Limpieza de Datos - Práctica Grupo 2

Endika Momoitio y Sergio Postigo

09/06/2020

Contents

1. Descripción del Dataset y pregunta que se pretende responder.

El dataset a analizar es un fichero que contiene información sobre los pasajeros del Titanic. El dataset contiene las siguientes variables (columnas):

- **PassengerId**: valor numérico que nos proporciona un ID del pasajero/pasajera.
- **Survived**: nos proporciona información sobre si el pasajero/pasajera sobrevivió o no. 0 = No y 1 = Si.
- **pclass**: indica la categoría del ticket. Hay 3 valores posibles: 1 = Primera clase, 2 = Segunda clase y 3 = Tercera clase.
- **name**: nos indica el nombre del pasajero/pasajera.
- **Sex**: nos indica el sexo del pasajero/pasajera.
- **Age**: nos indica la edad en años del pasajero/pasajera.
- **sibsp**: nos indica el número de hermanos/hermanas o pareja abordo del Titanic.

- **parch**: nos indica el número de padres/hijos abordo del Titanic.
- **Ticket**: nos indica el número de ticket del pasajero.
- **fare**: coste del ticket del pasajero/pasajera.
- **cabin**: número de camarote del pasajero/pasajera.
- **embarked**: puerto de embarque. 3 posibles valores: C = Cherbourg, Q = Queenstown y S = Southhampton.

Dicho dataset se puede encontrar en la siguiente dirección: https://www.kaggle.com/c/titanic/

La pregunta que queremos responder es si efectivamente, los niños y mujeres sobrevivieron en mayor medida debido a que tuvieron prioridad en el rescate sobre los varones adultos.

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

Lo primero que debemos realizar es la carga del fichero para proceder a seleccionar los datos de interés de cara a nuestro análisis. Para ello cargamos primero el CSV y después realizamos la eliminación de las variables que no son interesantes para nuestro análisis.

```
titanic <- read.csv("train.csv", header = TRUE)
summary(titanic)
```

```
##   PassengerId       Survived          Pclass
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000
##  Median :446.0   Median :0.0000   Median :3.000
##  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
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                   :  1   female:314   Min.   : 0.42
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   1st Qu.:20.12
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                Median :28.00
##  Abelson, Mr. Samuel                   :  1                Mean   :29.70
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
##  Adahl, Mr. Mauritz Nils Martin        :  1                Max.   :80.00
##  (Other)                               :885                NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
##                                   (Other) :852
##          Cabin     Embarked
##             :687    :  2
##  B96 B98    :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6         :  4   S:644
##  C22 C26    :  3
##  D          :  3
##  (Other)    :186
```

Podemos ver que el fichero contiene 891 registros y 12 columnas, que corresponden a PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

Hacemos la selección de variables a utilizar en base a tres pasos:

- **Variables que sólo proporcionan información de identificación del pasajero, no importante para analizar el índice de supervivencia (ejemplo: nombre)**
- **PassengerId**: no necesitamos el ID del pasajero ya que queremos después discernir en dos grupos, los supuestamente preferentes y los no preferentes.
- **name**: al igual que la columna "PassengerId" no nos proporciona información relevante para el análisis.
- **ticket**: el identificador del ticket no es importante para analizar tasa de supervivencia en nuestro análisis.
- **cabin**: descartamos el número de cabina ya que no presenta información relevante para nuestro análisis.
- **Embarked**: descartamos el puerto de procedencia del pasajero ya que una vez abordo no debería influir en nuestro análisis.
- **Para el resto de variables realizamos un test de correlación entre las variables cuantitativas frente a la variable que queremos explicar: Survived**

```
# Survived y Pclass
cor.test(titanic$Survived, titanic$Pclass, method = "pearson", use = "complete.obs") # Utilizamos compl
```

```
##
##  Pearson's product-moment correlation
##
## data:  titanic$Survived and titanic$Pclass
## t = -10.725, df = 889, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3953692 -0.2790061
## sample estimates:
##       cor
## -0.338481
```

```
# Survived y Age
cor.test(titanic$Survived, titanic$Age, method = "pearson", use = "complete.obs") # Utilizamos complete
```

```
##
##  Pearson's product-moment correlation
##
## data:  titanic$Survived and titanic$Age
## t = -2.0667, df = 712, p-value = 0.03912
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.149744955 -0.003870727
## sample estimates:
##         cor
## -0.07722109
```

```r
# Survived y SibSp
cor.test(titanic$Survived, titanic$SibSp, method = "pearson", use = "complete.obs") # Utilizamos comple
```

```
##
##  Pearson's product-moment correlation
##
## data:  titanic$Survived and titanic$SibSp
## t = -1.0538, df = 889, p-value = 0.2922
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.10076614  0.03042549
## sample estimates:
##        cor
## -0.0353225
```

```r
# Survived y Parch
cor.test(titanic$Survived, titanic$Parch, method = "pearson", use = "complete.obs") # Utilizamos comple
```

```
##
##  Pearson's product-moment correlation
##
## data:  titanic$Survived and titanic$Parch
## t = 2.442, df = 889, p-value = 0.0148
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01603798 0.14652128
## sample estimates:
##        cor
## 0.08162941
```

```r
# Survived y Fare
cor.test(titanic$Survived, titanic$Fare, method = "pearson", use = "complete.obs") # Utilizamos complet
```

```
##
##  Pearson's product-moment correlation
##
## data:  titanic$Survived and titanic$Fare
## t = 7.9392, df = 889, p-value = 6.12e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1949232 0.3176165
## sample estimates:
##       cor
## 0.2573065
```

Vemos que las variables tienen poca correlación con la variable que queremos explicar: Survived por los que la mantenemos. Por otro lado, podemos sospechar una fuerte correlación entre la variable Pclass y Fare ya que, por lógica, los de primera tendrán los tickets más caros.

```r
# Fare y Pclass
cor.test(titanic$Pclass, titanic$Fare, method = "pearson", use = "complete.obs") # Utilizamos complete.
```

```
##
##  Pearson's product-moment correlation
##
## data:  titanic$Pclass and titanic$Fare
## t = -19.61, df = 889, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5937488 -0.5019370
## sample estimates:
##        cor
## -0.5494996
```

No hay una correlación tan fuerte como podíamos sospechar por lo que mantenemos la variable también.

```
titanic2 = subset(titanic, select = -c(PassengerId,Name,Ticket,Cabin,Embarked))
head(titanic2)
```

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

3. Limpieza de los datos

Primeramente vamos a analizar que las distintas columnas tengan tipos de datos y valores razonables.

```
# Comprobamos que cada columna tiene el tipo de dato apropiado.
str(titanic2)
```

```
## 'data.frame':    891 obs. of  7 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

```
# Comprobamos que pclass sólo contenga valores 1, 2 y 3.
unique(titanic2$Pclass)
```

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

```
# Comprobamos que Survived sólo tenga valores 0 y 1.
unique(titanic2$Survived)
```

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

```r
# Comprobamos que sólo hay valores "male" y "female" para el sexo.
unique(titanic2$Sex)
```

```
## [1] male   female
## Levels: female male
```

3.1 Identificación de ceros o elementos vacíos

En nuestro dataset podemos encontrar ceros y tiene sentido dicho dato por tanto vamos a centrarnos en analizar los elementos vacíos.

