Tipología y ciclo de vida de los datos

Práctica 2: Limpieza y análisis de datos

Titanic

Alumnos: Alberto Sánchez Mazarro
Sergio Romero Córdoba

Índice

1. Descripción del dataset.

El juego de datos elegido para la realización de esta práctica es el propuesto en el enunciado de la misma, sobre el conjunto de pasajeros del Titanic.

Desde la página de Kaggle (https://www.kaggle.com/c/titanic) es posible descargar los siguientes ficheros:

- train.csv. Contiene el conjunto de datos "a entrenar"
- test.csv. Contiene el conjunto de datos para validar los modelos generados durante el entrenamiento. Contiene las mismas columnas que el dataset train.csv pero sin la variable que indica si el pasajero sobrevivió o no.
- gender_submission.csv. Contiene un conjunto de datos en el que solo se informa del identificador del pasajero y si este ha sobrevivido o no.

A continuación se explican brevemente los campos del fichero "train.csv".

1. PassengerId -> Identificador numérico único del pasajero.
2. Survived -> 0 si no sobrevivió, 1 si sobrevivió.
3. Pclass -> 1 si es primera categoría, 2 si segunda, 3 si tercera.
4. Name -> Nombre del pasajero
5. Sex -> "Male" si era hombre, "Female" si mujer
6. Age -> Atributo numérico con la edad del pasajero. Si es desconocida y, por tanto, estimada, se informa en el formato "XX.5".
7. SibSp -> Atributo numérico con el número de hermanos más esposo/a a bordo.
8. Parch -> Atributo numérico con el número de padres o hijos a bordo.
9. Ticket -> Identificador alfanumérico del billete.
10. Fare -> Tarifa del billete
11. Cabin -> Número de cabina ocupada
12. Emabarked -> Puerto donde embarcó (C = Cherbourg, Q = Queenstown, S = Southampton)

El principal objetivo de este dataset es responder a la pregunta ¿qué características hacían más probable que un pasajero sobreviviera?

Es posible que este ejemplo no sea el mejor para hacer predicciones sobre la posibilidad de sobrevivir en otro posible naufragio. Las características del barco pueden ser diferentes y, por ejemplo, que las cabinas de cierta clase estén en una posición que dificulte la supervivencia mientras que en este caso la facilitaran. No obstante, algunas propiedades sí que pueden resultar de interés y se tratarán de estudiar en capítulos posteriores.

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

Dado que disponemos de tres ficheros, vamos a realizar una tarea de integración para tener todos los datos en un solo dataset.

Como hemos indicado, en el conjunto de test no se proporciona el atributo “survived”. Sin embargo, se puede obtener del fichero “gender_submission”.

Por tanto, será posible crear un único juego de datos con todos los pasajeros, sus características y su información de si sobrevivió o no.

Se puede ver el código y la ejecución en el fichero adjunto, pero se muestran algunas capturas a continuación.

```
#Añadimos la variable "Survived" al dataset de Test
datosTestSurv<-merge(datosTest, datosSurvived, by.x="PassengerId", by.y="PassengerId")

#Mezclamos el nuevo dataset de test y el de train
datos<-rbind(datosTestSurv, datosTrain)

#Comprobamos que no hay duplicados
sum(duplicated(datos$PassengerId))
```

En el dataset resultante, tenemos 1309 observaciones con 12 propiedades.

Por trabajar un poco más con la manipulación de datos, vamos a agrupar las columnas de “SibSp” y “Parch” en una nueva columna “Relatives”, que sea la suma de los valores de las columnas anteriores.

```
#Agrupamos todos los parentescos familiares en una nueva columna
datos$Relatives=datos$SibSp+datos$Parch

#Borramos las columnas de SibSp y Parch
datos<-select(datos, -SibSp, -Parch)
```

Finalmente, éste es el dataset que obtenemos y con el que vamos a trabajar.

```
'data.frame':   1309 obs. of  11 variables:
 $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
 $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
 $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles,
Mr. Thomas Francis" "Wirz, Mr. Albert" ...
 $ Sex        : chr  "male" "female" "male" "male" ...
 $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
 $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
 $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
 $ Cabin      : chr  "" "" "" "" ...
 $ Embarked   : chr  "Q" "S" "Q" "S" ...
 $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
 $ Relatives  : int  0 1 0 0 2 0 0 2 0 2 ...
```

3. Limpieza de los datos.

3.1 Elementos vacíos

Vamos a revisar si existen valores nulos para cada columna y ver el tratamiento que le debemos dar en cada caso.

```
# Números de valores desconocidos por campo
sapply(datos, function(x) sum(is.na(x)|x==""))
```

```
PassengerId      Pclass        Name         Sex         Age      Ticket        Fare       Cabin
          0           0           0           0         263           0           1        1014
   Embarked    Survived   Relatives
          2           0           0
```

Tras realizar la comprobaciones anteriores vemos que tenemos:

- 263 casos en los que no disponemos de la edad
- 1014 casos en los que no disponemos del número de cabina
- 1 caso en el que no disponemos de la tarifa
- 2 casos en los que desconocemos el puerto de embarque.

