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?

El hundimiento del Titanic es uno de los naufragios más conocidos de la historia.

El 15 de abril de 1912, durante su viaje inaugural, el considerado "insumergible" RMS Titanic se hundió después de chocar con un iceberg en el océano Atlántico. Desafortunadamente, no había suficientes botes salvavidas para todos los pasajeros a bordo, lo que resultó en la muerte de 1502 de 2224 pasajeros y tripulación.

Aunque había un elemento de suerte en la supervivencia, parece que algunos grupos de personas tenían más probabilidades de sobrevivir que otros.

Con el dataset que proporciona la plataforma Kaggle se pide construir un modelo predictivo que responda a la pregunta: "¿qué tipo de personas tenían más probabilidades de sobrevivir?" usando los datos de los pasajeros (es decir, nombre, edad, sexo, clase socioeconómica, etc.). Además, con este dataset también intentaremos ver qué relaciones existen entre las distintas variables.

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

La integración o fusión de los datos consiste en la combinación de datos procedentes de múltiples fuentes, con el fin de crear una estructura de datos coherente y única que contenga mayor cantidad de información.

Los datos de los que disponemos son:

- Train.csv: contiene los detalles de un subconjunto de los pasajeros a bordo (891 para ser exactos) y, además revela si sobrevivieron o no.
- Test.csv: contiene información similar al train.csv pero no revela si cada pasajero sobrevivió o no al hundimiento.

Las variables presentes en el dataset son:

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	

sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Para construir un modelo predictivo que nos informe sobre qué tipo de personas tenían más probabilidades de sobrevivir usaremos el dataset train.csv que es el único que tiene información al respecto de la supervivencia de cada pasajero.

Pero si queremos realizar otro tipo de test estadístico, cómo por ejemplo, correlaciones, comparaciones, etc., podemos integrar los dos dataset train.csv y test.csv en uno.

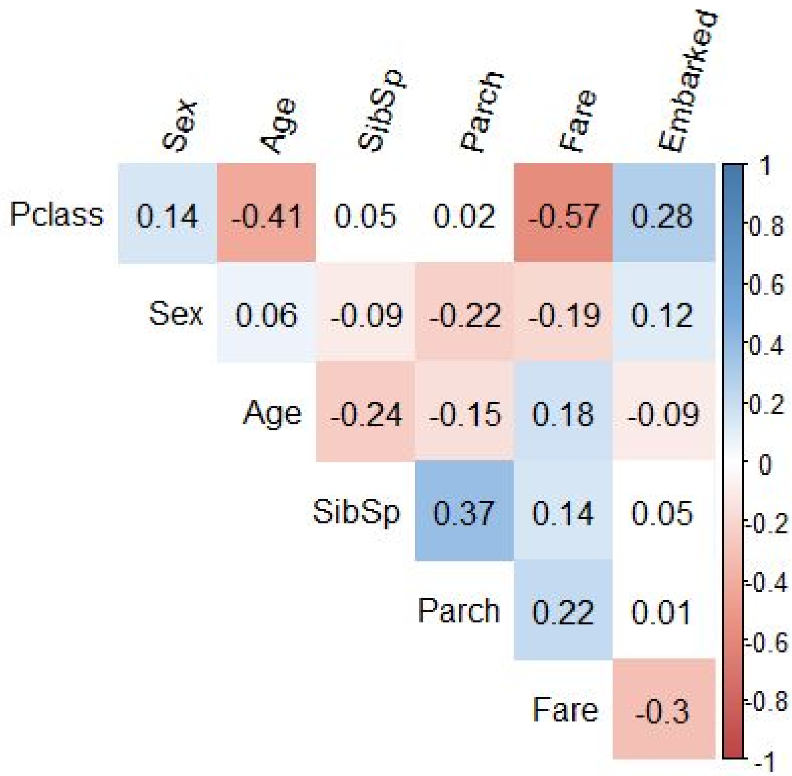
```
train <- read.csv('C:/Users/ester/Desktop/Kaggle/train.csv')
test <- read.csv('C:/Users/ester/Desktop/Kaggle/test.csv')
datos <- rbind(train[,-2], test)
```

Una de las primeras etapas en el preprocesado de los datos es el filtrado o selección de datos de interés.

Realizamos una exploración de los datos con el objetivo de analizar globalmente sus características e identificar fuertes correlaciones entre atributos, de modo que se pueda prescindir de aquella información más redundante.

Para estudiar la correlación transformamos aquellas variables que nos interesa en numéricas y miramos el coeficiente de correlación que presentan.

```
datos_ <- datos
datos_$Pclass <- as.numeric(as.factor(datos$Pclass))
datos_$Sex <- as.numeric(as.factor(datos$Sex))
datos_$Embarked <- as.numeric(as.factor(datos$Embarked))
datos_cor <- datos_[,c(2,4:7,9,11)]
```



A diferencia de una matriz de correlación que indica los coeficientes de correlación entre pares de variables, la prueba de correlación se utiliza para comprobar si la correlación (denominada ρ) entre 2 variables es significativamente diferente de 0 o no.

En realidad, un coeficiente de correlación diferente de 0 no significa que la correlación sea significativamente diferente de 0. Esto debe probarse con una prueba de correlación.

Las hipótesis nula y alternativa para la prueba de correlación son las siguientes:

H0: $\rho=0$

H1: $\rho \neq 0$

```
##          Pclass  Sex   Age SibSp Parch Fare Embarked
## Pclass       NA 0.00 0.000 0.028 0.508    0    0.000
## Sex       0.000   NA 0.040 0.000 0.000    0    0.000
## Age       0.000 0.04    NA 0.000 0.000    0    0.004
## SibSp     0.028 0.00 0.000    NA 0.000    0    0.014
## Parch     0.508 0.00 0.000 0.000    NA    0    0.089
## Fare      0.000 0.00 0.000 0.000 0.000   NA    0.000
## Embarked  0.000 0.00 0.004 0.014 0.089    0       NA
```

Vemos que hay correlación entre la variable `Pclass` y `Fare`, y entre las variables `SibSp` y `Parch`. Con lo cual prescindiremos de alguna de estas variables en los estudios posteriores.

3. Limpieza de los datos.

Realizamos una primera limpieza de los datos transformando algunas variables en factores:

```
datos$Pclass <- factor(datos$Pclass, levels = c(1,2,3), labels = c('First',
'Second', 'Third'))
datos$Sex  <-  factor(datos$Sex,  levels  =  c('male','female'),  labels  =
c('Male', 'Female'))
datos$Embarked <- factor(datos$Embarked, levels = c('C','Q','S'), labels =
c('Cherbourg', 'Queenstown', 'Southampton'))
```

Además, creamos nuevas variables que puede ser interesante para los estudios estadísticos posteriores.

A partir de la variable `Name` creamos la variable `Formula` y `Nombre_Familia`, la primera hace referencia a la fórmula que usan para dirigirse a la persona (Mr, Miss, etc) lo cual nos da información sobre si es un niño o no, la segunda hace referencia al apellido de la familia.

A partir de las variables `SibSp` y `Parch`, sumandolas creamos una nueva variable `Num_Familiares_Totales` ya que estas dos variables hemos visto que presentaban cierta correlación.

```
formula  <-  unlist(sapply(strsplit(datos$Name,  ",  "),  function(x)  x[2],
simplify=FALSE))
datos$Formula  <-  unlist(sapply(strsplit(formula,  ".  "),  function(x)  x[1],
simplify=FALSE))
datos$Nombre_Familia <- unlist(sapply(strsplit(datos$Name, ". "), function(x)
x[1], simplify=FALSE))
datos$Num_Familiares_Totales <- datos$SibSp + datos$Parch
```

Las variables `Pclass` y `Fare` también presentan correlación, por tanto, en los estudios posteriores prescindiremos de la variable `Fare`.

Las variables `Ticket` y `Cabin` prescindimos de ellas, ya que presentan demasiados registros vacíos, y por tanto, no nos aportan información para el estudio.

```
datos_clean <- datos[,c(1,2,4:7,9,11:14)]
```

El dataset queda del siguiente modo:

```
str(datos_clean)

## 'data.frame':    1309 obs. of  11 variables:
##  $ PassengerId           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Pclass                : Factor w/ 3 levels "First","Second",..: 3 1 3 1
3 3 1 3 3 2 ...
##  $ Sex                   : Factor w/ 2 levels "Male","Female": 1 2 2 2 1 1
```

```
1 1 2 2 ...
##  $ 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 ...
##  $ Embarked              : Factor w/ 3 levels "Cherbourg","Queenstown",..: 
3 1 3 3 3 2 3 3 3 1 ...
##  $ Formula               : chr  "Mr" "Mrs" "Miss" "Mrs" ...
##  $ Nombre_Familia        : chr  "Braund" "Cumings" "Heikkinen" "Futrelle" 
...
##  $ Num_Familiares_Totales: int  1 1 0 1 0 0 0 4 2 1 ...
```

```
head(datos_clean)
```

```
##   PassengerId Pclass    Sex Age SibSp Parch    Fare    Embarked Formula
## 1           1  Third   Male  22     1     0  7.2500 Southampton      Mr
## 2           2  First Female  38     1     0 71.2833   Cherbourg     Mrs
## 3           3  Third Female  26     0     0  7.9250 Southampton    Miss
## 4           4  First Female  35     1     0 53.1000 Southampton     Mrs
## 5           5  Third   Male  35     0     0  8.0500 Southampton      Mr
## 6           6  Third   Male  NA     0     0  8.4583  Queenstown      Mr
##   Nombre_Familia Num_Familiares_Totales
## 1         Braund                      1
## 2        Cumings                      1
## 3      Heikkinen                      0
## 4       Futrelle                      1
## 5          Allen                      0
## 6          Moran                      0
```

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Vemos que algunas variables contienen <u>ceros y/o elementos vacíos</u>.

Las variables `SibSp`, `Parch` y `Fare` contienen ceros.

