

Práctica 2

github de la práctica

<https://github.com/mishuvale91/titanic-dataset>

Alumnos:

Michaelle Estefanía Valenzuela Sangoquiza

Juan Manuel Penalta Rodríguez.

Índice

M2.851 - Tipología y ciclo de vida de los datos.....	1
1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?.....	3
2. Integración y selección de los datos de interés a analizar.....	4
3. Limpieza de los datos.....	4
3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?.....	5
3.2. Identificación y tratamiento de valores extremos.....	5
4. Análisis de los datos.....	6
4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).....	6
4.2. Comprobación de la normalidad y homogeneidad de la varianza.....	8
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.....	9
5. Representación de los resultados a partir de tablas y gráficas.....	12
Gráfica 5.1 Boxplot de los campos Age y Fare.....	12
Gráfica 5.2 Histograma de las edades de 5 en 5 años.....	12
Gráfica 5.3 Boxplot del campo Fare filtrando por los tres valores de Pclass.....	12
Gráfica 5.4 Relación entre las variables “Sex” y “Survived”.....	13
Gráfica 5.5 Relación entre "Survived" como función de "Embarked".....	13
Gráfica 5.6 Relación entre “Survived” y “Family Size”.....	14
Gráfico 5.7 Relación entre "Survived" en función de "Age".....	15
Gráfico 5.8 Modelo 1.....	15
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?.....	16
7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.....	16

Este documento responde a las preguntas de la segunda práctica de la materia "Tipología y ciclo de vida de los datos" del Master en Ciencia de Datos de la UOC.

Todo el código, los datos y este documento se encuentran en el repositorio de github creado para la práctica: <https://github.com/mishuvale91/titanic-dataset>

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset está compuesto por datos de pasajeros del Titanic, hundido en el océano atlántico en el año 1912 tras chocar con un iceberg. En su momento fue una de las tragedias náuticas más importantes en tiempos de paz. Murieron 1496 de los 2208 pasajeros.

Los datos de este dataset fueron obtenidos originalmente de la página Kaggle (<https://www.kaggle.com/c/titanic/data>). Los datos con los que se realiza este trabajo se pueden consultar en el siguiente enlace:

[\[https://github.com/mishuvale91/titanic-dataset/tree/main/data/train.csv\]](https://github.com/mishuvale91/titanic-dataset/tree/main/data/train.csv)

En total el dataset dispone de **891 entradas**, con **12 campos** para cada entrada. Los campos disponibles se describen a continuación:

- **Survived:** Variable numérica que indica si el pasajero supervivió o murió en el hundimiento.
 - 1 = Superviviente
 - 0 = No superviviente
- **Pclass:** Variable numérica que indica el tipo de ticket que tenía el pasajero.
 - 1 = Primera Clase
 - 2 = Segunda Clase
 - 3 = Tercera Clase
- **Sex:** Variable alfanumérica que indica el género del pasajero: female o male
- **Age:** Variable numérica que indica la edad en años del pasajero.
- **Sibsp:** Variable numérica que indica el número de hermanos o esposos del pasajero que viajan con él.
- **Parch:** Variable numérica que indica el número de padres o hijos del pasajero que viajan con él.
- **Ticket:** Variable alfanumérica con el número de ticket del pasajero.
- **Fare:** Variable numérica que indica la tarifa pagada por el pasajero.

- **Cabin:** Variable alfanumérica que indica la cabina o cabinas que ocupaba el pasajero y sus parientes.
- **Embarked:** Variable alfanumérica que indica el puerto de embarque del pasajero.
 - C = Cherbourg
 - Q = Queenstown
 - S = Southampton

Con estos datos vamos a intentar analizar si existe alguna relación entre las variables disponibles y la supervivencia de los pasajeros. Principalmente intentaremos averiguar en esta práctica:

- Si se cumplió el protocolo de salvamento que rige el proceso de evacuación del barco de «mujeres y niños primero». Es decir, si existe una relación entre el sexo y la edad, y, y el sexo y la supervivencia de los pasajeros.
- Si existió una relación entre la clase social del pasaje y la supervivencia.
- Si existe una relación entre tener familiares embarcados y la supervivencia

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

Se procede a la lectura del conjunto de datos “train.csv”, a partir del nombre “titánica-data”, que contiene 891 registros con 12 columnas.

Decidimos excluir dos campos, “Name” y “Cabin”, que no nos parecen interesantes para los análisis que pensamos realizar. Como resultado nos quedan 10 campos prioritarios para realizar el análisis de predicción de supervivencia.

