Tipología y ciclo de vida de los datos: Práctica 2

Jorge Marchán Gutiérrez Rafael Jiménez Sarmentero

mayo 2022

Contents

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

El dataset elegido para la realización de la práctica ha sido el de Titanic que contiene una serie de datos sobre los pasajeros del Titanic, entre otras cosas, si finalmente sobrevivieron o no, los datos se dividen en varios ficheros `train.csv` y `test.csv`, además de un tercer fichero `gender_submission.csv` que para la realización de esta práctica no es necesario, ya que es un ejemplo de fichero de envío para la competición, de Kaggle. A nosotros nos interesa el fichero de `train.csv`, sobre el cual vamos a realizar las tareas de limpieza y análisis.

Con este *dataset* se podrían encontrar relaciones entre supervivencia y edad, o supervivencia y género, entre otras, o se podría utilizar para entrenar un modelo capaz de predecir, si una persona con unas características determinadas sobrevivió al accidente o no.

```
data <- read.csv("./input_files/train.csv", header = TRUE, stringsAsFactors = FALSE)
dim(data)
```

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

```
head(data)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                                   Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

Podemos observar que el *dataset* contiene 891 filas y 12 atributos, a continuación vamos a ver los tipos de atributos y su significado

```
str(data)
```

```
## '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" ...
```

Los atributos que encontramos son:

- **PassengerId**: Es el identificador interno del pasajero, de tipo entero
- **Survived**: Es un valor de tipo entero que nos indica si el pasajero ha sobrevivido o no (0 o 1)
- **Pclass**: El tipo de billete que ha adquirido el pasajero, tipo entero (1 = Primera, 2 = Segunda, 3 = Tercera)
- **Name**: El nombre del pasajero, tipo char
- **Sex**: El género del pasajero, tipo char (male o female)
- **Age**: La edad del pasajero, tipo number
- **SibSp**: El numero de hermanos y conyuges que hay abordo en el Titanic, tipo entero
- **Parch**: El número de padres e hijos que hay abordo en el Titanic, tipo entero
- **Ticket**: El identificador del billete, tipo char

- **Fare**: El precio del billete, tipo number
- **Cabin**: El código del camarote, tipo char
- **Embarked**: El puerto donde embarco el pasajero, tipo char (C = Cherbourg, Q = Queenstown, S = Southampton)

Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

De los atributos presentados, sin realizar ningún trabajo previo, consideramos que la siguiente lista de atributos no es relevante para el análisis estadístico que queremos llevar a cabo:

- **PassengerId**: La podemos eliminar del conjunto de datos ya que no contribuye a la supervivencia del pasajero
- **Ticket**: Por los mismos de PassengerId, consideramos que los identificadores internos no afectan a la supervivencia
- **Name**: Por si solo el nombre del pasajero creemos que no aporta nada a la supervivencia del mismo, sin embargo observamos que todos los nombres siguen un formato determinado y que todos contienen el titulo que se aplica a la persona, por lo tanto podríamos extraer esta característica para contar con un *dataset* con más información con la que trabajar.
- **Cabin**: Del camarote podemos llegar a saber qué pasajeros viajaban en el mismo y si han sobrevivido o no, por lo tanto podemos saber si el camarote o el tipo de camarote están relacionados con una mayor supervivencia.
- **SibSp y Parch**: Estas dos variables podemos condensarlas en una sola, que hace referencia al número de familiares que el pasajero tenía a bordo.

Limpieza de los datos.

¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Del análisis del fichero que contiene el *dataset* `train.csv` podemos extraer la siguiente información:

1. Algunas cadenas de caracteres tienen espacios en blanco al inicio y/o final.
2. Los valores decimales están separados por el carácter ".".
3. La edad puede contener valores decimales al ser de tipo number y no entero.
4. El separador de columnas es el carácter ",".

Para la limpieza de los datos resulta interesante conocer qué atributos contienen valores vacíos:

```
colSums(data == "")
```

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

Con esta información y con lo que conocemos del *dataset* podemos concluir que las siguientes transformaciones serían interesantes con el objetivo de facilitar el análisis:

1. El atributo `Survived` debería ser un factor debido a que es cualitativa con valores `1` y `0`
2. El atributo `Pclass` debería ser un factor debido a que es cualitativa con valores `1`, `2` y `3`
3. El atributo `Sex` debería ser un factor debido a que es cualitativa con valores `male` y `female`
4. El atributo `Embarked` debería ser un factor debido a que es cualitativa con valores `C`, `Q` y `S`, además de que deberíamos cambiar los valores vacíos por `NA`
5. El atributo `Cabin` contiene valores vacíos por lo que hay que reemplazarlos por `NA`.

En primer lugar, deberíamos reemplazar los valores que consideramos vacíos por `NA`:

```
data$Cabin[data$Cabin == ""] <- NA
data$Embarked[data$Embarked == ""] <- NA
```

Comprobamos ahora cuántos datos vacíos (NA) tiene cada atributo:

```
colSums(is.na(data))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

Podemos prescindir del atributo `PassengerId`, que no tiene utilidad de cara a análisis estadísticos:

```
data$PassengerId <- NULL
```