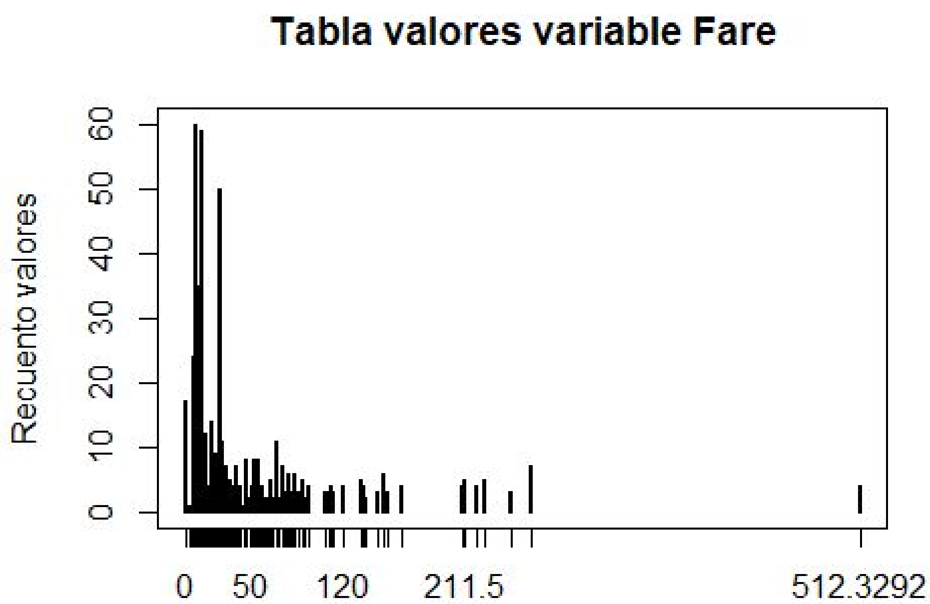
La variable `SibSp` hace referencia al número de hermanos/cónyuges a bordo del Titanic, con que los ceros tienen sentido y entran dentro del rango de la variable.

La variable `Pacrh` hace referencia al número de padre e hijos a bordo del Titanic, con lo cual los ceros también tienen sentido y entran dentro del rango de valores admisibles para la variable.

La variable `Fare` hace referencia a la tarifa que pagan los pasajeros por su ticket. Aparecen 17 valores 0, no sabemos si es un error o esos pasajeros viajaron gratis.

Entre la gente que viajaba en el Titanic había tripulación y pasajeros, podemos suponer que esos 17 0 que aparecen son debidos a la tripulación que aparece en el dataset.

```r
plot(table(datos$Fare), ylab = 'Recuento valores', main = 'Tabla valores variable Fare')
```




```r
datos_0 <- datos_clean[datos_clean$Fare == 0,]
head(datos_0)
```

```
##     PassengerId Pclass  Sex Age SibSp Parch Fare    Embarked Formula
## 180         180  Third Male  36     0     0    0 Southampton      Mr
## 264         264  First Male  40     0     0    0 Southampton      Mr
## 272         272  Third Male  25     0     0    0 Southampton      Mr
## 278         278 Second Male  NA     0     0    0 Southampton      Mr
## 303         303  Third Male  19     0     0    0 Southampton      Mr
## 414         414 Second Male  NA     0     0    0 Southampton      Mr
##     Nombre_Familia Num_Familiares_Totales
## 180        Leonard                      0
## 264       Harrison                      0
## 272      Tornquist                      0
## 278        Parkes                       0
## 303        Johnson                      0
## 414     Cunningham                      0
```

Vemos que todos son varones, mayores de edad y embarcaron en el puerto de Southampton. Cuando buscamos información de estos pasajeros vemos que algunos pertenecían al Titanic Guarantee Group (El equipo de Belfast enviado por los constructores de barcos Harland & Wolff para acompañar al Titanic en su viaje inaugural), con lo cual podemos suponer que los 0 son correctos, era gente que estaba viajando por su trabajo.

Vemos que las variables `Age`, `Fare` y `Embarked` contienen valores perdidos.

Las ventajas de imputar son que logramos obtener un conjunto de datos completo sin datos faltantes, se puede reducir el sesgo debido a la no respuesta y la imputación opera sobre los datos, de forma que los resultados obtenidos por los diferentes análisis son mutuamente consistentes.

Por otra parte, la imputación también tiene desventajas ya que hay que tener en cuenta que el futuro análisis no distingue entre las imputaciones y los datos reales. Además los valores imputados pueden ser buenas estimaciones pero no son datos reales y no podemos asegurar una mejora en el sesgo respecto del sistema de datos incompletos. Al fin y al cabo la imputación es un procedimiento para generar datos.

Si el método de imputación no es el adecuado, posiblemente aumente el sesgo y sobreestime la varianza, obteniendo datos imputados inconsistentes produciendo una base de datos no confiables, llevando a la interpretación errónea de los resultados por parte de los usuarios.

Las variables `Fare` y `Embarked` tienen 1 y 2 valores perdidos respectivamente, como la muestras es bastante grande no hace falta imputar datos. Pero en la variable `Age` faltan 263 valores del 1309, representa un 20% de los datos, además un 20-30% es el máximo de valores perdidos para los que algunos autores recomiendan la imputación de datos.

Realizaremos una imputación simple por regresión de la variable `Age`:

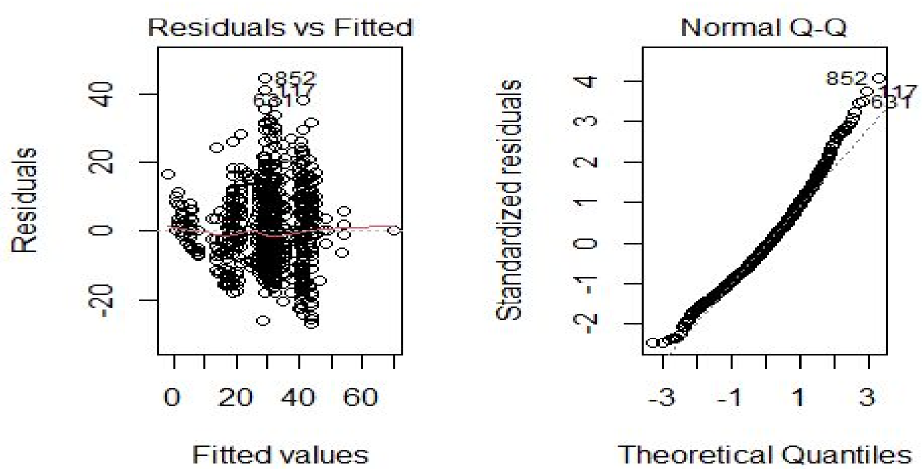
```
library(arm)
datos_clean$Age_lm <- datos_clean$Age
lmod <- lm(Age ~ Sex + Pclass + Num_Familiares_Totales + Formula, data =
datos_clean)
summary(lmod)

##
## Call:
## lm(formula = Age ~ Sex + Pclass + Num_Familiares_Totales + Formula,
##     data = datos_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.978  -7.704  -1.047   5.981  44.811
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               71.7526    11.0498   6.494 1.30e-10 ***
## SexFemale                  2.3258    11.9284   0.195  0.84545
## PclassSecond              -9.3936     0.9722  -9.662  < 2e-16 ***
## PclassThird              -12.2857     0.8566 -14.343  < 2e-16 ***
## Num_Familiares_Totales    -0.8763     0.2717  -3.225  0.00130 **
## FormulaCol               -17.5335    12.3483  -1.420  0.15594
## FormulaDon               -31.7526    15.6173  -2.033  0.04229 *
## FormulaDona              -35.0784    19.6439  -1.786  0.07444 .
```

```
## FormulaDr                 -25.0784    11.9280  -2.102  0.03575 *
## FormulaJonkheer           -33.7526    15.6173  -2.161  0.03091 *
## FormulaLady               -25.2021    19.6421  -1.283  0.19976
## FormulaMajor              -23.2526    13.5277  -1.719  0.08594 .
## FormulaMaster             -52.7012    11.1707  -4.718 2.71e-06 ***
## FormulaMiss               -43.1464    16.2856  -2.649  0.00819 **
## FormulaMlle               -50.0784    18.0272  -2.778  0.00557 **
## FormulaMme                -50.0784    19.6439  -2.549  0.01094 *
## FormulaMr                 -30.2777    11.0713  -2.735  0.00635 **
## FormulaMrs                -29.2242    16.2875  -1.794  0.07306 .
## FormulaMs                 -36.6848    19.6835  -1.864  0.06265 .
## FormulaRev                -20.6708    11.7522  -1.759  0.07889 .
## FormulaSir                -21.8763    15.6102  -1.401  0.16139
## Formulath                 -41.0784    19.6439  -2.091  0.03676 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.04 on 1024 degrees of freedom
##   (263 observations deleted due to missingness)
## Multiple R-squared:  0.4255, Adjusted R-squared:  0.4137
## F-statistic: 36.11 on 21 and 1024 DF,  p-value: < 2.2e-16
```

```
datos_clean$Age_lm[is.na(datos_clean$Age)] <- predict(lmod, newdata = subset(datos_clean, is.na(Age)))
```

Comprobamos que se cumplen los supuestos del modelo lineal:

```
par(mfrow = c(1,2))
plot(lmod,1:2)
```



Transformamos la variable `Age_lm` (`Age` con datos faltantes imputados) en un factor con tres niveles, Nino (0-15), Adulto Joven (16-40) y Adulto Mayor (> 41), ya que nos parece más interesante para estudios posteriores:

```
datos_clean$Age_grupo <- datos_clean$Age_lm
datos_clean$Age_grupo[datos_clean$Age_lm <= 15] <- "Nino"
datos_clean$Age_grupo[datos_clean$Age_lm > 15 & datos_clean$Age_lm <= 40] <-
"Adulto Joven"
datos_clean$Age_grupo[datos_clean$Age_lm > 40] <- "Adulto Mayor"
datos_clean$Age_grupo <- factor(datos_clean$Age_grupo)
```

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos o outliers son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Al ser observaciones que se desvían del resto levantan sospechas sobre si fueron generadas mediante el mismo mecanismo.

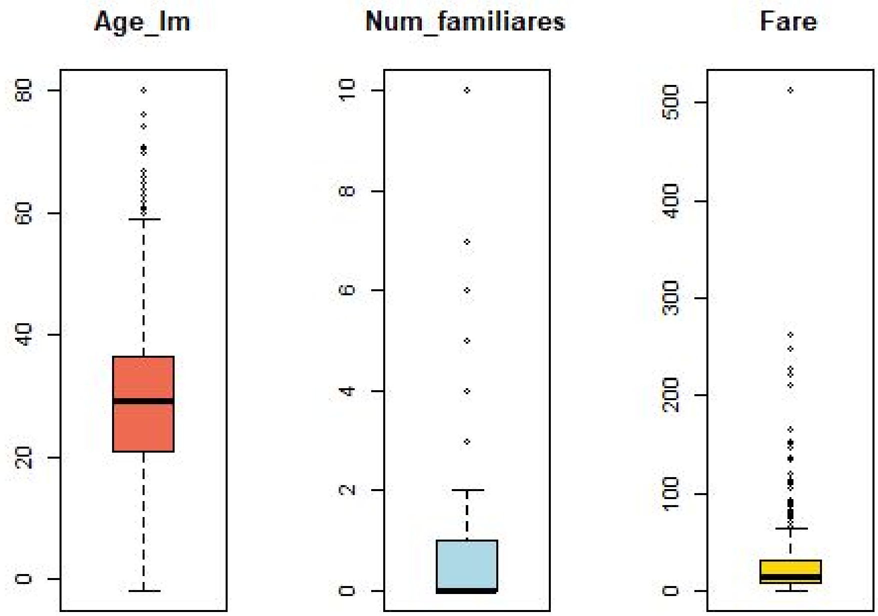
La decisión sobre qué se considera un valor extremo puede resultar controvertida, pero generalmente se considera que un valor es extremo cuando el valor se encuentra alejado 3 desviaciones estándar con respecto a la media del conjunto de datos. Por ello, normalmente se utiliza la representación de los datos mediante gráficos de cajas (boxplots) con el objetivo de detectar dichos outliers. Otros métodos que permiten detectar los valores extremos se basan en la distancia de Mahalanobis o la distancia de Cook.