Se ha decidido crear una nueva variable llamada “FamilySize”, la cual es calculada a partir de la suma de las variables SibSp y Parch, lo que nos da una idea del tamaño de la familia que viaja junta.

```
titanic_data$FamilySize <- titanic_data$SibSp + titanic_data$Parch +1;
```

Con lo que nos quedan sólo 9 campos:

1. PassengerId
2. Survived
3. Pclass
4. Sex
5. Age
6. FamilySize
7. Ticket
8. Fare
9. Embarked

3. Limpieza de los datos.

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

Veamos que campos tienen columnas con valores no definidos:

```
colSums(is.na(titanic_data))
PassengerId    Survived    Pclass      Sex      Age      Ticket      Fare
            0            0            0            0      177            0            0
      Embarked  FamilySize
            0            0
```

Se tiene 177 registros vacíos en la variable “Age”, para proceder con la imputación de valores se utiliza la media de esta variable.

```
titanic_data$Age[is.na(titanic_data$Age)] <- mean(titanic_data$Age, na.rm=T)
```

Veamos ahora que campos tienen valores vacíos:

```
colSums(titanic_data=="")
PassengerId    Survived    Pclass      Sex      Age      Ticket      Fare
            0            0            0            0            0            0            0
      Embarked  FamilySize
            2            0
```

Se tiene 2 registros vacíos en la variable “Embarked”, que para saber qué valor imputar se realiza un análisis para examinar qué pasajero ha desaparecido, una vez que se ha identificado y se puede evidenciar que mencionados pasajeros están en clase 1 y han pagado la tarifa de \$80, se concluye que la tarifa mediana para el pasajero de primera clase que sale de C (Charbourg) Embarcado coincide muy bien con los \$80 pagados por los pasajeros cuyo Embarcado falta.

Entonces se procede a reemplazar con seguridad el NA con C.

```
titanic_data$Embarked[c(62, 830)] <- "C"
```

Miramos que variables se puede discretizar mirando cuantos posibles valores toman:

```
apply(titanic_data, 2, function(x) length(unique(x)))
```

Vemos que las variables *Survived*, *Pclass*, *Sex*, *Embarked*, son candidatas a ser variables discretas, lo que coincide con la definición de los campos que teníamos en el punto 1.

```
cols<-c("Survived", "Pclass", "Sex", "Embarked")
for (i in cols){
  titanic_data[,i] <- as.factor(titanic_data[,i])
}
```

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos o outliers son aquellos que parecen no ser congruentes si los comparamos con el resto de los datos. Para identificarlos se utilizar la función `boxplots.stats()` de R, la cual se emplea a continuación. Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
boxplot.stats(titanic_data$Survived)$out
# Levels: 0 1
boxplot.stats(titanic_data$Pclass)$out
# Levels: 1 2 3
boxplot.stats(titanic_data$Sex)$out
# Levels: female male
boxplot.stats(titanic_data$Embarked)$out
# Levels:  C  Q  S
```

Estas cuatro variables, *Survived*, *Pclass*, *Sex* y *Embarked* son variables discretas, ya lo vimos antes, podemos comprobar que ninguna tiene valores se salgan de los posibles según su definición dada al principio del documento, por lo que las damos por válidas.

Para las otras variables podemos ver una representación gráfica. Como se puede observar en la Gráfica 5.1, numéricamente los resultados son:

```
boxplot.stats(titanic_data$Age)$out
# 2 58 55 2 66 65 0 59 71 70 2 55 1 61 1 56 1 58 2 59 62 58 63 65 2 0 61 2 60 1
1 64 65 56 0 2 63 58
# 55 71 2 64 62 62 60 61 57 80 2 0 56 58 70 60 60 70 0 57 1 0 2 1 62 0 74 56
```

Con estos valores, para el campo Age, se puede observar que hay 66 pasajeros cuya edad excede los valores más comunes, es decir, los valores superiores a 64 o inferiores a 3, por lo que se puede deducir que hay personas mayores a bordo del barco. Pero ninguno parece que tenga un valor que no coincida con una edad valida. Se puede comprobar con la Gráfica 5.2, que las edades se reparten de una manera lógica y esos valores superiores a 64 no son raros.