Convertimos a `Factor` las variables categóricas:

```
data$Survived <- as.factor(data$Survived)
data$Pclass <- as.factor(data$Pclass)
data$Sex <- as.factor(data$Sex)
data$Embarked <- as.factor(data$Embarked)

str(data)
```

```
## 'data.frame':    891 obs. of  11 variables:
##  $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name    : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "H
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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  NA "C85" NA "C123" ...
##  $ Embarked: Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

Podemos incluir un atributo nuevo que nos indique el tamaño de la familia que viaja a bordo de cada pasajero. De cara al análisis posterior es más interesante tener el dato agrupado en un único atributo que en varios.

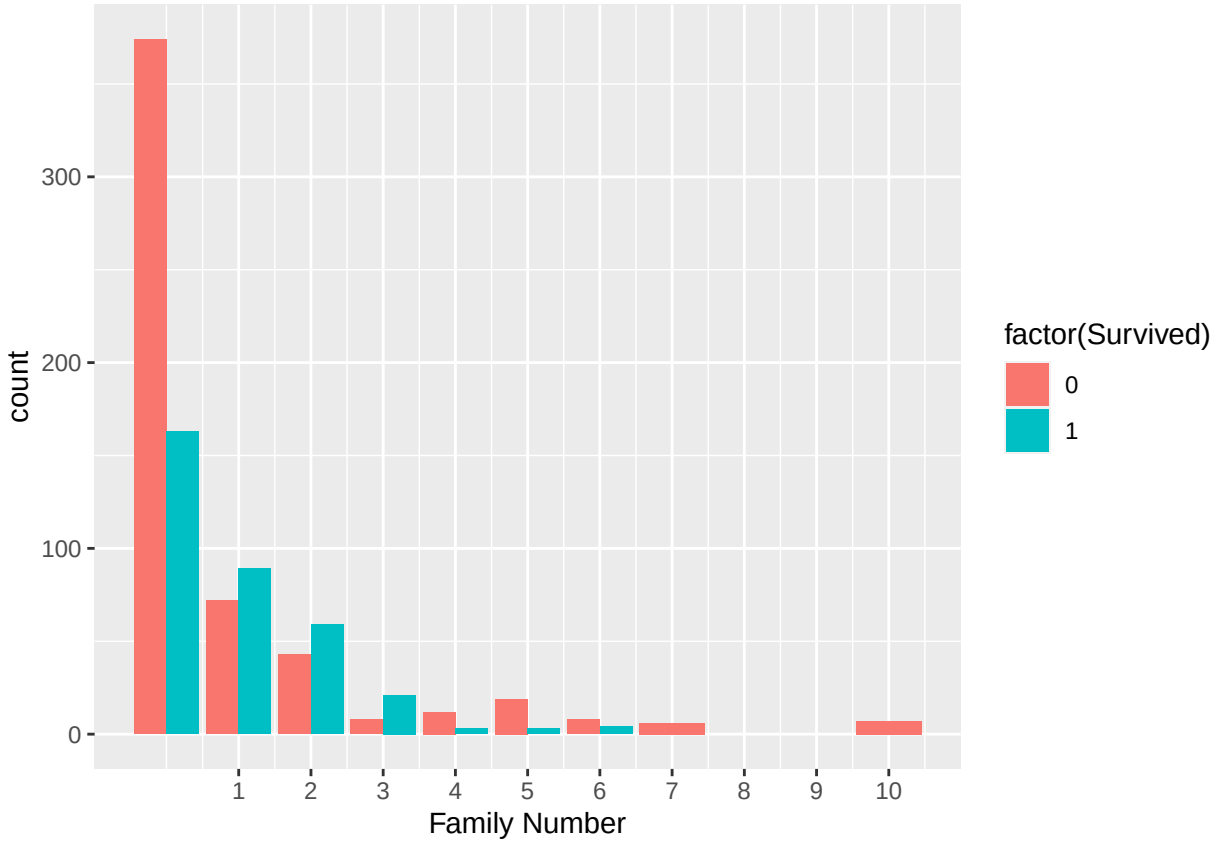
```
data$Fnumber <- data$SibSp + data$Parch
data$SibSp <- NULL
data$Parch <- NULL
```

Vamos a comprobar visualmente si puede existir una relación entre la variable `Survived` y el número de familiares:

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
```

```
## Loading required package: ggplot2
```

```
ggplot(data, aes(x = Fnumber, fill = factor(Survived))) +
  geom_bar(stat='count', position='dodge') +
  scale_x_continuous(breaks=c(1:11)) +
  labs(x = 'Family Number')
```



Observando los resultados nos damos cuenta de que los pasajeros que viajaban solos tenían más probabilidades de no sobrevivir que de sobrevivir, así como también ocurre con los pasajeros de más de 3 familiares a bordo, por lo que podemos crear tres categorías:

