Práctica 2 Limpieza y resolución de datos

Jone Aliri Lazcano, Nadia Nathaly Sánchez Pozo

07/01/2020

Índice

1. DETALLES DE LA ACTIVIDAD

1.1. Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

1.3. Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

2. RESOLUCIÓN

2.1. Descripción del dataset

La base de datos que se ha analizado en esta práctica se título **Titanic: Machine Learning from Disaster** (https://www.kaggle.com/c/titanic) [!https://www.kaggle.com/c/titanic].

Los datos de este archivo se encuentran divididos en dos bases de datos, uno de ellos es la base de datos de entrenamiento (*train.csv*) y la otra base de datos es la de prueba (*test.csv*), para este caso de estudio vamos a utilizar el conjunto train.

A partir de esta base de datos se plantea la problemática de determinar qué variables son más influyentes en la probabilidad de sobrevivencia de los pasajeros, mediante el estudio individual y colectivo de las variables, aplicando pruebas estadisticas.

Resumiendo, el 14 de Abril de 1992 el Titanic chocó con un iceberg y se llevo aproximadamente a 1500 de sus pasajeros y tripulación a las profundidades del oceano. Este incidente ha sido considerado uno de los desastres marinos más importantes en tiempos de paz, y a causa de dicho indicente se actualizaron o renovaron numerosas políticas de seguridad. Sin embargo, existen numerosas voces que dicen que hubo circunstancias que hicieron que hubiera un desproporcionada cantidad de muertos. El objetivo de analizar esta base de datos es explorar los factores que tuvieron relación con el hecho de que una persona sobreviviera o no a la catastrofe del Titanic.

Descríbamos la base de datos.

Esta base de datos tiene 891 observaciones y 12 variables.

- **PassengerId**:Variable que aporta el código de identificación de los pasajeros.
- **Survived**: Variable dicotómica que indica si el pasajero sobrevivión (1) o no sobrevivió (0).
- **Pclass**: Variable categórica que indica si los pasajeros tenían tickets de primera clase (1), segunda clase (2) o tercera clase (3). Obviamente los tickets más caros eran los de primera clase, seguidos de los de segunda clase y finalmente los de tercera clase.
- **Name**: Variable de tipo cadena con el nombre y apellidos de los pasajeros.
- **Sex**: Variable dicotómica que indica si el pasajero era un hombre (1) o una mujer (2).
- **Age**: Variable numérica que indica la edad de los pasajeros en años. En el caso de ser personas con menos de un años se indica la fracción (con un décimal), en caso de tener más de un año se utilizan números enteros.
- **SibSp**: Variable numérica (números enteros) que indicaba el número de familiares/cónyuges que tenían los pasajeros a bordo del Titanic.
- **Parch**: Variable numérica (números enteros) que indicaba el número de hijos/padres que tenían los pasajeros a bordo del Titanic.
- **Ticket**: Código/Número del tiket (podía haber más de un pasajero con el mismo número).
- **Fare**: Variable numérica que indica el precio del pase del pasajero.
- **Cabin**: Código que identifica la cabina del pasajero.
- **Embarked**: Variable categórica con tres niveles que indica el puerto en el cual embarcaron (puerto "C", "Q", o "S") que indican C = Cherbourg, Q = Queenstown y S = Southampton.

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

En primer lugar cargamos el dataset train ya que este será la muestra para generar el modelo.

```
train <- read.csv("train.csv",stringsAsFactors = FALSE)
```

Veamos qué es lo que tenemos en la base de datos completa. Para tener una idea general de los datos podemos utilizar las funciones **str()**, **summary()** y **dim()**.

```
dim(train)
```

```
## [1] 891  12
```

```
summary(train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
```

```
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Bien, vemos que el conjunto de datos completo contiene 891 registros y 12 variables

2.3. Limpieza de los datos.

Una de las cosas importantes a la hora de importar los datos es ver si R ha asignado correctamente la categoría a cada varaiable. Por ejemplo, si nos fijamos en la variable *Survived* podemos ver que para R es una variable de tipo integer (numérico), pero quizás sería mejor que fuera un factor con dos niveles, donde 0 indique que no sobrevivió y 1 indique que sí sobrevivió. Así mismo, la variable *Pclass* también sería un factor, en este caso un factor ordenado ya que es una variable ordinal.

Hagamos estos cambios para luego poder trabajar mejor.

```
train <- train %>%
  mutate(Survived = as.factor(Survived),
         Embarked =as.factor(Embarked),
         Sex= as.factor(Sex),
         Pclass = as.factor(Pclass),
         SibSp = as.factor(SibSp),
         Parch = as.factor(Parch))
```

2.3.1. Ceros o elementos vacíos.

```
# Estadísticas de valores vacíos
colSums(is.na(train))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```
colSums(train=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

Tenemos tres variables con datos perdidos o elementos vacios, las variables: *Age*, *Cabin* y *Embarked*. La variable *Age* es una variable cuantitativa donde tenemos 177 valores perdidos que están indicados con "NA"; las variables *Cabin* y *Embarked*, están caracterizados como factores y tienen un nivel que indica un valor perdido. Así, en la variable *Cabin* tenemos 687 observaciones con valores perdidos, y en la variable *Embarked* 2 observaciones con valores perdidos. Las variables *Survived*, *SibSp* y *Parch* también contienen ceros, pero esos valores no son valores perdidos y tienen un sentido claro (que no sobrevivió, que no tenía hermanos/as ni esposo/a a bordo, y que no tenía padres ni hijos/as a bordo).

En primer lugar, observamos que la variable *PassengerId* es una variable de tipo identificador, pero que no aporta nada al estudio desde el conjunto de datos por tanto procedemos a eliminarla.

```
train$PassengerId <- NULL
```

A continuación estudiamos cabin ya que es la que tiene mayor número de valores perdidos.

```
camarote <- train[train$Cabin == "",]
dim(camarote)
```

```
## [1] 687  11
```

Podemos ver que tenemos 687 sujetos sin el dato del camarote. Realmente este dato es un dato categórico que no tiene mucho sentido imputar, ya que no tiene lógica que pongamos que unos sujetos estaban en un camarote X utilizando para ello el dato de en qué camarote estaban otros sujetos. Por lo tanto no vamos a imputar nada en estos casos. Por lo tanto eliminamos dicha variable