```
boxplot.stats(titanic_data$Fare)$out
# 71.2833 263.0000 146.5208 82.1708 76.7292 80.0000 83.4750 73.5000 263.0000 77.2875
247.5208 73.5000
# 77.2875 79.2000 66.6000 69.5500 69.5500 146.5208 69.5500 113.2750 76.2917 90.0000
83.4750 90.0000
# 79.2000 86.5000 512.3292 79.6500 153.4625 135.6333 77.9583 78.8500 91.0792 151.5500
247.5208 151.5500
# 110.8833 108.9000 83.1583 262.3750 164.8667 134.5000 69.5500 135.6333 153.4625 133.6500
66.6000 134.5000
# 263.0000 75.2500 69.3000 135.6333 82.1708 211.5000 227.5250 73.5000 120.0000 113.2750
90.0000 120.0000
# 263.0000 81.8583 89.1042 91.0792 90.0000 78.2667 151.5500 86.5000 108.9000 93.5000
221.7792 106.4250
# 71.0000 106.4250 110.8833 227.5250 79.6500 110.8833 79.6500 79.2000 78.2667 153.4625
77.9583 69.3000
# 76.7292 73.5000 113.2750 133.6500 73.5000 512.3292 76.7292 211.3375 110.8833 227.5250
151.5500 227.5250
# 211.3375 512.3292 78.8500 262.3750 71.0000 86.5000 120.0000 77.9583 211.3375 79.2000
69.5500 120.0000
# 93.5000 80.0000 83.1583 69.5500 89.1042 164.8667 69.5500 83.1583
```

La variable Fare (precio del billete) tiene muchos valores que podrían ser posibles valores extremos. La primera idea es que podrían ser debido a la existencia de tres clases distintas de pasajeros, que lógicamente pagarían precios muy distintos por los billetes. Filtrando por los tres valores del campo Pclass vemos que la cantidad se reduce, pasando de 112 a 79 posibles valores extremos. Aun así mirando la Gráfica 5.3, llama la atención uno de los valores extremos que se produce con la clase 1.

```
boxplot.stats(filter(titanic_data,Pclass==1)$Fare)$out
```

[completar]

4. Análisis de los datos.

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

Relación entre las variables "Sex" y "Survived":

Una de las preguntas que nos hicimos al principio es si existe una relación entre el sexo y la supervivencia de los pasajeros. En la Gráfica 5.4, se puede observar fácilmente la cantidad de mujeres que viajaban respecto a los hombres y a su vez observar los que no sobrevivieron. Numéricamente el número de mujeres supervivientes es mayor que el de hombre, pero porcentualmente es mucho mayor.

Parece que tiene sentido preguntarnos si existe una relación entre el sexo y la supervivencia.

Relación entre "Survived" como función de "Embarked":

Nos preguntamos si podría existir una relación entre el puerto de embarque y la supervivencia. En la Gráfica 5.5, de forma porcentual, se observa los puertos de embarque y los porcentajes de supervivencia en función del puerto. Con el siguiente script obtenemos matriz de porcentaje de frecuencias:

```
t <-table(titanic_data1[1:filas,]$Embarked,titanic_data1[1:filas,]$Survived)
for (i in 1:dim(t)[1]){
  t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

Cuyo resultado es:

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

Hay una pequeña diferencia con el puerto C (Cherburgo), con respecto a los otros dos puertos, para explicar la diferencia en los datos se podría trabajar con estos y preguntarnos si.

- ¿Quizás porcentualmente embarcaron más mujeres o niños?
- O más gente de primera clase?

Relación entre "Survived" como función de "Family Size":

En la Gráfica 5.6, vemos la relación entre el tamaño de la familia, "Family Size" y la supervivencia. Mirando esa gráfica parece que existe alguna relación, las familias muy grandes parece que tienen una menor supervivencia ¿será cierto?

Relación entre "Survived" en función de "Age":

También nos preguntamos al principio si los niños se salvarían antes que los adultos. Mirando la Gráfica 5.7, parece que esto puede ser posible.

En esta gráfica se aprecia un valor "raro", se debe a nuestra decisión de completar las edades que faltaban con la edad media del resto de los tripulantes, lo que da lugar a ese pico que se ve para los 26 años.

Al final decidimos seleccionar para el análisis los campos:

- Sex
- Embarked,

- Age,
- FamilySize,
- Survived

```
# SELECCIÓN DE GRUPOS DE DATOS
titanic_analisis <- titanic_data1 %>%
  select(Sex,
         Embarked,
         Age,
         FamilySize,
         Survived)
```

Los datos limpios se guardan en el fichero [titanic-cleaning.csv](#) en la carpeta data del repositorio github de la práctica.

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

Vamos a ver ahora varios métodos basados en el análisis estadístico de los datos, para comprobar la normalidad y la homocedasticidad.

