Practica 2: Limpieza y análisis de datos

Juan Rodríguez Vega, Alejandro Gallardo Alberola

Enero 2021

Contents

Enunciado

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).

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

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

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.

   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?

Solución

Descripción del dataset

El RMS Titanic fue un transatlántico británico, el mayor barco de pasajeros del mundo al finalizar su construcción, que se hundió durante la noche del 14 y la madrugada del 15 de abril de 1912 durante su viaje inaugural desde Southampton a Nueva York. En el hundimiento del Titanic murieron 1496 personas de las 2208 que iban a bordo, lo que convierte a esta catástrofe en uno de los mayores naufragios de la historia ocurridos en tiempo de paz. Construido entre 1909 y 1912 en los astilleros de Harland & Wolff en Belfast, el Titanic era el segundo de los tres buques que formaban la clase Olympic, propiedad de la naviera White Star Line, junto al RMS Olympic y, posteriormente, el HMHS Britannic (Wikipedia).

El conjunto de datos está dividido en dos subconjuntos de datos:

- ***train.csv***: datos de entrenamiento. Contiene toda la información sobre los pasajeros, incluyendo si finalmente murieron o sobrevivieron. Esta será nuestra variable objetivo a predecir y nos permitirá elaborar modelos de aprendizaje supervisado.
- ***test.csv***: datos para test. Contiene la misma información salvo si el pasajero sobrevivió. Sobre este subconjunto de datos se podrá aplicar el modelo elaborado mediante el subconjunto de entrenamiento para predecir si los pasajeros sobrevivieron o no en base a sus atributos o variables independientes.

El número total de pasajeros del RMS Titanic fue de 1496, sin embargo, en el subconjunto de datos de entrenamiento se tienen 891 entradas, mientras que para el subconjunto de datos de test se tienen 418 registros. La suma total de registros asciende a 1309, por lo que es importante señalar que el dataset no contiene información de todos los pasajeros del barco. El dataset tiene 12 atributos, 11 serán variables independientes y la restante será la variable dependiente, aunque como ya se comentaba esta última no estará presente en el juego de datos para test. A continuación se describen los diferentes atributos:

- ***PassengerID***: Variable de tipo numérica. ID unívoco del pasajero.
- ***Survived***: Variable numérica. Indica si el pasajero sobrevive o no (0 = No, 1 = Yes). Se trata de la variable dependiente, que se pretende predecir.
- ***Pclass***: Variable numérica. Identifica la clase en la que viajaba el pasajero o clase del billete (1 = 1st, 2 = 2nd, 3 = 3rd).
- ***Name***: Variable tipo texto. Contiene el nombre del pasajero.
- ***Sex***: Variable categórica. Refleja el género del pasajero (“male” o “female”).
- ***Age***: Variable numérica. Determina la edad del pasajero.
- ***SibSp***: Variable numérica. Indica el número de hermanos/cónyuges del pasajero abordo del barco.
- ***Parch***: Variable numérica. Informa sobre el número de padres/hijos del pasajero abordo del barco.
- ***Ticket***: Variable de tipo texto. Hace referencia al número o identificador del billete del pasajero.

- ***Fare***: Variable numérica. Muestra la tarifa del billete del pasajero.
- ***Cabin***: Variable categórica/texto. Contiene el número de cabina en la que viajaba el pasajero.
- ***Embarked***: Variable categórica. Indica el puerto de embarque del pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

A partir de este conjunto de datos se pretende dar respuesta a muchas de las cuestiones que siempre han rodeado a la supervivencia de los pasajeros del Titanic. De esta forma, se puede ver qué condiciones (variables) influyeron principalmente a la hora de que una persona sobreviese o no. Por ejemplo:

- ¿Se priorizó a los pasajeros de primera clase sobre el resto?
- ¿Fueron las mujeres y los niños los primeros que embarcaron en los botes salvavidas?
- ¿Qué tipos de pasajeros tuvieron más probabilidades de sobrevivir?
- ¿Qué tipo de pasajero procedía de cada punto de embarque?
- Etc.

Integración y selección de los datos

En este apartado se procede a cargar el dataset *train* y *test*, ya que con el primero se construirán los modelos y sobre el segundo se realizará la fase de test. Dado que ambos subconjuntos tienen un propósito distinto se mantendrán en 2 dataframes distintos.

En caso de querer unificarlos, esto se podría hacer mediante la función `rbind`, para lo que sería necesario crear la variable dependiente en el subconjunto de test y dejarla vacía. Posteriormente, se podría comprobar que no existen duplicidades de registros con la función `duplicated` o `unique`.

Existe un identificador único para cada uno de los pasajeros como veíamos en el apartado anterior, que lo identifica de forma unívoca independientemente de si se encuentra en el subconjunto de entrenamiento o de test.

A continuación, se muestran algunos registros e información general de los datos que servirá para posteriormente proceder a la limpieza y adaptación de los datos.

```
# Carga de librerías
library(car)
library(VIM)
library(psych)
library(Hmisc)
library(psych)
library(corrplot)
library(gridExtra)
library(C50)


# Carga de los datasets
df_train <- read.csv("train.csv", header=T, sep=",")
df_test <- read.csv("test.csv", header=T, sep=",")

# Visualización de algunos registros del dataset
head(df_train,5)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
```

```
## 4           4        1      1
## 5           5        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
##             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
```

```
tail(df_train,5)
```

```
##     PassengerId Survived Pclass                                      Name    Sex
## 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
## 887  27     0     0     211536 13.00              S
## 888  19     0     0     112053 30.00   B42        S
## 889  NA     1     2 W./C. 6607 23.45              S
## 890  26     0     0     111369 30.00  C148        C
## 891  32     0     0     370376  7.75              Q
```

```
# Datos estadísticos básicos
summary(df_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
##
```

```
# Estructura y tipo de los datos
str(df_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)"
##  $ 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" ...
```

Para nuestro caso de estudio no se van a descartar registros porque carezcan de interés, ya que no nos vamos a centrar en un segmento de edad concreto, sexo o puerto de embarque, sino que se quieren considerar todos los pasajeros con sus correspondientes características para extraer conclusiones en base a ellas.

En esta primera inspección de los datos ya se puede intuir que habrá ciertos atributos que tendrán más relevancia que otros, e incluso otros, como el nombre del pasajero, que carecerán de interés alguno.

Limpieza de los datos

Análisis de valores vacíos y/o nulos

En primer lugar, se realiza un análisis de los valores nulos o vacíos y, posteriormente, se procede a eliminar variables con poco valor significativo para el análisis de datos.

Nota: Estos valores están representados en el dataset por NA o "". No aparecen otros típicos como" " o "?".

```
# Análisis de valores nulos o vacíos
colMeans(is.na(df_train))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000   0.1986532
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000
```

```
colMeans(df_train =="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
## 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
## 0.000000000 0.000000000 0.000000000 0.000000000 0.771043771 0.002244669
```

A la vista de los datos se pueden extraer las siguientes conclusiones:

- La variable *Cabin* posee un 77% de valores nulos o vacíos. Por tanto, se prescindirá de dicha variable ya que no tiene sentido inferir tal elevada cantidad de valores.
- La variable *Age* posee casi un 20% de valores nulos; en este caso concreto, si bien es una cifra considerable, la significancia de este atributo puede ser relevante y, por tanto, se mantiene. Por ejemplo, una opción podría ser sustituir los valores por la media/mediana en tanto que exista un comportamiento de normalidad.

- Por último, se observa que la variable *Embarked* posee un 0.2% de valores nulos. En este sentido, al tratarse de una variable cualitativa, se procede a inferir dichos valores con el valor más representado.