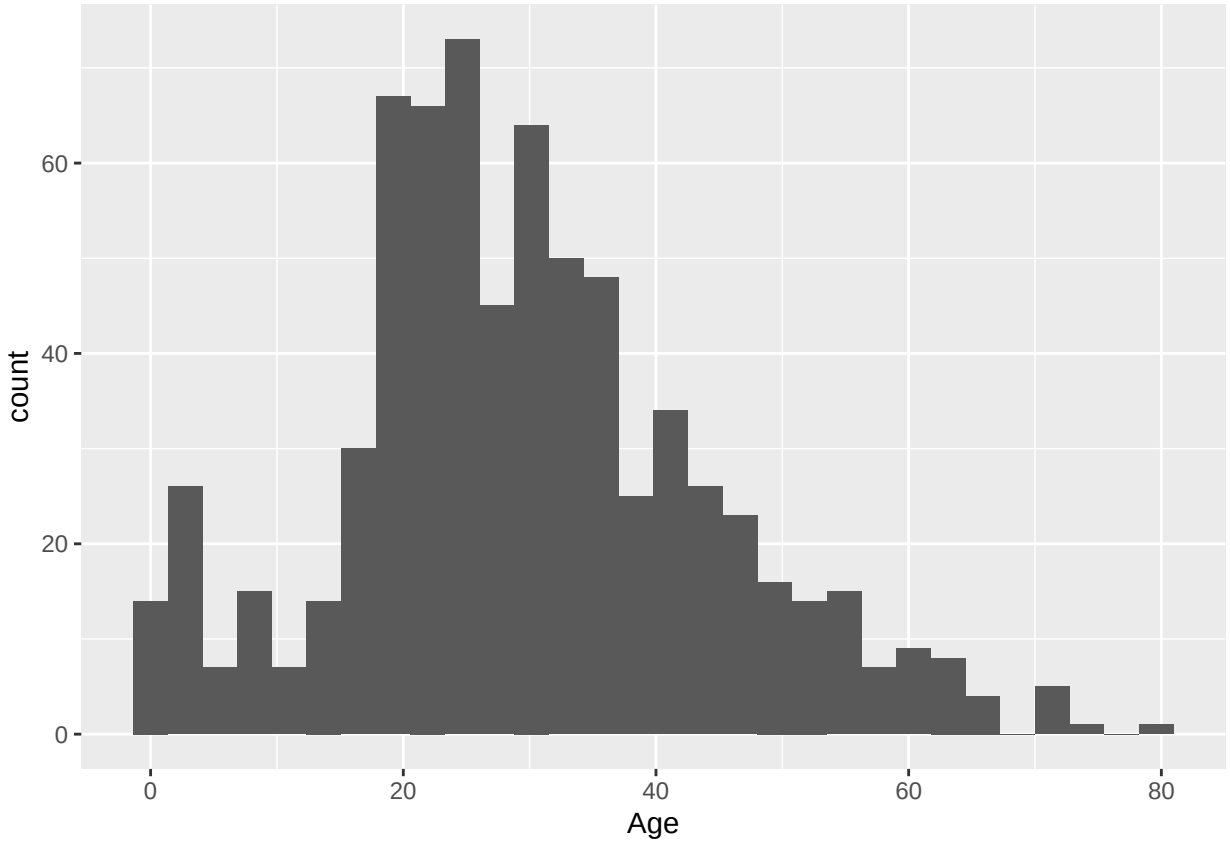
```
train$Cabin<- NULL
```

Ahora estudiamos la Variable *Age*. Esta variable es una variable cuantitativa e indica la edad de los pasajeros. Empecemos por explorar la distribución de esta variable en nuestra muestra.

```
train %>%
  ggplot(aes(Age)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

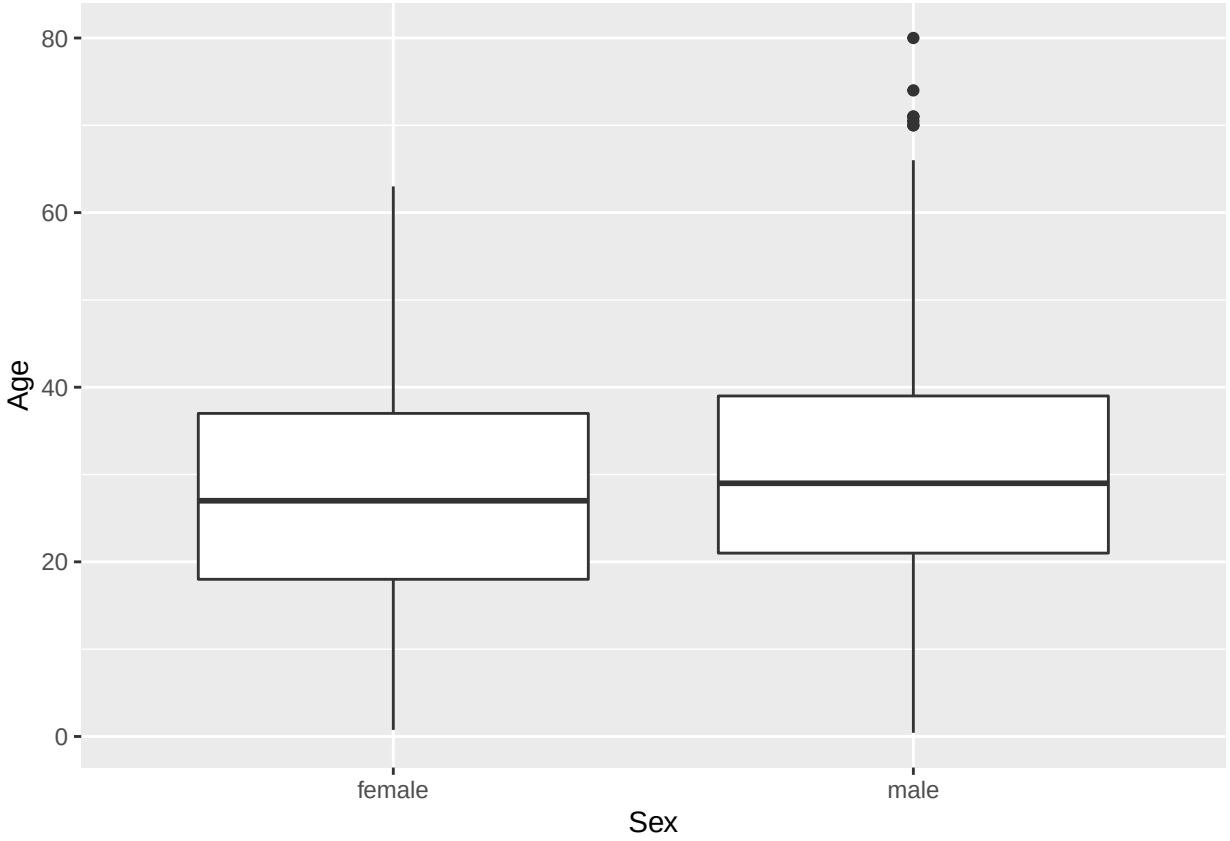
```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



Bien, podemos observar que la mayoría de los sujetos que tenemos se encuentran entre los 20 y los 40 años. Una de las opciones de imputar los datos de esta variable sería utilizar la media del grupo. Pero antes de hacer esto, veamos si por ejemplo la distribución es muy diferente en función del género, ya que en ese caso en cada género podríamos imputar la media de su grupo.

```
train %>%
  ggplot(aes(Sex, Age)) +
  geom_boxplot()
```

```
## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```



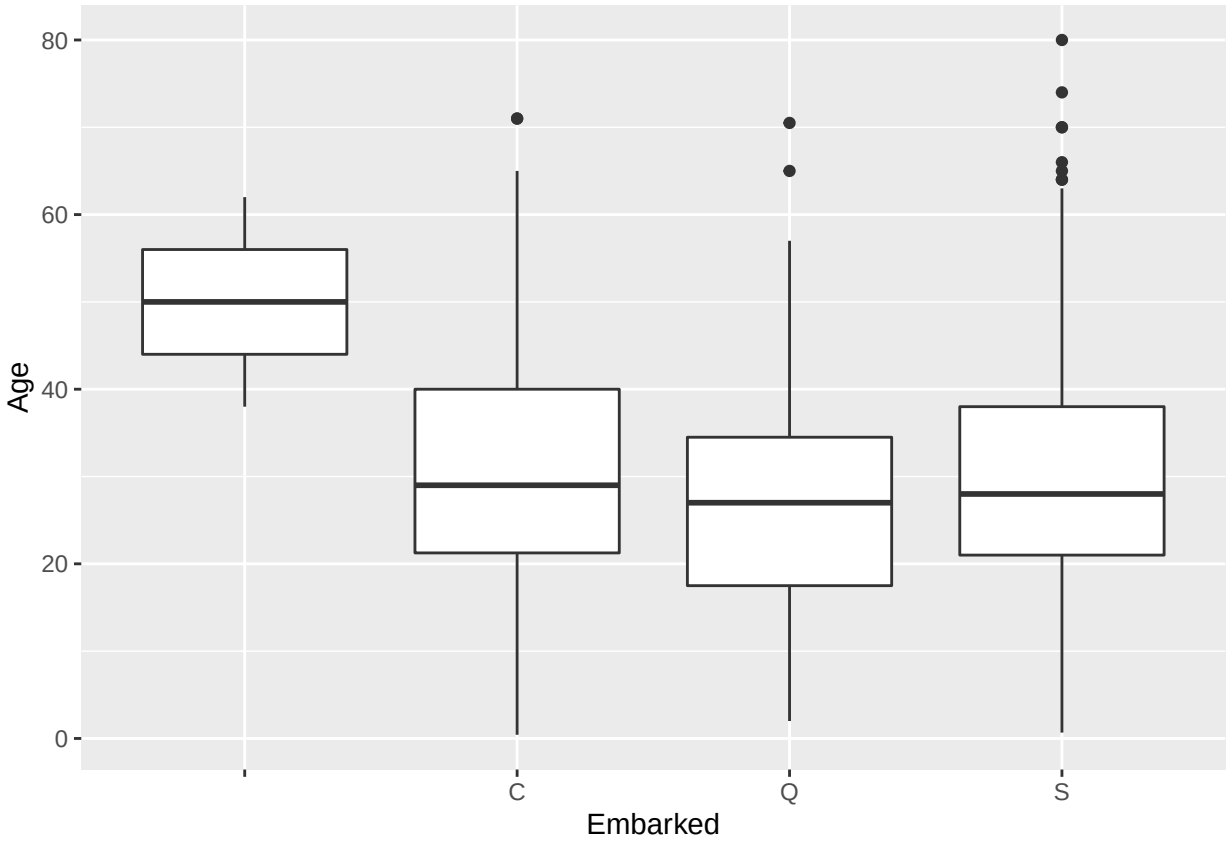