Para la comprobación de la normalidad, se utilizará la prueba de normalidad de Anderson- Darling. Se comprueba que para cada prueba se obtiene un p-valor superior al nivel de significación establecido de 0.05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

Utilizamos el siguiente script que recorre los 5 campos que tenemos y realiza el test para aquellos campos que sean de tipo numeric o integer, ya que para los otros campos no tiene sentido realizar este tipo de análisis.

```
alpha = 0.05
col.names = colnames(titanic_analisis)
for (i in 1:ncol(titanic_analisis)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(titanic_analisis[,i]) | is.numeric(titanic_analisis[,i])) {
    p_val = ad.test(titanic_analisis[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Formato de salida
      if (i < ncol(titanic_analisis) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

El resultado es que **ninguno** de los dos campos para los que se realiza el análisis, “Age” y “Family Size”, y “Fare” sigue una distribución normal.

Para estudiar la homogeneidad de varianzas se utilizará el Test Fligner-Killeen, ya que permite comparar las varianzas basándose en la mediana. Es también una alternativa cuando no se cumple la condición de normalidad en las muestras.

Empezamos con “Age” y “Survived”

```
fligner.test(Age ~ Survived, data = titanic_analisis)
```

Cuyo resultado es


```
# Fligner-Killeen test of homogeneity of variances
# data: Age by Survived
# Fligner-Killeen:med chi-squared = 5.4693, df = 1, p-value = 0.01935
```

Como se obtiene un p-valor inferior a 0.05, se rechaza la hipótesis de que las varianzas de ambas muestras son homogéneas.

Hacemos lo mismo para “FamilySize” y “Survived”

```
fligner.test(FamilySize ~ Survived, data = titanic_analisis)
# Fligner-Killeen test of homogeneity of variances
# data: FamilySize by Survived
# Fligner-Killeen:med chi-squared = 19.647, df = 1, p-value = 9.317e-06
```

Tiene un p-valor inferior a 0.05, por lo que se rechaza la hipótesis de que las varianzas de ambas muestras son homogéneas.

Y lo mismo con “Fare” y “Survived”

```
fligner.test(Fare ~ Survived, data = titanic_analisis)
# Fligner-Killeen test of homogeneity of variances
# data: Fare by Survived
# Fligner-Killeen:med chi-squared = 96.253, df = 1, p-value < 2.2e-16
```

Tiene un p-valor inferior a 0.05, por lo se rechaza la hipótesis de que las varianzas de ambas muestras son homogéneas.

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.

Empezamos calculando los coeficientes de correlación entre las variables cuantitativas “...y” y con respecto al campo “Survived”

```
calculo_numericas <- titanic_analisis %>%
  select(PassengerId,
         Age,
         Fare,
         FamilySize)

corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa con respecto al
campo "Survived"

for (i in 1:(ncol(calculo_numericas) - 1)) {
  if (is.integer(calculo_numericas[,i]) | is.numeric(calculo_numericas[,i])) {
    spearman_test = cor.test(calculo_numericas[,i],
                             calculo_numericas[,length(calculo_numericas)],
                             method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
```

```

# A?adir la fila de matriz
pair = matrix(ncol = 2, nrow = 1)
pair[1][1] = corr_coef
pair[2][1] = p_val
corr_matrix <- rbind(corr_matrix, pair)
rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(calculo_numericas)[i]
}
}

print (corr_matrix)

```

Que nos da como resultado:

```

# estimate      p-value
# PassengerId -0.05041556 1.326511e-01
# Age         -0.18421248 3.052168e-08
# Fare         0.52890733 2.269544e-65

```

Con la matriz de correlación se puede identificar cuáles son las variables más correlacionadas con la supervivencia en función de su proximidad con los valores -1 y +1.

Vamos ahora a aplicar un modelo de regresión logística. Se utiliza el modelo de regresión logística con el conjunto de datos del Titanic para predecir si cada uno de los pasajeros sobrevivirá o no.

```

modelo <- glm(Survived~., family=binomial(link='logit'), data=titanic_analisis)
summary(modelo)

```

Ejecutamos el modelo

```

fitted.proBABILITIES <- predict(log.model, titanic_data1, type='response')

```

Comprobamos su eficacia

```

library(caTools)
sample <- sample.split(titanic_analisis, 0.7)
train_final <- subset(titanic_analisis, sample=TRUE)
test_final <- subset(titanic_analisis, sample=FALSE)
final_model <- glm(Survived ~., family=binomial(link='logit'), train_final)
summary(final_model)

```

De los resultados miramos que

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.2470169  0.5127043   8.284 < 2e-16 ***
PassengerId  0.0001287  0.0003473   0.371  0.71095
Sexmale     -2.7387864  0.2007974 -13.640 < 2e-16 ***
EmbarkedQ   -0.0760647  0.3796199  -0.200  0.84119
EmbarkedS   -0.4646937  0.2390633  -1.944  0.05192 .
Age         -0.0383377  0.0078438  -4.888 1.02e-06 ***
Pclass2     -0.8988690  0.2969430  -3.027  0.00247 **
Pclass3     -2.1298144  0.2974504  -7.160 8.05e-13 ***
Fare         0.0024862  0.0024848   1.001  0.31703
FamilySize  -0.2206660  0.0683752  -3.227  0.00125 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 1186.7 on 890 degrees of freedom  
Residual deviance: 785.8 on 881 degrees of freedom  
AIC: 805.8
```

```
Number of Fisher Scoring iterations: 5
```

Vemos que igual que en el modelo inicial se puede observar que tanto pertenecer a la clase 2 como a la clase 3 parece estar relacionado con la variable survived, así como ser hombre.