```r
# Comprobamos con sapply y haciendo un sum(is.na) cuántos valores nulos hay para cada variable.
sapply(titanic2, function(x) sum(is.na(x)))
```

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

Encontramos en Age diversos valores nulos. Procedemos a sustituir dichos valores con el algoritmo kNN (k-nearest neighbors algorithm) utilizando la librería VIM. Es importante recalcar que rellenaremos los valores nulos con el dataset inicial (titanic) ya que, aunque hayamos eliminado valores para responder a nuestra pregunta sobre si sobrevivieron más niños y mujeres que hombres, estos datos pueden ser relevantes para hacer una estimación de la edad.

```r
# Cargamos librería VIM omitiendo warnings y mensajes en la carga de la librería.
suppressWarnings(suppressMessages(library(VIM)))
# Utilizamos la función kNN sobre el dataset inicial completo.
titanic2$Age <- kNN(titanic)$Age
# Volvemos a comprobar que ya no haya valores nulos.
sapply(titanic2, function(x) sum(is.na(x)))
```

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

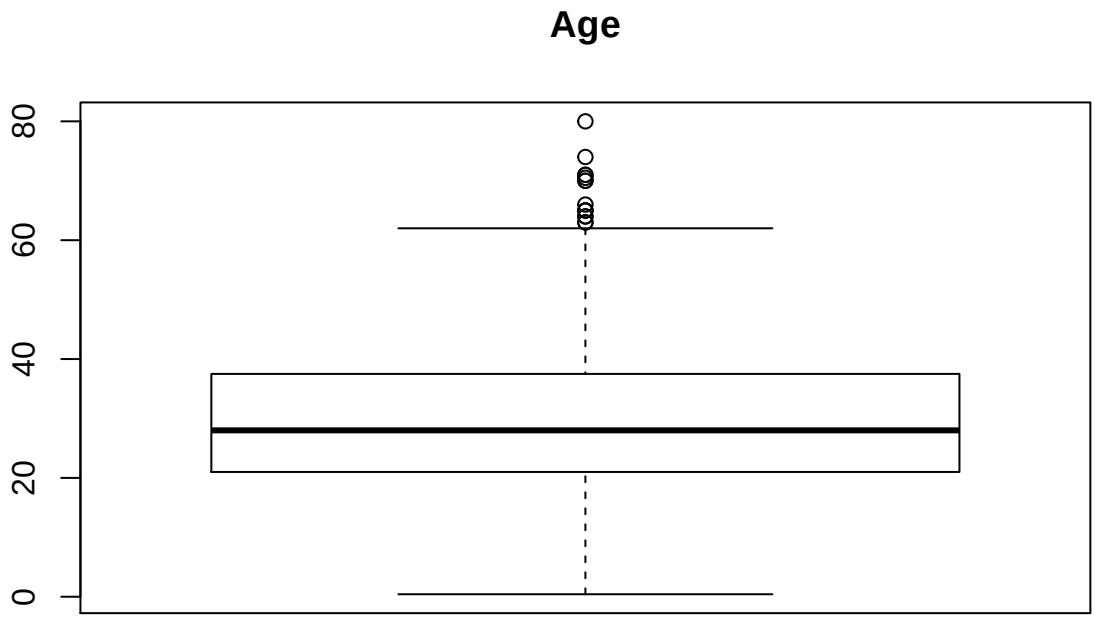
3.2 Identificación de valores extremos (outliers)

Hay variables como Survived, Pclass y Sex que hemos analizado que todos los valores que contienen sean válidos por tanto vamos a analizar las variables: Age, SibSp y Parch.

```r
# Realizamos inicialmente un listado de outliers y después lo representamos gráficamente para cada una
# Edad (Age)
boxplot.stats(titanic2$Age)$out
```

```
##  [1] 66.0 65.0 71.0 70.5 63.0 65.0 64.0 65.0 63.0 71.0 64.0 80.0 70.0 70.0 74.0
```

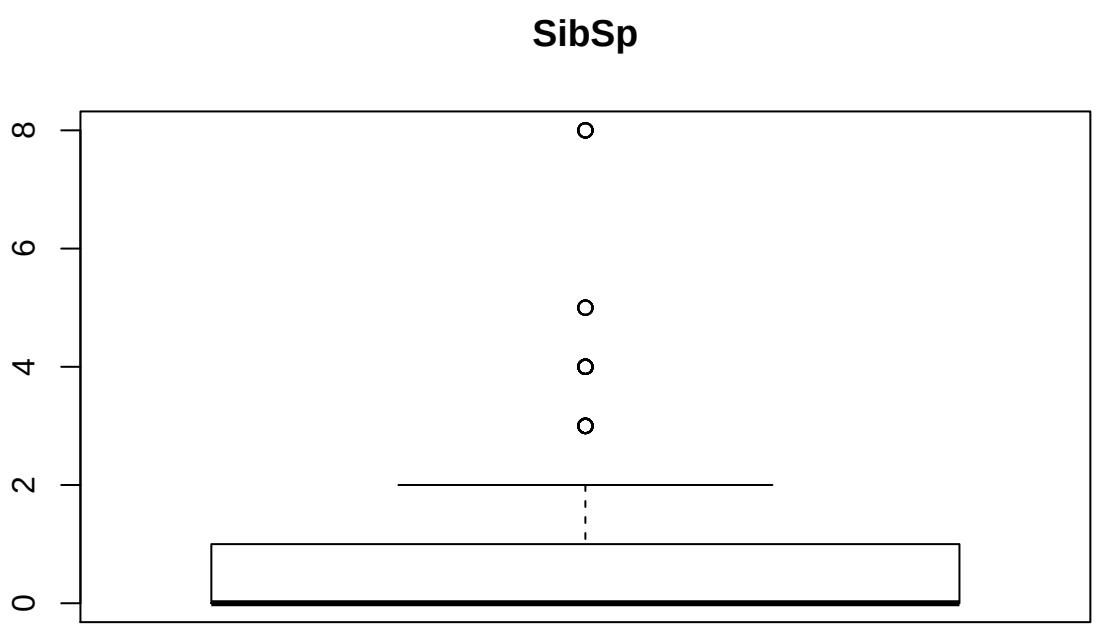
```r
boxplot(titanic2$Age, main = "Age", width = 100)
```




