 25% de los que viajaban, mientras que el de mujeres es de casi el 75%.

Visualización de la relación de la clase en la que viajaba cada pasejero y la supervivencia.

```
ggplot(data = datos_master_Anls[1:filas,], aes(x = Pclass, fill = Survived)) + geom_bar(position = "fill") + ylab("Frecuencia")
```

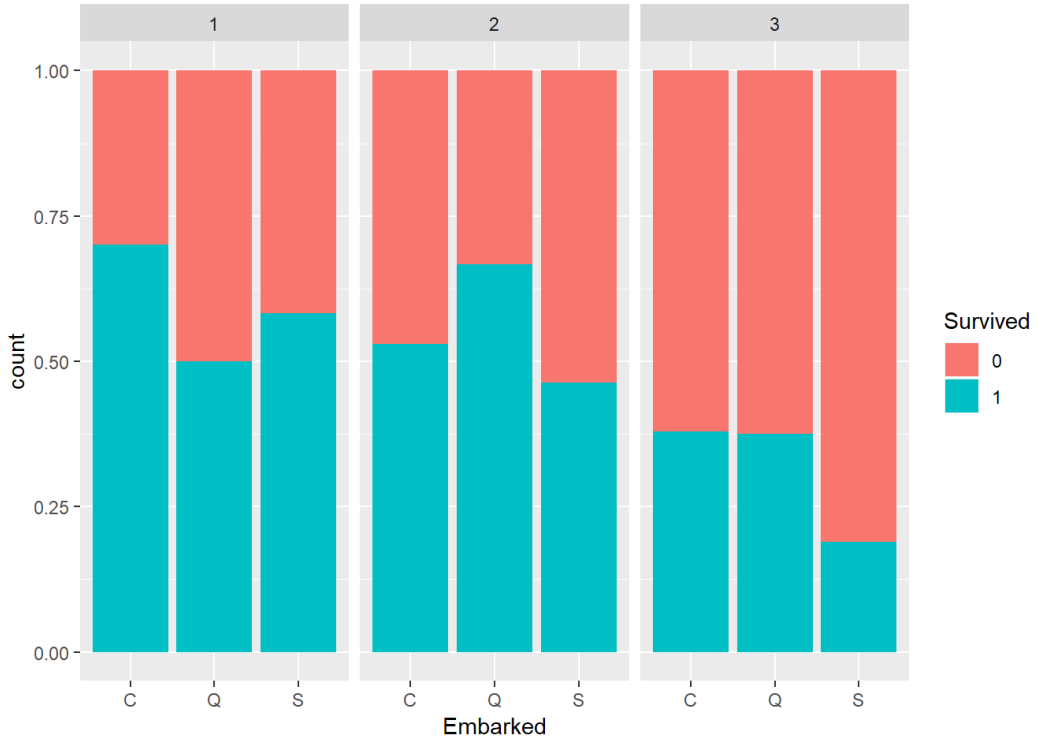



Como es de esperar, dado el contexto historico-social y las condiciones de actuación en caso de catástrofe, el porcentaje de supervivientes aumenta según la clase en la que se viajaba. De manera que los pasajeros de la tercera clase (la más humilde), tenían una porcentaje de sobrevivir del 25%, cuando para los pasajeros de la primera clase era casi del 70%.

Visualización en un mismo gráfico de frecuencias trabajando con 3 atributos: **Embarked**, **Survived** y **Pclass**.

```
ggplot(data = datos_master_Anls[1:filas,], aes(x = Embarked, fill = Survived)) + geom_bar(position = "fill") + facet_wrap(~Pclass)
```

```
## Warning in 1:filas: numerical expression has 2 elements: only the first
## used
```




Profundizando en este *topic*, se observa que, relacionando el análisis planteado con el puerto de embarque, aparentemente los pasajeros

que embarcaron desde el puerto de **Cherbourg** tuvieron más posibilidades de sobrevivir.

Para poder confirmar esto.