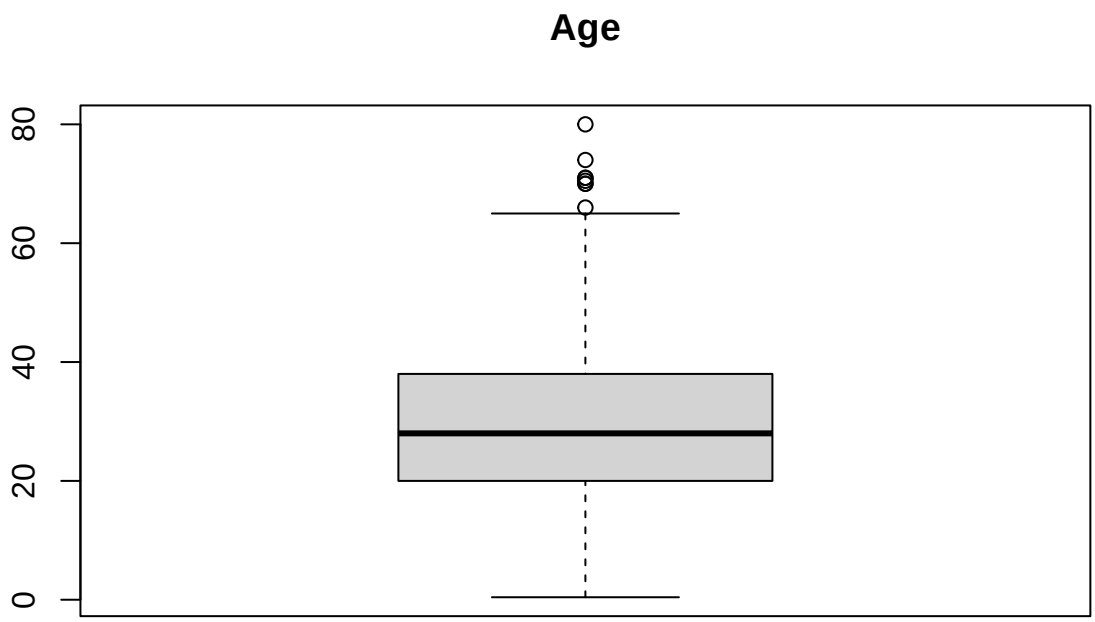
```
data$Ftype[data$Fnumber == 0] <- 'single'
data$Ftype[data$Fnumber < 4 & data$Fnumber > 0] <- 'small'
data$Ftype[data$Fnumber >= 4] <- 'large'

data$Ftype <- as.factor(data$Ftype)
```

Identifica y gestiona los valores extremos.

Tenemos tres atributos numéricos: `Age`, `Fare` y `Fnumber`. A continuación vamos a visualizar con boxplot cada una de las variables y a realizar su análisis para determinar si los valores extremos son correctos o son fallos:

```
boxplot(data$Age, main="Age")
```




```
boxplot.stats(data$Age)$out
```

```
## [1] 66.0 71.0 70.5 71.0 80.0 70.0 70.0 74.0
```

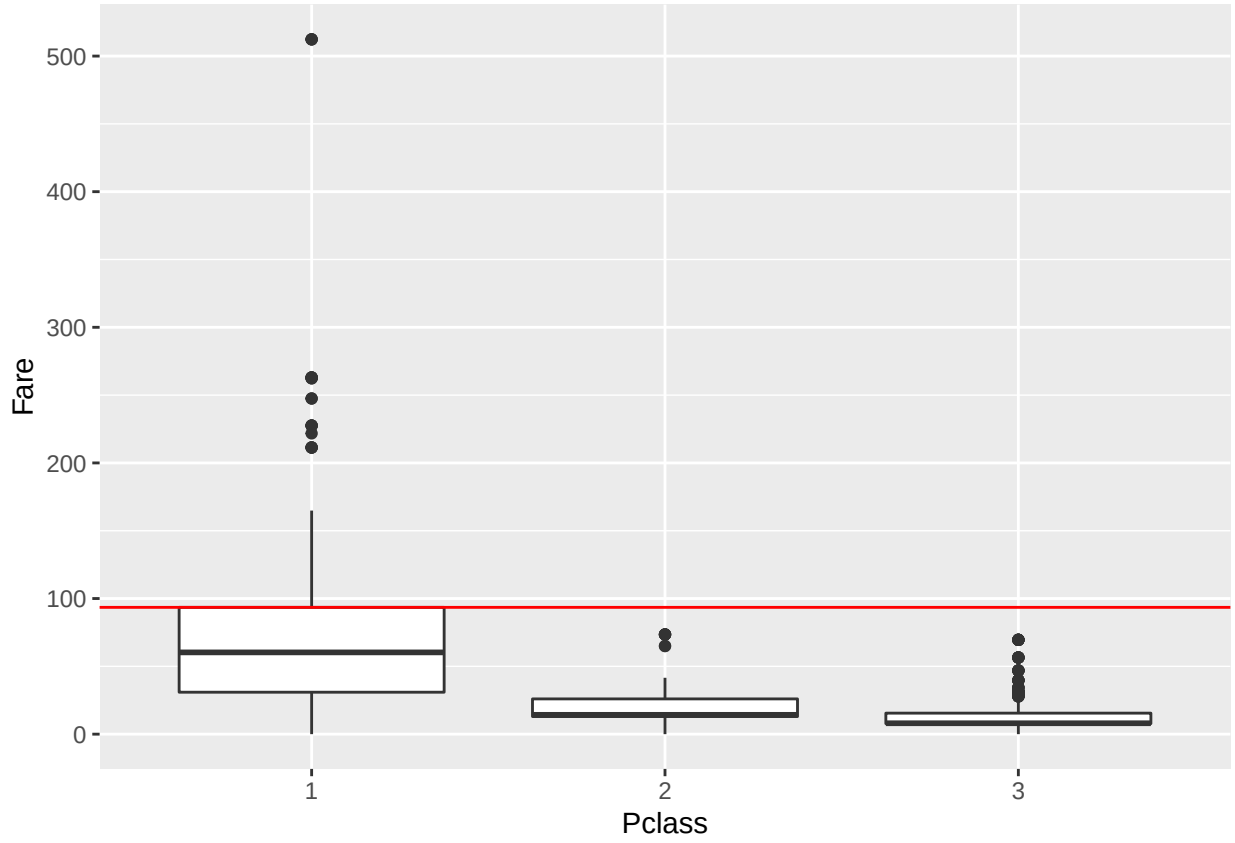
El atributo `Age` nos muestra que los valores extremos son aquellos que están por encima de 66 años. Sin embargo, no observamos ningún valor que aparentemente sea incorrecto. Podemos extraer la conclusión de que era raro ver pasajeros de más de 66 años, pero estos datos no necesitan ser tratados.