```
# Número de hermanos/hermanas y cónyuge abordo.
boxplot.stats(titanic2$SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3 5 4 3
## [39] 4 8 4 3 4 8 4 8
```

```
boxplot(titanic2$SibSp, main = "SibSp", width = 100)
```




```
# Número de padres o hijos abordo.
boxplot.stats(titanic2$Parch)$out
```

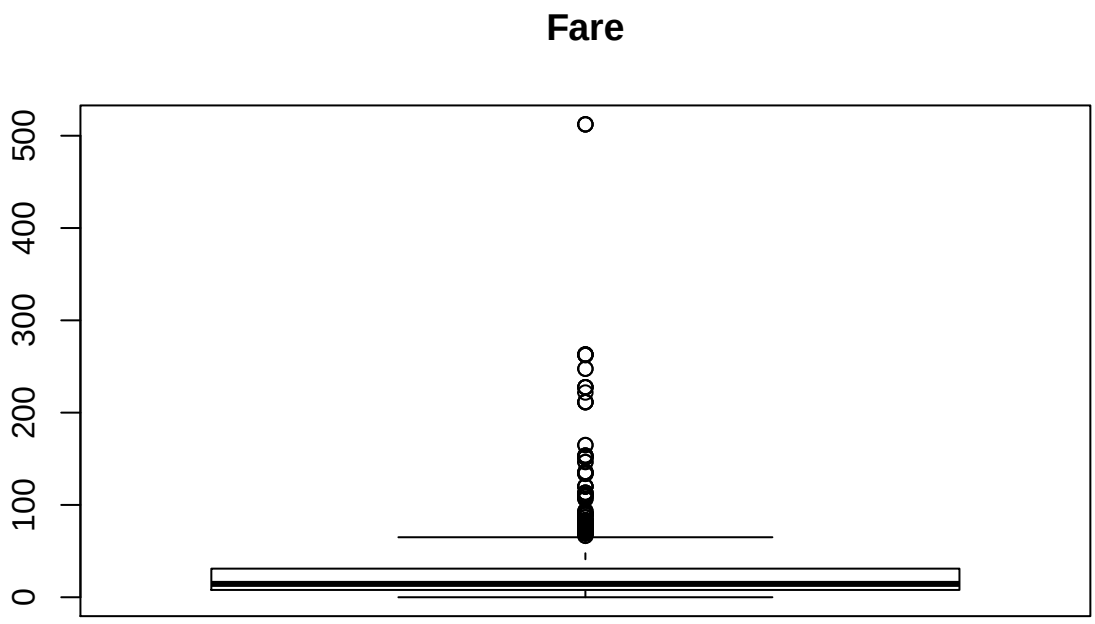
```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1 2 1
##  [38] 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1 2 1 2
##  [75] 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2 2 3 4 1 2 1
## [112] 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1 2 1 1 2 5 2 1 1
## [149] 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1 1 2 1 2 3 1 2 1 2 2
## [186] 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1 1 5 2
```

```
boxplot(titanic2$Parch, main = "Parch", width = 100)
```


```
# Coste del ticket (Fare)
boxplot.stats(titanic2$Fare)$out
```

```
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
##   [9] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500
##  [17]  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000
##  [25]  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500
##  [33]  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750
##  [41] 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000
##  [49] 263.0000  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
##  [57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792
##  [65]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
##  [73]  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000
##  [81]  78.2667 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500
##  [89]  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250
##  [97] 211.3375 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
## [105] 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
## [113]  89.1042 164.8667  69.5500  83.1583
```

```
boxplot(titanic2$Fare, main = "Fare", width = 100)
```

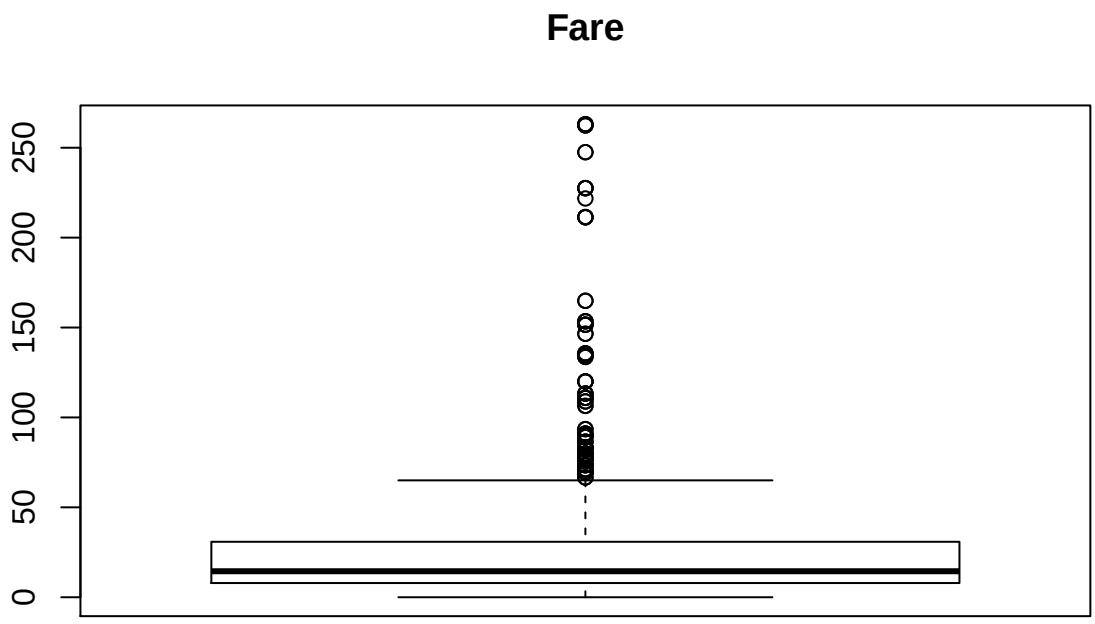


De estos análisis de identificación de valores extremos, podemos ver que:

- Age. Aunque sí aparecen valores extremos, son totalmente correctos, ya que el máximo, como hemos visto previamente es 80 años.
- SibSp. Aunque la mayoría de los casos es 0 hermanos, era normal en esa época tener 8 hermanos.
- Parch. Al igual que con los hermanos, también era normal que un crucero fuesen hasta 6 familiares.

Viendo los datos que se están analizando, aunque puedan parecer outliers (valores extremos), son valores razonables por lo cual no es necesario realizar ninguna limpieza adicional salvo a excepción de Fare, en donde encontramos un valor superior a 500 el cual podría no tener sentido.