Modelo para testear

```
fitted.proBABilities.final <- predict(final_model, test_final, type='response')  
fitted.results.final <- ifelse(fitted.proBABilities.final>0.5, 1,0)
```

Eficacia del modelo

```
1-mean(fitted.results.final != test_final$Survived)  
# 0.8013468
```

El modelo acertó en un 80% de los datos

Aplicamos un modelo de predicción, RandomForest

```
library(randomForest)  
train <- titanic_analisis[1:474,]  
test <- titanic_analisis[475:891,]  
set.seed(754)  
modelo1 <- randomForest(factor(Survived) ~ Pclass + Sex + Age + FamilySize + Fare +  
Embarked, data = train)
```

Error del modelo

```
plot(modelo1, ylim = c(0, 0.36))  
legend("topright", colnames(modelo1$err.rate), col = 1:3, fill = 1:3)
```

En el gráfico correspondiente, Gráfico 5.8, se puede observar que la línea negra indica la tasa de error global que cae por debajo del 20%. Las líneas roja y verde muestran la tasa de error para el que "murió" y "sobrevivió" respectivamente.

Predicción del modelo

```
prediccion <- predict(modelo1, test)  
solucion <- data.frame(PassengerID = test$PassengerID, Survived = prediccion)  
table(solucion$Survived)
```

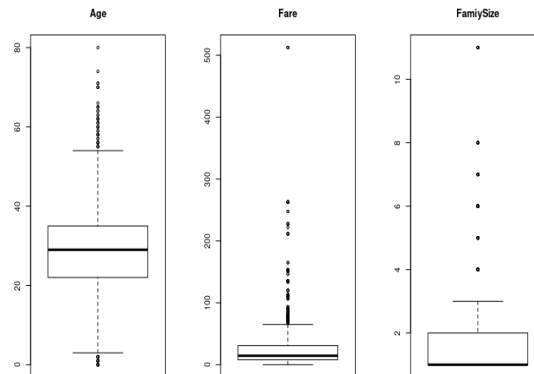
```
# 0 1  
# 296 121
```

Se puede concluir que al analizar el conjunto de datos Titanic en el que se predice que 296 murieron de 417 pasajeros en el conjunto de prueba.

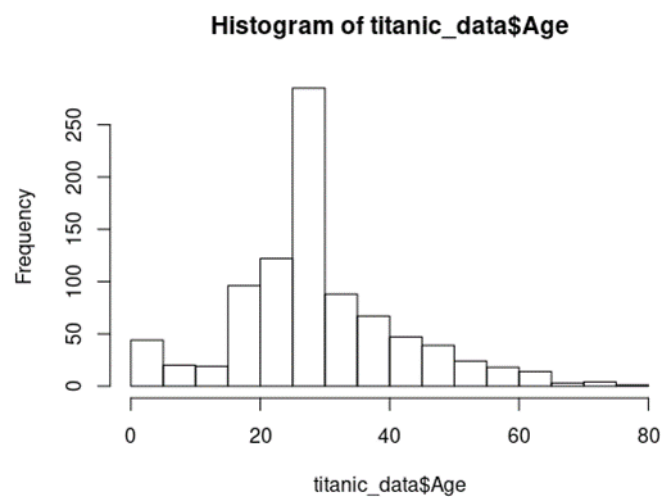
Entonces, el total de pasajeros que murieron en el Titanic son 632 de 891 pasajeros.