```
table(train$Sex)
```

```
##
## female   male
##    314    577
```

Podemos observar que la distribución es realmente parecida en ambos géneros, por lo que no mejoraría la imputación el hecho de utilizar las medias de cada grupo de género. Podemos analizar por ejemplo la distribución de esta variable en función del puerto de embarque:

```
train %>%
  ggplot(aes(Embarked, Age)) +
  geom_boxplot()
```

```
## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```



Sin embargo, sí que vemos algunas diferencias en función del puerto de embarque. Sobre todo en el caso de las observaciones que tienen un valor perdido en el puerto de embarque. Sin embargo, antes de confiar en este gráfico analicemos cuántas personas forman cada grupo, ya que si el número de personas es reducido, el utilizar esa media no sería la mejor opción (ya que dicho dato estaría sobredimensionado y no sería representativo).

```
table(train$Embarked)
```

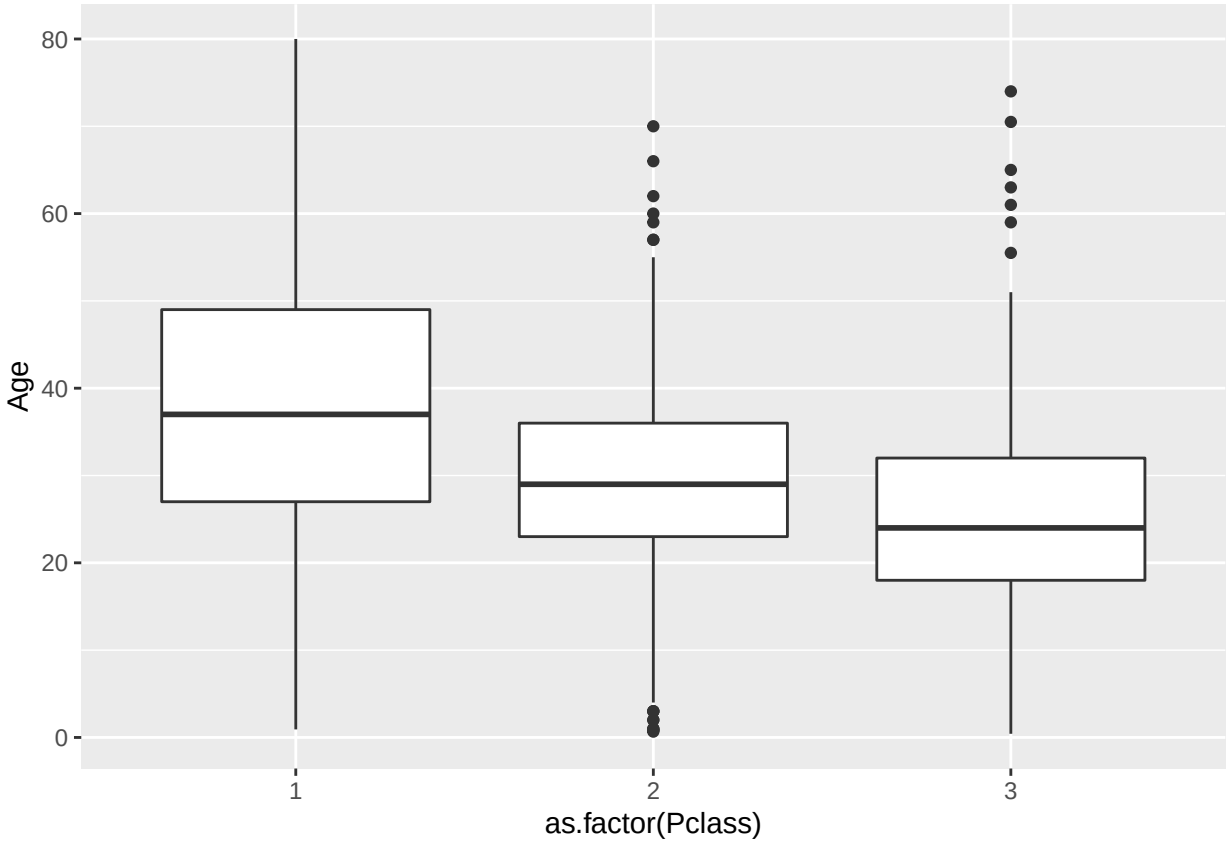
```
##
##       C   Q   S
##   2 168  77 644
```

Efectivamente, vemos que son tan solo dos observaciones las que conforman el "grupo" de las personas que no tienen dato sobre el puerto de embarque, por lo que la media de edad de este "grupo" no sería el ideal a la hora de realizar imputaciones. En cuanto a los demás grupos vemos que el puerto de embarque S es el que más observaciones tiene, y viendo que los otras dos puertos tienen relativamente pocas observaciones y las diferencias no son excesivas, no parece que mejoraríamos mucho la imputación de la edad utilizando las medias de cada uno de esos grupos.

Sigamos analizando la edad en función de otras variables como la clase.

```
train %>%
  ggplot(aes(as.factor(Pclass), Age)) +
  geom_boxplot()
```

```
## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```




```
table(train$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

Podemos observar que el rango de edad es algo más amplio, y existe mayor desviación en los pasajeros de primera clase, y que los pasajeros de segunda y tercera clase tienden a ser algo más jóvenes (aunque no sabemos si esas diferencias son significativas o no). Sin embargo de nuevo, el número de sujetos es bastante más grande en tercera clase (y ya se sabe que a mayor número de observaciones la tendencia es que la varianza sea más pequeña). Por ello, se vuelve a concluir que imputar las medias de edad en función de la clase no mejoraría enormemente la calidad de la imputación.

Podríamos analizar la distribución de la edad en función de las variables *Parch* y *SibSp*, pero éstas tienen unas distribuciones muy desiguales, y las medias obtenidas de grupos no muy numerosos no son estables y por lo tanto no son útiles para realizar imputaciones.

```
table(train$Parch)
```

```
##
##   0   1   2   3   4   5   6
## 678 118  80   5   4   5   1
```

```r
table(train$SibSp)
```

```
##
##   0   1   2   3   4   5   8
## 608 209  28  16  18   5   7
```

Sí, tal y como podemos apreciar en ambos casos, las frecuencias son muy elevadas en una categoría, y bastante más reducidas en las demás, por lo que no realizaremos imputaciones en función de estas variables.

Antes de realizar las imputaciones en la varaible *Age* echaremos un vistazo a los valores perdidos de la variable *Embarked*.