```
titanic2 <- titanic2[which(titanic2$Fare < 500),]
boxplot(titanic2$Fare, main = "Fare", width = 100)
```



4. Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar y comparar.

Realizamos para este punto, una segregación en dos grupos.

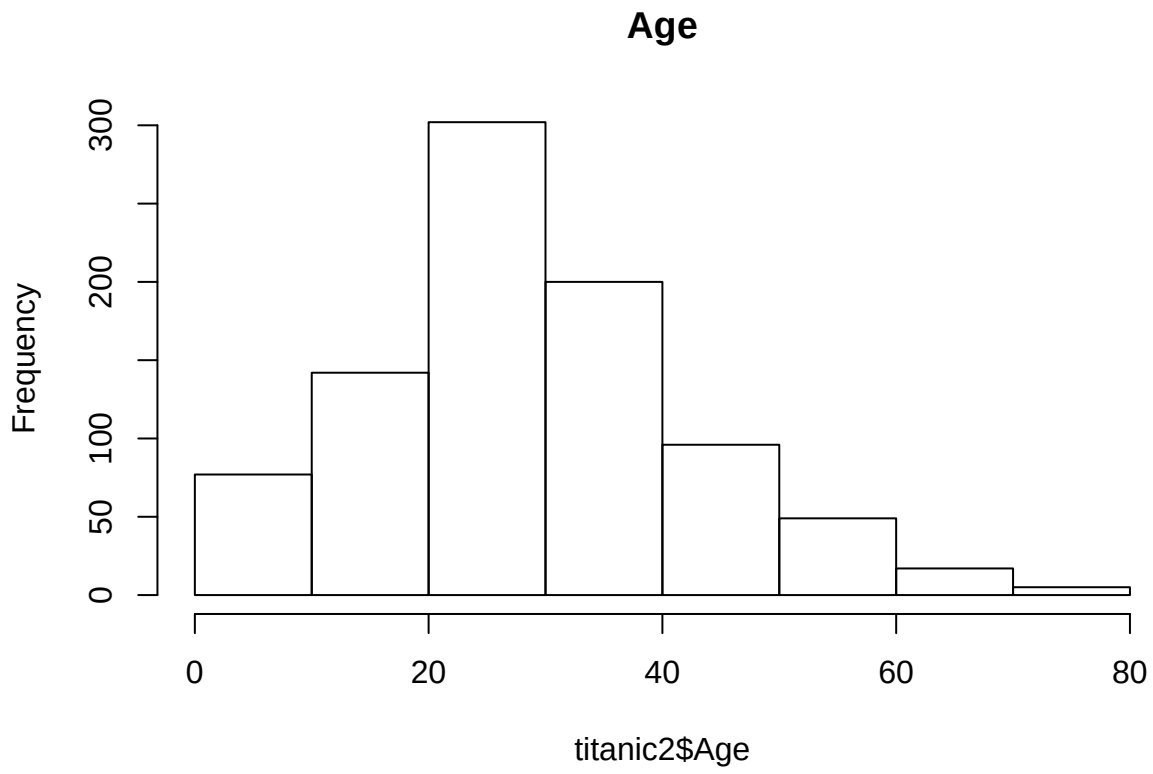
- **Prioritarios (P)**: para mujeres y niños.
- **No prioritarios (NP)**: para varones adultos.

```
# Cargamos librería DPLYR omitiendo warnings y mensajes en la carga de la librería.
# Creamos el grupo P para mujeres y niños (Priority) y el grupo NP para varones adultos.
suppressMessages(library(dplyr))
titanic_clean <- titanic2 %>%
  mutate(Priority = case_when(titanic2$Age < 18 ~ 'P',
                              titanic2$Sex == "female" ~ 'P',
                              TRUE ~ 'NP'))
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Realizamos de cara a comprobar la normalidad de las variables, un test Shapiro Wilk sobre las variables numéricas y que sean cuantitativas al igual que histogramas para analizar visualmente si sigue una distribución normal (Campana de Gauss). Compararemos el valor p obtenido frente al valor 0,05.