Vamos a analizar ahora los valores extremos del atributo `Fare`:

```
boxplot(data$Fare, main="Fare")
```

En un primer lugar observamos los valores extremos del atributo `Fare` y vemos que por encima de 90 se consideran *outliers*. Sin embargo, el precio del billete depende de la clase del mismo, y hay menos de primera clase que del resto; por lo tanto, es probable que los billetes de primera clase se consideren valores extremos, de modo que analizaremos los boxplot por clase y marcaremos el límite para considerarlos *outliers* en el atributo `Fare`:

```
q3 <- quantile(x=data$Fare[data$Pclass == 1], 0.75)
ggplot(data, aes(x=Pclass, y=Fare)) +
  geom_boxplot() +
  geom_hline(aes(yintercept=q3), colour='red')
```



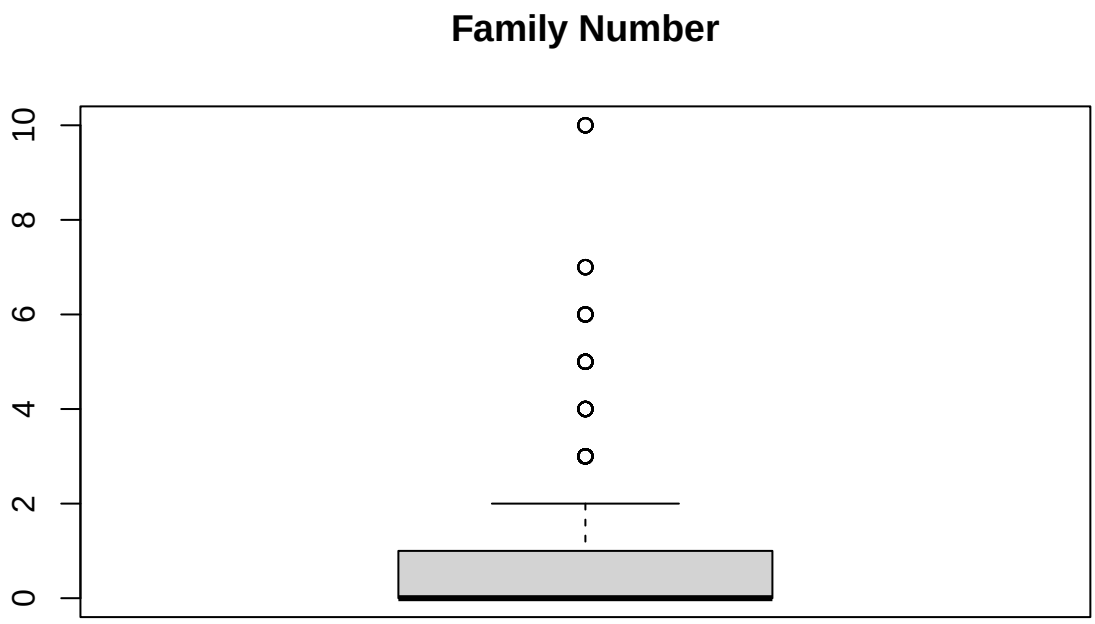

```
print(data[data$Fare > 200,])
```

```
##     Survived Pclass                                                Name
## 28         0      1                     Fortune, Mr. Charles Alexander
## 89         1      1                         Fortune, Miss. Mabel Helen
## 119        0      1                           Baxter, Mr. Quigg Edmond
## 259        1      1                                    Ward, Miss. Anna
## 300        1      1       Baxter, Mrs. James (Helene DeLaudeniere Chaput)
## 312        1      1                         Ryerson, Miss. Emily Borie
## 342        1      1                      Fortune, Miss. Alice Elizabeth
## 378        0      1                           Widener, Mr. Harry Elkins
## 381        1      1                              Bidois, Miss. Rosalie
## 439        0      1                                  Fortune, Mr. Mark
## 528        0      1                                Farthing, Mr. John
## 558        0      1                               Robbins, Mr. Victor
## 680        1      1                 Cardeza, Mr. Thomas Drake Martinez
## 690        1      1                  Madill, Miss. Georgette Alexandra
## 701        1      1     Astor, Mrs. John Jacob (Madeleine Talmadge Force)
## 717        1      1                      Endres, Miss. Caroline Louise
## 731        1      1                      Allen, Miss. Elisabeth Walton
## 738        1      1                             Lesurer, Mr. Gustave J
## 743        1      1                Ryerson, Miss. Susan Parker "Suzette"
## 780        1      1 Robert, Mrs. Edward Scott (Elisabeth Walton McMillan)
##        Sex Age   Ticket     Fare           Cabin Embarked Fnumber  Ftype
## 28    male  19    19950 263.0000     C23 C25 C27        S       5  large
## 89  female  23    19950 263.0000     C23 C25 C27        S       5  large
```

```
## 119   male  24 PC 17558 247.5208         B58 B60        C       1  small
## 259 female  35 PC 17755 512.3292            <NA>        C       0 single
## 300 female  50 PC 17558 247.5208         B58 B60        C       1  small
## 312 female  18 PC 17608 262.3750 B57 B59 B63 B66        C       4  large
## 342 female  24    19950 263.0000     C23 C25 C27        S       5  large
## 378   male  27   113503 211.5000             C82        C       2  small
## 381 female  42 PC 17757 227.5250            <NA>        C       0 single
## 439   male  64    19950 263.0000     C23 C25 C27        S       5  large
## 528   male  NA PC 17483 221.7792             C95        S       0 single
## 558   male  NA PC 17757 227.5250            <NA>        C       0 single
## 680   male  36 PC 17755 512.3292     B51 B53 B55        C       1  small
## 690 female  15    24160 211.3375              B5        S       1  small
## 701 female  18 PC 17757 227.5250         C62 C64        C       1  small
## 717 female  38 PC 17757 227.5250             C45        C       0 single
## 731 female  29    24160 211.3375              B5        S       0 single
## 738   male  35 PC 17755 512.3292            B101        C       0 single
## 743 female  21 PC 17608 262.3750 B57 B59 B63 B66        C       4  large
## 780 female  43    24160 211.3375              B3        S       1  small
```