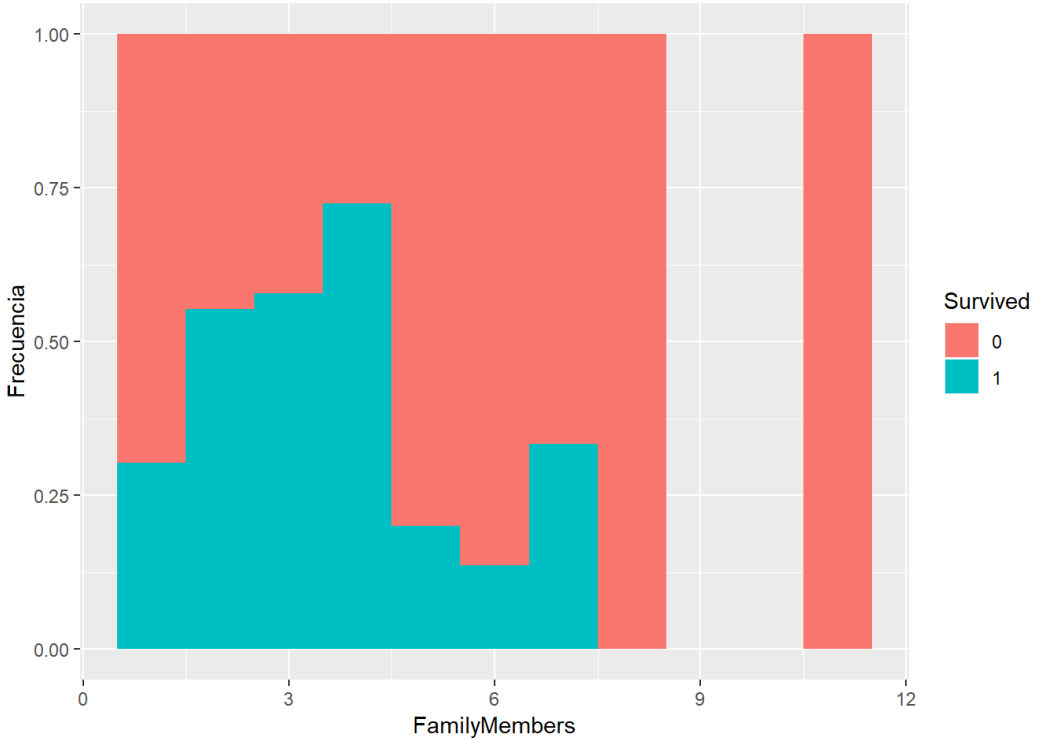
```
table_Survived_Embarked <- table(datos_master_Anls[1:filas,]$Embarked,
datos_master_Anls[1:filas,]$Survived)
for (i in 1:dim(table_Survived_Embarked)){
    table_Survived_Embarked[i,] <- table_Survived_Embarked[i,] / sum(table_Survived_Embarked[i,])*100
}
table_Survived_Embarked
```

```
##
##            0        1
##
##   C 44.11765 55.88235
##   Q 61.03896 38.96104
##   S 66.30435 33.69565
```

Se analizan los números y se comprueba lo comentado: los pasajeros de **Cherbourg** tuvieron más posibilidades de sobrevivir, un 56%, frente a un 39% de los pasajeros de **Queenstown** y un 34% de los pasajeros de **Southampton**, aproximadamente.