No obstante, se va a optar por imputar los valores perdidos de las variables *Age* y *Embarked* por medio de la función *KNN* del paquete *VIM*. Este método realiza la imputación basándose en los k vecinos más próximos, en este caso se toma el valor `k = 10`. En variables cualitativas, el atributo más frecuente en los k registros más próximos se usa para hacer la imputación. Si la variable es cuantitativa, es el valor de la mediana de los k registros más próximos

```
# Se sustituye la cadena vacía por NA antes de aplicar KNN
df_train$Embarked[df_train$Embarked == ""]= NA
df_train$Age <- kNN(df_train, k = 10)$Age
df_train$Embarked <- kNN(df_train, k = 10)$Embarked
```

Se va prescindir de los atributos **Name**, **Ticket** y **cabin**, pues no tienen relevancia a la hora de extraer conclusiones de los datos, ya que identifican al pasajero en cierto modo y para esta función se mantiene el atributo **PassengerID**. Además, como se acaba de comentar, en el caso de la variable **cabin** se tiene un porcentaje de valores perdidos muy elevado.

```
# Se eliminan las variables Name, Ticket y Cabin
df_train <- df_train[,-c(4, 9, 11)]
```

Conversión y adaptación de los datos

Para trabajar correctamente con los datos, se van a realizar algunas conversiones de los tipos de algunos de ellos. Esto nos permitirá realizar análisis de forma más eficiente y obtener resultados más interpretables. Comenzamos convirtiendo a tipo factor las variables *Pclass*, *Sex*, *Embarked* y *Survive*.

```
#Factorización de las variables categóricas
df_train$Survived <- factor(df_train$Survived, levels = c(0,1),
                            labels= c("No", "Yes"))
df_train$Pclass <- factor(df_train$Pclass, levels = c(1,2,3),
                            labels= c("1st", "2nd", "3rd"))
df_train$Sex <- factor(df_train$Sex, levels= c("female", "male"),
                            labels = c("Female", "Male"))
df_train$Embarked <- factor(df_train$Embarked, levels= c("C","Q","S"),
                            labels = c("Cherbourg","Queenstown", "Southampton"))
```

Existen otro tipo de conversiones que se podrían hacer de los datos, como la normalización de las variables numéricas entre los valores [0,1] mediante transformaciones *min-max* o la normalización *z-score* que resta la media a la variable y la divide por su desviación estándar. Utilizaremos la normalización *z-score* por medio de la función *scale* para normalizar las variables cuantitativas.

```
# Index. de variables cuantitativas
v_var_cuant <- c(5:8)

# Normalizamos las variables cuantitativas
df_norm <- scale(df_train[,v_var_cuant])
```

Más adelante será necesario utilizar los datos normalizados, no obstante, se van a mantener sin normalizar ya que para mostrar los resultados resulta más intuitivo verlos en su escala natural.

En el caso de que las variables no presenten una distribución normal, sería interesante realizar transformaciones de tipo *Box-Cox* para poder mejorar su normalidad y su homocedasticidad.

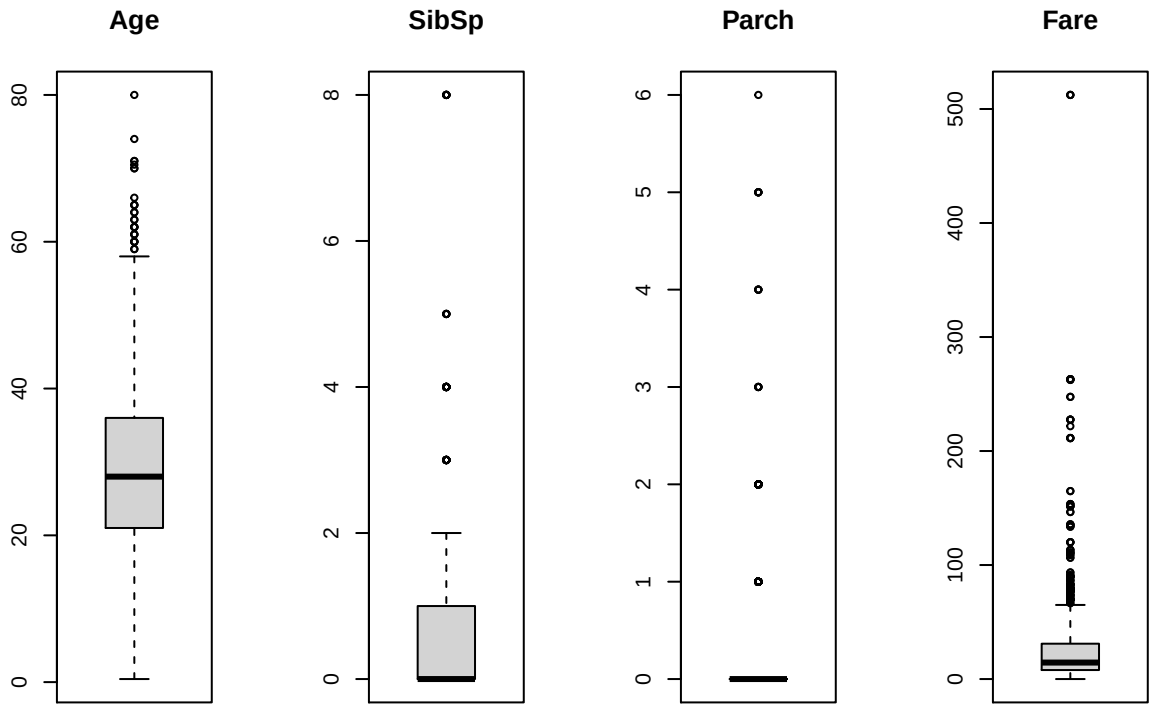
Por último, algunas variable como por ejemplo *age* parecen candidatas para realizar sobre ellas un proceso de discretización. Esto nos permitiría hacer análisis sobre ellas que aportarán mayor información y se podría optar por diferentes niveles de precisión. Algunas posibles discretizaciones en el caso de la variable *age* podría

ser en rangos de edad de 10 años; por "etapas" del tipo *niños*, *adolescentes*, *jóvenes*, *adultos* y *jubilados*; o sencillamente limitarnos a distinguir entre *niños* y *adultos*.

Análisis de valores extremos

A continuación, se procede a visualizar y analizar los posibles *outliers* asociados a las variables continuas.

```
# Estadísticas de valores nulos o vacíos
par(mfrow=c(1,4))
for(i in v_var_cuant){
   boxplot(df_train[,i], main=colnames(df_train)[i])
}
```



Observando los datos considerados como *outliers* vemos que son valores que destacan considerablemente sobre la media, pero vamos a analizar si se tratan de valores que pueden ser válidos:

- *Age*: a bordo del Titanic se encontraban pasajeros de avanzada edad, viendo el resumen de los datos el máximo de la variable es 80 años. Este valor está por encima de la media pero se trata de un valor perfectamente válido, al igual que el resto de valores que aparecen como valores extremos.
- *SibSp* y *Parch*: a la vista de los resultados se puede ver que la mayoría de pasajeros del Titanic no viajaron con toda la familia a bordo, o si así era, se trataban de familias de tamaño reducido. Sin embargo, se dan algunos casos donde viajaban familias con un número importante de hermanos y/o hijos a bordo. Por tanto, estos valores que aparecen como *outliers*, aunque son menos frecuentes sí que son valores válidos.
- *Fare*: haciendo un análisis de los datos, se puede observar que los precios más altos se corresponden con los billetes de primera clase, incluso los más elevados tienen más de una cabina. A priori no hay nada

que nos indique que estos precios son erróneos y que vayan a introducir errores en los resultados de nuestro análisis, por lo que de momento se van a dejar sin modificar.

Es importante considerar que estas observaciones pueden afectar a los estadísticos y, por tanto, hacer un análisis sesgado de los datos (por ejemplo, incrementan significativamente la varianza de los datos). No obstante, como se ha comentado, no parecen cifras muy desproporcionadas.

Generar archivo con datos tratados

Se va a generar a continuación el fichero con los datos ya tratados tal y como se solicita en la práctica.

```
# El dataframe se llamará df_output
df_output <- df_train
# Se incluyen las variables cuantitativas normalizadas
df_output[, v_var_cuant] <- df_norm
# Se exporta a formato csv
write.csv(df_output, file = "clean_data.csv", row.names = FALSE,
                                    col.names = TRUE)
```

Análisis de datos

En este apartado se va a estudiar en más detalle cómo son los datos y qué relación existe entre ellos.

Planificación de los análisis a realizar

Comentamos a continuación algunas de las premisas que se quieren comprobar en los siguientes apartados:

- ¿Tuvieron los pasajeros de primera clase más probabilidades de salvarse que los de tercera clase?
- ¿Se intentó salvar antes a mujeres y niños que a hombres?
- ¿Existió alguna preferencia a la hora de salvar a los pasajeros según su puerto de embarque?