Sus posibles efectos son:

- incrementar el error en la varianza de los datos

- sesgar los cálculos y estimaciones.

```
par(mfrow = c(1,3))
boxplot(datos_clean$Age_lm, main= "Age_lm", col = 'coral2')
boxplot(datos_clean$Num_Familiares_Totales, main = "Num_familiares", col =
'lightblue')
boxplot(datos_clean$Fare, main = "Fare", col = 'gold')
```



Los valores extremos (outliers) pueden aparecer por distintos motivos.

- son valores válidos que forman parte de la muestra

- son valores debidos a una desviación sistemática en el grupo de valores extremos

- son errores en los datos

En este caso parece que los outliers son valores válidos y entran dentro del rango de las variables, por tanto, forman parte de la muestra, por lo que no se deben modificar ni eliminar, y se deben tener en cuenta en el análisis de los datos.

4. Análisis de los datos.

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

- Mediante un algoritmo de clasificación predecir que pasajeros sobrevivieron al hundimiento del Titanic según sus características. En este caso parece interesante utilizar un árbol de decisión (decision trees), ya que nos da la información de cómo se clasifican los pasajeros según sus características y nos devuelve un diagrama del mismo.

- Mediante un modelo de regresión logística predecir que probabilidad hay de que un pasajero sobreviva en base a sus características.

- Diferencia de la supervivencia entre las distintas clases del Titanic.

- Diferencias de la mediana edad entre hombres y mujeres a bordo del Titanic.
- Diferencia de las medianas de edad entre las distintas clases de pasajeros del Titanic.
- Diferencia de la mediana de número total de familiares entre hombres y mujeres en el Titanic.

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

<u>Comprobación de la normalidad:</u>

Este contraste se realiza para comprobar si se verifica la hipótesis de normalidad necesaria para que el resultado de algunos análisis sea fiable, como por ejemplo para el t-test o Anova.

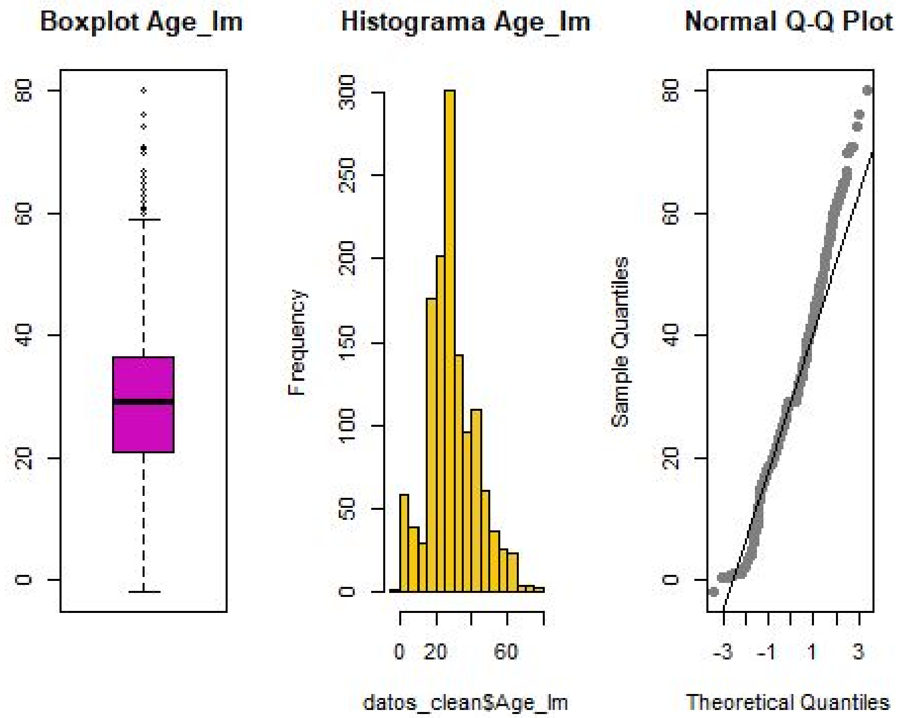
Para comprobar la hipótesis nula de que la muestra ha sido extraída de una población con distribución de probabilidad normal se puede realizar un estudio gráfico y/o analítico.

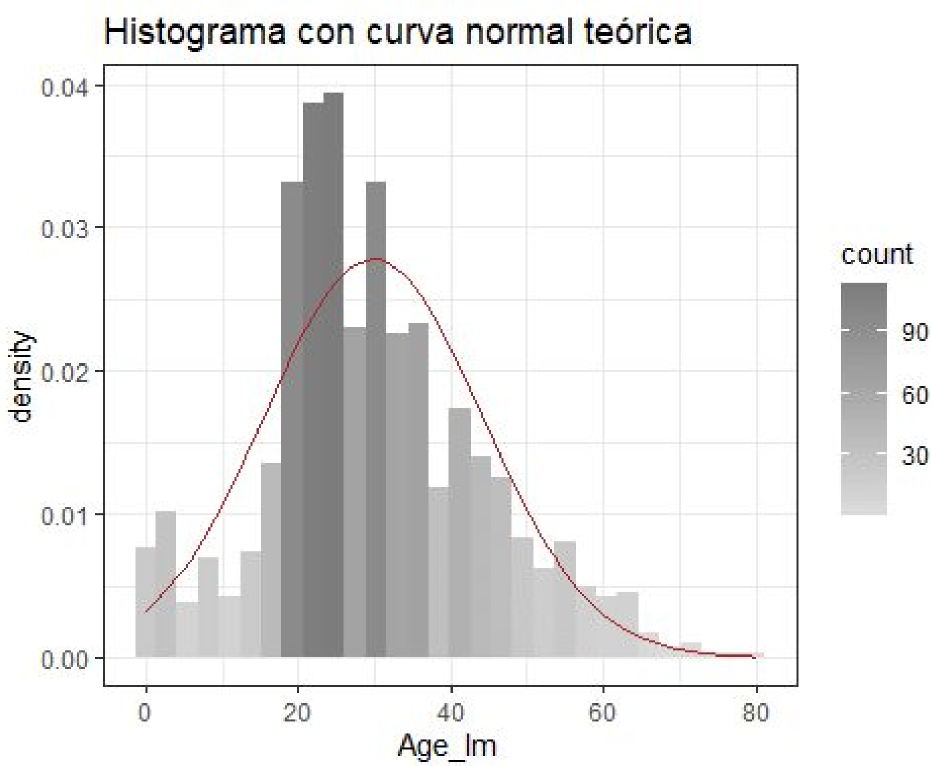
La hipótesis nula y alternativa son:

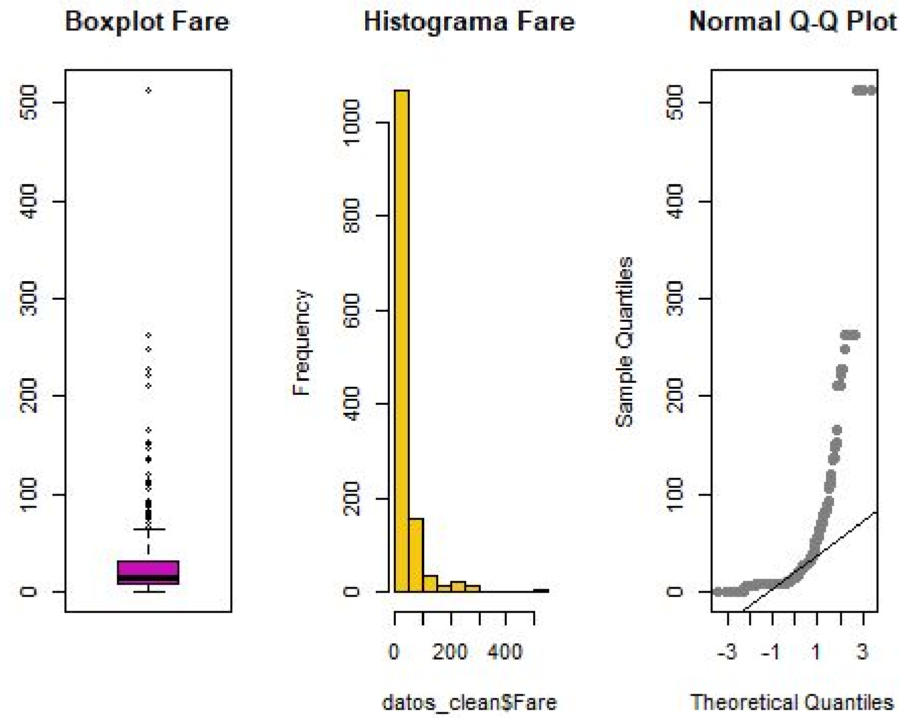
H0: los datos provienen de una distribución normal.

H1: los datos no provienen de una distribución normal.