Para operar con los atributos de pasajeros que viajaban con parientes de primer y segundo grado, **Parch** y **SibSp**, se decide crear un atributo nuevo que los engoble para estudiarlos como una variable definida como pasajeros que viajaban con familia, **FamilyMembers**, y observar de que manera la canitadad de miembros pudo ser un *handicap* o no.

```
datos_master_Anls$FamilyMembers <- datos_master_Anls$SibSp + datos_master_Anls$Parch + 1;
datos_master_Anls_1 <- datos_master_Anls[1:filas,]
ggplot(data = datos_master_Anls_1[!is.na(datos_master_Anls[1:filas,]$FamilyMembers),], aes(x = FamilyM
embers, fill = Survived)) + geom_histogram(binwidth = 1, position = "fill") + ylab("Frecuencia")
```

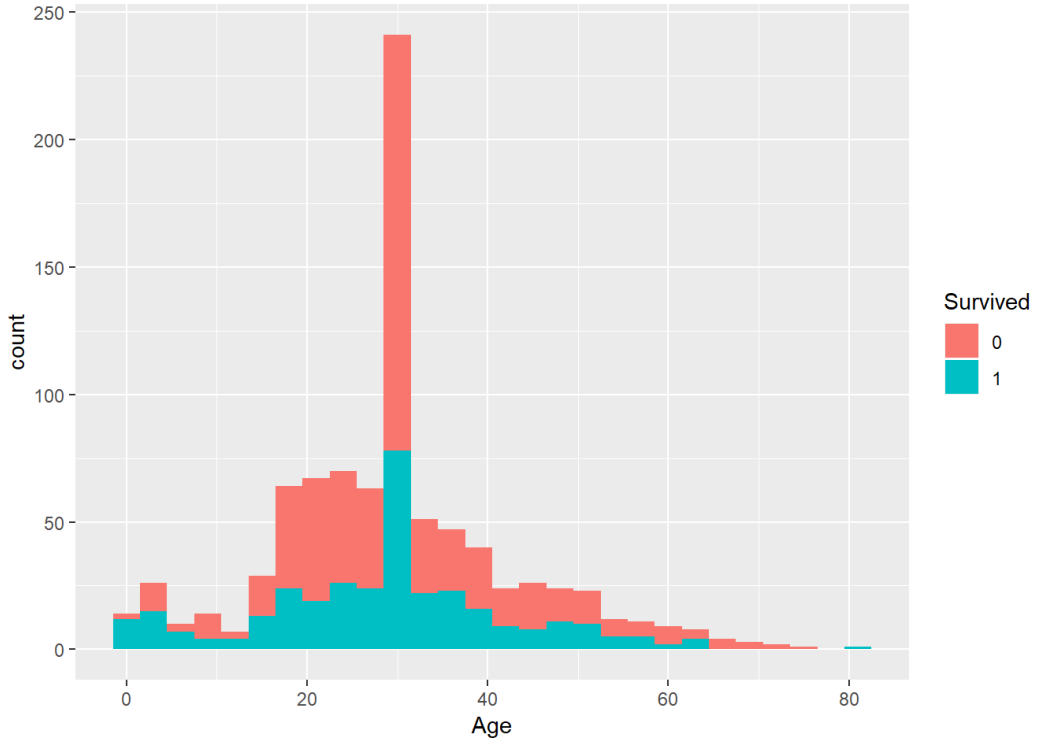



Las familias compuestas entre 2 y 4 miembros tenían más del 50% de posibilidades de supervivencia.

Dado que el atributo **Age** tiene un rango de valores muy amplio, se decide generar rangos de edades para, manteniendo la visibildiad de la variable, hacerla más representativa en un nuevo atributo denominado **AgeRange**.

De esta manera, en una visualización previa.

```
ggplot(data = datos_master_Anls[!(is.na(datos_master_Anls[1:filas,]$Age)),], aes(x = Age, fill = Survi
ved)) + geom_histogram(binwidth = 3)
```




Aparentemente, los pasajeros entre los 0 y los 15 años presentaban un mayor probabilidad de sobrevivir, algo lógico ante unas condiciones de catastrofe en pro de la priorización de la supervicenia humana.