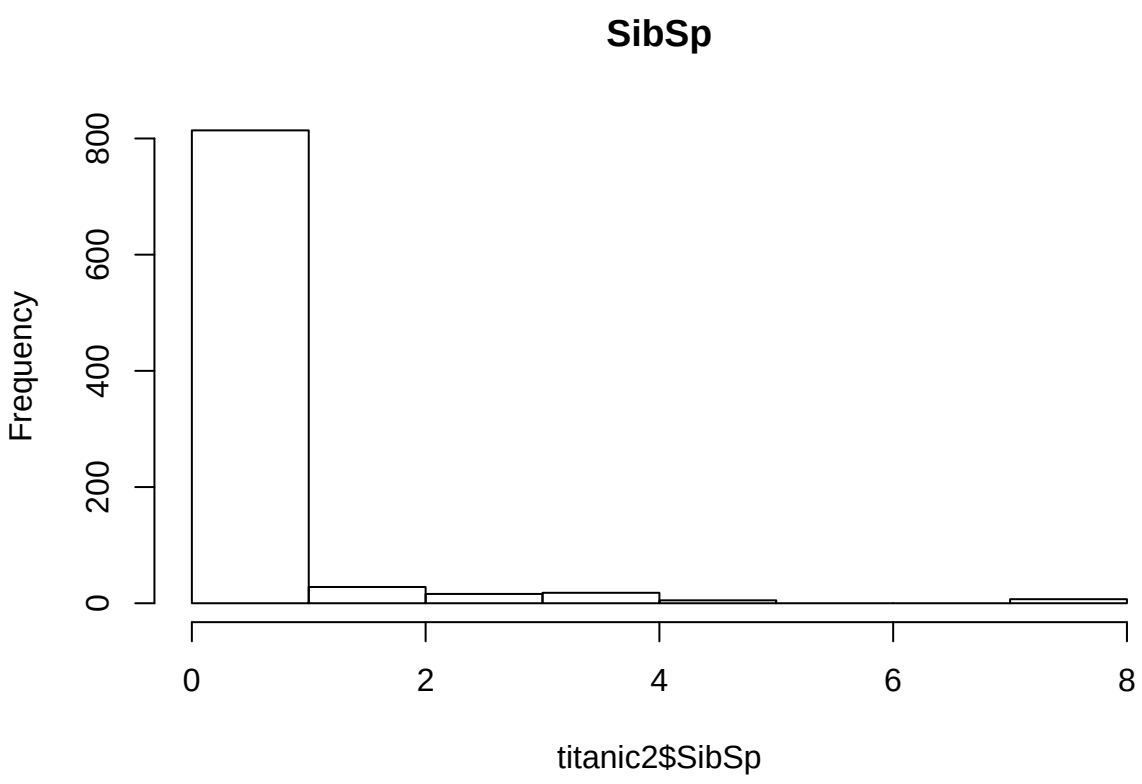
```r
# Edad (Age)
hist(titanic2$Age, main = "Age")
```




```r
shapiro.test(titanic_clean$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_clean$Age
## W = 0.98063, p-value = 1.805e-09
```

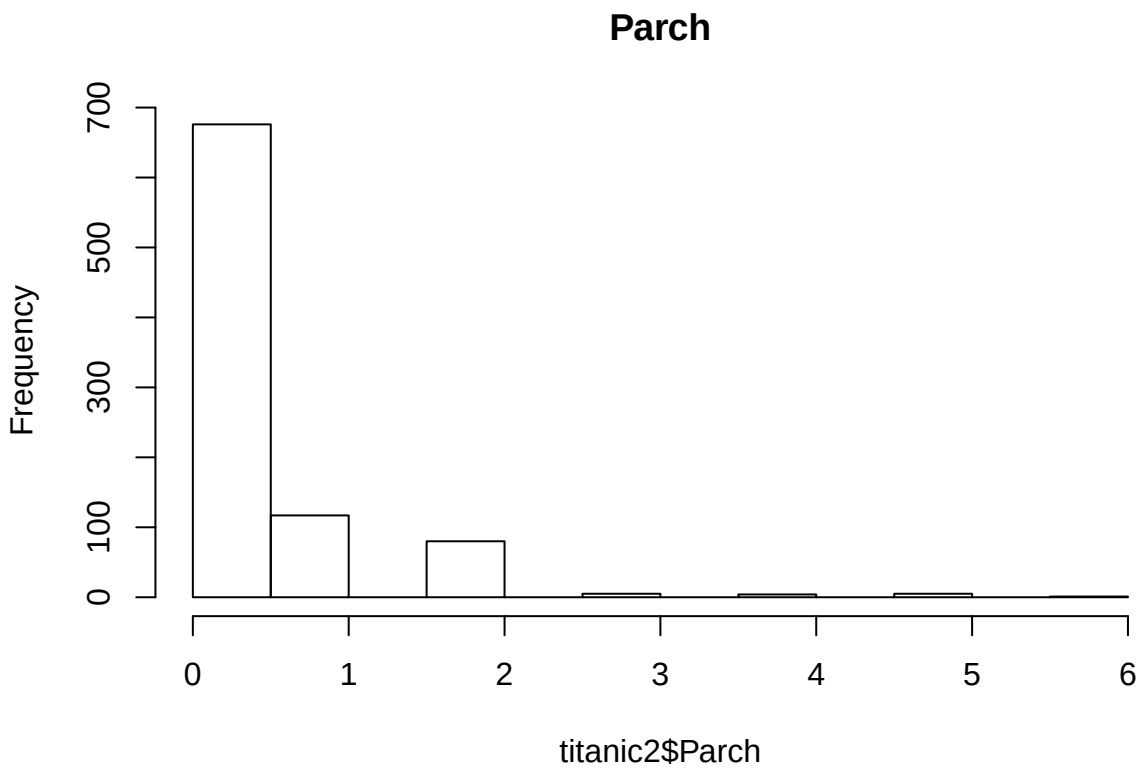
```r
# Número de hermanos/hermanas y cónyuge abordo (SibSp).
hist(titanic2$SibSp, main = "SibSp")
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_clean$SibSp
## W = 0.51381, p-value < 2.2e-16
```