Medidas de dispersión

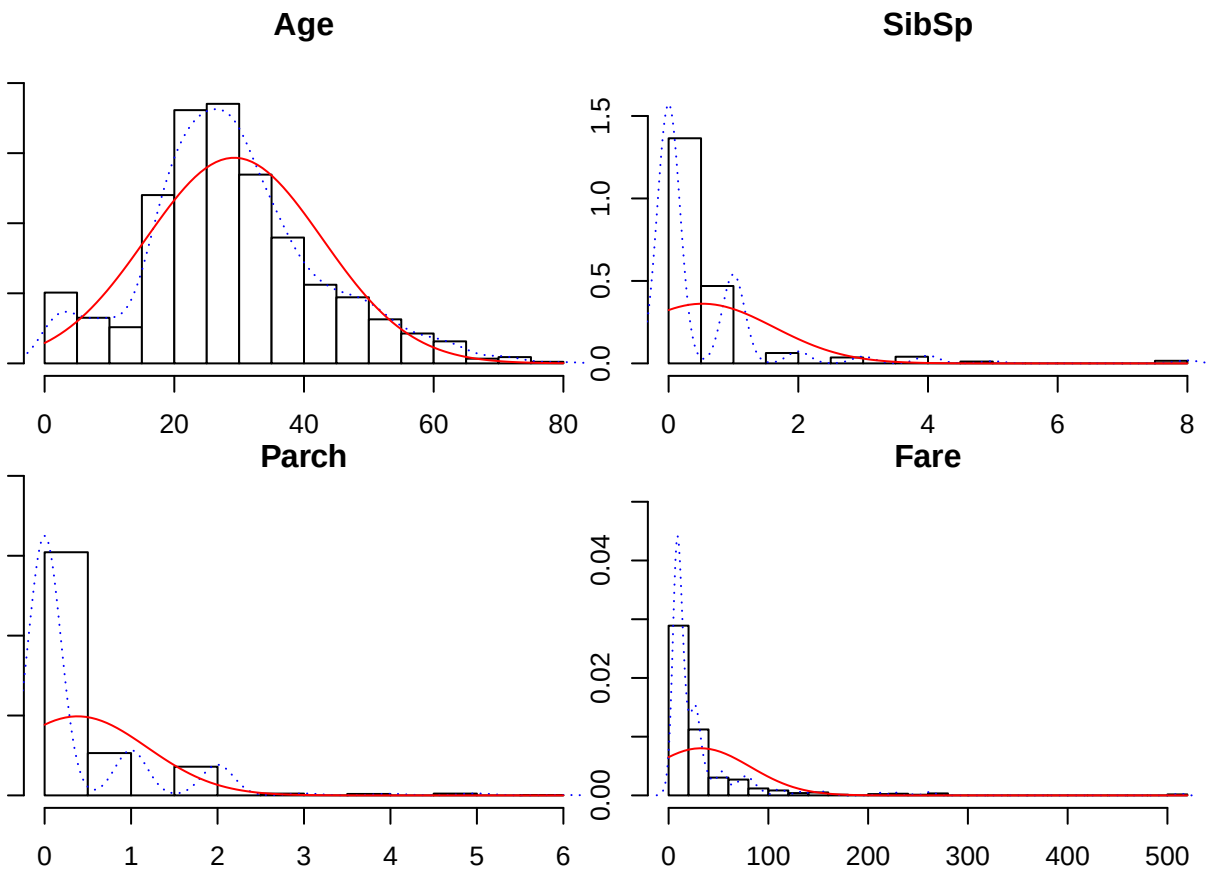
En primer lugar, se analizan otros estadísticos como la varianza o la desviación estándar y se aplica el test de *Shapiro-Wilk* para comprobar la normalidad de las variables a través de contraste de hipótesis y, a partir del cual, si el p-value es menor al nivel de significancia (0.05), se rechaza la hipótesis nula (distribución normal).

```
# Cálculo de la var, sd y p-value para ver normalidad
v_var <- vector(length = length(v_var_cuant))
v_sd <- vector(length = length(v_var_cuant))
v_pvalue_shapiro <- vector(length = length(v_var_cuant))
for (i in seq_along(v_var_cuant)){
  v_var[i] <- var(df_train[,v_var_cuant[i]],na.rm = TRUE)
  v_sd[i] <- sd(df_train[,v_var_cuant[i]] ,na.rm = TRUE)
  v_pvalue_shapiro[i] <- as.double(shapiro.test(
                                    df_train[,v_var_cuant[i]])["p.value"])
}
# Se muestran los datos en formato dataframe
medidas_dispersion_shap <- data.frame(colnames(df_train[,v_var_cuant]),
                              round(v_var,4), round(v_sd,4),
                              round(v_pvalue_shapiro,4))

colnames(medidas_dispersion_shap) <- c("attr", "varianza",
                                       "desviación estándar", "shapiro p-value")
medidas_dispersion_shap
```

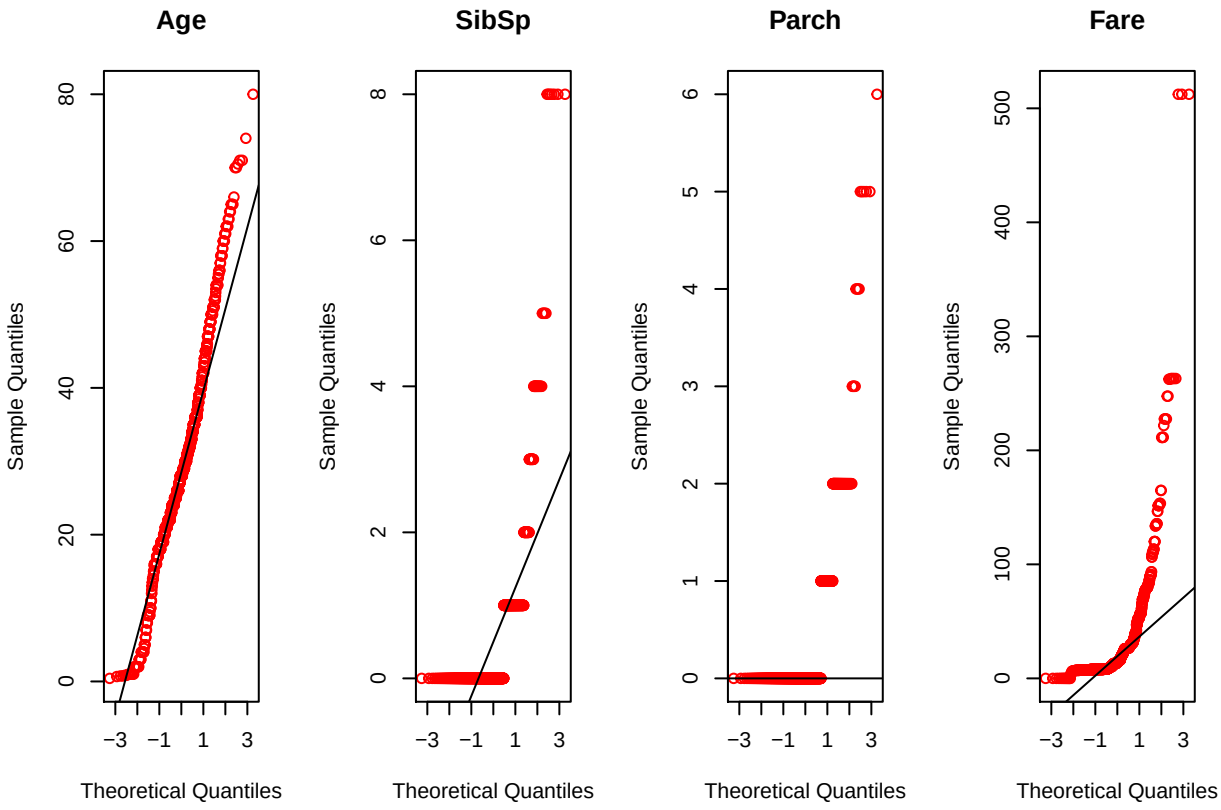
```
##    attr  varianza desviación estándar shapiro p-value
## 1   Age  184.8419             13.5957               0
## 2 SibSp    1.2160              1.1027               0
## 3 Parch    0.6497              0.8061               0
## 4  Fare 2469.4368             49.6934               0
```

Por tanto, se puede concluir que las variables no atienden a una distribución normal. A continuación, se muestra la distribución de las variables en comparación con la normal donde se verifica lo anteriormente expuesto. Además, se muestra un *Q-Q plot* para comprobar si los cuantiles siguen o no una distribución lineal.

```
# Histograma vs. normal
multi.hist(x = df_train[,v_var_cuant], dcol = c("blue", "red"),
                              dlty = c("dotted", "solid"))
```




```
# Gráfico Q-Q
par(mfrow = c(1, 4))
for (i in v_var_cuant){
  qqnorm(df_train[,i], main=colnames(df_train)[i], col = "red")
  qqline(df_train[,i])
}
```



No obstante, a la vista de las gráficas y puesto que hay un número considerable de muestras, en el caso de la variable *Age* vamos a considerar normalidad a través del *teorema central del límite* dado que tenemos un número de muestras superior a 30.