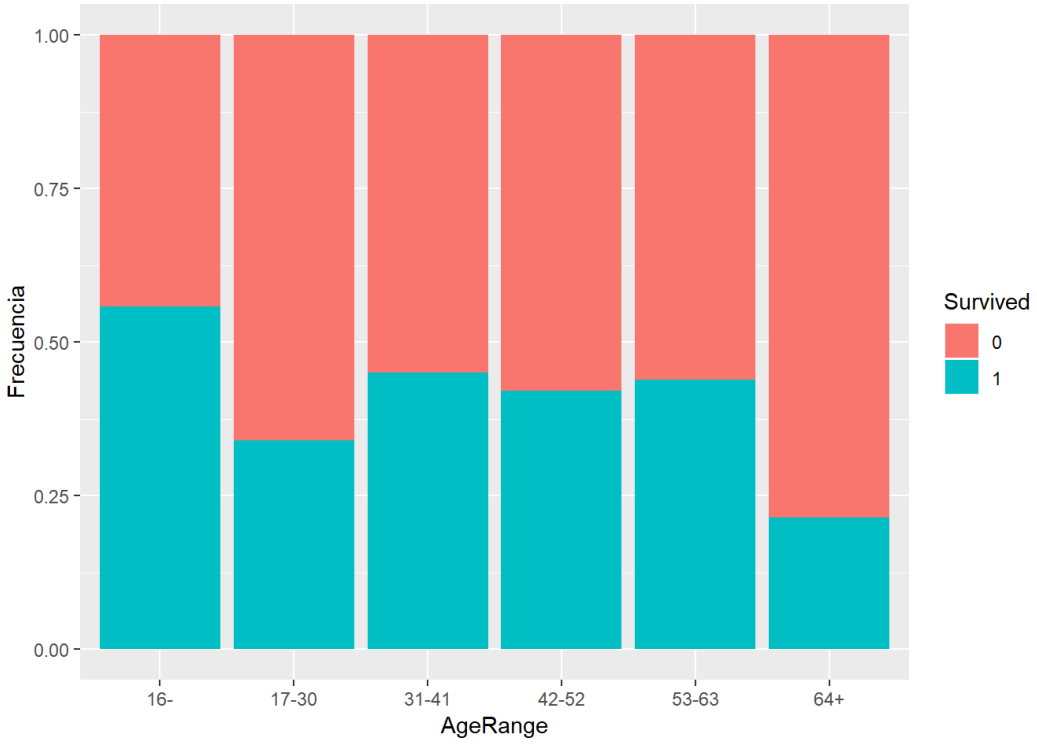
Pero volviendo a lo comentado, se definen unos rangos de edades para el nuevo atributo **AgeRange** en función de age:

- Hasta 15 años = 16
- De 16 a 22 años = 17-30
- De 16 a 22 años = 31-41
- De 34 a 45 años = 42-52
- De 46 a 65 años = 53-63
- Desde 66 años = +64

```
datos_master_Anls$AgeRange <- datos_master_Anls$Age
for (i in 1:length(datos_master_Anls$AgeRange))
{
  if (datos_master_Anls[i,]$AgeRange <= 16)
  {
    datos_master_Anls[i,]$AgeRange <- "16-"
  }
  else
  {
    if (datos_master_Anls[i,]$AgeRange > 16 && datos_master_Anls[i,]$AgeRange <= 30)
    {
      datos_master_Anls[i,]$AgeRange <- "17-30"
    }
    else
    {
      if (datos_master_Anls[i,]$AgeRange > 30 && datos_master_Anls[i,]$AgeRange <= 41)
      {
        datos_master_Anls[i,]$AgeRange <- "31-41"
      }
      else
      {
        if (datos_master_Anls[i,]$AgeRange > 41 && datos_master_Anls[i,]$AgeRange <= 52)
        {
          datos_master_Anls[i,]$AgeRange <- "42-52"
        }
        else
        {
          if (datos_master_Anls[i,]$AgeRange > 52 && datos_master_Anls[i,]$AgeRange <= 63)
          {
            datos_master_Anls[i,]$AgeRange <- "53-63"
          }
          else
          {
            if (datos_master_Anls[i,]$AgeRange > 63)
            {
              datos_master_Anls[i,]$AgeRange <- "64+"
            }
          }
        }
      }
    }
  }
}
```