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

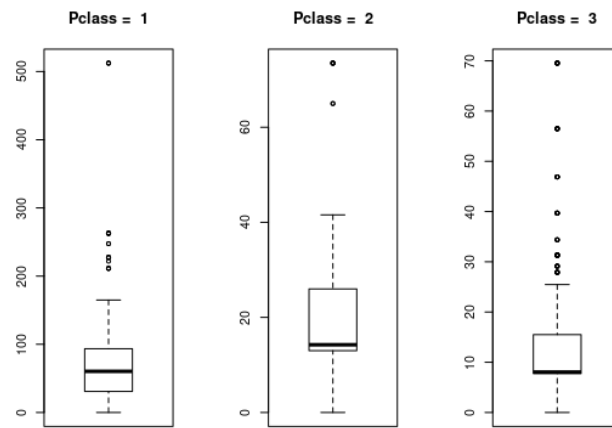
Gráfica 5.1 Boxplot de los campos Age y Fare



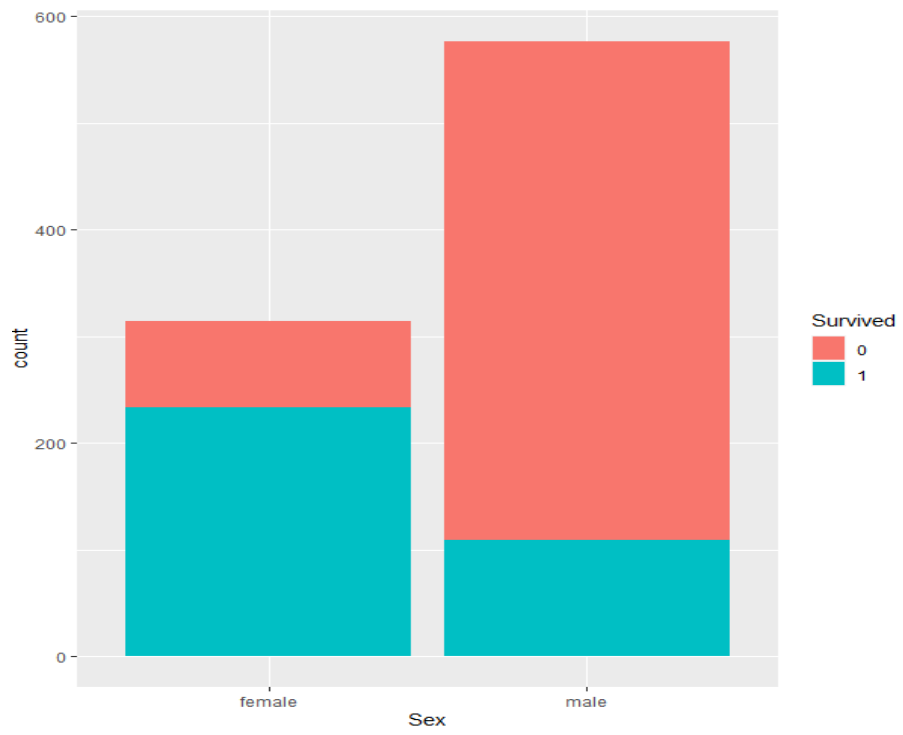
Gráfica 5.2 Histograma de las edades de 5 en 5 años.



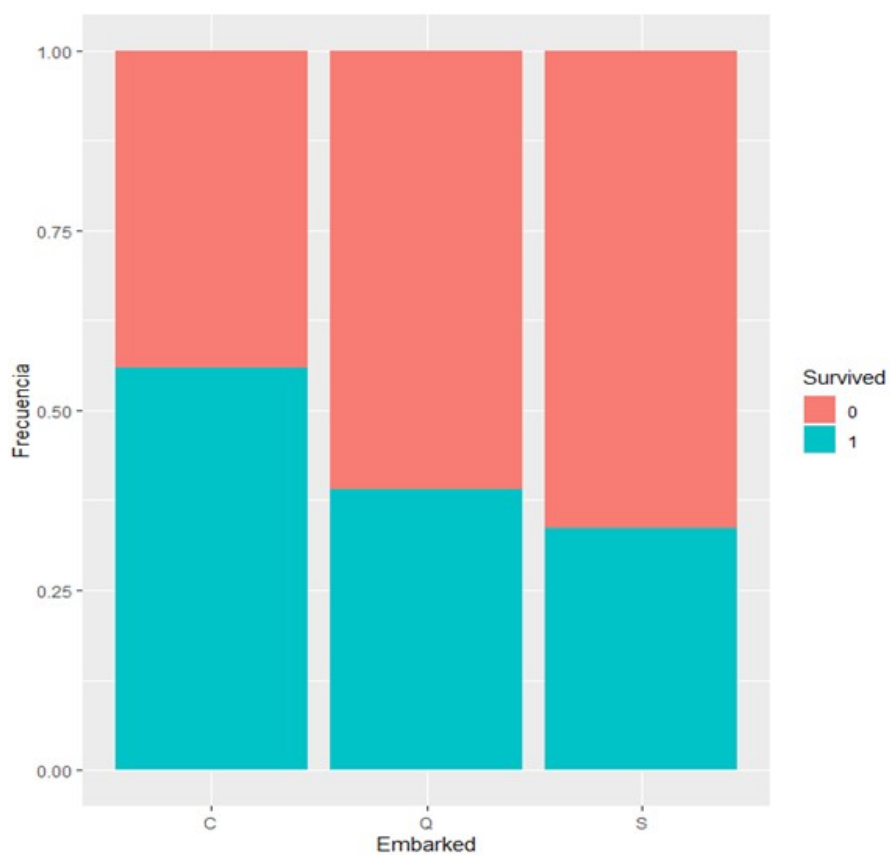
Gráfica 5.3 Boxplot del campo Fare filtrando por los tres valores de Pclass