Esta variable es un factor y tiene como hemos dicho antes solo dos valores perdidos. Concretamente la variable *Embarked* indica el puerto en el que se han embarcado los sujetos y puede tener cuatro valores. Por lo tanto calcularemos una tabla de frecuencias y veremos si hay un puerto que tenga una mayor frecuencia, en cuyo caso imputaremos dicho valor a los perdidos.

```r
table(train$Embarked)
```

```
##
##       C   Q   S
##   2 168  77 644
```

Podemos observar que la mayoría de las personas embarcaron en Southampton por lo tanto imputaríamos S a las dos observaciones que tienen valores perdidos.

```r
train$Embarked[train$Embarked == ""] <- "S"
```

```r
# Estadísticas de valores vacíos
colSums(is.na(train))
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch   Ticket
##        0        0        0        0      177        0        0        0
##     Fare Embarked
##        0        0
```

```r
colSums(train=="")
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch   Ticket
##        0        0        0        0       NA        0        0        0
##     Fare Embarked
##        0        0
```

Ya solo nos quedaría realizar las imputaciones necesarias en la variable *Age*. Para ello, usaremos el método *mice* que realiza múltiples imputaciones, lo que reduce el sesgo y mejora los valores faltantes predichos. En esencia, este paquete utiliza el algoritmo de maximización de expectativas bootstrapped.

```r
impute <- mice(train,  parallel = 'multicore', idvars =c(), noms=c())
```

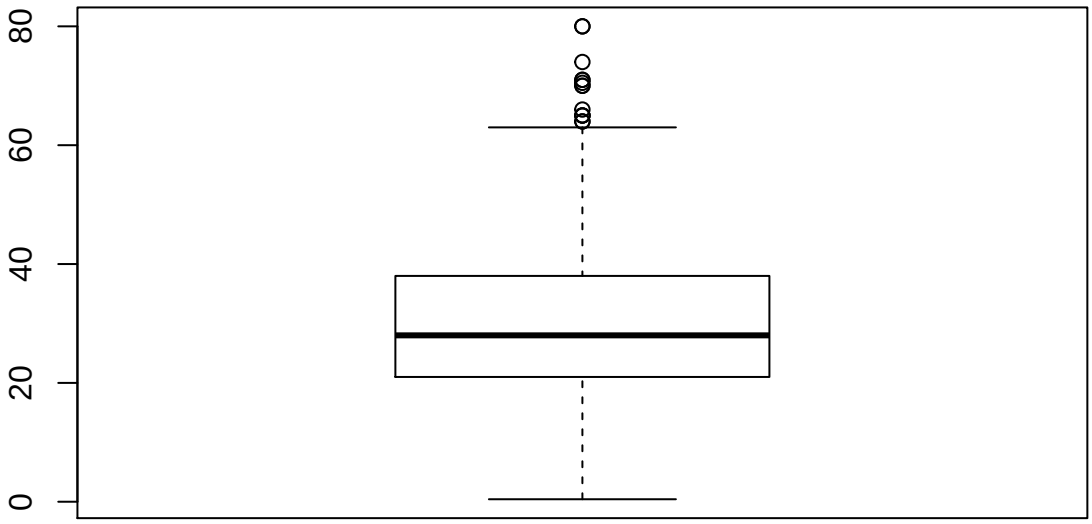
```
##
##  iter imp variable
##   1   1  Age
##   1   2  Age
##   1   3  Age
##   1   4  Age
##   1   5  Age
##   2   1  Age
##   2   2  Age
##   2   3  Age
##   2   4  Age
##   2   5  Age
##   3   1  Age
##   3   2  Age
##   3   3  Age
##   3   4  Age
##   3   5  Age
##   4   1  Age
##   4   2  Age
##   4   3  Age
##   4   4  Age
##   4   5  Age
##   5   1  Age
##   5   2  Age
##   5   3  Age
##   5   4  Age
##   5   5  Age

## Warning: Number of logged events: 27
```

```
complete2_df <- complete(impute)
```

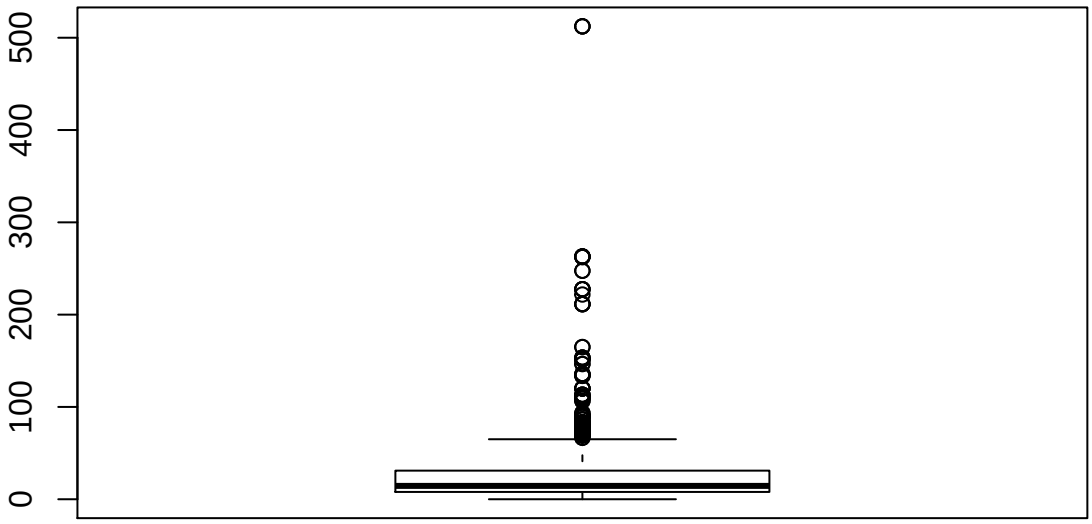
2.3.2. Identificación y tratamiento de valores extremos.

Tal y como hemos observado antes, tenemos valores extremos en la variable *Fare* y en la variable *Age*.

```
boxplot(complete2_df$Age)
```




```
boxplot(complete2_df$Fare)
```



Aunque son valores extremos, debido a que se alejan de la media y de los valores del resto del grupo, son valores que son posibles (están dentro del rango de posibles valores en esas variables), por lo tanto, a priori no los vamos a eliminar. Sin embargo, tendremos que ver en los análisis posteriores si podemos seguir usándolos o debemos prescindir de ellos en algún caso.

2.3.3. Exportación de los datos preprocesados

Luego de realizar los procedimientos de integración, validación y limpieza, procedemos a guardar nuestros datos en un nuevo fichero denominado *titanic_clean.csv*

```
# Exportación de los datos limpios en .csv
write.csv(complete2_df, "titanic_clean.csv")
```

2.4. Análisis de los datos.

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

En este caso práctico, trabajaremos con cinco variables de entrada y una variable de respuesta. Las cuales han sido más sobresalientes en los estudios realizados en apartados anteriores. Las variables de entrada son: *Age*, *PClass*, *Sex*, *Fare* y *Embarked*. La variable de respuesta es *Survived* que indica si los sujetos sobrevivieron o no.

Las variables *Name*, *SibSp*,*Parch*,*Title* no aportan información significativa para realizar predicciónes y es por ello que no los vamos a utilizar.

Seleccionamos el conjunto de datos de interés, para luego predecir la posibilidad de supervivencia.

```
df = complete2_df[,c(1,2,4,5,9,10)]
```

A continuación, se especifica los análisis que vamos a realizar en torno a nuestro conjunto de datos:

Análisis 1: comparar cuántos han sobrevivido en función del género y en función de la clase (primera, segunda o tercera) en la que estaban. Este análisis requeriría realizar una tabla de contingencia y una chi cuadrado, para ver si ambas variables están relacionadas. Este análisis es no parámetrico por lo tanto no habría que comprobar los supuestos de normalidad y homogeneidad de la varianza para realizar este análisis.

Análisis 2: Analizar si existen diferencias en la edad entre los sujetos que han sobrevivido y los que no han sobrevivido. Para ello realizaríamos una prueba de diferencia de medias. Podríamos hacer una prueba t (si se cumplen los supuestos de homogeneidad de varianzas y de normalidad de la variable dependiente), o si no se cumplen los supuestos podríamos utilizar una prueba no paramétrica como por ejemplo la prueba U de Mann-Whitney.