Homocedasticidad

Dado que para *Age* se ha supuesto normalidad por el *teorema central del límite*, se usará el *test de Levene* para comprobar si existe homocedasticidad. Puesto que el resto de datos no siguen una distribución normal, se aplicará el *test de Fligner-Killeen* como alternativa no paramétrica para evaluar la igualdad de varianzas basada también en el contraste de hipótesis. Para ambos test, la hipótesis nula asume igualdad de varianzas en los diferentes grupos de datos, por lo que un p-valor inferior al nivel de significancia indicará heterocedasticidad.

```
# Se comprueba si existe homocedasticidad
leveneTest(Age ~ Pclass, data = df_train)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  11.917 7.815e-06 ***
##       888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Age ~ Sex, data = df_train)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.6461 0.4217
##       889
```

```r
leveneTest(Age ~ Survived, data = df_train)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  3.6474 0.05648 .
##       889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Se comparan de forma conjunta las variables SibSp y Parch
fligner.test(SibSp+Parch ~ Pclass, data = df_train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp + Parch by Pclass
## Fligner-Killeen:med chi-squared = 0.041221, df = 2, p-value = 0.9796
```

```r
fligner.test(SibSp+Parch ~ Survived, data = df_train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp + Parch by Survived
## Fligner-Killeen:med chi-squared = 19.647, df = 1, p-value = 9.317e-06
```

Las conclusiones que se pueden extraer de estos test son los siguientes:

- La variables *Age* presenta heterocedasticidad con *Pclass* y homocedasticidad con *Sex* y *Survived*. En las diferentes clases no habrá una varianza constante en la edad de los pasajeros; mientras que sí que la habrá en cuanto al sexo de los pasajeros y el hecho de si sobrevivieron o no.
- La suma de las variables *SibSp+Parch*, que nos da una idea del tamaño de la familia con que viajaba a bordo el pasajero, presenta homocedasticidad con *Pclass*, por lo que la varianza del tamaño familiar no varía con la clases; sin embargo, vemos que sí que varía la varianza de esta suma de variables cuando se tiene en cuenta si los individuos sobreviven o no.

Relación de la variable Survived con las variables numéricas

Dado que se puede asumir normalidad para la variable *Age* y hemos visto que presenta homocedasticidad con la variable *Survived* (a posteriori será la variable dependiente en nuestros modelos), vamos a aplicar la prueba de *t de student*, donde la hipótesis nula asume que las medias de los grupos de datos son las mismas.

```r
# test t de student Age vs Survived
t.test(Age ~ Survived, data = df_train)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Survived
## t = 2.4048, df = 684.5, p-value = 0.01645
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4191229 4.1480451
## sample estimates:
##  mean in group No mean in group Yes
##          30.11157          27.82798
```

Para el caso de la variable *Age* se puede ver que en media no existen diferencias significativas.

Vamos a comprobarlo para el resto de variables, pero dado que estas variables no presentan normalidad se realizará mediante el test de *Wilcoxon* o *Mann-Whitney*, donde la hipótesis nula asume igualdad en la distribución para los diferentes grupos de la variable categórica. Ambos test se aplican indistintamente con la función *wilcox.test*.

```
# Se comparan las distribuciones del resto de variables
# cuantitativas con la variable dependiente Survive

wilcox.test(SibSp+Parch ~ Survived, data = df_train)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SibSp + Parch by Survived
## W = 77659, p-value = 7.971e-07
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Fare ~ Survived, data = df_train)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Fare by Survived
## W = 57807, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Para ambos casos vemos que no se puede determinar que la distribución de las variables es la misma en los diferentes grupos de la variable dependiente *Survived*.

Relación entre variables categóricas

Para comprobar si existen diferencias significativas entre las variables categóricas de nuestro dataset, se va a utilizar el test de χ^2. La hipótesis nula que asume este test es que no existen diferencias significativas entre los grupos de ambas variables.

```
# Se comprueba si murieron igualmente hombres y mujeres
table(df_train$Sex, df_train$Survived)
```

```
##
##           No Yes
##   Female  81 233
##   Male   468 109
```

```
chisq.test(table(df_train$Sex, df_train$Survived))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df_train$Sex, df_train$Survived)
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

```
# Si tuvo alguna influencia la clase en la que viajaba el pasajero
table(df_train$Pclass, df_train$Survived)
```

```
##
##        No Yes
##   1st  80 136
```

```
##   2nd  97  87
##   3rd 372 119
```

```
chisq.test(table(df_train$Pclass, df_train$Survived))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df_train$Pclass, df_train$Survived)
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

```
# Existió alguna relación con el puerto de embarque
table(df_train$Embarked, df_train$Survived)
```

```
##
##                No Yes
##   Cherbourg    75  93
##   Queenstown   47  30
##   Southampton 427 219
```

```
chisq.test(table(df_train$Embarked, df_train$Survived))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df_train$Embarked, df_train$Survived)
## X-squared = 25.964, df = 2, p-value = 2.301e-06
```

A la vista de los datos, vemos que tanto el sexo del pasajero, la clase en la que viajó, como el puerto desde el que embarcó tuvieron más o menos repercusión en si finalmente consiguió salvarse o no, pues que no se cumple la hipótesis nula en ninguno de los 3 casos analizados y no se puede decir que no existen diferencias en la variable objetivo respecto a estos hechos.

Correlación de variables cuantitativas

A continuación, se procede a analizar la correlación entre las variables cuantitativas, para lo que se hace necesario utilizar las variables normalizadas, recurrimos a ellas ya que se normalizaron anteriormente.

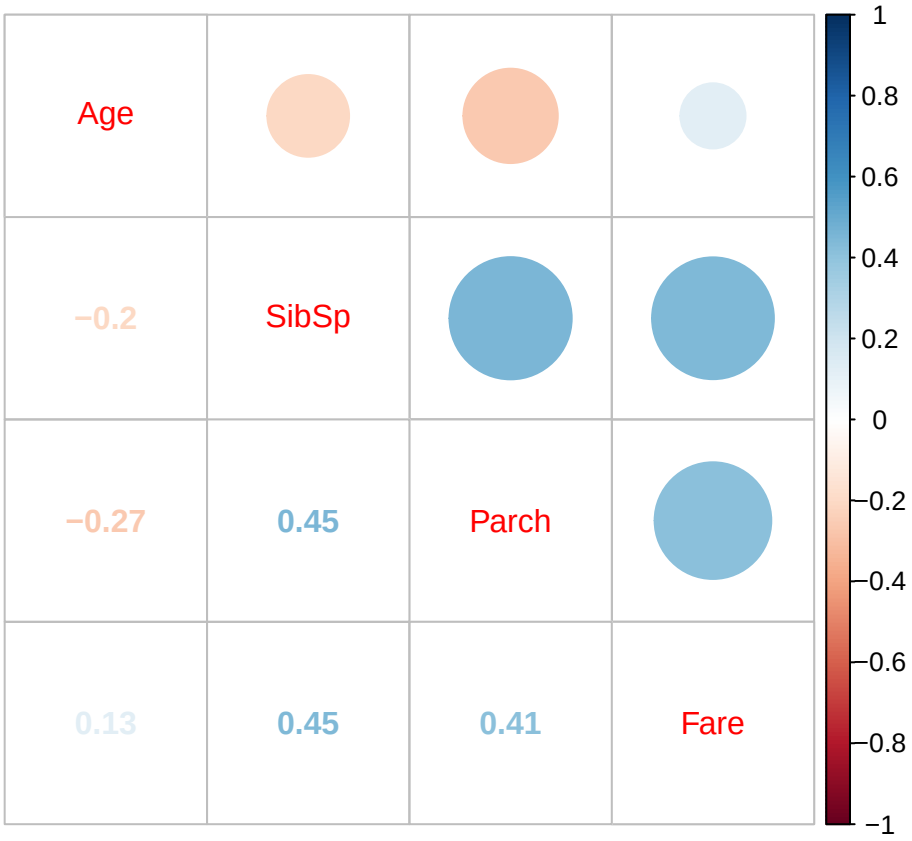
Dado que por el *teorema central del límite* se ha supuesto normalidad para la variable *Age* se podría pensar en utilizar el *coeficiente de correlación de Pearson*, pero dado que ninguna de las otras variables cuantitativas presenta normalidad se aplicará el *coeficiente de Spearman*, que no asume ninguna distribución de los datos.