Vamos a intentar completar nuestro dataset con valores estimados para la tarifa y la edad en los registros con datos desconocidos. En todos los casos vamos a interpretar que simplemente son valores desconocidos, es decir, no disponer de tarifa, no necesariamente implica que la tarifa era 0 (situación para la que también existen algunos registros), sino que no disponemos del dato.

En el caso de la tarifa únicamente tenemos un registro desconocido. Podemos intentar estimarlo en función de la clase y el puerto de embarque, ya que parece que la tarifa y la clase guardan relación. Por tanto, vemos que es un registro con clase 3 y embarque Southampton. Obtenemos la media de los datos de todos los registros con esas características y lo asignamos.

```
casosS3<-filter(datos, Pclass==3 & datos$Embarked=="S")
mediaTarifaS3<-mean(casosS3$Fare,na.rm=TRUE)

#Asignamos el valor medio al dato perdido.
datos[153,"Fare"]<-mediaTarifaS3
```

Para la edad tenemos 263 por lo que ir uno a uno implicaría demasiado tiempo. En este caso, vamos a agrupar los datos en función de la clase y el número de familiares embarcados y vamos a calcular la media de edad de estos grupos. Esta media será la que asignemos a los valores desconocidos en función del grupo al que pertenecen.

```
# Agrupamos por clase y familiares y asignamos la media a los valores
# desconocidos
datos %>%
  group_by(Pclass, Relatives) %>%
  mutate(Age = ifelse(is.na(Age),
                      median(Age, na.rm = TRUE),
                      Age))
```

El número de cabina y el puerto de embarque no tienen una relación directa con ninguna de las columnas restantes, por lo que en principio no es posible estimar su valor. Por lo tanto, no haremos nada con ellos y los dejaremos con NA.