Análisis 3: Calcular un modelo de regresión logística para predecir la supervivencia, para contrastar con la predicción de supervivencia utilizando un modelo de árbol de decisión:

2.4.2. Comprobación de la normalidad y homogeneidad de las varianzas.

Tal y como hemos comentado, el Análisis 2 implica realizar una comparación de medias. Para poder utilizar una prueba t hemos de comprobar los supuestos de normalidad y homogeneidad de varianzas de la variable dependiente, en este caso edad. Existen muchas opciones de comprobar la normalidad, una opción viable sería utilizar la prueba de Kolmogorov-Smirnov con la corrección de Lilliefors.

```
lillie.test(df$Age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Age
## D = 0.071761, p-value = 1.031e-11
```

```
lillie.test(df$Age[df$Survived == 0])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Age[df$Survived == 0]
## D = 0.091659, p-value = 6.721e-12
```

```
lillie.test(df$Age[df$Survived == 1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Age[df$Survived == 1]
## D = 0.070615, p-value = 0.0003088
```

Los resultados, muestran que la distribución de la variable edad no es normal ni en la muestra en su conjunto (K-S(891)= 0,16; p < 0,001), ni en el subconjunto de los no supervivientes (KS(549)=0,18; p < 0,001) ni en el subconjunto de supervivientes (KS(342)=0,10; p < 0,001).

** Notese que la imputación realizada en la variable AGE es diferente cada vez que se ejecuta el código, por lo que la interpretación de los datos no encajará con los resultados obtenidos **

Comprobemos ahora la homogeneidad de las varianzas de la edad en los dos grupos que queremos comparar. Para ello podemos utilizar la prueba F de Levene.

```
leveneTest(df$Age, df$Survived)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.6312 0.2019
##       889
```

Tal y como se puede ver en los resultados [F(1, 889) = 0,029; p = 0.86] sí se cumple el supuesto de homogeneidad de varianzas.

** Notese que la imputación realizada en la variable AGE es diferente cada vez que se ejecuta el código, por lo que la interpretación de los datos no encajará con los resultados obtenidos **

A la vista de estos resultados tenemos dos opciones, utilizar la prueba U de Mann Whitney que como es una prueba no paramétrica no necesita que se cumplan los supuestos; o utilizar la prueba t, que aunque es cierto que no se cumple la normalidad, teniendo una muestra tan amplia podríamos no tener en cuenta este hecho ya que la prueba t es robusta ante el inclumplimiento de este supuesto, sobre todo con muestras grandes.

2.4.3. Aplicación de pruebas estadísticas

2.4.3.1. ¿Cuántos han sobrevivido en función del género y en función de la clase?

El objetivo de este primer análisis es ver si existe relación entre la supervivencia y las variables de sexo y de clase.

Survived vs Sex

Para contrastar la hipótesis de independencia entre género y supervivencia, realizaremos una tabla de contingencia y calcularemos el estadístico de chi cuadrado, para ello usaremos la función *chisq.test* de R.

```
summary(df$Sex)
```

```
## female   male
##    314    577
```

```
cuadro<-table(df$Survived, df$Sex)
# Frecuencias esperadas
chisq.test(cuadro)$expected
```

```
##
##       female     male
##   0 193.4747 355.5253
##   1 120.5253 221.4747
```

```
# Residuos estandarizados
chisq.test(cuadro)$residuals
```

```
##
##        female      male
##   0 -8.086170  5.965128
##   1 10.245095 -7.557757
```

```
# Chi cuadrado
chisq.test(cuadro)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cuadro
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Tal y como podemos ver en los resultados (X-squared(1)=260,72; p < 0,001) existe una relación estadísticamente significativa entre las variables género y supervivencia. Es decir, el género y la supervivencia no son independientes.

Concretamente si nos fijamos en los residuos estandarizados veremos que hay más mujeres de las esperadas por azar que sobrevivieron, mientras que hay más hombres de los esperados por azar que no sobrevivieron. Es decir, observamos que las supervivientes femeninas eran mucho más numerosas que los hombres.

Survived vs Pclass

A continuación realizaremos el mismo análisis para contrastar la hipótesis de independencia entre clase y sobrevivencia, por lo tanto, volveremos a utilizar la función *chisq.test* de R.

```
summary(df$Pclass)
```

```
##   1   2   3
## 216 184 491
```

```
cuadro<-table(df$Survived, df$Pclass)
# Frecuencias esperadas
chisq.test(cuadro)$expected
```

```
##
##             1         2        3
##   0 133.09091 113.37374 302.5354
##   1  82.90909  70.62626 188.4646
```

```
# Residuos estandarizados
chisq.test(cuadro)$residuals
```

```
##
##             1         2         3
##   0 -4.601993 -1.537771  3.993703
##   1  5.830678  1.948340 -5.059981
```

```r
# Estadístico de chi cuadrado
chisq.test(cuadro)
```

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

Tal y como podemos ver en los resultados (chi-squared(2)=102.89; p <0,001) la relación entre las variables clase y supervivencia es estadísticamente significativa. Es decir, la clase y la supervivencia no son independientes.

Concretamente, si observamos los residuos estandarizados podemos ver que hay mas sujetos de los esperados por azar que sobrevivieron y estaban en primera y segunda clase, asimismo hay más sujetos de los esperados por azar que no sobrevivieron y eran de tercera clase. Es decir, observamos que los pasajeros que viajarón en primera y segunda clase tuvieron mayor probabilidad de supervivencia.

2.4.3.2.¿Existen diferencias significativas de edad entre los supervivientes?

Esta prueba consistirá en un contraste de hipótesis sobre dos muestras para determinar si existen diferencias de edad entre los sujetos que han sobrevivido y los que no.

A continuación se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias,

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

donde μ es igual a la media de edad de edad sujetos que han sobrevivido. μ_0 es igual a la media de edad de edad sujetos que no han sobrevivido.

Al tener una muestra de tamaño superior a 30, consideramos aplicar el t test.

```r
a <- t.test(complete2_df$Age[complete2_df$Survived == 0],
       complete2_df$Age[complete2_df$Survived == 1],
       paired=FALSE,
       var.equal=TRUE,
       conf.level=0.95,
       alternative = "two.sided")

a
```

```
##
##  Two Sample t-test
##
## data:  complete2_df$Age[complete2_df$Survived == 0] and complete2_df$Age[complete2_df$Survived == 1]
## t = 1.7202, df = 889, p-value = 0.08575
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2332675  3.5430297
## sample estimates:
## mean of x mean of y
##  30.43169  28.77681
```

```r
supervivientes <- complete2_df$Age[complete2_df$Survived == 0]
no_supervivientes <- complete2_df$Age[complete2_df$Survived == 1]

# Tamaño del efecto
a$statistic*sqrt(1/length(supervivientes) + 1/length(1/no_supervivientes))
```

```
##         t
## 0.1184977
```

Tal y como podemos ver en los resultados de la prueba t si nos fijamos en la prueba de hipótesis diríamos que existen diferencias estadísticamente significativas en la media de edad de las personas que han sobrevivido y las personas que no han sobrevivido. Concretamente podemos ver que la media de edad de las personas que han sobrevivido es de 28,04 años, mientras que la media de edad de los que no han sobrevivido es de 30,51 años. Sin embargo, si nos fijamos en el tamaño del efecto de esta diferencia podemos ver que el valor de la d de Cohen es reducida, lo que implica que aunque la diferencia resulte estadísticamente significativa es muy pequeña.

** Notese que la imputación realizada en la variable AGE es diferente cada vez que se ejecuta el código, por lo que la interpretación de los datos no encajará con los resultados obtenidos **

2.4.3.2. Modelo de regresión logística

Finalmente se calculará un modelo de regresión logística utilizando regresores tanto cuantitativos como cualitativos, para analizar los regresores con mayor influencia al realizar las predicciones de los supervivientes.

Definimos una partición de los datos (entrenamiento/test) para ver cómo clasifica los datos de test.