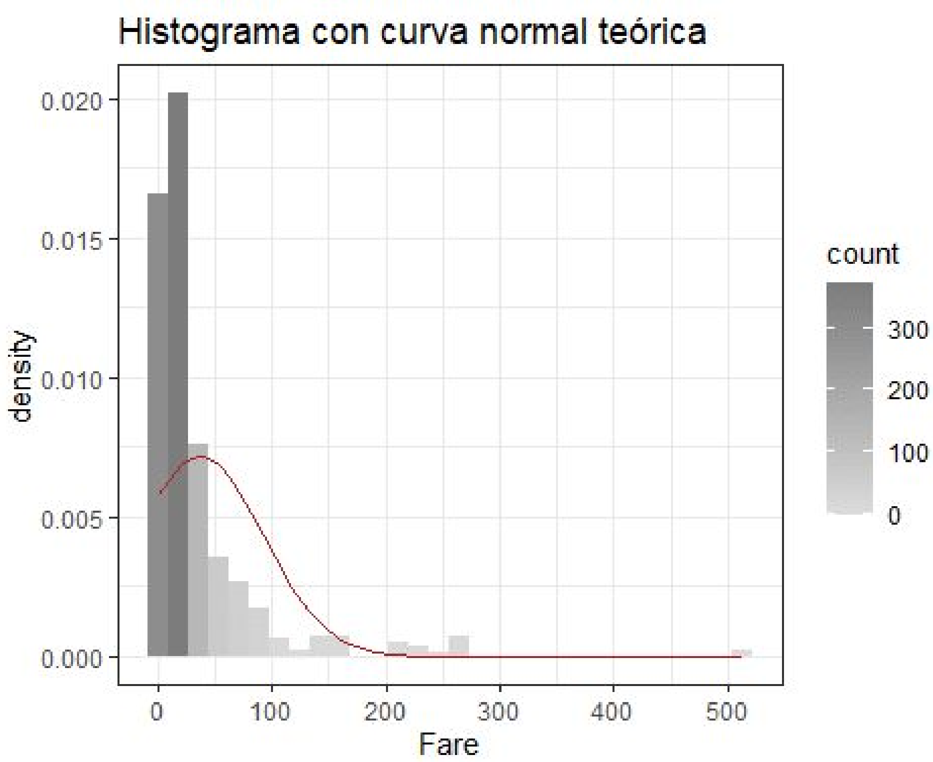
Estudio gráfico de la normalidad de las variables Age_lm, Fare y Número de familiares totales:


Boxplot Age_Im
80
60
40
20
0
Histograma Age_Im
Frequency
300
250
200
150
100
50
0
0 20 60
datos_clean$Age_Im
Normal Q-Q Plot
Sample Quantiles
80
60
40
20
0
-3 -1 1 3
Theoretical Quantiles

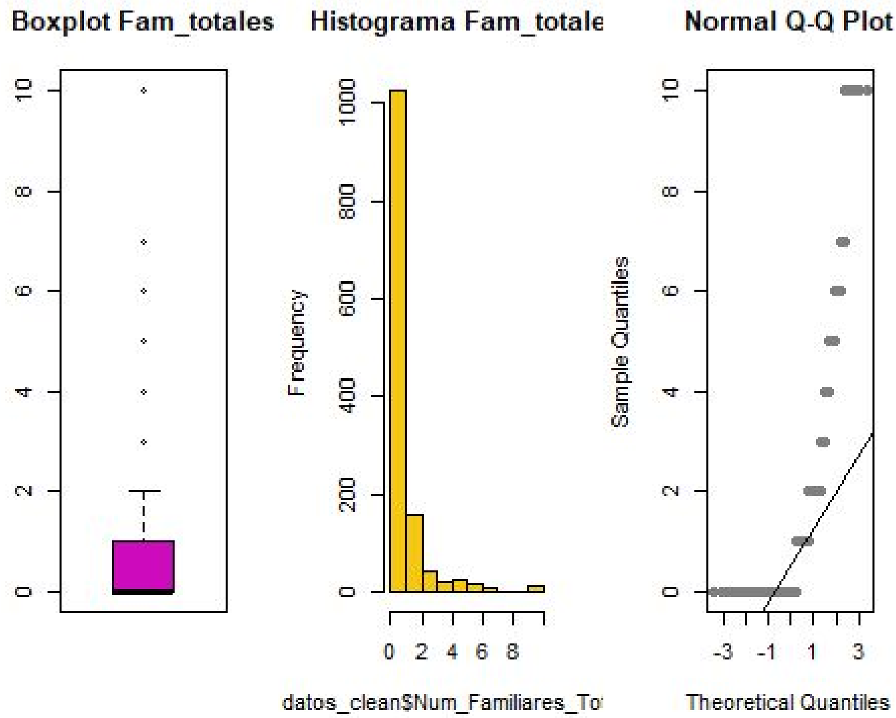

Histograma con curva normal teórica
density
0.04
0.03
0.02
0.01
0.00
0 20 40 60 80
Age_Im
count
90
60
30



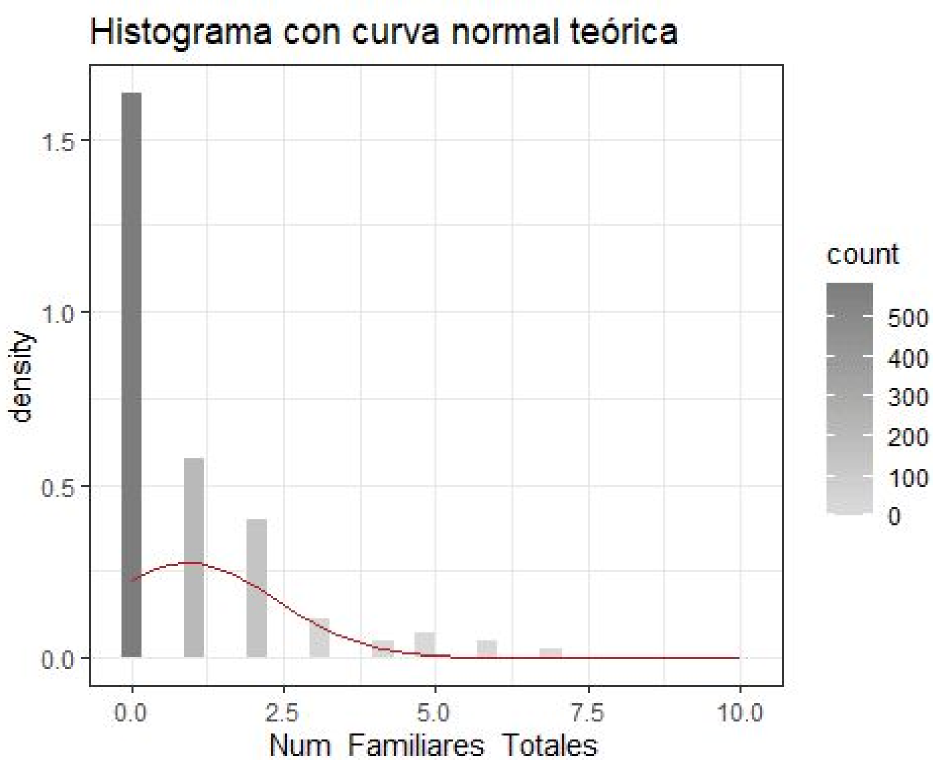
Boxplot Fare
Histograma Fare
Normal Q-Q Plot
Frequency
datos_clean$Fare
Sample Quantiles
Theoretical Quantiles



Histograma con curva normal teórica
density
Fare
count



Boxplot Fam_totales
Histograma Fam_totale
Normal Q-Q Plot
Frequency
datos_clean$Num_Familiares_To
Sample Quantiles
Theoretical Quantiles



Histograma con curva normal teórica
density
Num_Familiares_Totales
count

El test Lilliefors asume que la media y varianza son desconocidas, estando especialmente desarrollado para contrastar la normalidad. Es la alternativa al test de Shapiro-Wilk cuando el número de observaciones es mayor de 50, como es nuestro caso.

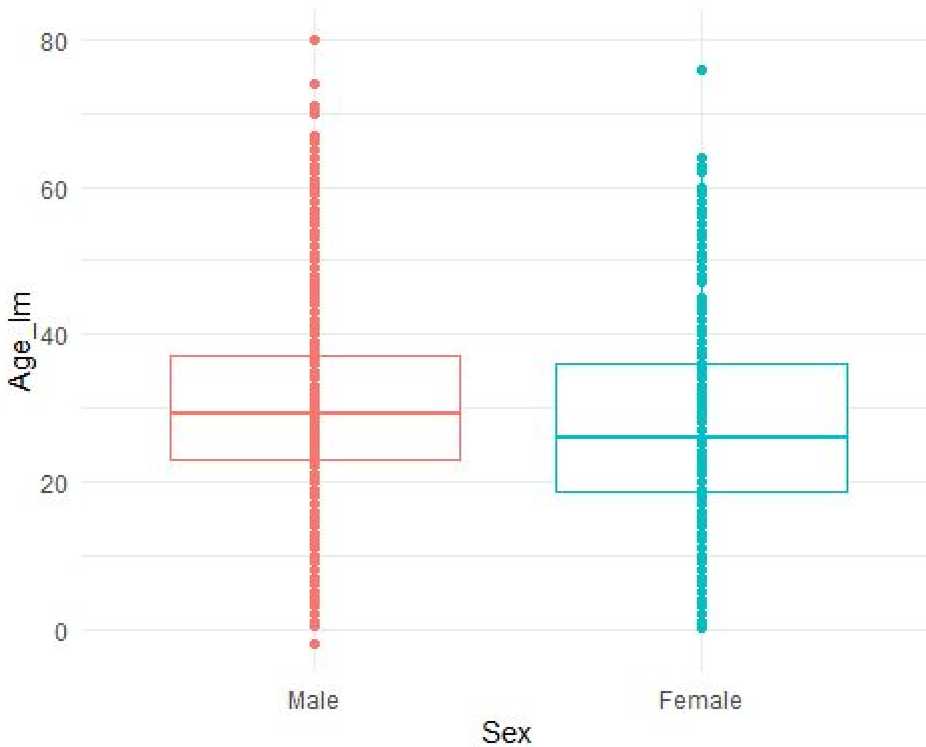
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_clean$Age_lm
## D = 0.10229, p-value < 2.2e-16

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_clean$Fare
## D = 0.28586, p-value < 2.2e-16