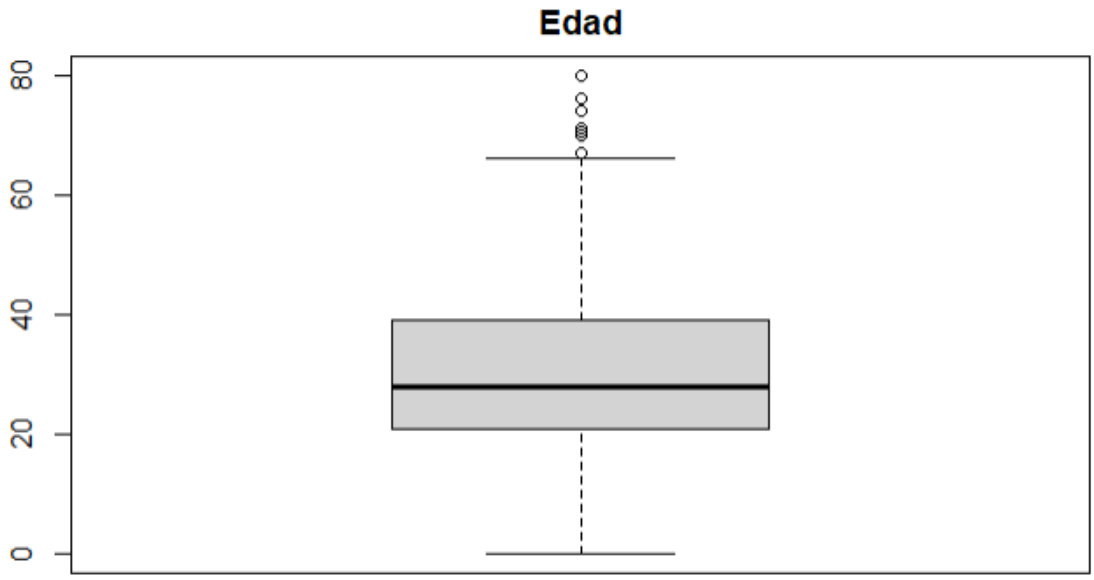
3.2 Outliers

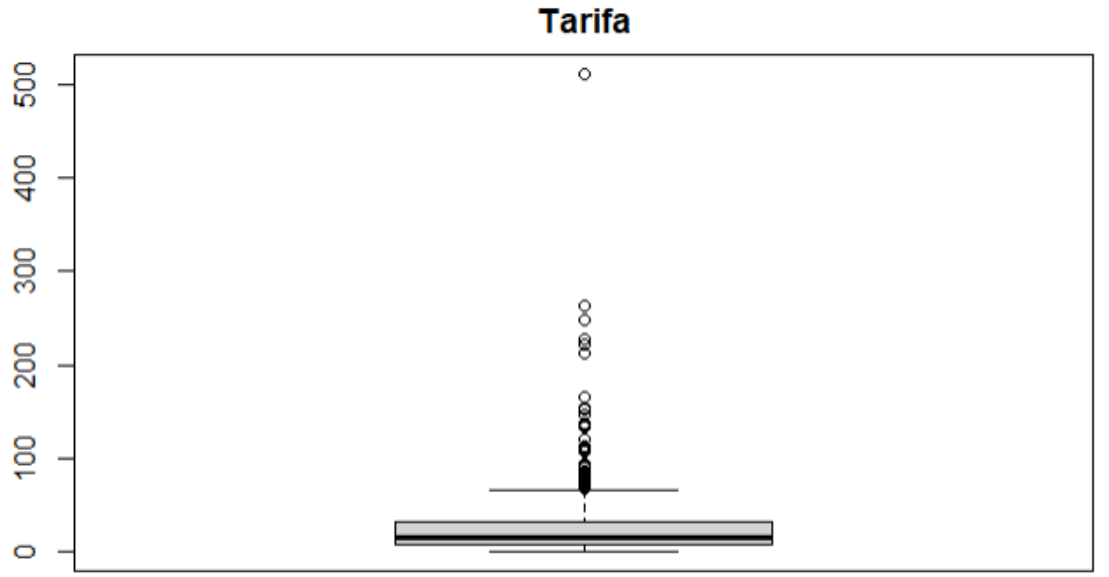
Los outliers o valores extremos son aquellos que llaman la atención por su evidente diferencia con respecto a la mayoría de datos de otros registros. Vamos a analizar con diagramas de cajas los atributos para intentar encontrar los valores outlier de cada propiedad numérica (tarifa, edad y familiares a bordo).

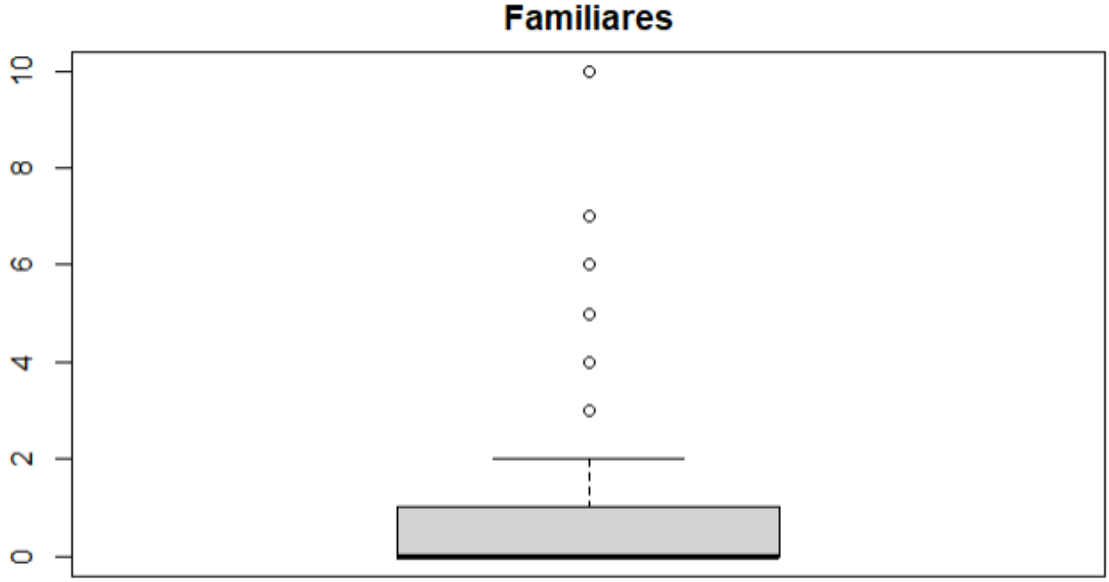
```
# Visualización de la edad
edad<-datos.bp<-boxplot(datos$Age,main="Edad")
edad$out

# Visualización de la tarifa
tarifa<-datos.bp<-boxplot(datos$Fare,main="Tarifa")
tarifa$out

# Visualización del número de famimliares
familiares<-datos.bp<-boxplot(datos$Relatives,main="Familiares")
familiares$out
```







Como podemos comprobar, aunque sí que hay valores que no están dentro de los valores más "normales" de estos atributos, no existen datos realmente anómalos. Para las edades y el número de familiares, los valores que vemos más alejados de los valores medios son razonables y no tendrían por qué ser incorrectos. Respecto a las tarifas, sí que vemos valores mucho más elevados de los normal. Sin embargo, analizando los datos, vemos que corresponden a pasajes de primera clase por lo que es razonable pensar que también son correctos. Además, hay algunos valores a 0 que asumiremos como invitaciones.