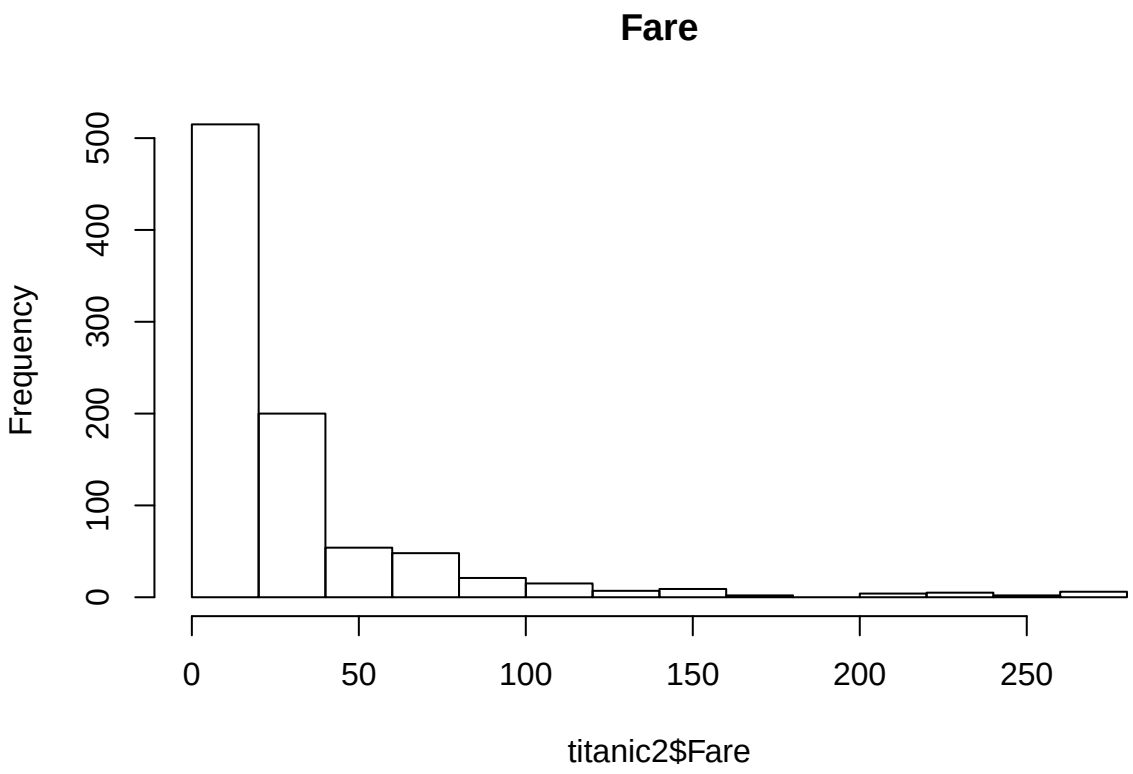
```
# Número de hijos o padres abordo (Parch)
hist(titanic2$Parch, main = "Parch")
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_clean$Parch
## W = 0.53254, p-value < 2.2e-16
```

```
# Coste del ticket (Fare)
hist(titanic2$Fare, main = "Fare")
```




```r
shapiro.test(titanic_clean$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_clean$Fare
## W = 0.60472, p-value < 2.2e-16
```

Todos los valores P que obtenemos para cada una de las variables es inferior al coeficiente 0.05 por lo que podemos rechazar la hipótesis nula y afirmar que dichas variables no siguen una distribución normal.

4.3 Aplicación de las pruebas estadísticas para comprobar los grupos de datos.

4.3.1 Contraste de hipótesis entre dos muestras (prioritarios - niños y mujeres contra no prioritarios - varones adultos)

Creamos inicialmente dos dataset para diferenciar de la gente prioritaria de la que no lo es teóricamente.

```r
# Subset para prioritarios (menores de 18 y mujeres)
titanic_clean.p <- titanic_clean[titanic_clean$Priority == "P",]
# Subset para no prioritarios (varones mayores de 18)
titanic_clean.np <- titanic_clean[titanic_clean$Priority == "NP",]
head(titanic_clean.p)
```

```
##    Survived Pclass    Sex Age SibSp Parch    Fare Priority
## 2         1      1 female  38     1     0 71.2833        P
## 3         1      3 female  26     0     0  7.9250        P
## 4         1      1 female  35     1     0 53.1000        P
## 8         0      3   male   2     3     1 21.0750        P
## 9         1      3 female  27     0     2 11.1333        P
## 10        1      2 female  14     1     0 30.0708        P
```

```
head(titanic_clean.np)
```

```
##    Survived Pclass  Sex Age SibSp Parch    Fare Priority
## 1         0      3 male  22     1     0  7.2500       NP
## 5         0      3 male  35     0     0  8.0500       NP
## 6         0      3 male  40     0     0  8.4583       NP
## 7         0      1 male  54     0     0 51.8625       NP
## 13        0      3 male  20     0     0  8.0500       NP
## 14        0      3 male  39     1     5 31.2750       NP
```

Realizamos inicialmente un contraste de hipótesis para analizar sobre las dos muestras para comprobar si las mujeres y niños tuvieron, efectivamente, un índice de supervivencia más alto. Los datasets diferenciados ya los tenemos, ahora seleccionamos la variable a explicar para poder realizar el contraste de hipótesis de las dos muestras sobre la diferencia de medidas de manera unilateral.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

En este caso tenemos

$$\mu_1$$

como la media para la primera muestra (gente no prioritaria) frente a

$$\mu_2$$

como la media para la segunda muestra, gente prioritaria.

```
titanic_clean.np.Survived <- titanic_clean.np$Survived
titanic_clean.p.Survived <- titanic_clean.p$Survived
t.test(titanic_clean.np.Survived, titanic_clean.p.Survived, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic_clean.np.Survived and titanic_clean.p.Survived
## t = -17.631, df = 692.53, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.4660987
## sample estimates:
## mean of x mean of y
## 0.1640625 0.6781915
```

Como resultado, obtenemos un p-valor menor que el valor de significación fijado (0.05) por lo que rechazamos la hipótesis nula y podemos afirmar que efectivamente, los pasajeros prioritarios tuvieron un índice de supervivencia más alto que el de no prioritarios.

4.3.2 Contraste de hipótesis entre dos muestras (Pasajeros que viajaban en primera y pasajeros que viajaban en segunda y tercera)

Al realizar una comprobación de la correlación entre la variable que queremos explicar (Survived) y el resto de variables, hemos visto que existe cierta correlación entre la variable Pclass y Survived por lo que sería también interesante comprobar si hubo un mayor índice de supervivencia entre gente que viajaba en primera clase y viajeros con pasajes en segunda y tercera clase.

```
titanic_clean.nonfirst.Survived <- titanic_clean[titanic_clean$Pclass > 1,]$Survived
titanic_clean.first.Survived <- titanic_clean[titanic_clean$Pclass == 1,]$Survived

head(titanic_clean.nonfirst.Survived)
```

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

```
head(titanic_clean.first.Survived)
```

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

```
t.test(titanic_clean.nonfirst.Survived, titanic_clean.first.Survived, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic_clean.nonfirst.Survived and titanic_clean.first.Survived
## t = -8.4689, df = 341.05, p-value = 3.742e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2570579
## sample estimates:
## mean of x mean of y
## 0.3051852 0.6244131
```

Obtenemos, al igual que en el contraste anterior, que al tener un p-valor menor que 0,05 rechazamos la hipótesis nula y podemos afirmar que también, en el caso de los pasajeros de primera clase, tuvieron un índice de supervivencia más alto que los pasajeros que viajaban en segunda clase o en tercera independientemente del sexo y edad.