Gráfica 5.4 Relación entre las variables “Sex” y “Survived”



Gráfica 5.5 Relación entre "Survived" como función de "Embarked"



Gráfica 5.6 Relación entre “Survived” y “Family Size”

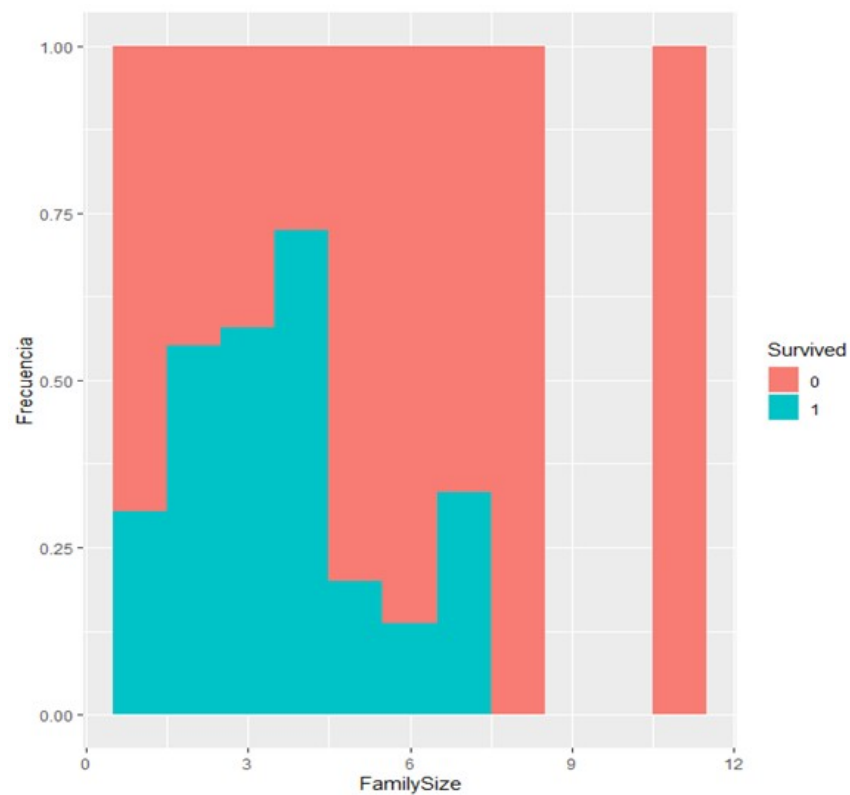


Gráfico 5.7 Relación entre "Survived" en función de "Age"