4. Análisis de los datos

4.1 Selección de grupos de datos

Lo primero que vamos a realizar es una discretización de los datos relativos al sexo, el puerto de embarque, la clase del billete y la supervivencia, que, aunque vienen representados en enteros o cadenas, en realidad pertenecen a grupos que deben ser factorizados.

```r
datos$Sex <- as.factor(datos$Sex)
datos$Embarked <- as.factor(datos$Embarked)
datos$Pclass <- as.factor(datos$Pclass)
datos$Survived <- as.factor(datos$Survived)
# Después de los cambios, analizamos la nueva estructura del conjunto de datos
str(datos)
```

La estructura de los datos queda de la siguiente manera.

```
'data.frame':   1309 obs. of  11 variables:
 $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
 $ Pclass     : Factor w/ 3 levels "1","2","3": 3 3 2 3 3 3 3 2 3 3 ...
 $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis"
"Wirz, Mr. Albert" ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
 $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
 $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
 $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
 $ Cabin      : chr  "" "" "" "" ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 3 4 3 4 4 4 3 4 2 4 ...
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 2 1 2 1 ...
 $ Relatives  : int  0 1 0 0 2 0 0 2 0 2 ...
```

En general, lo que vamos a querer analizar es la relación que existe entre los pasajeros que sobreviven con el resto de atributos. Queremos comprobar si existen ciertas características que hacen que los pasajeros tengan más probabilidades de sobrevivir. En concreto, en los próximos apartados vamos a analizar si el sexo, la clase del billete, la edad o la tarifa son determinantes para decir que un pasajero sobrevivió.

4.2. Comprobación de normalidad y homogeneidad de la varianza

La normalidad y la homogeneidad de la varianza únicamente tienen sentido con variables numéricas, por lo que vamos a utilizar la edad y la tarifa.

Con el objetivo de verificar la suposición de normalidad, vamos a realizar el test de Shapiro-Wilk.

```r
shapiro.test(datos$Age)
shapiro.test(datos$Fare)
```

```
        Shapiro-Wilk normality test

data:  datos$Age
W = 0.97955, p-value = 5.747e-11


        Shapiro-Wilk normality test

data:  datos$Fare
W = 0.52766, p-value < 2.2e-16
```

Asumimos como hipótesis nula que la población está distribuida normalmente. En ambos casos, obtenemos p-values muy bajos, mucho menores que 0.05 que podríamos tomar como nivel de significancia, por lo que vamos a rechazar la hipótesis nula, es decir, no podemos asumir normalidad.

Para la homocedasticidad (igualdad de varianza entre dos grupos), vamos a comprobar la edad y la tarifa para los grupos que sobrevivieron y los que no por medio del test de Fligner-Kileen, dado que los atributos no presentan normalidad.

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

```
        Fligner-Killeen test of homogeneity of variances

data:  Age by Survived
Fligner-Killeen:med chi-squared = 2.7432, df = 1, p-value = 0.09767


        Fligner-Killeen test of homogeneity of variances

data:  Fare by Survived
Fligner-Killeen:med chi-squared = 129.22, df = 1, p-value < 2.2e-16
```

Vemos que, tomando alfa=0.05 como valor aceptado, la varianza en los grupos de supervivencia presenta homocedasticidad para el atributo edad pero no para la tarifa.

4.3 Aplicación de técnicas estadísticas

Sexo-Supervivencia

Vamos a ver en primer lugar la relación entre el sexo y la supervivencia. Para ello, aplicamos el test de χ^2 que nos permite comparar dos variables categóricas. En primer lugar creamos la tabla con las frecuencias de cada grupo.

```
numHombresSuperV=sum(datos$Sex=='male' & datos$Survived==1)
numHombresNoSuperv=sum(datos$Sex=='male' & datos$Survived==0)
numMujeresSuperv=sum(datos$Sex=='female' & datos$Survived==1)
numMujeresNoSuperv=sum(datos$Sex=='female' & datos$Survived==0)
hombres=c(numHombresSuperV, numHombresNoSuperv)
mujeres=c(numMujeresSuperv, numMujeresNoSuperv)
sexoSuperv=as.data.frame(rbind(hombres, mujeres))
names(sexoSuperv) = c('Sobrevive', 'NoSobrevive')
sexoSuperv
```

	Sobrevive <int>	NoSobrevive <int>
hombres	109	734
mujeres	385	81

2 rows

Y seguidamente aplicamos el test.

```
chisq.test(sexoSuperv)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  sexoSuperv
X-squared = 617.31, df = 1, p-value < 2.2e-16
```

Obtenemos un p-value muy bajo que indica diferencias significativas entre ambos grupos.

Tarifa, Edad-Supervivencia

El test de Kruskal-Wallis es la alternativa no paramétrica a los contrastes de hipótesis de más de dos grupos cundo no se cumple la condición de normalidad. Vamos a aplicar este test para ver si la tarifa y la edad influyen en la supervivencia de los pasajeros.

```
kruskal.test(Fare ~ Survived, data = datos)
kruskal.test(Age ~ Survived, data = datos)
```

```
        Kruskal-Wallis rank sum test

data:  Fare by Survived
Kruskal-Wallis chi-squared = 106.27, df = 1, p-value < 2.2e-16


        Kruskal-Wallis rank sum test

data:  Age by Survived
Kruskal-Wallis chi-squared = 1.7982, df = 1, p-value = 0.1799
```

Para la tarifa, obtenemos un p-value muy bajo, lo que nos indica que sí que se encuentra significancia en evaluar la probabilidad de supervivencia de la persona.