Observamos cómo los valores considerados *outliers* para `Fare` están asociados con la clase en la que viajan: cuanto más alta es la clase y mayor número de pasajeros comparten billete, más alta es la tarifa. Por lo tanto, los valores extremos en este atributo son valores que consideramos válidos.

```
boxplot(data$Fnumber, main="Family Number")
```




```
boxplot.stats(data$Fnumber)$out
```

```
##  [1]  4  6  5  4  6  5  3  5  3  7  5  6  7  3  4  5  3  6  4 10  5  5  5  4 10
```

```
## [26]  6  3 10  4  6  6  5  5  3  3  4 10  5  5  4  7  3  4  3  4  5  5  3  3  3
## [51]  3  7  4  3  3  6  6  4  3  3  6  3  3  5  5  5  3  7  7  3  5  3  4  4  3
## [76]  3  4  3  5  3 10  3  6  5  5 10  6  3 10  5  3
```

Observamos cómo el boxplot nos cataloga como valores extremos todos aquellos pasajeros que viajasen con 3 familiares más. Sin embargo, no parece ser un dato incorrecto. Quizá 10 familiares es un poco sospechoso, por lo que veamos los pasajeros con `Fnumber` = 10 existentes en el *dataset*:

```
data[data$Fnumber == 10,]
```

```
##     Survived Pclass                                     Name    Sex Age   Ticket  Fare
## 160        0      3                Sage, Master. Thomas Henry   male  NA CA. 2343 69.55
## 181        0      3              Sage, Miss. Constance Gladys female  NA CA. 2343 69.55
## 202        0      3                        Sage, Mr. Frederick   male  NA CA. 2343 69.55
## 325        0      3                   Sage, Mr. George John Jr   male  NA CA. 2343 69.55
## 793        0      3                    Sage, Miss. Stella Anna female  NA CA. 2343 69.55
## 847        0      3                   Sage, Mr. Douglas Bullen   male  NA CA. 2343 69.55
## 864        0      3 Sage, Miss. Dorothy Edith "Dolly" female  NA CA. 2343 69.55
##     Cabin Embarked Fnumber Ftype
## 160  <NA>        S      10 large
## 181  <NA>        S      10 large
## 202  <NA>        S      10 large
## 325  <NA>        S      10 large
## 793  <NA>        S      10 large
## 847  <NA>        S      10 large
## 864  <NA>        S      10 large
```

Aquí podemos observar cómo todos los pasajeros que viajaban con 10 familiares eran familia, compartían billete y tarifa, por lo que los valores *outliers* de `Fnumber` son correctos.

Antes de proceder con el análisis, nos interesa quitar aquellos registros de los cuales no tenemos la edad, ya que pensamos que ésta es un dato que va a resultarnos de mucha utilidad en los análisis futuros.

```
data <- data[!is.na(data$Age),]
```

Análisis de los datos.