4.3.3 Modelo de regresión lineal

De cara a realizar el modelo de regresión lineal haremos los siguientes pasos:

- Crearemos variables para cada una de las columnas con el fin de evitar problemas a la hora de pasar la función lm()
- Iremos creando modelos añadiendo cada vez más variables y analizando su coeficiente de determinación ajustado R2 adjusted.
- Es importante centrarnos en el coeficiente de determinación ajustado puesto que al tener un modelo con más de una variable explicativa, al añadir más variables, podríamos obtener un coeficiente de determinación mayor aunque no sea cierto ya que se produce al simplemente añadir variables nuevas

```r
head(titanic_clean)
```

```
##   Survived Pclass    Sex Age SibSp Parch    Fare Priority
## 1        0      3   male  22     1     0  7.2500       NP
## 2        1      1 female  38     1     0 71.2833        P
## 3        1      3 female  26     0     0  7.9250        P
## 4        1      1 female  35     1     0 53.1000        P
## 5        0      3   male  35     0     0  8.0500       NP
## 6        0      3   male  40     0     0  8.4583       NP
```

```r
# Para evitar problemas a la hora de pasar variables en la función lm (regresión lineal), creamos varia

Survived <- titanic_clean$Survived
Age <- titanic_clean$Age
Pclass <- titanic_clean$Pclass
Sex <- titanic_clean$Sex
SibSp <- titanic_clean$SibSp
Parch <- titanic_clean$Parch
Fare <- titanic_clean$Fare

# Generamos distintos modelos añadiendo cada vez más variables y sacamos el coeficiente R cuadrado para

m1 <- lm(Survived ~ Age)
m2 <- lm(Survived ~ Age + Pclass)
m3 <- lm(Survived ~ Age + Pclass + Sex)
m4 <- lm(Survived ~ Age + Pclass + Sex + SibSp)
m5 <- lm(Survived ~ Age + Pclass + Sex + SibSp + Parch)
m6 <- lm(Survived ~ Age + Pclass + Sex + SibSp + Parch + Fare)

summary(m1)$adj.r.squared
```

```
## [1] 0.007473963
```

```r
summary(m2)$adj.r.squared
```

```
## [1] 0.1727532
```

```r
summary(m3)$adj.r.squared
```

```
## [1] 0.3868991
```

```r
summary(m4)$adj.r.squared
```

```
## [1] 0.4005017
```

```r
summary(m5)$adj.r.squared
```

```
## [1] 0.4004243
```

```r
summary(m6)$adj.r.squared
```

```
## [1] 0.3997438
```

Mediante esta comparación de modelos, vemos que el que tiene una mejor bondad de ajuste es el modelo 4. Dicho modelo contiene las variables explicativas Age, Pclass, Sex y SibSp.

La ecuación de regresión lineal sería tal como: **Survived** = 1.4025496 -0.0071439 **Age** -0.1998553 **Pclass** -0.4944821 **Sex** + 0.0611048 **SibSp**

Se puede hacer una predicción, por ejemplo, de la siguiente forma usando el modelo ya definido:

```r
predict(m4, data.frame(Age = 21, Pclass = 1, Sex = "male", SibSp = 0), interval = "prediction")
```

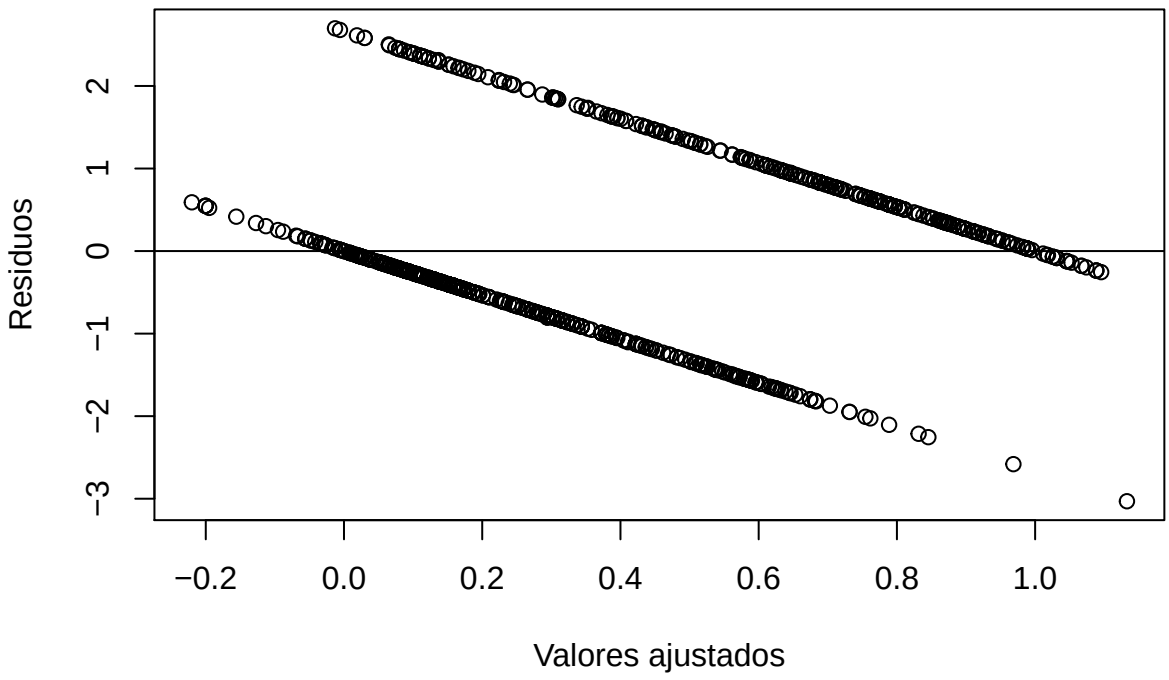
```
##         fit        lwr      upr
## 1 0.5581902 -0.1835928 1.299973
```