Para la edad, sin embargo, el p-value es superior al nivel de significancia por lo que no parece que sea un factor que influya en determinar si el pasajero sobrevivió o no.

Regresión lineal y logística

A fin de utilizar la regresión lineal, vamos a intentar ver si podemos obtener alguna relación entre el precio y otros atributos, pese a que ya hemos comentado que los principales estudios sobre este dataset tienen que ver con el factor "Survived".

Vamos a comenzar el estudio con la edad, que es también una variable cuantitativa. De ser así, lo que estaríamos indicando es que cuanto mayores son las personas, tarifas más altas pagan, lo cual no tiene por qué ser así.

```
m1 = lm(Age~Fare,data=datos)
summary(m1)
```

```
Call:
lm(formula = Age ~ Fare, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-34.238  -8.630  -1.559   8.359  50.425

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.19711    0.52537  53.671  < 2e-16 ***
Fare         0.04593    0.00788   5.829 7.44e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.19 on 1044 degrees of freedom
  (263 observations deleted due to missingness)
Multiple R-squared:  0.03152,   Adjusted R-squared:  0.03059
F-statistic: 33.97 on 1 and 1044 DF,  p-value: 7.439e-09
```

Obtenemos un R-Squared realmente bajo, que nos indica que en este caso no podemos obtener una relación sólida entre la edad del pasajero y la tarifa que paga, lo cual tiene bastante sentido.

Vamos a utilizar el modelo de regresión lineal para predecir la tarifa en base a otras variables cualitativas.

Para ello, vamos a aplicar regresión sobre diferentes variables y vamos a tratar de encontrar las que mejor se ajustan.

```
mClass = lm(Fare~Pclass,data=datos)
summary(mClass)

mEmbarked = lm(Fare~Embarked,data=datos)
summary(mEmbarked)
```

```
Call:
lm(formula = Fare ~ Pclass, data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-87.51  -8.18  -5.41   2.82 424.82

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   87.509      2.298   38.07   <2e-16 ***
Pclass2      -66.330      3.383  -19.61   <2e-16 ***
Pclass3      -74.205      2.773  -26.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.31 on 1306 degrees of freedom
Multiple R-squared:  0.3636,    Adjusted R-squared:  0.3627
F-statistic: 373.1 on 2 and 1306 DF,  p-value: < 2.2e-16


Call:
lm(formula = Fare ~ Embarked, data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-58.32 -19.51 -12.83   1.60 449.99

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    80.00      34.93   2.290   0.0222 *
EmbarkedC     -17.66      35.06  -0.504   0.6145
EmbarkedQ     -67.59      35.21  -1.920   0.0551 .
EmbarkedS     -52.60      34.97  -1.504   0.1328
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.4 on 1305 degrees of freedom
Multiple R-squared:  0.09065,   Adjusted R-squared:  0.08856
F-statistic: 43.37 on 3 and 1305 DF,  p-value: < 2.2e-16
```

De estos resultados vemos que la clase es el factor que más influencia tiene con la tarifa, pese a que un valor de R-Squared de 0.36 tampoco puede considerarse demasiado bueno.

Por último, vamos a aplicar regresión logística sobre la variable dicotómica objetivo "Survived" para intentar encontrar mediante esta técnica qué parámetros influyen más en la supervivencia.

```
modeloLogistico=glm(formula = Survived ~ Pclass+ Sex + Age+ Fare + Embarked, data = datos,
family=binomial(link=logit))
summary(modeloLogistico)
```

```
Call:
glm(formula = Survived ~ Pclass + Sex + Age + Fare + Embarked,
    family = binomial(link = logit), data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7457  -0.5379  -0.3469   0.4952   2.5733

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.615e+01  6.167e+02   0.026 0.979102
Pclass2     -1.134e+00  2.966e-01  -3.823 0.000132 ***
Pclass3     -2.146e+00  3.067e-01  -6.998 2.60e-12 ***
Sexmale     -3.550e+00  1.980e-01 -17.929  < 2e-16 ***
Age         -3.086e-02  7.151e-03  -4.315 1.59e-05 ***
Fare        -2.248e-05  1.997e-03  -0.011 0.991017
EmbarkedC   -1.208e+01  6.167e+02  -0.020 0.984374
EmbarkedQ   -1.229e+01  6.167e+02  -0.020 0.984098
EmbarkedS   -1.234e+01  6.167e+02  -0.020 0.984032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1406.80  on 1045  degrees of freedom
Residual deviance:  794.83  on 1037  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 812.83

Number of Fisher Scoring iterations: 13
```

De los datos anteriores se ve claramente que el factor de corte es la clase (el modelo toma como referencia la clase 1 y vemos como para pClass2 y pClass3 obtenemos una estimación negativa, que nos indica que es menos probable la supervivencia). Esto puede tener cierto sentido (a mejores condiciones en la clase del billete mayores probabilidades de supervivencia)

Del mismo modo, un valor negativo (y muy significativo en valor absoluto) de la variable "Sexmale" nos indica que los hombres tienen menos probabilidad de supervivencia que las mujeres.