```
# Matriz de correlación y p-value
rcorr(df_norm, type = "spearman")
```

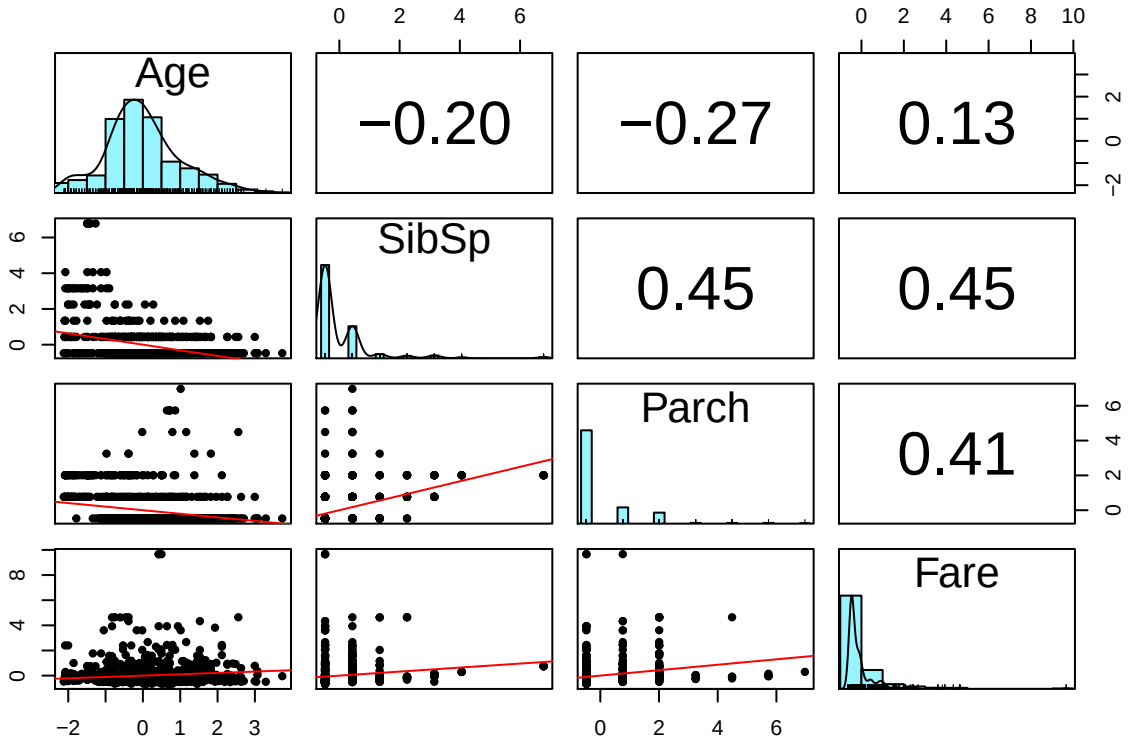
```
##          Age SibSp Parch Fare
## Age    1.00 -0.20 -0.27 0.13
## SibSp -0.20  1.00  0.45 0.45
## Parch -0.27  0.45  1.00 0.41
## Fare   0.13  0.45  0.41 1.00
##
## n= 891
##
##
## P
##       Age   SibSp Parch Fare
## Age         0e+00 0e+00 1e-04
## SibSp 0e+00       0e+00 0e+00
```

```
## Parch 0e+00 0e+00       0e+00
## Fare  1e-04 0e+00 0e+00
```

```r
# Visualización de correlación y los diagramas de dispersión
corrplot.mixed(cor(df_norm, method = "spearman"))
```




```r
pairs.panels(x = df_norm, ellipses = FALSE, lm = TRUE,
                          method = "spearman", hist.col = "cadetblue1")
```



Se observa que no existen fuertes correlaciones entre las variables cuantitativas (si se establece como umbral |0.5|). Además, se comprueba que son estadísticamente significativos y, por tanto, improbable que este resultado se haya debido al azar.

Selección de datos

Puesto que la variable clasificadora *Survived* es dicotómica, la cual determina si la persona sobrevivió o no, se va a realizar un modelo logístico para analizar la influencia de cada una de las variables de forma que se pueda observar cuáles son las más significativas. Esto se hace de forma complementaria a los análisis realizados hasta ahora y que ya nos dan una idea bastante acertada de la relación que existe entre los distintos atributos sobre la variable dependiente.

```
attach(df_train)

set.seed(1234)
# Modelo de regresión logarítmico
summary(glm(Survived~., df_train, family=binomial(link=logit)))
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = logit),
##     data = df_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7366  -0.5961  -0.4008   0.6195   2.5127
##
```

```
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             4.356469   0.509224   8.555  < 2e-16 ***
## PassengerId             0.000136   0.000350   0.388 0.697675
## Pclass2nd              -1.109160   0.306715  -3.616 0.000299 ***
## Pclass3rd              -2.407963   0.313882  -7.672 1.70e-14 ***
## SexMale                -2.684227   0.202830 -13.234  < 2e-16 ***
## Age                    -0.046726   0.008163  -5.724 1.04e-08 ***
## SibSp                  -0.378960   0.109594  -3.458 0.000544 ***
## Parch                  -0.095703   0.121565  -0.787 0.431133
## Fare                    0.002149   0.002494   0.862 0.388846
## EmbarkedQueenstown     -0.018702   0.393288  -0.048 0.962072
## EmbarkedSouthampton -0.338256   0.243856  -1.387 0.165407
## ---
## 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:  774.89  on 880  degrees of freedom
## AIC: 796.89
##
## Number of Fisher Scoring iterations: 5
```

Tal y como se puede observar, las variables más significativas son *Age, Pclass y Sex*. Por tanto, y teniendo presente la información recabada hasta ahora, será sobre estas variables sobre las que se centrará el análisis de datos.

Regresión

A continuación, se procede a analizar a través de regresión la relación existente entre las variables comentadas y se verá cómo afectan estas a la variable clasificadora *Survived* (variable dicotómica). Para ello, se aplicarán modelos de regresión logística.

Nos centraremos en el sexo, la edad y la clase del billete. De esta forma, se infiere un modelo considerando la variable *Survived* como variable dependiente y los atributos *PClass*, *Sex* y *Age* como variables regresoras o independientes en combinación unas de otras.

```
set.seed(1234)

# Generación del modelo
glm1 <- glm(Survived~Pclass, data = df_train, family=binomial(link=logit))
glm2 <- glm(Survived~Pclass+Sex, data = df_train, family=binomial(link=logit))
glm3 <- glm(Survived~Pclass+Sex+Age, data = df_train,
                                family=binomial(link=logit))
glm4 <- glm(Survived~Pclass+Sex+Age+Embarked, data = df_train,
                                family=binomial(link=logit))

# Tabla de coeficientes IC
tabla.coeficientes <- data.frame(c(1:4), c(glm1$aic,glm2$aic,glm3$aic,glm4$aic))
colnames(tabla.coeficientes) <- c("Modelo", "AIC")
tabla.coeficientes
```

```
##   Modelo       AIC
## 1      1 1089.1080
## 2      2  834.8884
```

```
## 3      3  808.9372
## 4      4  807.1694
```

El índice AIC (Akaike Information Criterion) relaciona la bondad junto con la complejidad del modelo y será de utilidad de cara a comparar con otros modelos (cuanto menor sea el índice, mejor se comportará el modelo). Tal y como se puede observar, la mejora cuando se incorpora la variable *embarked* es prácticamente inexistente, por lo que se optará por un modelo *glm3*:

```
summary(glm3)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = binomial(link = logit),
##     data = df_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7193  -0.6630  -0.4033   0.6396   2.4843
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.748175   0.373923  10.024  < 2e-16 ***
## Pclass2nd   -1.220637   0.261921  -4.660 3.16e-06 ***
## Pclass3rd   -2.519204   0.255748  -9.850  < 2e-16 ***
## SexMale     -2.556278   0.187104 -13.662  < 2e-16 ***
## Age         -0.038040   0.007439  -5.114 3.16e-07 ***
## ---
## 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:  798.94  on 886  degrees of freedom
## AIC: 808.94
##
## Number of Fisher Scoring iterations: 5
```

```
# Odds Ratio
exp(coefficients(glm3))
```

```
## (Intercept)   Pclass2nd   Pclass3rd     SexMale         Age
## 42.44355253  0.29504208  0.08052371  0.07759303  0.96267464
```

Según los resultados, se observa que todas las variables son altamente significativas ($\Pr(>|z|) < 0.05$). Puesto que hay índices de OR inferiores a uno, se hace necesario calcular su inversa para poder valorar la contribución relativa de las distintas variables en el modelo. Por tanto:

```
1/exp(coefficients(glm3))[2:5]
```

```
## Pclass2nd Pclass3rd   SexMale       Age
##  3.389347 12.418703 12.887756  1.038773
```

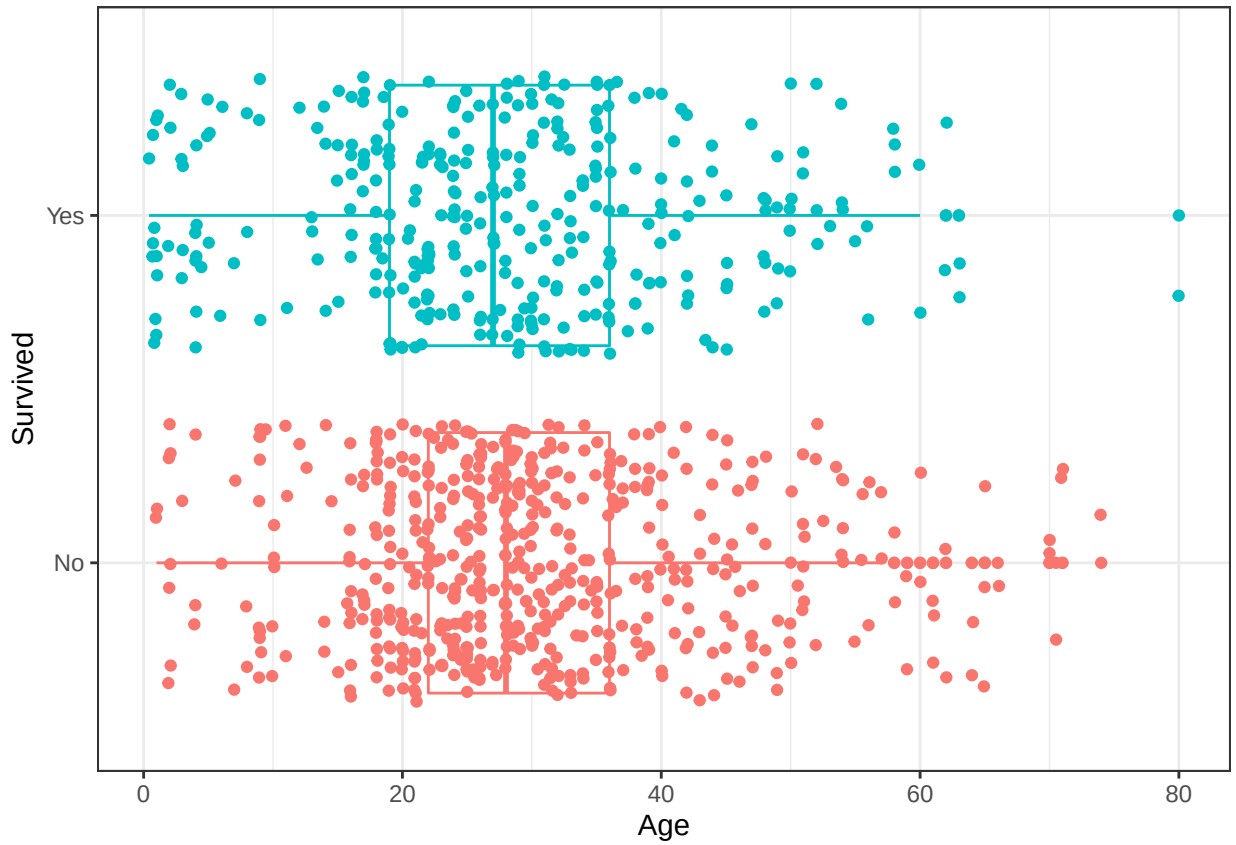
De esta forma, se observa que la variable que más impacto genera en este modelo sobre la variable dependiente *Survived* es la variable *Sex* (un indicativo adicional de que era uno de los principales criterios para abordar el salvamento y que, por tanto, tuvo que ver en el tipo de personas que lograron salvarse).

Si interpretamos los coeficientes parciales de la variable *Sex* (-2.5366) se observa que es negativo. Esto quiere decir que la probabilidad de que una persona sea sexo masculino influye indirectamente en que esta misma

persona sea salvada; o lo que es lo mismo, reduce la probabilidad de ser salvada. Lo mismo sucede con ser pasajeros de 2º y 3º clase. Esto verifica en parte las suposiciones de que se priorizaba los miembros de 1º clase, así como las mujeres y los niños.

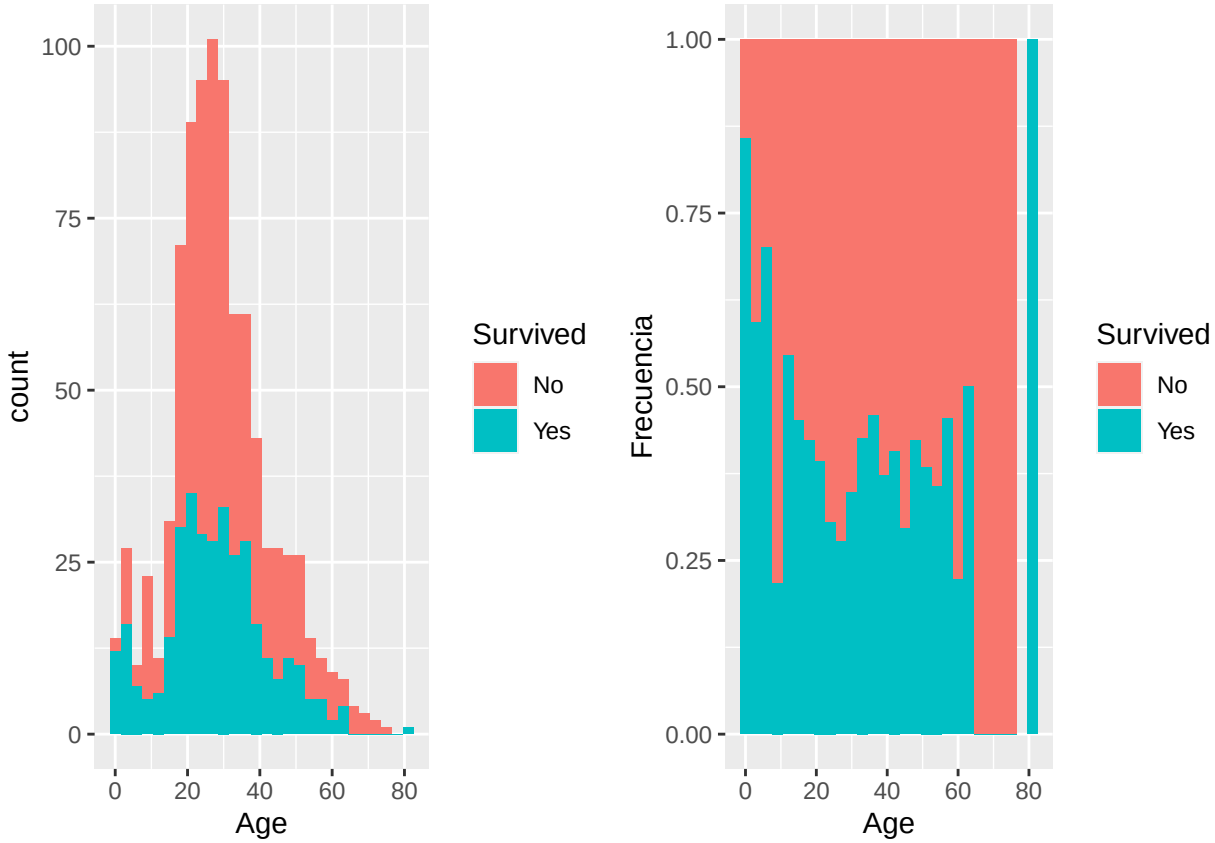
Sin embargo, según este modelo, se observa que la variable *Age* posee un OR próximo a 1; esto es un indicativo que prácticamente no existe asociación entre la variable respuesta y la covariable en dicho modelo. Esto puede verse en la siguiente visualización:

```
ggplot(data = df_train, mapping=aes(x = Age, y = Survived, color=Survived)) +
   geom_boxplot() +
   geom_jitter(width = 0.1) +
   theme_bw() + theme(legend.position = "none")
```



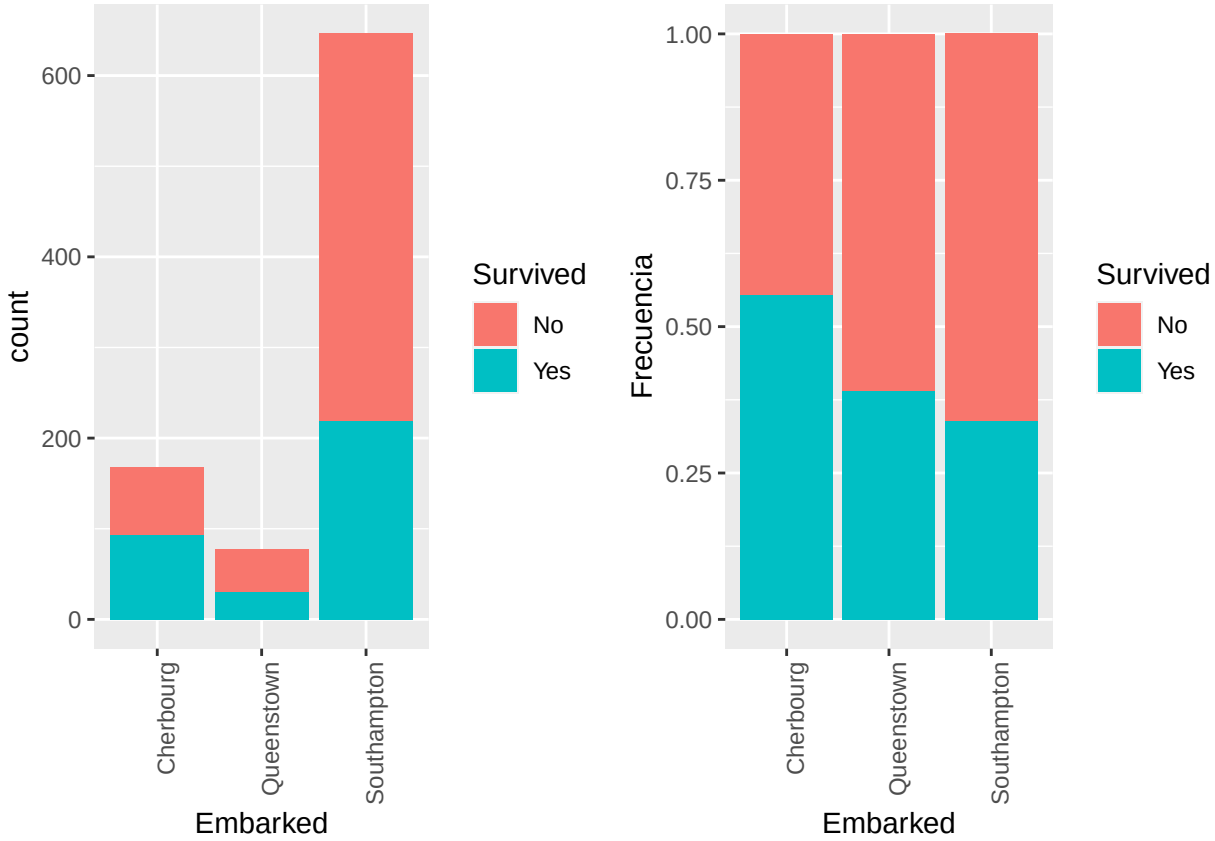
No obstante, si se realiza un análisis comparativo de las variables en términos absolutos y de frecuencia, se observa que la proporción de niños salvados es más elevada que en el resto de edades:

```
g1 <- ggplot(data = df_train, aes(x=Age, fill=Survived)) +
                      geom_histogram(binwidth =3)

g2 <- ggplot(data = df_train, aes(x=Age, fill=Survived)) +
                      geom_histogram(binwidth = 3,position="fill") +
                                            ylab("Frecuencia")

grid.arrange(g1, g2,  nrow = 1)
```



En este punto, se puede decir que se priorizó el salvamento de niños frente a adultos. Además, también se ha visto la relación inversa entre el hecho de ser hombre y ser salvado, al igual que ocurría con la clase del billete, especialmente si se trataba de tercera clase.

```
g1 <- ggplot(data = df_train, aes(x=Embarked, fill=Survived)) +
            geom_bar() +
            theme(axis.text.x = element_text(angle = 90, hjust = 1))

g2 <- ggplot(data = df_train, aes(x=Embarked, fill=Survived)) +
            geom_bar(binwidth = 3,position="fill") +
            ylab("Frecuencia") +
            theme(axis.text.x = element_text(angle = 90, hjust = 1))

grid.arrange(g1, g2,  nrow = 1)
```



Vemos que el porcentaje fue un poco mayor en el caso de Cherbourg con respecto a Queenstown y a Southampton, pero como ya se ha visto este hecho no es relevante.

Nota: Este sería otro de los puntos donde se hace retrospectiva, para ver cómo afecta al modelo y su calidad la inferencia de los valores nulos y el tratamiento de outliers.

Árbol de decisión

El modelo de análisis a aplicar se basa en el algoritmo C5.0, una evolución del C4.5 y el ID3, el cual utiliza una pospoda por promoción automática. Este modelo permite sólo variables de salida categórica, mientras que las de entrada pueden ser de naturaleza continua o categórica. Este modelo se obtiene a partir de la función C5.0() y permite su conversión a reglas. Se aplica sobre las mismas variables anteriormente evaluadas:

```
set.seed(1234)

# Modelo C5.0 en árbol
model_C50t <- C5.0(Survived~Age+Pclass+Sex , df_train)

# Detalle del modelo
summary(model_C50t)
```

```
##
## Call:
## C5.0.formula(formula = Survived ~ Age + Pclass + Sex, data = df_train)
##
##
## C5.0 [Release 2.07 GPL Edition]      Wed Dec 30 11:14:47 2020
## -----------------------------
```

```
##
## Class specified by attribute `outcome'
##
## Read 891 cases (4 attributes) from undefined.data
##
## Decision tree:
##
## Sex = Female:
## :...Pclass in {1st,2nd}: Yes (170/9)
## :   Pclass = 3rd:
## :   :...Age <= 38.5: Yes (132/61)
## :       Age > 38.5: No (12/1)
## Sex = Male:
## :...Age > 8: No (549/92)
##     Age <= 8:
##     :...Pclass in {1st,2nd}: Yes (11)
##         Pclass = 3rd: No (17/6)
##
##
## Evaluation on training data (891 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       6  169(19.0%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     479    70    (a): class No
##      99   243    (b): class Yes
##
##
##  Attribute usage:
##
##  100.00% Sex
##   80.92% Age
##   38.38% Pclass
##
##
## Time: 0.0 secs
```

Tal y como se puede observar este modelo consigue una tasa de error de 19%. La variable de más peso en este tipo modelo es *Sex*, al igual que sucedía en el caso anterior.

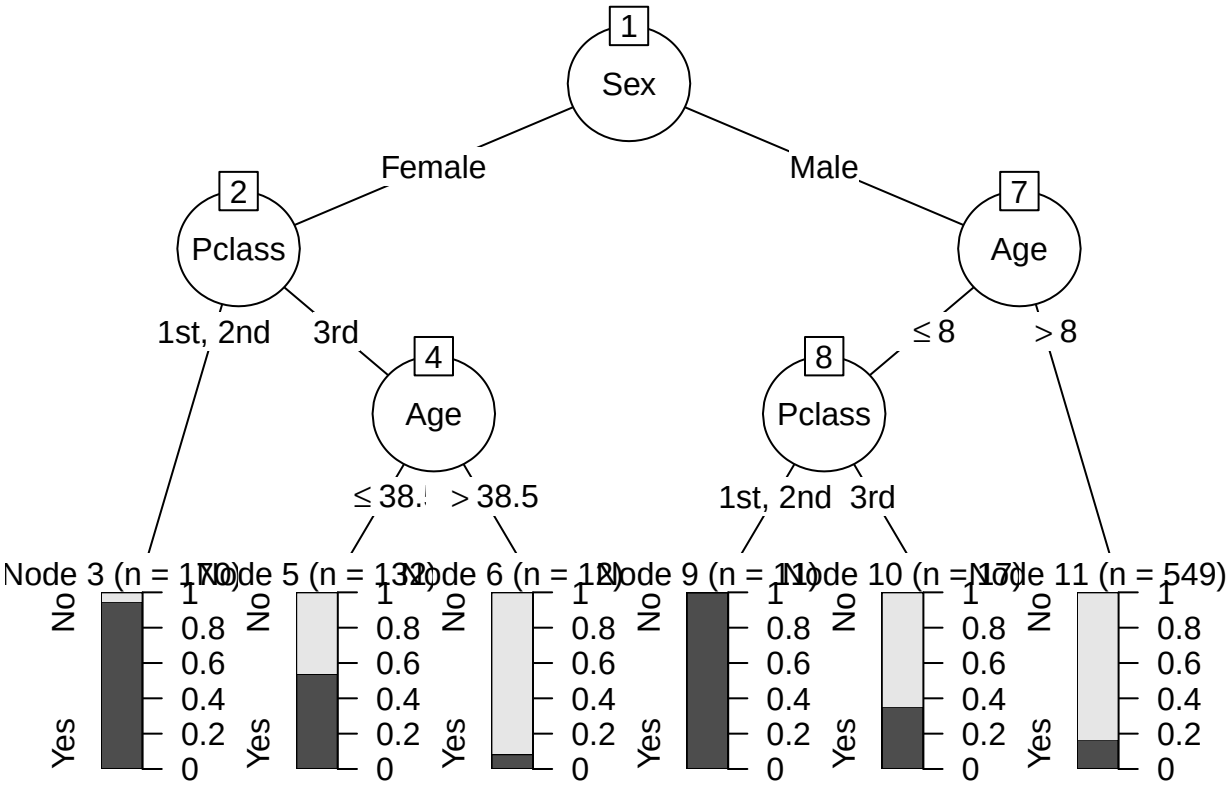
A continuación, generamos el conjunto de reglas y se visualiza el árbol generado para su posterior análisis:

```
set.seed(1234)