Selección de los grupos de datos que se quieren analizar/comparar (p. e., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

Como se ha comentado al principio, vamos a trabajar con el conjunto de datos de entrenamiento y vamos a analizar la relación existente entre la supervivencia y los atributos `Pclass`, `Sex`, `Embarked` y `Ftype`.

Relacion entre Survived y Pclass

```
frequency_table <- table(data$Survived, data$Pclass, dnn = c("Survived", "Pclass"))
proportions_table <- prop.table(frequency_table)
percentages_table <- round((proportions_table * 100), 2)
t <- addmargins(percentages_table)
t
```

```
##         Pclass
## Survived      1      2      3    Sum
##      0     8.96  12.61  37.82  59.39
##      1    17.09  11.62  11.90  40.61
```

```
##      Sum  26.05  24.23  49.72 100.00
```

De esta tabla de porcentajes llegamos a la conclusión de que los pasajeros que viajaban en tercera clase tenían menos posibilidades de supervivencia que los que iban en segunda y estos, menos que los que iban en primera, siendo los pasajeros de primera clase los únicos que tenían una probabilidad mayor de sobrevivir que de no sobrevivir. Por lo tanto, podemos afirmar que hay una relación entre la clase en la que se viajaba y la supervivencia.

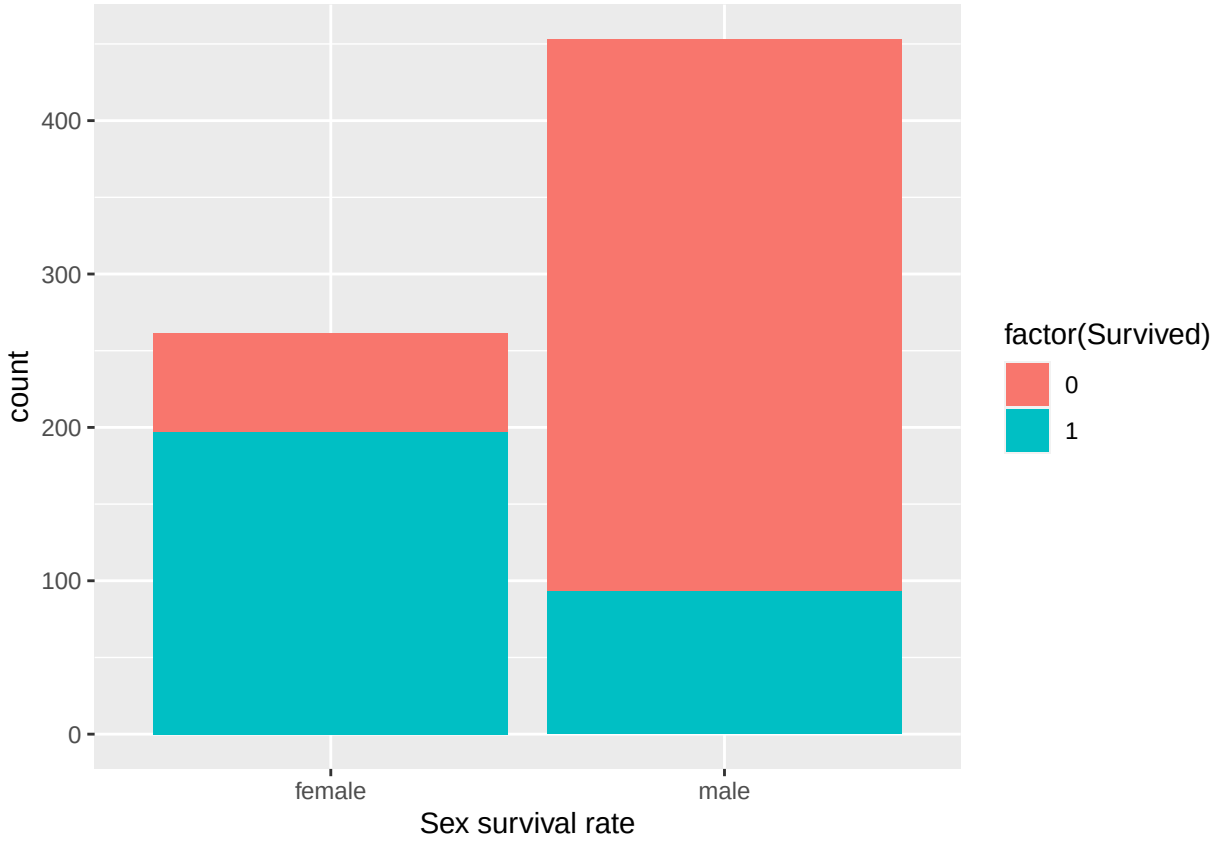
Relacion entre Survived y Sex

```
frequency_table <- table(data$Survived, data$Sex, dnn = c("Survived", "Sex"))
proportions_table <- prop.table(frequency_table)
percentages_table <- round((proportions_table * 100), 2)
t <- addmargins(percentages_table)
t
```

```
##         Sex
## Survived female   male    Sum
##      0     8.96  50.42  59.38
##      1    27.59  13.03  40.62
##      Sum  36.55  63.45 100.00
```