Por último, sobre el modelo de regresión anterior vamos a realizar un ejercicio de validación cruzada sobre los datos de entrenamiento y test.

Para ello usamos el conjunto de entrenamiento (datosTrain) haciendo 10 folds y aplicamos regresión logística sobre las dos variables más significativas: la clase y el sexo.

```
folds <- createFolds(datosTrain$Survived, k = 10)

cvRegresionLogistica <- lapply(folds, function(x){

training_fold <- datosTrain[-x, ]
test_fold <- datosTrain[x, ]

clasifLogistico=glm(formula = Survived ~ Sex + Pclass ,  data = training_fold, family=binomial(link=logit))
y_pred <- predict(clasifLogistico, type = 'response', newdata = test_fold)

cm <- table(test_fold$Survived, y_pred)
precision <- (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] +cm[1,2] + cm[2,1])
return(precision)
})
precisionRegresionLogistica <- mean(as.numeric(cvRegresionLogistica))
precisionRegresionLogistica
```

```
[1] 0.700793
```

Y obtenemos una precisión cercana al 70%

Ahora extrapolamos este modelo a los datos de test

```
clasifLogisticoTest=glm(formula = Survived ~ Pclass+ Sex,  data = datosTrain, family=binomial(link=logit))
y_pred <- predict(clasifLogisticoTest, type = 'response', newdata = datosTestSurv)

cm <- table(datosTestSurv$Survived, y_pred)
precisionFinal <- (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] +cm[1,2] + cm[2,1])
precisionFinal
```

```
[1] 0.6985646
```

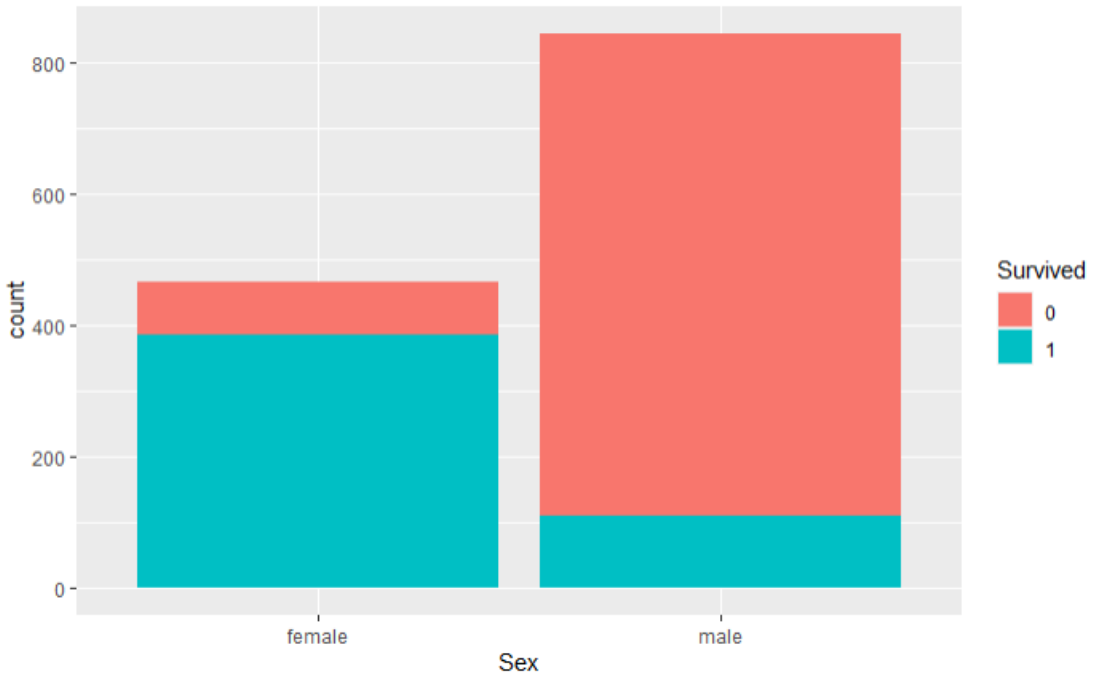
Y podemos apreciar que el porcentaje de acierto se mantiene en torno al 70% tal y como obtuvimos con el conjunto de entrenamiento.

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

A continuación, vamos a mostrar una serie de gráficas que nos permitan visualizar gráficamente los datos. Vamos a utilizar gráficos de barras apiladas utilizando la librería ggplot.

Un primer análisis que puede resultar interesante es ver la relación entre el sexo del pasajero y su capacidad de supervivencia.