```r
ind<-sample(1:dim(df)[1],500) # Sample of 500 out of 891
train1<-df[ind,] # conjunto train del modelo
test1<-df[-ind,] # conjunto test
```

Una vez que tenemos la partición realizada crearemos el modelo de regresión logística utilizando la función del modelo lineal general *glm*. En esta función se pueden specificar diferentes tipos de distribuciones y funciones, especificaremos la función binomial, que por defecto asume la función logit. Trataremos de predecir la probabilidad de que una persona sea superviviente basándonos en las variables Pclass, Sex, Age, Fare y Embarked.

```r
modelo <- glm(Survived~.,family=binomial(link='logit'),data=train1)
summary(modelo)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3011  -0.7251  -0.4294   0.7535   2.3445
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.1578551  0.5720045   5.521 3.38e-08 ***
```

```
## Pclass2     -0.8966573  0.3791266  -2.365   0.0180 *
## Pclass3     -2.0736507  0.3829618  -5.415 6.14e-08 ***
## Sexmale     -2.2937339  0.2433475  -9.426  < 2e-16 ***
## Age         -0.0181415  0.0087863  -2.065   0.0389 *
## Fare         0.0007555  0.0029797   0.254   0.7998
## EmbarkedQ    0.3075330  0.4939587   0.623   0.5336
## EmbarkedS   -0.6621069  0.3031940  -2.184   0.0290 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 664.06  on 499  degrees of freedom
## Residual deviance: 478.67  on 492  degrees of freedom
## AIC: 494.67
##
## Number of Fisher Scoring iterations: 4
```

```
sel <- which(summary(modelo)$coefficients[-1,4] < 0.05)
sel <- sel + 1
```

Con los resultados obtenidos en nuestro modelo de regresión logística podemos observar que no todas las predictoras son estadísticamente significativas. Así, vemos que ha sido significativo el test parcial sobre los coeficientes de Pclass2, Pclass3, Sexmale, Age, EmbarkedS. Siendo la estimacion de su coeficiente -0.897, -2.074, -2.294, -0.018, -0.662.

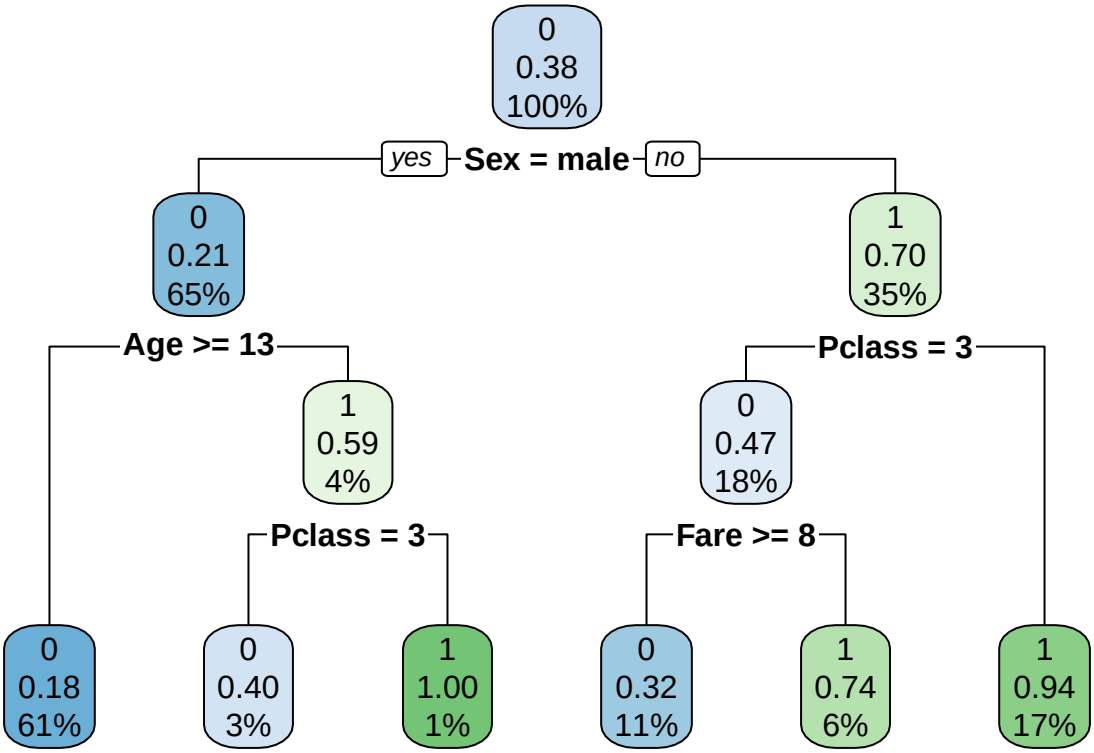
```
pred.train <- predict(modelo,test1)
pred.train <- ifelse(pred.train > 0.5,1,0)

# Media de la predicción verdadera
mean(pred.train==test1$Survived)
```

```
## [1] 0.8209719
```

Por lo tanto podemos ver que el porcentaje de predicciones correctas es muy elevada.

Ahora vamos a predecir la supervivencia utilizando un modelo de árbol de decisión. De nuevo utilizaremos la misma partición y las mismas variables.

```
model_arbol<- rpart(Survived ~.,data=train1, method="class")
rpart.plot(model_arbol)
```




```
pred.train.dt <- predict(model_arbol,test1,type = "class")
mean(pred.train.dt==test1$Survived)
```