De esta tabla de porcentajes podemos deducir que había una relación fuerte entre el género del pasajero y su probabilidad de supervivencia: el género `female` tenía muchas más posibilidades de sobrevivir que el género `male`. Aquí las proporciones:

```
ggplot(data, aes(x = Sex, fill = factor(Survived))) +
  geom_bar(stat='count')+
  labs(x = 'Sex survival rate')
```




```
female_surviving_prob <- (t["1", "female"] * 100) / t["Sum", "female"]
sprintf("Female surviving probability = %f",female_surviving_prob)
```

```
## [1] "Female surviving probability = 75.485636"
```

```
male_surviving_prob <- (t["1", "male"] * 100) / t["Sum", "male"]
sprintf("Male surviving probability = %f",male_surviving_prob)
```

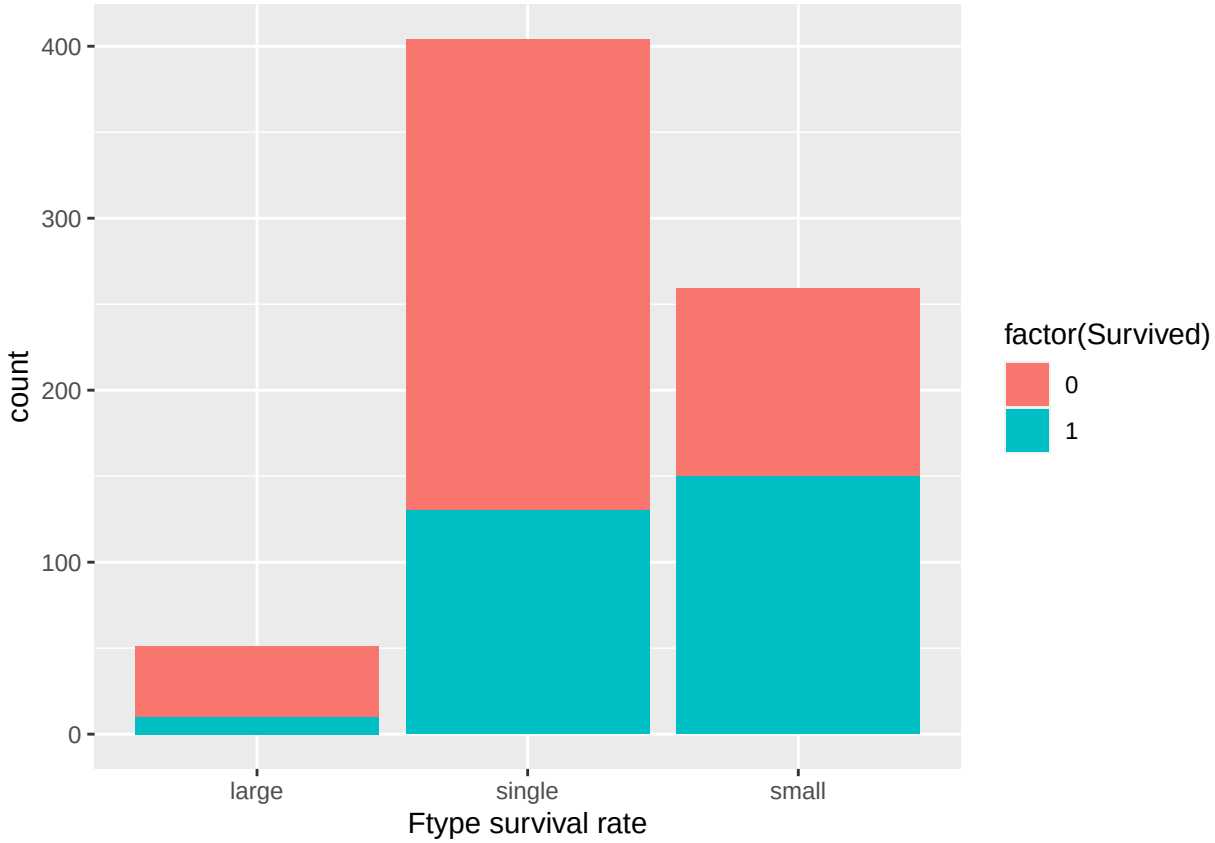
```
## [1] "Male surviving probability = 20.535855"
```

Relación entre Survived y Ftype

```
frequency_table <- table(data$Survived, data$Ftype, dnn = c("Survived", "Ftype"))
proportions_table <- prop.table(frequency_table)
percentages_table <- round((proportions_table * 100), 2)
t <- addmargins(percentages_table)
t
```

```
##          Ftype
## Survived  large single  small    Sum
##      0     5.74  38.38  15.27  59.39
##      1     1.40  18.21  21.01  40.62
##      Sum   7.14  56.59  36.28 100.01
```

```
ggplot(data, aes(x = Ftype, fill = factor(Survived))) +
  geom_bar(stat='count')+
  labs(x = 'Ftype survival rate')
```




```
single_surviving_prob <- (t["1", "single"] * 100) / t["Sum", "single"]
sprintf("Single surviving probability = %f",single_surviving_prob)
```

```
## [1] "Single surviving probability = 32.178830"
```

```
small_surviving_prob <- (t["1", "small"] * 100) / t["Sum", "small"]
sprintf("Small surviving probability = %f",small_surviving_prob)
```

```
## [1] "Small surviving probability = 57.910695"
```

```
large_surviving_prob <- (t["1", "large"] * 100) / t["Sum", "large"]
sprintf("Large surviving probability = %f",large_surviving_prob)
```

```
## [1] "Large surviving probability = 19.607843"
```

Existe también una relación bastante visible entre el tamaño de la familia y la tasa de supervivencia, siendo un 57% la tasa de supervivencia para las familias denominadas `small` y un 16% para las familias `large`.