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_clean$Num_Familiares_Totales
## D = 0.31514, p-value < 2.2e-16
```

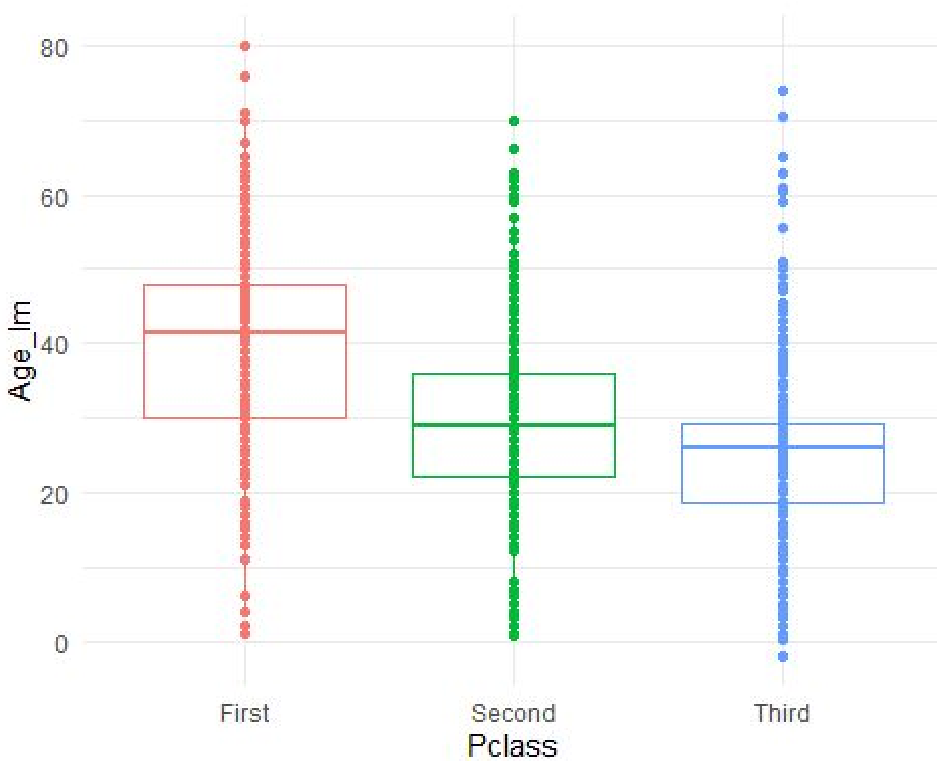
En los tres casos el p-valor es menor de 0.05, rechazamos la HO (los valores siguen una distribución normal), por tanto, no podemos afirmar que los valores de las variables `Age_lm`, `Fare` y `Num_Familiares_Totales` siguen una distribución normal.

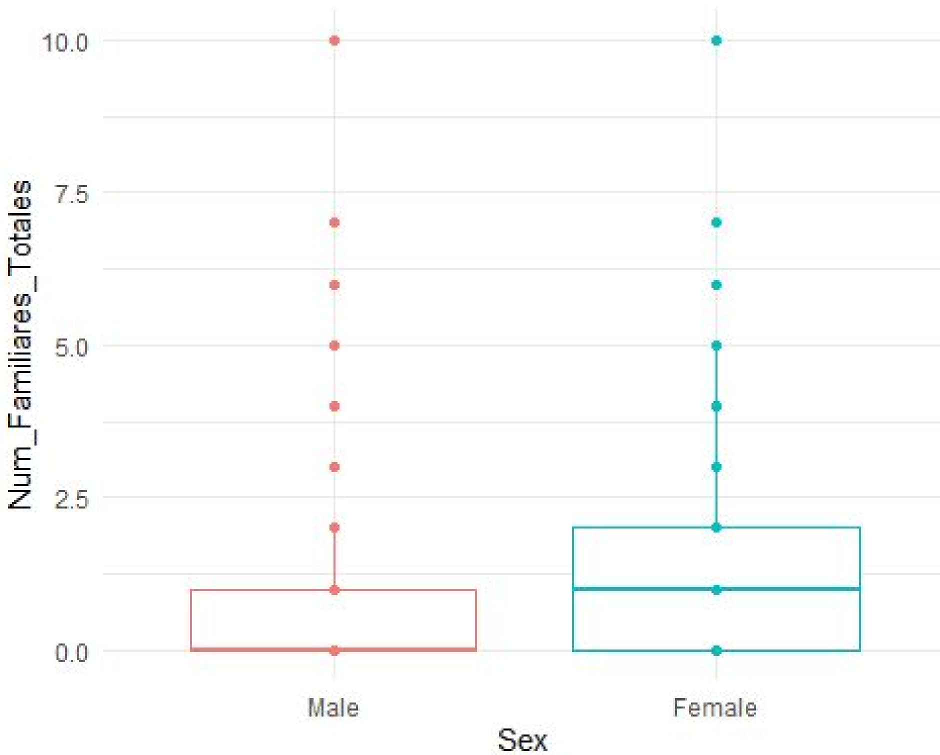
<u>Comprobación homogeneidad de la varianza:</u>

Si se tiene seguridad de que las muestras a comparar proceden de poblaciones que siguen una distribución normal, son recomendables el F-test y el test de Bartlet, pareciendo ser el segundo más recomendable ya que el primero es muy potente pero extremadamente sensible a desviaciones de la normal. Si no se tiene la seguridad de que las poblaciones de origen son normales, se recomienda el test de Leven utilizando la mediana o el test no paramétrico Fligner-Killeen que también se basa en la mediana.

Estudio gráfico de la varianza de las variables:


80
60
40
20
0
Age_lm
Male
Female
Sex


80
60
40
20
0
Age_lm
First
Second
Third
Pclass



Tablas varianza:

```
##      Sex   Age_lm
## 1   Male 170.9042
## 2 Female 195.7473

##   Pclass   Age_lm
## 1  First 187.5309
## 2 Second 176.1224
## 3  Third 114.4334

##      Sex Num_Familiares_Totales
## 1   Male               2.124285
## 2 Female               2.957077
```

La hipótesis nula y alternativa del test de Levene son:

-H0: las varianzas entre los grupos son iguales.

-H1: al menos la varianza de un grupo es distinta.

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    1  11.181 0.0008498 ***
##       1307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    2  16.664 7.139e-08 ***
##       1306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    1  31.154 2.895e-08 ***
##       1307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los p-valores obtenidos son menos de 0.05, rechazamos la HO (varianzas entre grupos son iguales) en los tres casos.

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.

1. Algoritmo de clasificación predecir que pasajeros sobrevivieron al hundimiento del Titanic segun sus caracteristicas: Decision tree

Utilizamos el dataset train.csv que es el que tiene la variable `Survived`.

Tabla supervivientes en el dataset train:

```
## 
##         0         1 
## 0.6161616 0.3838384
```

Partición dataset train en entrenamiento y prueba, proporciones de supervivientes en los datasets obtenidos:

```r
library(dplyr)
set.seed(123)
entrenamiento <- sample_frac(train_clean, .7)
prueba <- setdiff(train_clean, entrenamiento)
prop.table(table(entrenamiento$Survived))
```

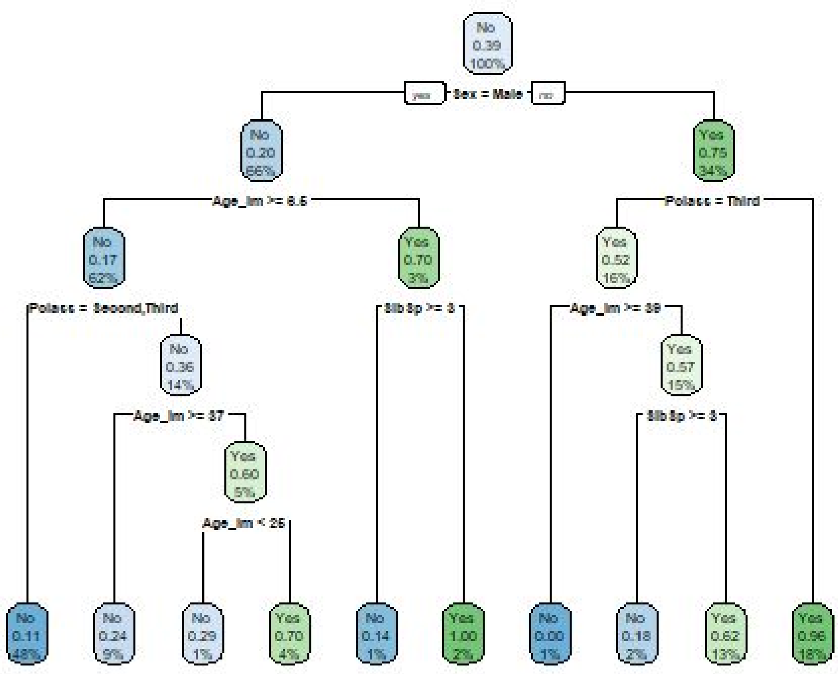
```
##
##         No       Yes
## 0.6121795 0.3878205
```

```r
prop.table(table(prueba$Survived))
```

```
##
##         No       Yes
## 0.6254682 0.3745318
```

Modelo 1 Árbol de decisión con las variables Age_lm, Pclass, Sex, SibSp y Parch:

```r
library(rpart)
library(rpart.plot)
library(caret)
library(e1071)
fit1 <- rpart(Survived~ Age_lm + Pclass + Sex + SibSp + Parch , data =
entrenamiento, method = 'class')
rpart.plot(fit1)
```

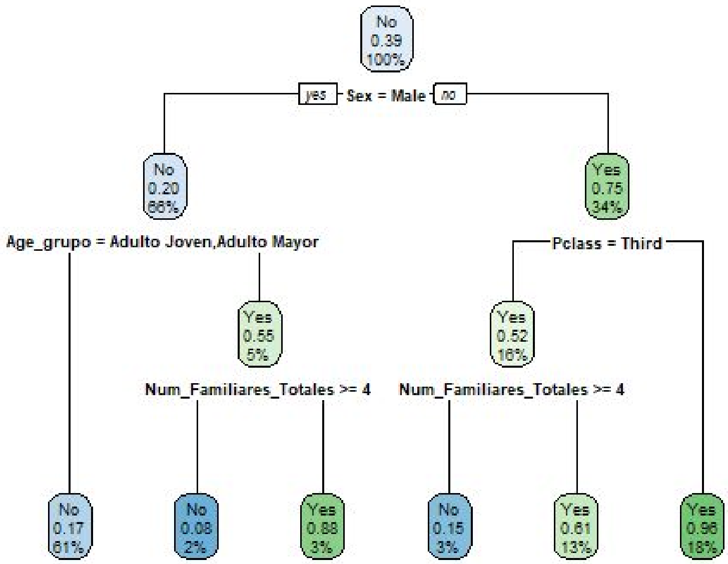


```r
prediccion_1 <- predict(fit1, newdata = prueba, type = "class")
confusionMatrix(prediccion_1, prueba$Survived)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction  No Yes
##        No  143  25
##        Yes  24  75
## 
##                Accuracy : 0.8165
##                  95% CI : (0.7647, 0.861)
##     No Information Rate : 0.6255
##     P-Value [Acc > NIR] : 8.865e-12
## 
##                   Kappa : 0.6075
## 
##  Mcnemar's Test P-Value : 1
## 
##             Sensitivity : 0.8563
##             Specificity : 0.7500
##          Pos Pred Value : 0.8512
##          Neg Pred Value : 0.7576
##              Prevalence : 0.6255
##          Detection Rate : 0.5356
##    Detection Prevalence : 0.6292
##       Balanced Accuracy : 0.8031
## 
##        'Positive' Class : No
## 
```

Modelo 2 Árbol de decisión con las variables Age grupo, Pclass, Sex y Num Familiares Totales:

```r
fit2 <- rpart(Survived~ Age_grupo + Pclass + Sex + Num_Familiares_Totales ,
data = entrenamiento, method = 'class')
rpart.plot(fit2)
```