```
## [1] 0.8235294
```

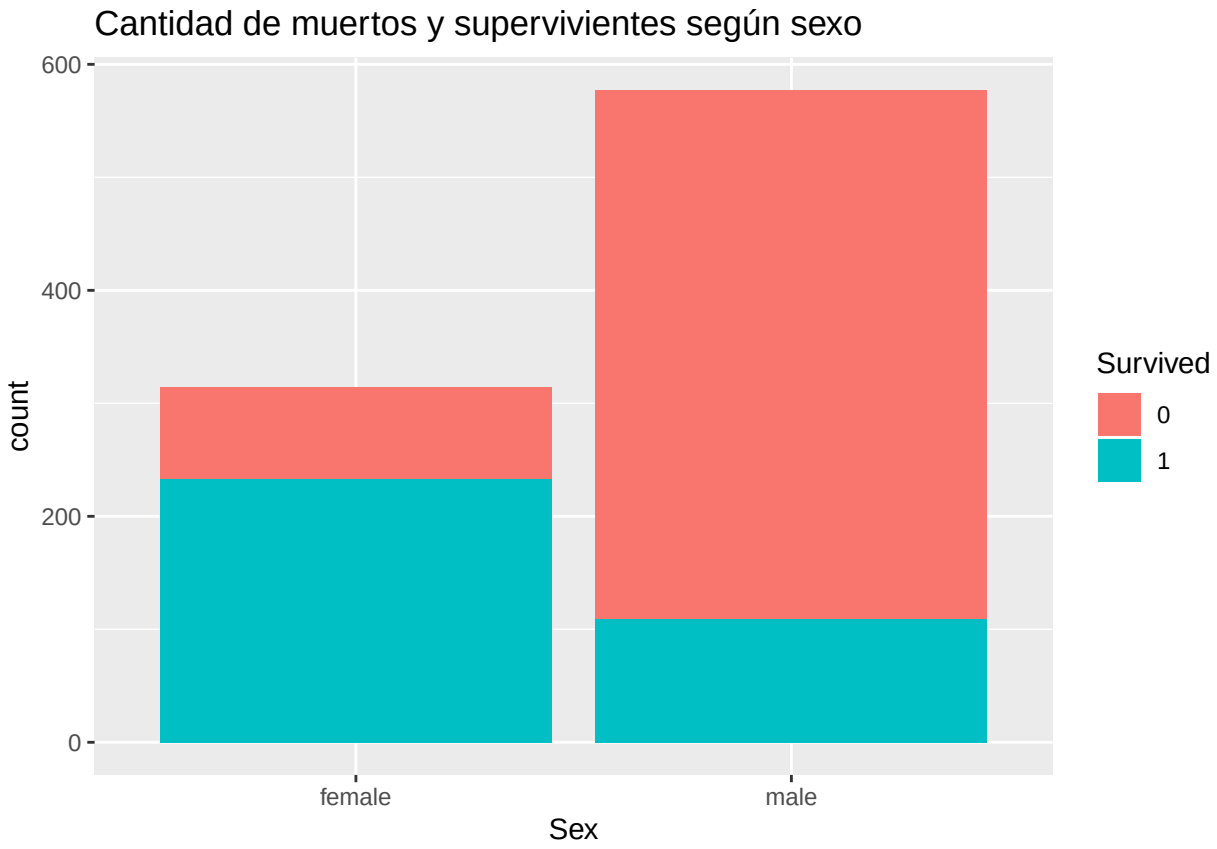
Vemos que con este modelo el porcentaje de los correctamente clasificados también es elevado.

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

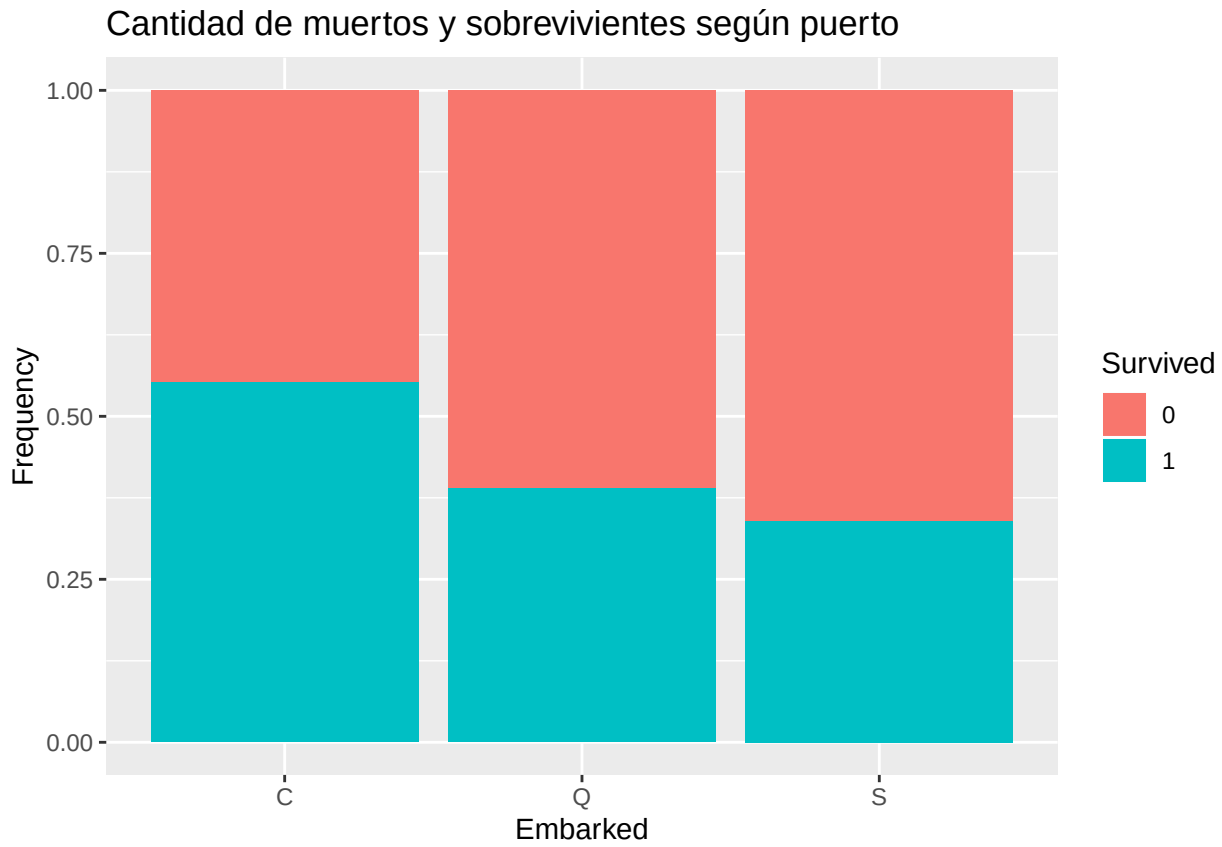
```
ggplot(data = df, mapping = aes(x = Pclass, fill = factor (Sex))) +
  geom_bar() +
  facet_wrap(~ Survived) +
  ggtitle("Cantidad de muertos y supervivientes según sexo y clase")
```

2.5.2.

```
ggplot(data=df,aes(x=Sex,fill=Survived)) +
  geom_bar() +
  ggtitle("Cantidad de muertos y supervivientes según sexo ")
```




```
ggplot(data=df, aes(x=Embarked,fill=Survived)) +
  geom_bar(position="fill") +
  ylab("Frequency") +
  ggtitle("Cantidad de muertos y sobrevivientes según puerto ")
```



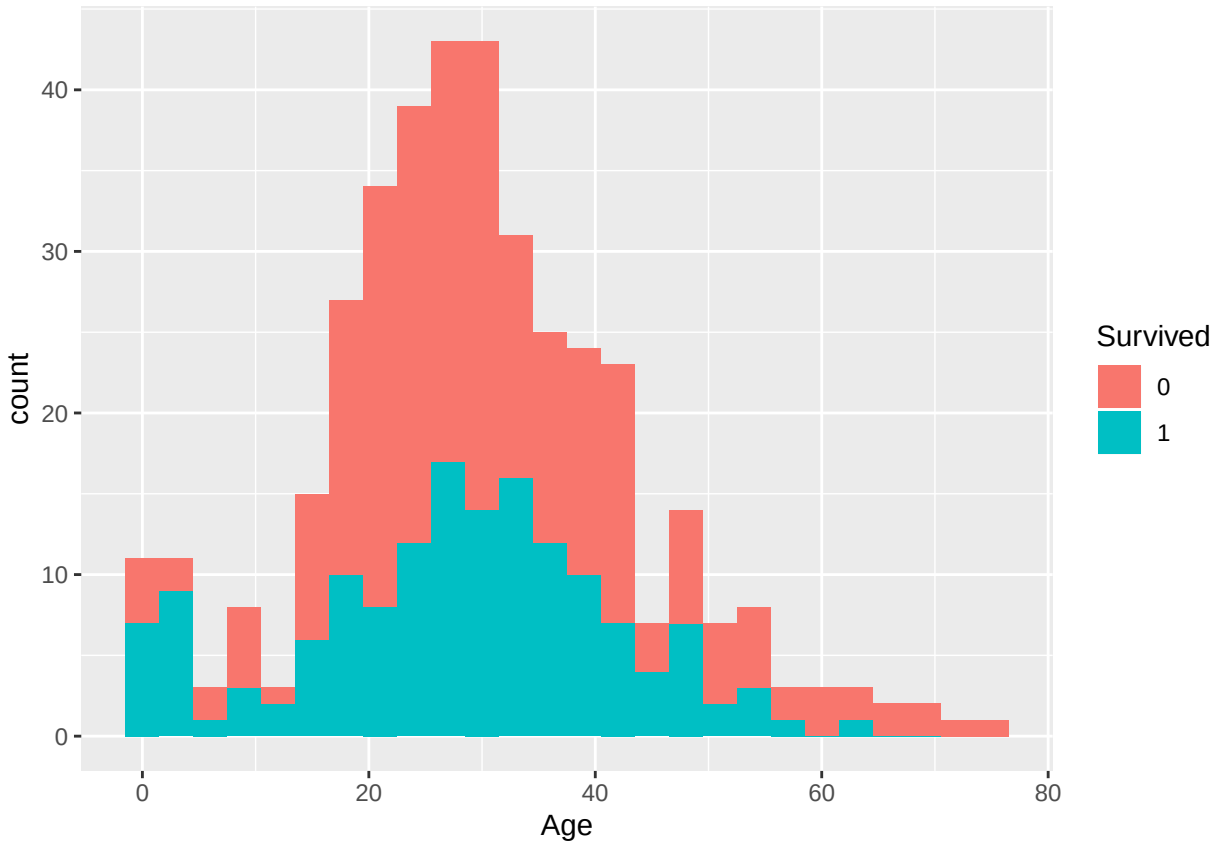
Vemos, por ejemplo que la probabilidad de sobrebivir si se embarcó en "C" es de un 54,90 %