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

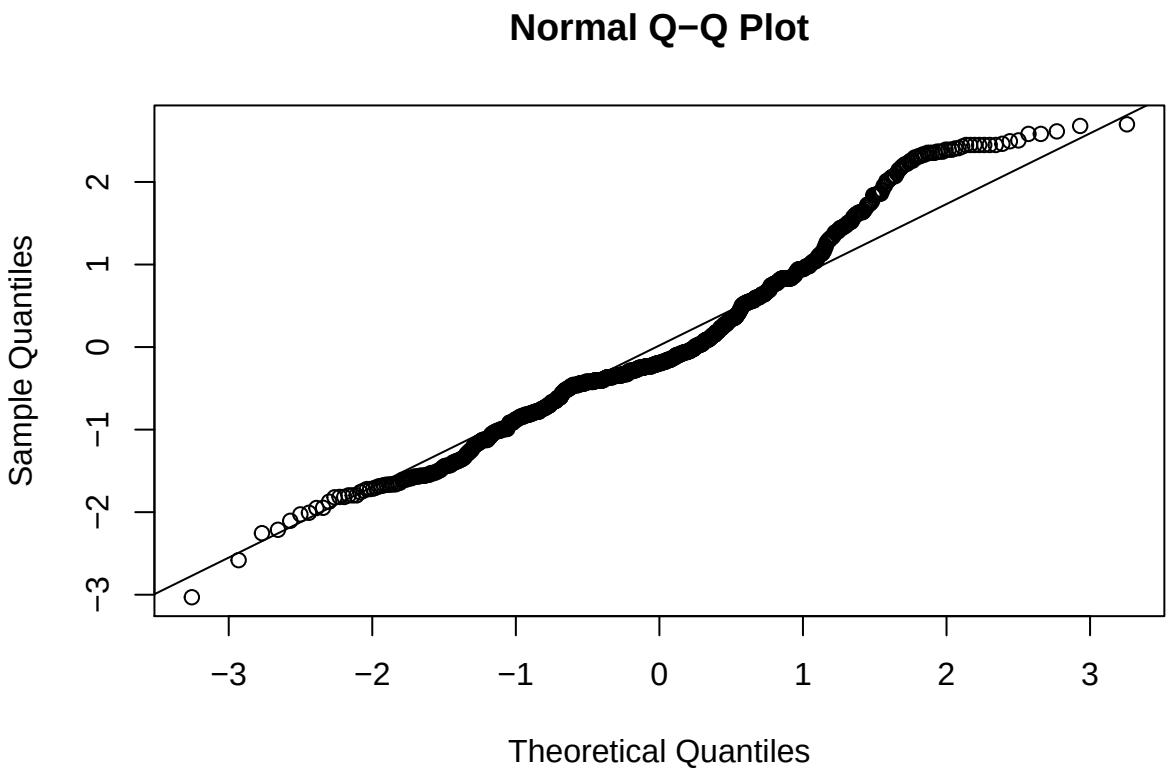
Para la diagnosis de este modelo se harán dos gráficos: uno con los valores ajustados frente a los residuos (que nos permitirá ver si la varianza es constante) y el gráfico cuantil-cuantil que compara los residuos del modelo con los valores de una variable que se distribuye normalmente (QQ plot).

```r
# Creamos una variable "residuos" con los residuos del modelo
residuos <- rstandard(m4)

# Creamos un gráfico para analizar los valores ajustados frente a los resultados
plot(fitted.values(m4),residuos, xlab="Valores ajustados", ylab="Residuos")
abline(h=0) #Dibujamos la línea en el valor 0
```




```
# Creamos un gráfico cuantil-cuantil que compara los residuos del modelo con los valores de la variable
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados
qqline(residuos)
```




```
# Obtenemos además de la bondad de ajuste, los valores beta para cada una de las variables, alfa, etc
summary(m4)
```

```
##
## Call:
## lm(formula = Survived ~ Age + Pclass + Sex + SibSp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13285 -0.20990 -0.07275  0.22422  1.01297
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.402550   0.061105  22.953  < 2e-16 ***
## Age         -0.007144   0.001064  -6.717 3.32e-11 ***
## Pclass      -0.199855   0.016993 -11.761  < 2e-16 ***
## Sexmale     -0.494482   0.027365 -18.070  < 2e-16 ***
## SibSp       -0.055561   0.012108  -4.589 5.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3764 on 883 degrees of freedom
## Multiple R-squared:  0.4032, Adjusted R-squared:  0.4005
## F-statistic: 149.1 on 4 and 883 DF,  p-value: < 2.2e-16
```

```r
# Tabla con los principales indicadores e información del reporte.

suppressMessages(library(kableExtra))
text_tbl2 <- data.frame(
  Variables = c("Media de la variable Survived","Media Survived para Prioritarios","Media Survived para
  Informacion = c(
    mean(titanic_clean$Survived),
    mean(titanic_clean.p.Survived),
    mean(titanic_clean.np.Survived),
    mean(titanic_clean.first.Survived),
    mean(titanic_clean.nonfirst.Survived),
    summary(m4)$adj.r.squared
  )
)

kable(text_tbl2) %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T, border_right = T) %>%
  column_spec(2, width = "30em")
```

Variables	Informacion
Media de la variable Survived	0.3817568
Media Survived para Prioritarios	0.6781915
Media Survived para no Prioritarios	0.1640625
Media Survived para Primera Clase	0.6244131
Media Survived no Primera Clase	0.3051852
R2 Ajustado modelo regresión	0.4005017

Exportamos también el dataset limpio y con los valores ya analizados a un fichero CSV:

```r
write.csv(titanic_clean, file = "salida.csv")
```

6. Resolución del problema

- Hemos creado el modelo *m4* cuya todas variables son significativas (véase el p-value de cada una de ellas en summary(m4)) y es capaz de predecir, pasando información de Edad, Categoría del pasaje, Sexo y número de hermanos/hermanas y pareja abordo.
- Podemos constatar que mujeres y niños tuvieron un índice de supervivencia más alto que el de varones adultos. Esto se puede observar mediante el contraste de hipótesis realizado al igual que comprobando la media de la variable Survived para ambos grupos.
- Podemos constatar que los pasajeros que viajaron en primera clase tuvieron un mayor índice de supervivencia que los pasajeros que viajaban en segunda y tercera clase. Esto se puede observar mediante el contraste de hipótesis realizado al igual que comprobando la media de la variable Survived para ambos grupos.

7. Participantes de la práctica y aportación

```
text_tbl2 <- data.frame(
  Contribuciones = c("Investigación previa","Redacción de las respuestas","Desarrollo del código"),
  Firma = c(
    "Endika Momoitio, Sergio Postigo",
    "Endika Momoitio, Sergio Postigo",
    "Endika Momoitio, Sergio Postigo"
  )
)

kable(text_tbl2) %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T, border_right = T) %>%
  column_spec(2, width = "30em")
```

Contribuciones	Firma
Investigación previa	Endika Momoitio, Sergio Postigo
Redacción de las respuestas	Endika Momoitio, Sergio Postigo
Desarrollo del código	Endika Momoitio, Sergio Postigo