```
prediccion_2 <- predict(fit2, newdata = prueba, type = "class")
confusionMatrix(prediccion_2, prueba$Survived)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  146  22
##        Yes  21  78
##
##                Accuracy : 0.839
##                  95% CI : (0.7893, 0.8809)
##     No Information Rate : 0.6255
##     P-Value [Acc > NIR] : 1.512e-14
##
##                   Kappa : 0.6556
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8743
##             Specificity : 0.7800
##          Pos Pred Value : 0.8690
##          Neg Pred Value : 0.7879
##              Prevalence : 0.6255
##          Detection Rate : 0.5468
##    Detection Prevalence : 0.6292
##       Balanced Accuracy : 0.8271
##
##        'Positive' Class : No
##
```

El segundo modelo presenta una precisión mayor, y además, es más simple esquemáticamente.

Vemos que en el segundo modelo las mujeres sobreviven un 75%, si no viajan en tercera clase sobreviven el 96%, si viajan en tercera clase sólo sobreviven el 52%. De las mujeres que viajan en tercera clase sobreviven más las que tienen un número de familiares totales menor a 4.

Los hombres tienen una probabilidad del 20% de sobrevivir, y si son hombres adultos la probabilidad es del 17%, mientras que si son niños la probabilidad es del 55%. Y de los niños, los que tienen más probabilidad de sobrevivir son los que tienen un número total de familiares menor a 4.

<u>2. Mediante un modelo de regresión logística predecir que probabilidad hay de que un pasajero sobreviva en base a sus características.</u>

<u>Modelo 1 de regresión logística con las variables Age_lm, Pclass, Sex, SibSp y Parch:</u>

```
modelo_logistico1 <- glm(Survived ~ Age_lm + Pclass + Sex + SibSp + Parch , data = entrenamiento, family = "binomial")
modelo_logistico1

##
## Call:  glm(formula = Survived ~ Age_lm + Pclass + Sex + SibSp + Parch,
##     family = "binomial", data = entrenamiento)
##
## Coefficients:
##  (Intercept)       Age_lm  PclassSecond   PclassThird     SexFemale
##      2.03669     -0.05619      -1.42623      -2.63043       2.81111
##        SibSp        Parch
##     -0.39799     -0.09766
##
## Degrees of Freedom: 623 Total (i.e. Null);  617 Residual
## Null Deviance:       833.4
## Residual Deviance: 544.8     AIC: 558.8

prediccion_3 <- predict(modelo_logistico1, newdata = prueba, type = "response")
prediccion_3[prediccion_3 > 0.5] <- "Yes"
prediccion_3[prediccion_3 <= 0.5] <- "No"
prediccion_3 <- factor(prediccion_3)
```

```r
confusionMatrix(prediccion_3, prueba$Survived)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  139  28
##        Yes  28  72
##
##                Accuracy : 0.7903
##                  95% CI : (0.7365, 0.8375)
##     No Information Rate : 0.6255
##     P-Value [Acc > NIR] : 4.958e-09
##
##                   Kappa : 0.5523
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8323
##             Specificity : 0.7200
##          Pos Pred Value : 0.8323
##          Neg Pred Value : 0.7200
##              Prevalence : 0.6255
##          Detection Rate : 0.5206
##    Detection Prevalence : 0.6255
##       Balanced Accuracy : 0.7762
##
##        'Positive' Class : No
##
```

<u>Modelo 2 de regresión logística con las variables Age grupo, Pclass, Sex y Num Familiares Totales:</u>

```r
modelo_logistico2 <- glm(Survived ~ Age_grupo + Pclass + Sex +
Num_Familiares_Totales , data = entrenamiento, family = "binomial")
modelo_logistico2
```

```
##
## Call: glm(formula = Survived ~ Age_grupo + Pclass + Sex +
Num_Familiares_Totales,
##     family = "binomial", data = entrenamiento)
##
## Coefficients:
##            (Intercept)   Age_grupoAdulto Mayor           Age_grupoNino
##                 0.2997                 -0.9439                  1.8932
##           PclassSecond             PclassThird               SexFemale
##                -1.3086                 -2.4277                  2.9033
## Num_Familiares_Totales
```

```
##                -0.3413
##
## Degrees of Freedom: 623 Total (i.e. Null);  617 Residual
## Null Deviance:       833.4
## Residual Deviance: 547.2     AIC: 561.2
```

```
prediccion_4 <- predict(modelo_logistico2, newdata = prueba, type =
"response")
prediccion_4[prediccion_4 > 0.5] <- "Yes"
prediccion_4[prediccion_4 <= 0.5] <- "No"
prediccion_4 <- factor(prediccion_4)

confusionMatrix(prediccion_4, prueba$Survived)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  142  27
##        Yes  25  73
##
##                Accuracy : 0.8052
##                  95% CI : (0.7526, 0.851)
##     No Information Rate : 0.6255
##     P-Value [Acc > NIR] : 1.534e-10
##
##                   Kappa : 0.5826
##
##  Mcnemar's Test P-Value : 0.8897
##
##             Sensitivity : 0.8503
##             Specificity : 0.7300
##          Pos Pred Value : 0.8402
##          Neg Pred Value : 0.7449
##              Prevalence : 0.6255
##          Detection Rate : 0.5318
##    Detection Prevalence : 0.6330
##       Balanced Accuracy : 0.7901
##
##        'Positive' Class : No
##
```

Nos quedamos con el segundo modelo de regresión logística, ya que presenta mayor precisión. Por los coeficientes de este modelo de regresión logística podemos ver que la variable mujeres y pertenecer al grupo de edad Niño aumenta la probabilidad de sobrevivir al hundimiento del Titanic. Mientras que pertenecer a tercera y segunda clase disminuye la probabilidad de supervivencia, asi como pertenecer al grupo Adulto Mayor, y en menor proporción el número de familiares totales.

3. Diferencia de la supervivencia entre las distintas clases del Titanic.

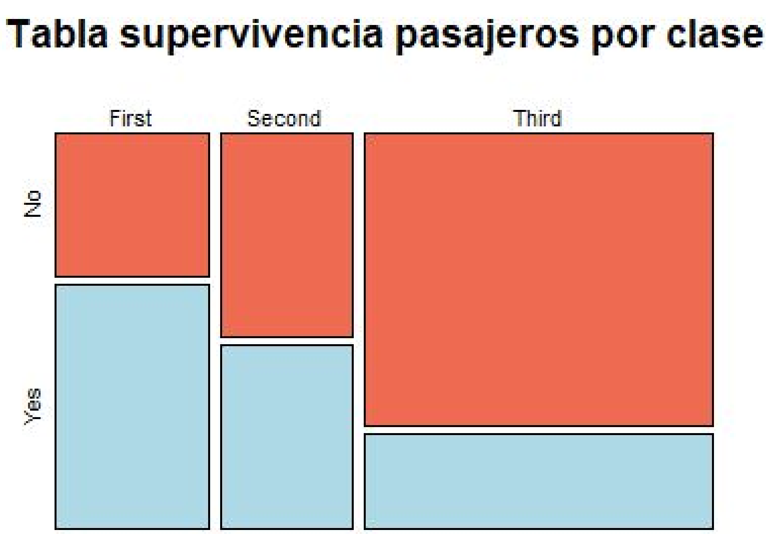
La prueba de chi cuadrado examina si las filas y columnas de una tabla de contingencia están asociadas de manera estadísticamente significativa.

Las hipótesis del test son:

- H0: las variables de fila y columna de la tabla de contingencia son independientes.

-H1: las variables de fila y columna son dependientes.

```
table1 <- table(train_clean$Pclass, train_clean$Survived)
table1

##
##           No Yes
##   First   80 136
##   Second  97  87
##   Third  372 119

plot(table1, color = c('coral2', 'lightblue'), main = 'Tabla supervivencia
pasajeros por clase')
```



```
chisq <- chisq.test(table1)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

El p-valor obtenido es menor que 0.05, por tanto, rechazamos H0 y concluimos que hay relación entre la supervivencia y la clase. Como vemos en el gráfico, sobrevivieron más porcentaje de primera clase y menos de tercera clase.

4. Diferencias de la mediana de edad entre hombres y mujeres.

Como hemos visto en el punto anterior los datos no siguen una distribución normal por tanto, usamos el test de Mann-Whitney-Wilcoxon para comparar las dos poblaciones.

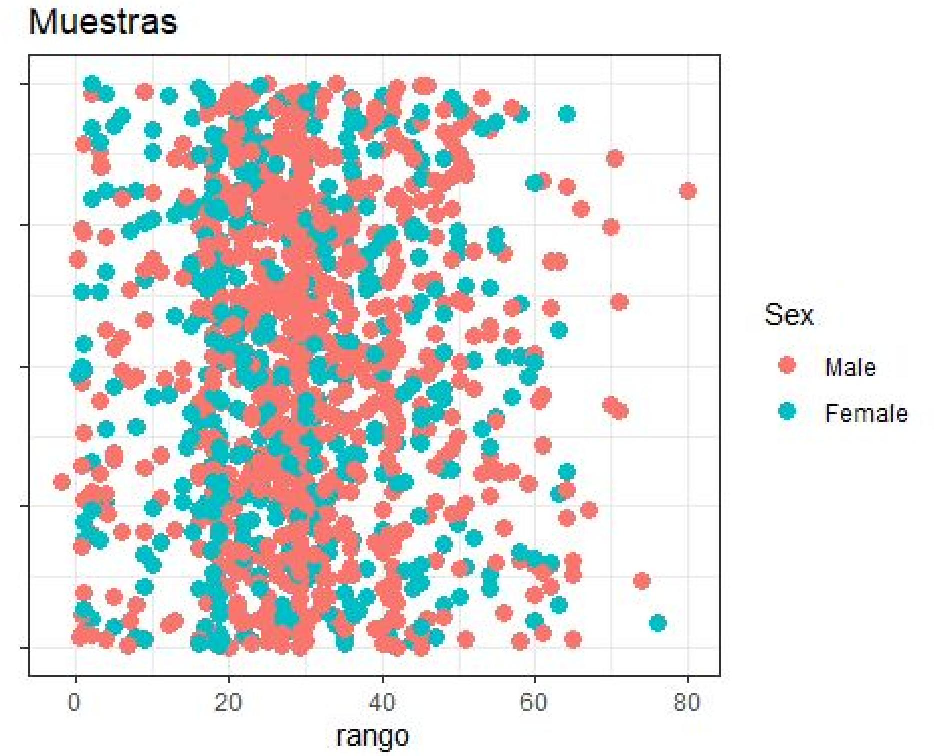
El test de Mann–Whitney–Wilcoxon (WMW), también conocido como Wilcoxon rank-sum test o u-test, es un test no paramétrico que contrasta si dos muestras proceden de poblaciones equidistribuidas.

La idea en la que se fundamenta este test es la siguiente: si las dos muestras comparadas proceden de la misma población, al juntar todas las observaciones y ordenarlas de menor a mayor, cabría esperar que las observaciones de una y otra muestra estuviesen intercaladas aleatoriamente. Por lo contrario, si una de las muestras pertenece a una población con valores mayores o menores que la otra población, al ordenar las observaciones, estas tenderán a agruparse de modo que las de una muestra queden por encima de las de la otra.

-H0: los miembros de un grupo no tienen mayor probabilidad a estar por encima de los del otro grupo (medianas ambos grupos son iguales).