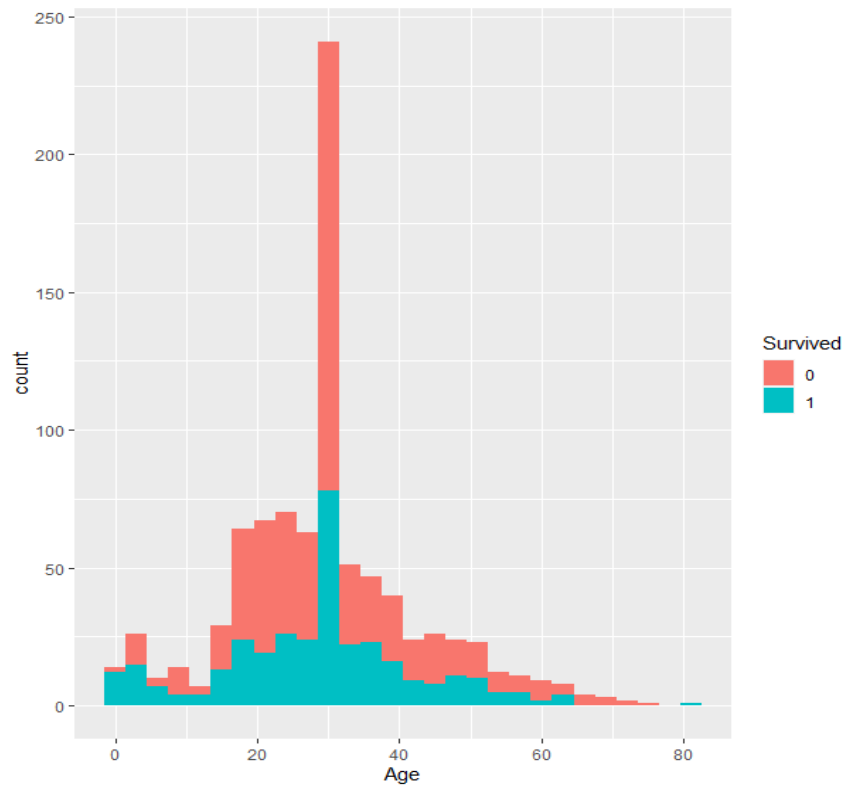
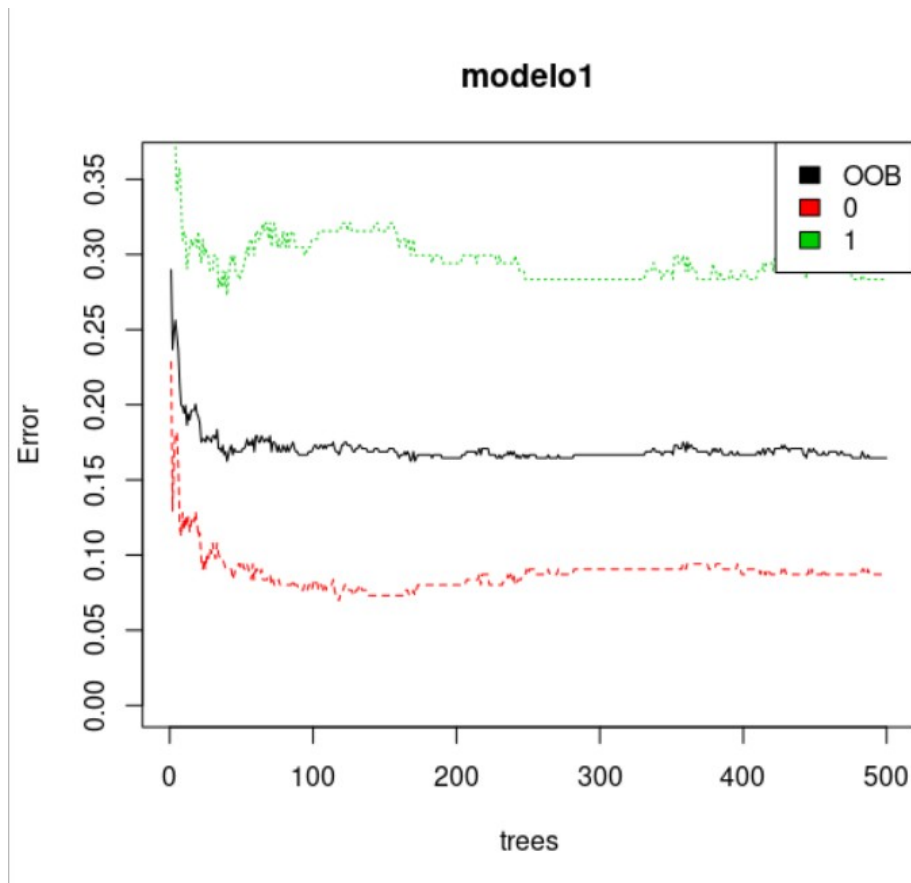


Gráfico 5.8 Modelo 1



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

De los resultados se puede concluir que existe una relación entre el sexo y la supervivencia, con una mayor supervivencia de las mujeres sobre los hombres.

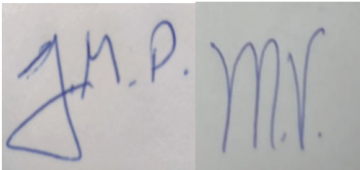
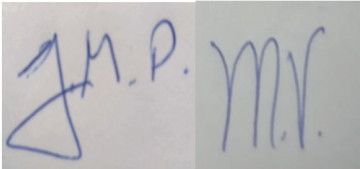
También existe relación entre pertenecer a una determinada clase y sobrevivir, con peor supervivencia los pasajeros de clase 2 y 3.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

El código R utilizado para esta práctica se encuentra en el repositorio github de la práctica, en los dos siguientes ficheros:

<https://github.com/mishuvale91/titanic-dataset/blob/main/src/titanic-cleaning.R>

<https://github.com/mishuvale91/titanic-dataset/blob/main/src/titanic-analisis.R>

Contribuciones	Firma
Investigación previa	
Redacción de las respuestas	
Desarrollo código	