Visualizando este nuevo atributo de rango de edades, **AgeRange**, en función de la supervivencia.

```
filas = dim(datos_master_Anls)
ggplot(data = datos_master_Anls[1:filas,], aes(x = AgeRange, fill = Survived)) + geom_bar(position = "
fill") + ylab("Frecuencia")
```




Porcentualmente se confirma lo comentado para la anterior gráfica, pero la visualización es más representativa al agrupar el rango de edades en seis clases. Es importante resaltar que, y quizá se debería profundizar en ello, el bajo porcetanje que se observa de supervivientes entre pasajeros de entre los 17 a los 30 años.

---

# 6.8 Conclusiones

El objetivo de este estudio era realizar un estudio demográfico para averiguar qué características comunes reúnen los pasajeros que no sobrevivieron al crucero Titanic observando que atributos de estos pudieron influir más en el hecho. Los datos están recogidos en dos archivos, *test.csv* y *train.csv*, que han sido unificados en uno solo, *master_clean.csv*, que se ha dimensionado con la estructura del dataset train.

En la limpieza de datos, se han sometido los datos a un pre-procesamiento para manejar los casos de ceros, elementos vacíos y valores extremos, *outliers*. Después del tratamiento, el resultado es un dataset master compuesto de 891 muestras y 9 atributos: *PassengerId*, *Survived*, *Pclass*, *Sex*, *Age*, *SibSp*, *Parch*, *Fare*, y *Embarked*.

Con los análisis estadísticos, se ha observado que, a través de la correlación, se han conocido los atributos que ejercen una menor influencia sobre que pasajeros sobreviven o no al crucero (*Age*, *SibSp*, *Parch*, *Fare*). Además, se ha evaluado y concluido que el precio del billete no está relacionado con la edad de los pasajeros. Con el contraste de hipótesis, que el valor del precio medio de un billete adquirido por la muestra mujeres es casi del doble del precio medio del billete adquirido por la muestra hombres. Una conclusión que sería interesante ampliar dando el marco socio-económico de la época. Y el modelo de regresión logística resulta de utilidad a la hora de realizar predicciones para el conocimiento de que pasajeros sobreviven o no dadas unas características concretas basadas en los atributos *Pclass*, *Sex* y *Embarked*. Siendo los dos primeros los que mejor resultados ofrecen a la hora de predecir, sabiendo que la probabilidad de sobrevivir siendo mujer y viajando en primera o segunda clase era muy alta.

Finalmente, y sabiendo que atributos tienen mayor influencia a la hora de conocer si un pasajero sobrevive o no al Titanic gracias a los análisis estadísticos, el conjunto de representaciones se ha centrado en estos atributos y su conocimiento. Con ello, se ha observado que:

- El porcentaje de hombres que sobreviven no alcanza el 25% de los que viajaban, mientras que el de mujeres es de casi el 75%.
- Los pasajeros de la tercera clase (la más humilde), tienen una porcentaje de sobrevivir del 25%, cuando para los pasajeros de la primera clase es casi del 70%.
- Los pasajeros que embarcan desde Cherbourg tienen más posibilidades de sobrevivir, 55,88%, que los pasajeros que lo hacen desde Queenstown, 38,96%, o Southampton, 36,70%.
- Las familias compuestas entre 2 y 4 miembros tienen más del 50% de posibilidades de supervivencia.
- Los pasajeros entre los 0 y los 15 años tienen mayor probabilidad de sobrevivir.

---

# 7 Tabla de contribuciones al trabajo

| Contribuciones | Firma |
| --- | --- |
| Investigación previa | Luis A. Bayo, Miguel A. Bermejo |
| Redacción de las respuestas | Luis A. Bayo, Miguel A. Bermejo |
| Desarrollo código | Luis A. Bayo, Miguel A. Bermejo |

Tabla de contribuciones