-H1: los miembros de un grupo tienen mayor probabilidad a estar por encima de los del otro grupo (medianas de ambos grupos son distintas).

```
##
##  Wilcoxon rank sum test with continuity correction
##
##   data:   datos_clean$Age_lm[datos_clean$Sex  ==  "Male"]  and
datos_clean$Age_lm[datos_clean$Sex == "Female"]
## W = 222445, p-value = 7.004e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  1.353664 4.312960
## sample estimates:
## difference in location
##               2.999973
```

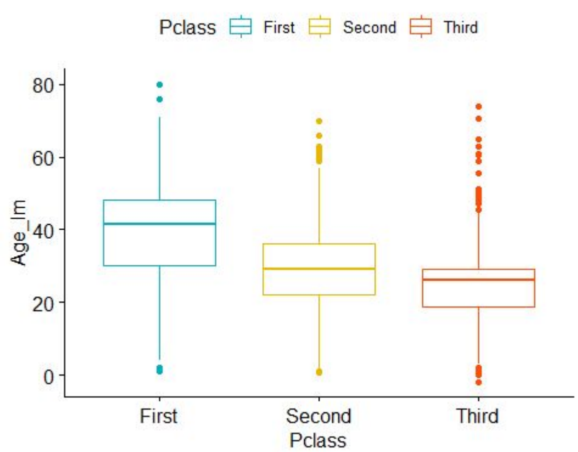


Rechazamos la HO (los miembros de un grupo no tienen mayor probabilidad a estar por encima de los del otro grupo), por tanto, podemos concluir que hay una diferencia significativa entre las medianas de edad de los dos grupos (hombres y mujeres).

<u>5. Diferencia de las medias de edad entre las distintas clases de pasajeros del Titanic.</u>

Chequear las condiciones para realizar un test ANOVA:

Las muestras deben tener una distribución aproximadamente normal, la variabilidad de todas las muestras debe ser similar y los tamaños de las muestras no deben ser muy dispares.



```
## 
##  Shapiro-Wilk normality test
## 
## data:  datos_clean$Age_lm[datos_clean$Pclass == "First"]
## W = 0.99519, p-value = 0.4153
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  datos_clean$Age_lm[datos_clean$Pclass == "Second"]
## W = 0.97396, p-value = 6.115e-05
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  datos_clean$Age_lm[datos_clean$Pclass == "Third"]
## W = 0.96023, p-value = 6.13e-13
```

Como las muestras no son normales, realizamos el test no paramétrico Kruskal-Wallis. Se usa para probar si un grupo de datos proviene de la misma población. Intuitivamente, es idéntico al ANOVA con los datos reemplazados por categorías. Es una extensión de la prueba de la U de Mann-Whitney para 3 o más grupos.

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  Age_lm by Pclass
## Kruskal-Wallis chi-squared = 243.38, df = 2, p-value < 2.2e-16
```

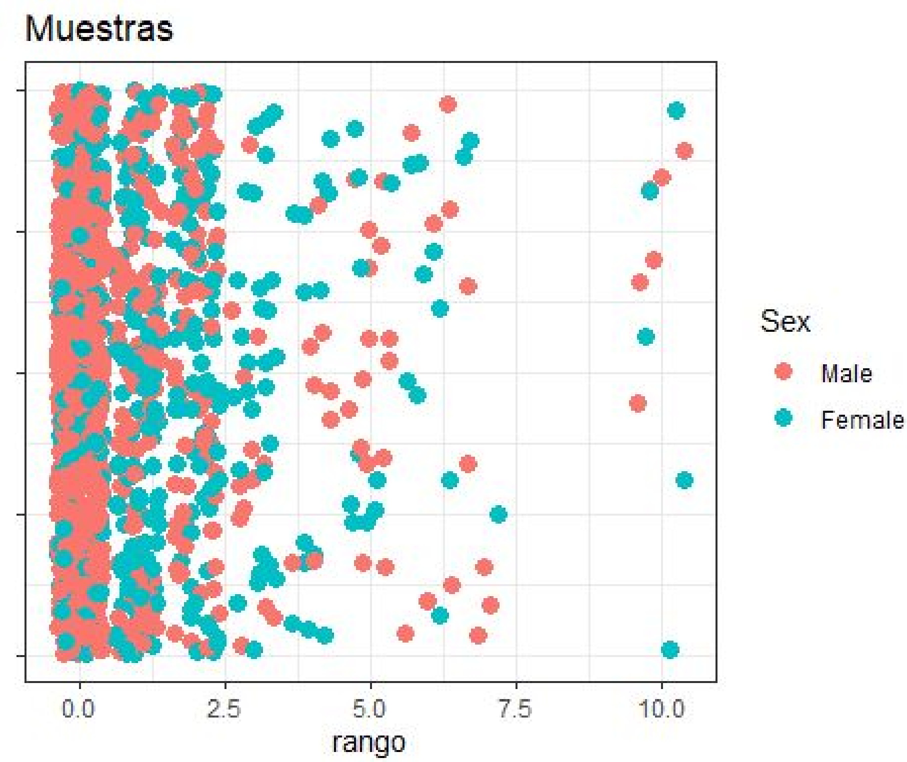
Como el p-valor es menor a 0.05, rechazamos H0, las medianas de edad entre las clases no son iguales.

6. Diferencia de la media de familiares entre hombres y mujeres en el Titanic.

Las muestras no siguen una distribución normal, por tanto, usamos el test de Mann-Whitney-Wilcoxon para comparar las dos poblaciones.

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  datos_clean$Num_Familiares_Totales[datos_clean$Sex == "Male"] and
datos_clean$Num_Familiares_Totales[datos_clean$Sex == "Female"]
## W = 139064, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
## 95 percent confidence interval:
##  -9.999724e-01 -4.320688e-05
## sample estimates:
## difference in location
##          -6.523438e-05
```



El p-valor obtenido es menor a 0.05, rechazamos la HO (los miembros de un grupo no tienen mayor probabilidad a estar por encima de los del otro grupo), por tanto, podemos concluir que hay una diferencia significativa entre las medianas del número de familiares totales entre los grupos de hombres y mujeres.

Gráficamente vemos que hay muchos más hombres que mujeres que viajan solos.

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

Resultados algoritmo árboles de decisión y modelo de regresión logística para predecir la supervivencia de los pasajeros:

Decision Tree

Modelo_1 : 0.8165

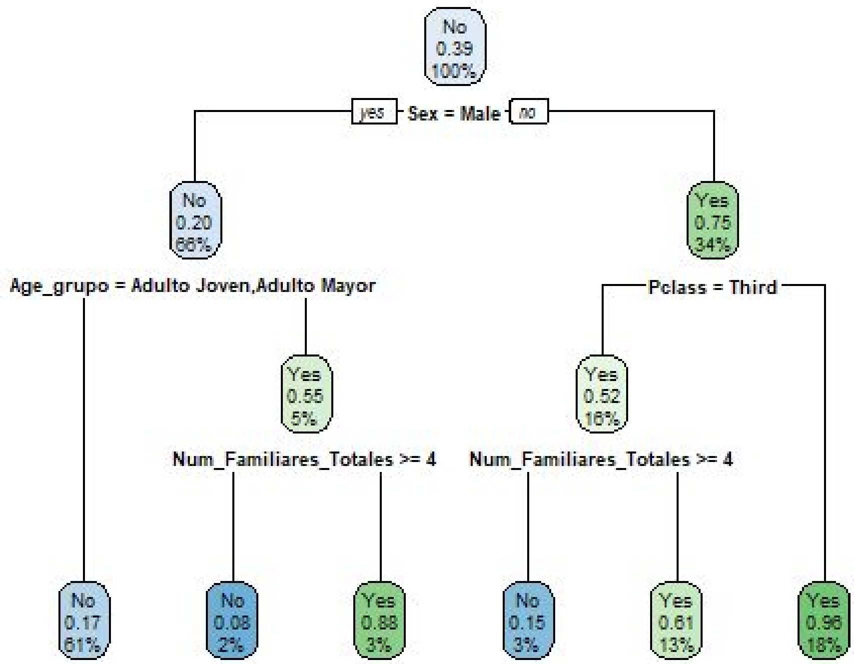
Modelo_2 : 0.834

Modelo regresión logística

Modelo_1 : 0.7903

Modelo_2 : 0.8052

El sistema de clasificación con un mayor nivel de precisión es el segundo modelo de árboles de decisión:

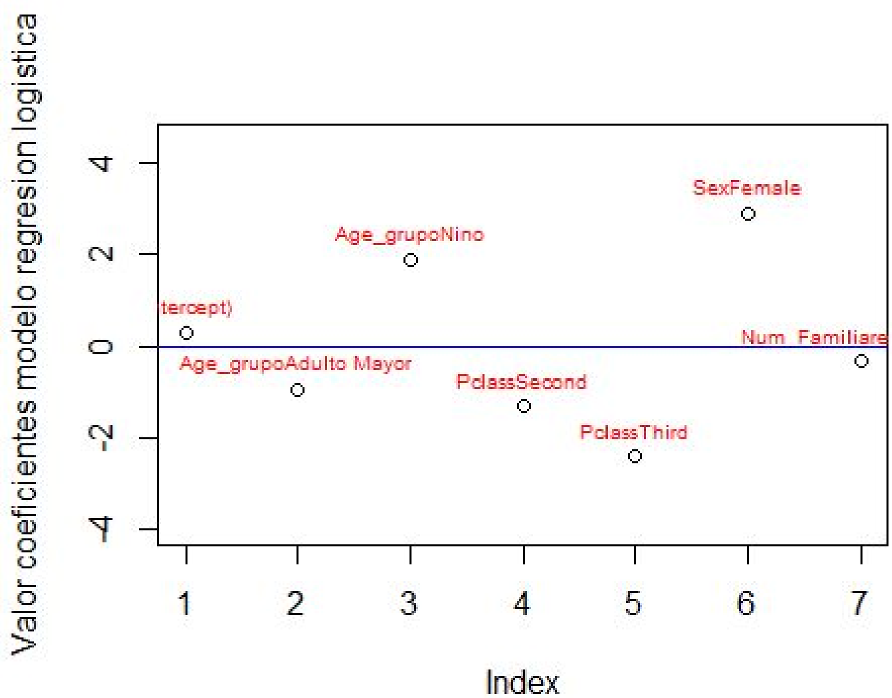


Las mujeres sobreviven un 75%, si no viajan en tercera clase sobreviven el 96%, si viajan en tercera clase sólo sobreviven el 52%. De las mujeres que viajan en tercera clase sobreviven más las que tienen un número de familiares totales menor a 4.

Los hombres tienen una probabilidad del 20% de sobrevivir, y si son hombres adultos la probabilidad es del 17%, mientras que si son niños la probabilidad es del 55%. Y de los niños, los que tienen más probabilidad de sobrevivir son los que tienen un número total de familiares menor a 4.

En cuanto a los modelos de regresión logística el segundo modelo también presenta mayor precisión.