```
filas=dim(train1)[1]
t<-table(train1[1:filas,]$Embarked,train1[1:filas,]$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

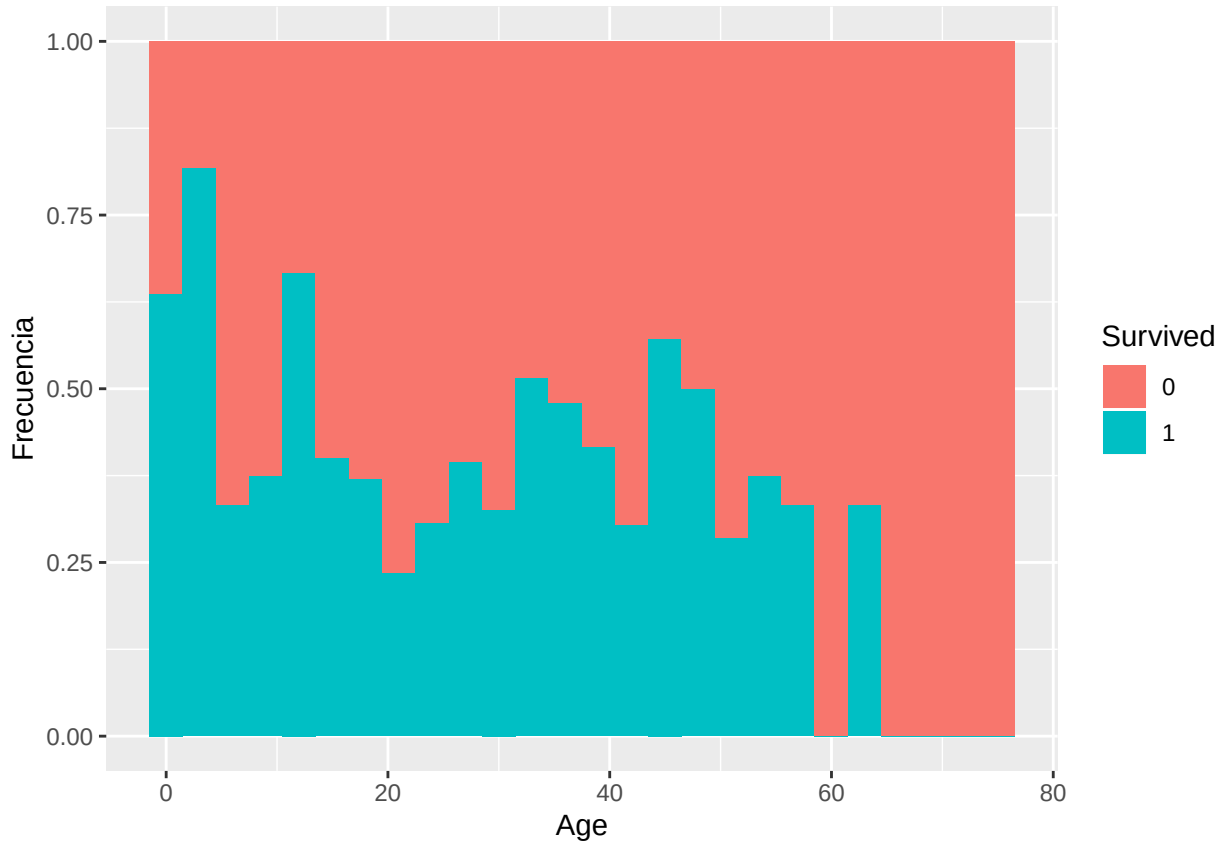
```
##
##            0        1
##
##   C 43.47826 56.52174
##   Q 51.42857 48.57143
##   S 67.56032 32.43968
```

Veamos ahora dos gráficos que nos compara los atributos Age y Survived.

```
# Survival como función de age:
ggplot(data = test1[!(is.na(test1[1:filas,]$Age)),],aes(x=Age,fill=Survived))+geom_histogram(binwidth =
```




```
ggplot(data = test1[!is.na(test1[1:filas,]$Age),],aes(x=Age,fill=Survived))+geom_histogram(binwidth = 3
```



Modelo de regresión logística

```
t1<-table(pred.train,test1$Survived)
# Presicion and recall del modelo
presicion<- t1[1,1]/(sum(t1[1,]))
recall<- t1[1,1]/(sum(t1[,1]))
presicion
```

```
## [1] 0.8050542
```

```
#
recall
```

```
## [1] 0.9330544
```

```
#F1 score
F1<- 2*presicion*recall/(presicion+recall)
F1
```

```
## [1] 0.8643411
```

El puntaje F1 en el conjunto de prueba inicial es 0.84, lo que es muy bueno.

Resultados árbol de decisión

```
t2<-table(pred.train.dt,test1$Survived)
presicion_ad<- t2[1,1]/(sum(t2[1,]))
recall_ad<- t2[1,1]/(sum(t2[,1]))
presicion_ad
```

```
## [1] 0.7972028
```

```
#
recall_ad
```

```
## [1] 0.9539749
```

```
#F1 score
F1<- 2*presicion_ad*recall_ad/(presicion+recall)
F1
```

```
## [1] 0.8751024
```

Como ejemplo probamos nuestro modelo de regresión logística para predecir la sobrevivencia de una mujer edad igual a 18 años que viajó en 1ra clase pagando 200.

```
newd=data.frame(Age= 18 , Sex="female" , Pclass= "1", Embarked="C" , Fare = 200 )

predict<- predict(modelo, newd, type= "response")
predict
```

```
##        1
## 0.951773
```

La predicción de la probabilidad supervivencia para este caso es 0.9518.

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

A partir de los ánalisis realizados podemos indicar que los resultados permiten responder a los planteamientos iniciales del problema, conocer las variables con mayor influencia para la predicción de la probabilidad de supervivencia de los pasajeros del Titanic.

Se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos, como se ha visto las variables no fueron seleccionadas a priori ya que a lo largo del estudio, limpieza y ánalisis se ha ido evaluando el comportamiento de las variables. Para cada una de ellas, hemos podido ver qué conocimientos aportan.

Hemos de tener en cuenta que por un lado la imputación de los valores perdidos de la edad y por otra la partición que hemos realizado en la base de datos son dinámicas, es decir cambian cada vez que ejecutamos el código por lo que la interpretación concreta que viene a continuación no encajará con los resultados al volver a ejecutar la sintaxis.

En el modelo logístico que hemos elaborado utilizando como variables predictoras la clase, el sexo, la edad, la tarifa y el puerto de embarque se ha podido apreciar que las variables clase, sexo y edad son las más importantes a la hora de determinar la probabilidad de que un sujeto sobreviva o no a la tragedia del Titanic. Concretamente se ha visto que estar en primera o segunda clase en comparación a tercera clase hacía que

aumentará la probabilidad de supervivencia, asimismo el ser mujer en comparación con ser hombre también aumentaba la probabilidad de supervivencia, y finalmente la edad decrecía dicha probabilidad, es decir a mayor edad menor era la probabilidad de sobrevivir.

La media de las predicciones correctas que obtuvimos en el conjunto de pruebas ha sido muy elevada tanto con el método del árbol de decisión como com el modelo de regresión logística.

2.7. Código en R

El código de resolución de la práctica y el pdf de respuestas se encuentran en el repositorio GitHub, pueden ser accedidos a través de este enlace

2.7.1. Tabla de contibuciones al trabajo

Contribuciones	Firma
Investigación previa	J.A.L, N.N.S.P
Redacción de las respuestas	J.A.L, N.N.S.P
Desarrollo código	J.A.L, N.N.S.P