# Modelo C5.0 en árbol
model_C50r <- C5.0(Survived~Age+Pclass+Sex , df_train, rules = TRUE)

# Detalle del modelo
summary(model_C50r)
```

```
##
## Call:
## C5.0.formula(formula = Survived ~ Age + Pclass + Sex, data = df_train, rules
##  = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Wed Dec 30 11:14:47 2020
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 891 cases (4 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (51/4, lift 1.5)
##  Age > 38.5
##  Pclass = 3rd
##  ->  class No  [0.906]
##
## Rule 2: (577/109, lift 1.3)
##  Sex = Male
##  ->  class No  [0.810]
##
## Rule 3: (11, lift 2.4)
##  Age <= 8
##  Pclass in {1st, 2nd}
##  Sex = Male
##  ->  class Yes  [0.923]
##
## Rule 4: (314/81, lift 1.9)
##  Sex = Female
##  ->  class Yes  [0.741]
##
## Default class: No
##
##
## Evaluation on training data (891 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##       4  169(19.0%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     479    70    (a): class No
##      99   243    (b): class Yes
##
##
##  Attribute usage:
##
```

```
##  100.00% Sex
##    6.96% Age
##    6.96% Pclass
##
##
## Time: 0.0 secs
```

```
# Visualización
plot(model_C50t)
```



Analizando las reglas, se puede hacer las siguientes observaciones:

- Una persona con edad superior a 38 años y que viajase en 3ª clase no sobreviviría con un 90,6% de validez.
- Un varón no sobreviviría con un 81,0% de validez.
- Un pasajero de menos de 9 años, que viajaba en 1ª o 2ª clase, de sexo masculino, se salvaba con 92,3% de validez.
- Una Mujer se salvaría con un 74,1% de validez.

Si se observa el gráfico del árbol, se pude observar que un pasajero de sexo femenino que viajaba en 1ª o 2ª clase tenía una alta probabilidad de sobrevivir. Lo mismo ocurre con un pasajero masculino menor de 9 años que se encuentre en 1º o 2º clase.

Representación de los resultados

Los resultados se han ido mostrando en las diferentes gráficas y datos numéricos calculados conforme se han ido realizando los diferentes análisis de datos. No obstante, se van a resumir a continuación con tablas y

gráficas a modo resumen.

Comenzamos viendo el porcentaje de pasajeros que sobrevivieron (o no) en función de las variables *Age*, *Pclass* y *Age*.

```
# Porcentaje de supervivencia por sexo
prop.table(table(df_train$Sex, df_train$Survived), margin=1)
```

```
##
##                 No       Yes
##   Female 0.2579618 0.7420382
##   Male   0.8110919 0.1889081
```

```
# Porcentaje de supervivencia por la clase en la que se viajaba
prop.table(table(df_train$Pclass, df_train$Survived), margin=1)
```

```
##
##              No       Yes
##   1st 0.3703704 0.6296296
##   2nd 0.5271739 0.4728261
##   3rd 0.7576375 0.2423625
```

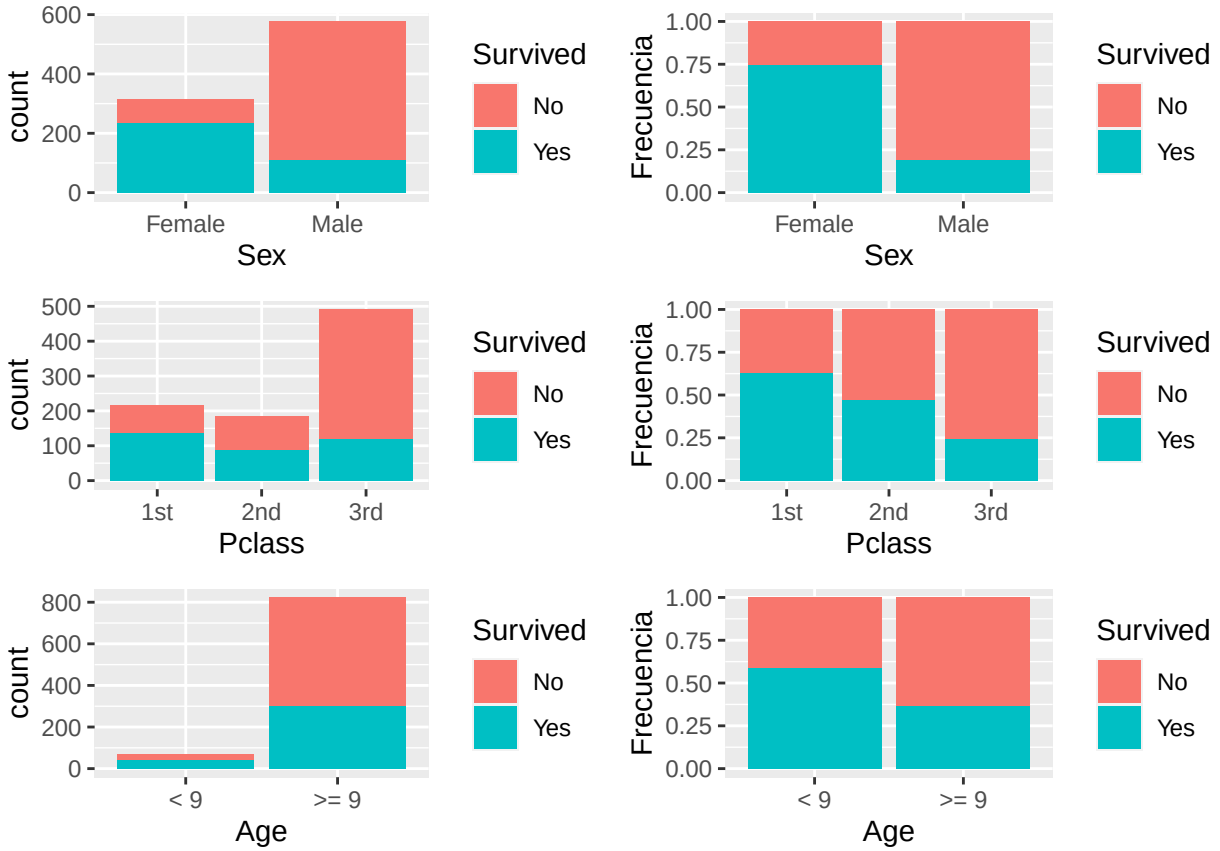
```
# Porcentaje de supervivencia por la edad
prop.table(table(cut(df_train$Age, breaks = c(-1,9,Inf),
                     labels = c("< 9",">= 9")), df_train$Survived), margin=1)
```

```
##
##               No       Yes
##   < 9  0.4117647 0.5882353
##   >= 9 0.6330498 0.3669502
```

A continuación se muestran algunas gráficas donde se representa de forma visual lo visto en datos numéricos.

```
g1 <- ggplot(data = df_train, aes(x=Sex, fill=Survived)) +
                              geom_bar()+xlab("Sex")

g2 <- ggplot(data = df_train, aes(x=Sex, fill=Survived)) +
                              geom_bar(position="fill") +
                              ylab("Frecuencia") +
                              xlab("Sex")

g3 <- ggplot(data = df_train, aes(x=Pclass, fill=Survived)) +
                              geom_bar() +
                              xlab("Pclass")

g4 <- ggplot(data = df_train, aes(x=Pclass, fill=Survived)) +
                              geom_bar(position="fill") +
                              ylab("Frecuencia") +
                              xlab("Pclass")

g5 <- ggplot(data = df_train, aes(x=cut(df_train$Age, breaks = c(-1,9,Inf),
                              labels = c("< 9",">= 9")), fill=Survived)) +
                              geom_bar() +
                              xlab("Age")

g6 <- ggplot(data = df_train, aes(x=cut(df_train$Age, breaks = c(-1,9,Inf),
                              labels = c("< 9",">= 9")), fill=Survived)) +
                              geom_bar(position="fill") +
```

```
                    ylab("Frecuencia")+xlab("Age")

grid.arrange(g1, g2, g3, g4, g5, g6, nrow = 3)
```



Conclusiones

A la vista de los resultados, se puede concluir lo siguiente:

- Las mujeres tenían una mayor probabilidad de sobrevivir que los hombres, por lo que se priorizó su evacuación en el momento del hundimiento.
- Los menores de 9 años tienen una probabilidad superior de salir con vida del naufragio que el resto de pasajeros.
- Los pasajeros con billete de 3º clase tuvieron una probabilidad de supervivencia mucho menor que los pasajeros de 1ª y 2ª clase, en especial si se tenía un billete de 1ª clase.

Con este estudio se da respuesta a las preguntas que se hacían al comienzo de este estudio, y se confirman las premisas que se podían ir intuyendo conforme se avanza en el análisis de los datos. En concreto, para validar estas conclusiones se han realizado contrastes de hipótesis, estudio de regresión logarítmica y un modelo supervisado basado en un árbol de decisión.

Contribuciones	**Firma**
Investigación previa	Juan Rodríguez Vega, Alejandro Gallardo Alberola
Redacción de respuestas	Juan Rodríguez Vega, Alejandro Gallardo Alberola
Desarrollo de código	Juan Rodríguez Vega, Alejandro Gallardo Alberola