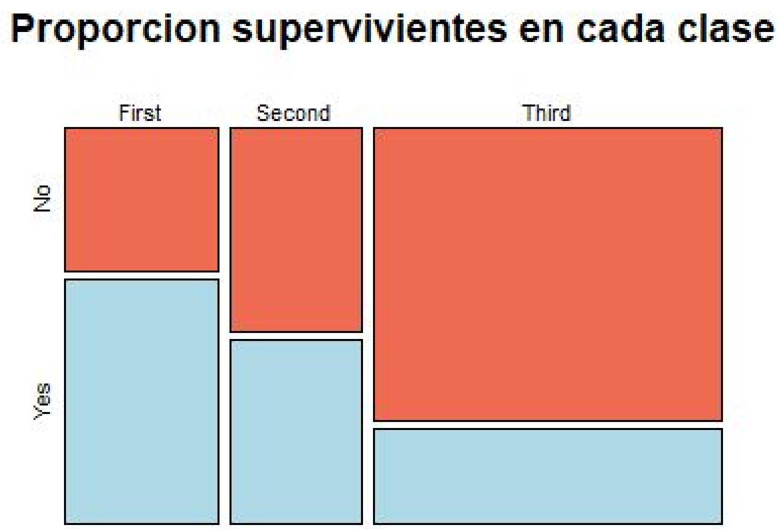
Por los coeficientes de este modelo de regresión logística podemos ver que la variable mujeres y pertenecer al grupo de edad Nino aumenta la probabilidad de sobrevivir al hundimiento del Titanic. Mientras que pertenecer a tercera y segunda clase disminuye la probabilidad de supervivencia, asi como pertenecer al grupo Adulto Mayor, y en menor proporción el número de familiares totales.



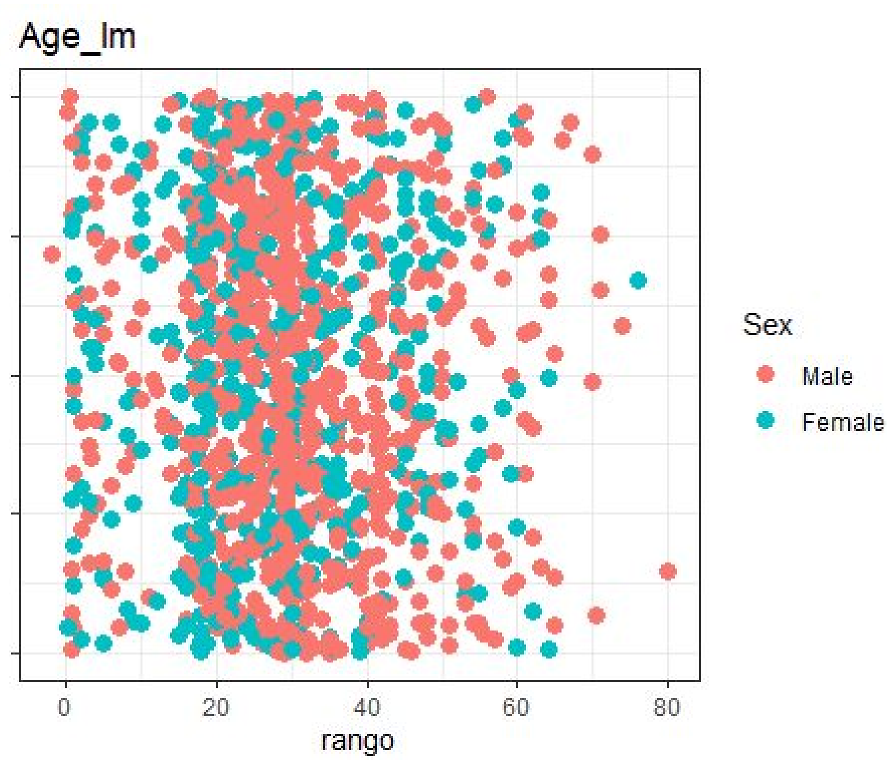
Los tests que hemos realizado podemos concluir que hay relación entre la clase y la supervivencia en el hundimiento, como muestra la tabla de valores absolutos, la tabla de proporciones y la gráfica los pasajeros de primera clase son los que sobrevivieron en mayor proporción.

```
##         
##           No Yes
##   First   80 136
##   Second  97  87
##   Third  372 119
```

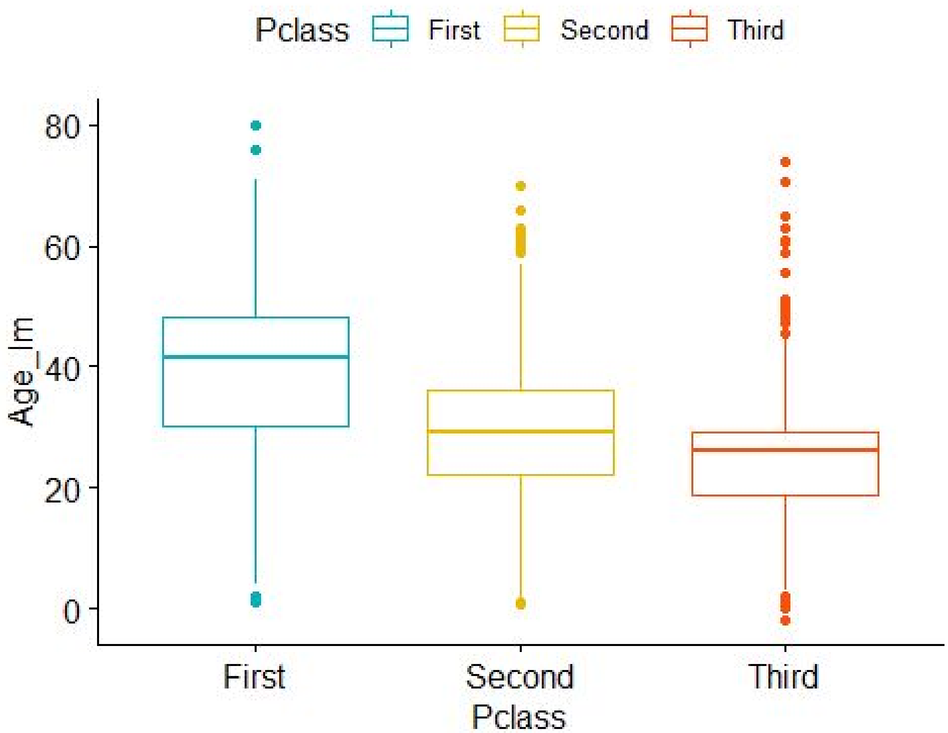
```
##         
##                  No        Yes
##   First  0.08978676 0.15263749
##   Second 0.10886644 0.09764310
##   Third  0.41750842 0.13355780
```



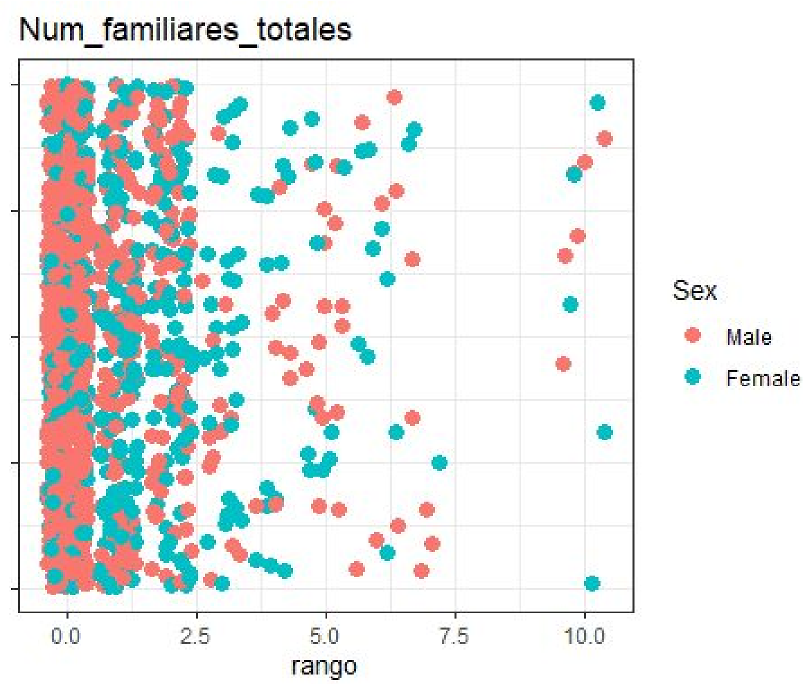
También podemos concluir que hay una diferencia significativa entre las medianas de edad entre hombres y mujeres, la mediana de los hombres es mayor que la de las mujeres.



Las medianas de edad entre las clases no son iguales, las clases más altas presentan edades más avanzadas.



Hay diferencias significativas entre las medianas del número de familiares totales entre hombres y mujeres, hay más hombres que viajan solos.



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

Tanto el modelo de clasificación como el modelo de regresión nos permiten predecir si un pasajero sobrevivió según sus características con una precisión alrededor del 80%, con lo que nos permite responder de manera satisfactoria al problema principal del estudio.

Del algoritmo de clasificación Decision Tree podemos ver que: las mujeres sobreviven un 75%, si no viajan en tercera clase sobreviven el 96%, si viajan en tercera clase sólo sobreviven el 52%. De las mujeres que viajan en tercera clase sobreviven más las que tienen un número de familiares totales menor a 4.

Los hombres tienen una probabilidad del 20% de sobrevivir, y si son hombres adultos la probabilidad es del 17%, mientras que si son niños la probabilidad es del 55%. Y de los niños, los que tienen más probabilidad de sobrevivir son los que tienen un número total de familiares menor a 4.

De los coeficientes del modelo de regresión logística podemos ver que la variable mujeres y pertenecer al grupo de edad Nino aumenta la probabilidad de sobrevivir al hundimiento del Titanic. Mientras que pertenecer a tercera y segunda clase disminuye la probabilidad de supervivencia, asi como pertenecer al grupo Adulto Mayor, y en menor proporción el número de familiares totales.

Por tanto, *podemos concluir que las variables Sexo y Edad (Nino), son en los dos modelos predictivos (arbol de decision y modelo regresion logistica) las variables mas importantes, seguida de la variable clase para predecir la supervivencia de los pasajeros.*

Por otra parte, por los tests que hemos realizado podemos concluir que hay una relación entre la variable clase y la supervivencia. Y las medianas de edad entre las clases no son iguales, las clases más altas presentan edades más avanzadas.

También podemos concluir que hay una diferencia significativa entre las medianas de edad entre hombres y mujeres, la mediana de los hombres es mayor que la de las mujeres.

Y hemos visto que hay diferencias significativas entre las medianas del número de familiares totales entre hombres y mujeres, hay más hombres que viajan solos.