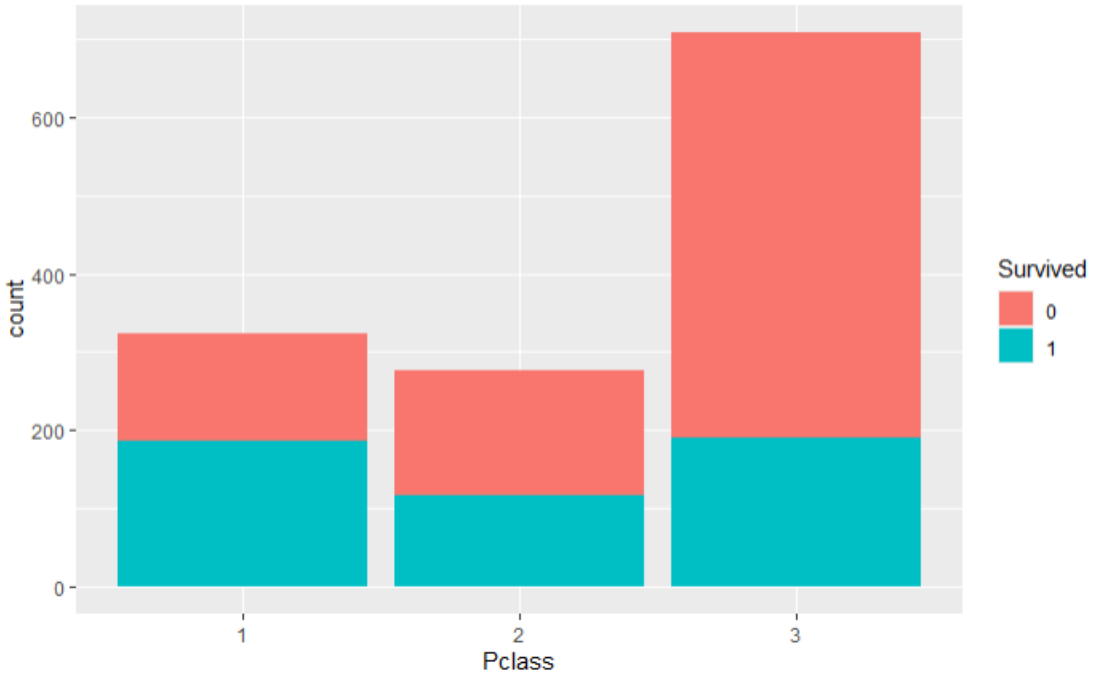
```
ggplot(data=datos,aes(x=Sex,fill=Survived))+geom_bar()
```



De esta primera consulta obtenemos visualmente una información bastante relevante: mientras que la gran mayoría de hombres falleció, la mayoría de las mujeres sobrevivieron.

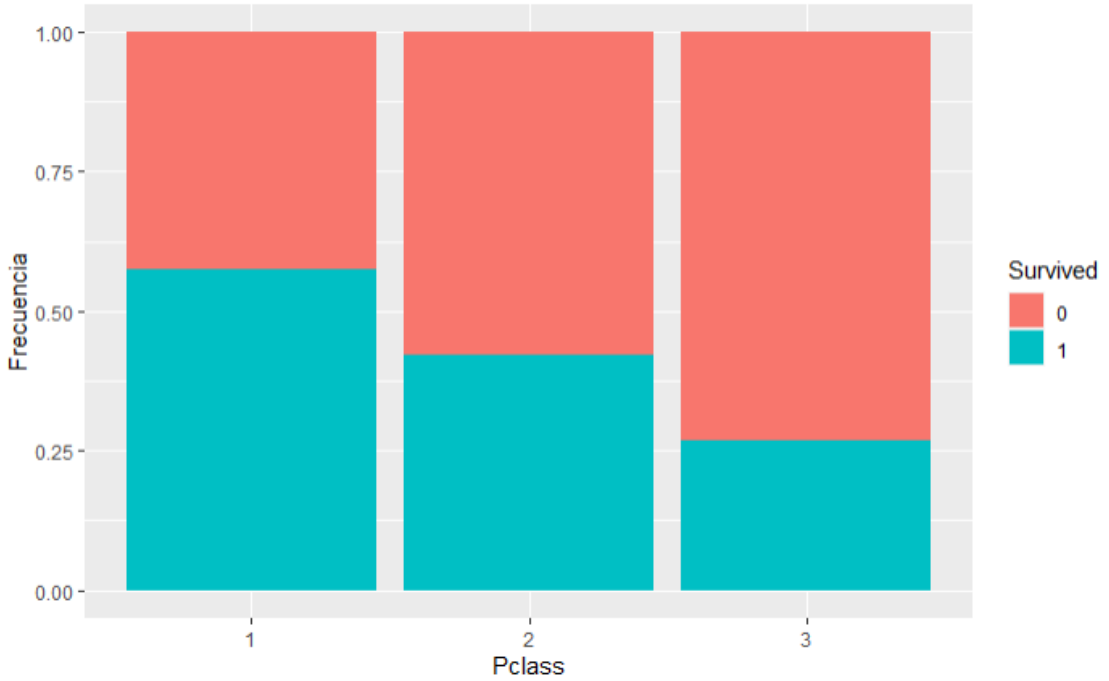
Veamos qué sucede comparando con la clase en la que viajaban.

```
ggplot(data = datos,aes(x=Pclass,fill=Survived))+geom_bar()
```



Vemos que el número de supervivientes es relativamente parecido en todas las categorías. Pero veamos qué sucede desde el punto de vista porcentual:

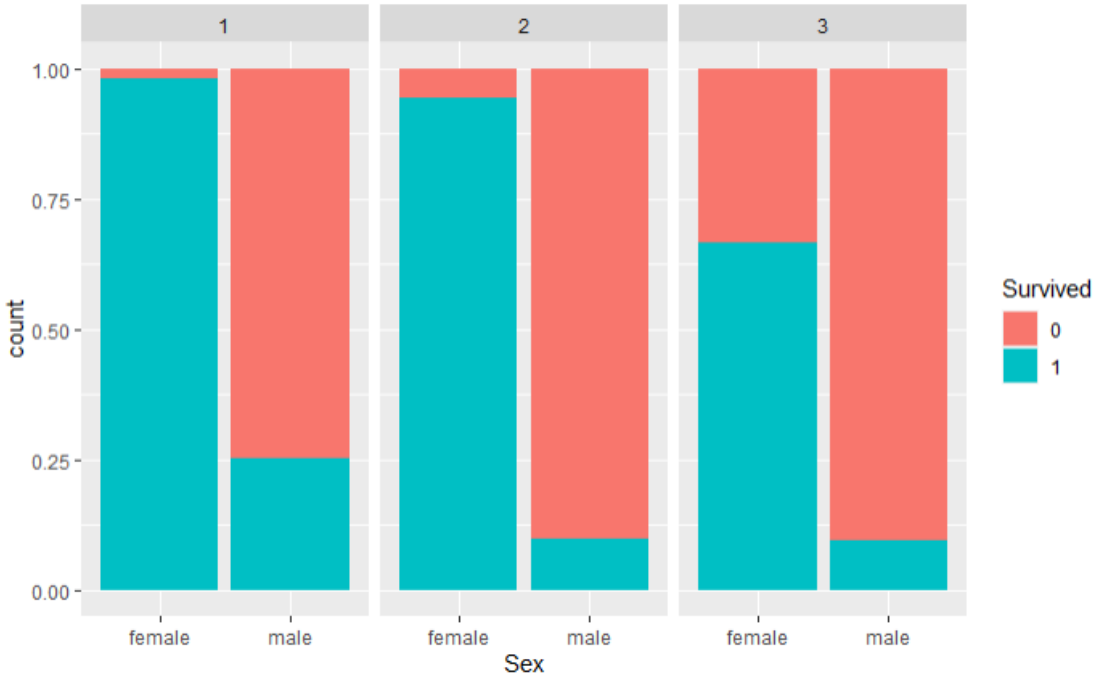
```
ggplot(data = datos,aes(x=Pclass,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```



Aquí comprobamos que los pasajeros de primera clase tenían más probabilidad de sobrevivir que los de segunda, y éstos más que los de tercera.

Por último, vamos a ver la relación entre las tres variables: Por cada clase (1, 2, 3) vemos el porcentaje de supervivientes.

```
ggplot(data = datos,aes(x=Sex,fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pclass)
```



Como dato llamativo, casi todas las mujeres de primera clase sobrevivieron, mientras que no lo hizo casi ningún hombre de segunda o tercera.

6. Resolución del problema

Como hemos comentado al principio, el principal objetivo del dataset es ver las características que hacían que un pasajero tuviera más probabilidades de sobrevivir en el Titanic.

Para ello, durante el trabajo hemos analizado la capacidad de supervivencia en función de distintas características, en concreto, en función de la edad, el sexo, la tarifa o la clase.

Tanto gráficamente como mediante el test de χ^2, hemos podido comprobar que una mujer a bordo del Titanic tiene más probabilidad de estar entre los supervivientes que un hombre y que, por tanto, el sexo es un factor determinante a la hora de poder averiguar si un pasajero sobrevivió.

Respecto a la clase en la que viajaba el pasajero, si bien el número de supervivientes es parecido en las tres categorías que había, en porcentaje hemos visto que los pasajeros de primera clase tenían más probabilidades de sobrevivir que los de segunda y a su vez, estos tenían más probabilidades de sobrevivir que los de tercera.

El pasajero prototipo que sobrevivió podemos decir que es una mujer que viajaba en primera clase.

Sobre la tarifa y la edad, hemos aplicado el test de Kruskal-Wallis para comprobar que mientras que la tarifa sí es un factor determinante para saber si el pasajero sobrevivió, la edad no parece tener un peso demasiado grande.

Otras comparaciones no relacionadas con la supervivencia se han realizado con fines más didácticos que prácticos para intentar abarcar los máximos tests y comprobaciones posibles para la realización de la práctica.

7. Participación

La práctica ha sido realizada conjuntamente. Si bien se ha realizado una lógica repartición de las tareas, ambos integrantes hemos participado en todos los puntos, así como en la decisión del dataset a utilizar.

CONTRIBUCIONES	FIRMA
Investigación previa	Sergio Romero y Alberto Sánchez
Desarrollo de código	Sergio Romero y Alberto Sánchez
Redacción de las respuestas	Sergio Romero y Alberto Sánchez