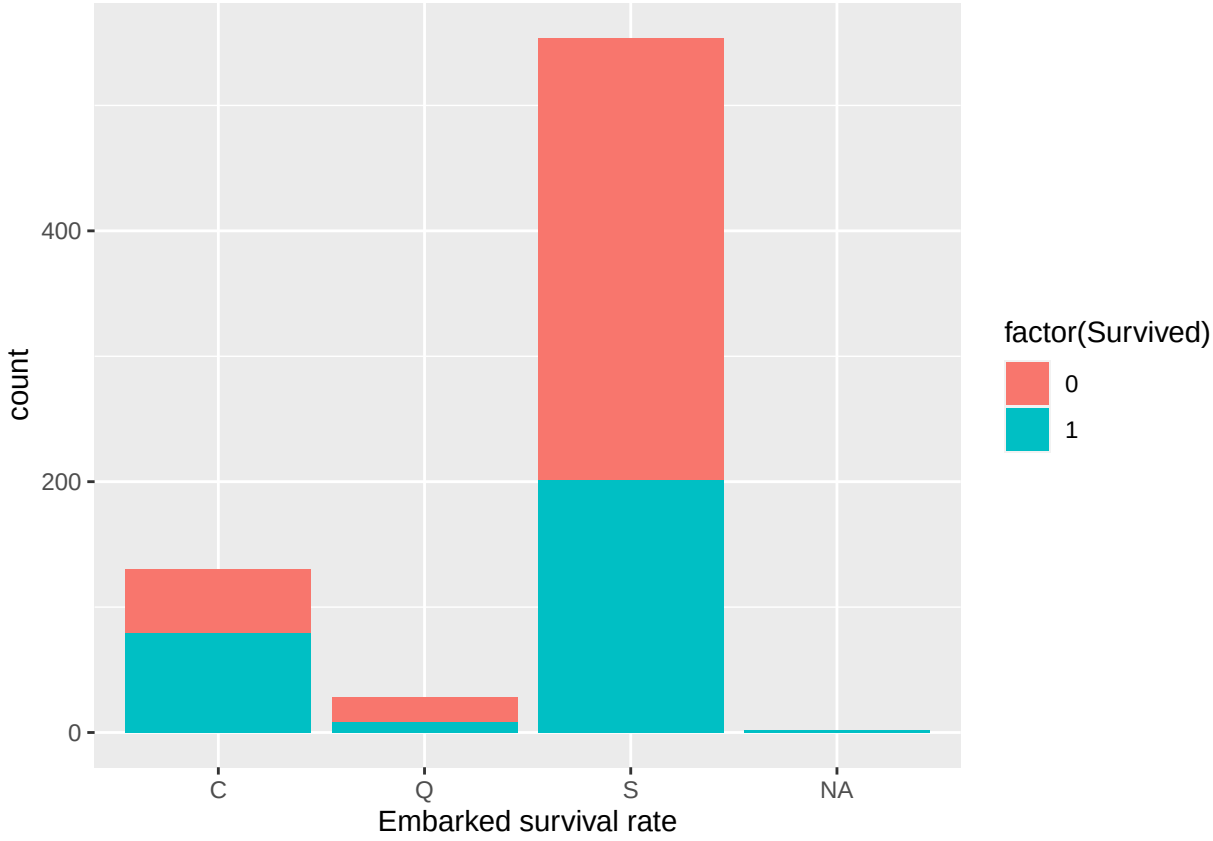
Relación entre Survived y Embarked

```
frequency_table <- table(data$Survived, data$Embarked, dnn = c("Survived", "Embarked"))
proportions_table <- prop.table(frequency_table)
percentages_table <- round((proportions_table * 100), 2)
t <- addmargins(percentages_table)
t
```

```
##          Embarked
## Survived      C      Q      S    Sum
##      0     7.16   2.81  49.58  59.55
```

```
##       1    11.10   1.12  28.23  40.45
##       Sum  18.26   3.93  77.81 100.00
```

```
ggplot(data, aes(x = Embarked, fill = factor(Survived))) +
  geom_bar(stat='count')+
  labs(x = 'Embarked survival rate')
```




```
C_surviving_prob <- (t["1", "C"] * 100) / t["Sum", "C"]
sprintf("C surviving probability = %f",C_surviving_prob)
```

```
## [1] "C surviving probability = 60.788609"
```

```
Q_surviving_prob <- (t["1", "Q"] * 100) / t["Sum", "Q"]
sprintf("Q surviving probability = %f",Q_surviving_prob)
```

```
## [1] "Q surviving probability = 28.498728"
```

```
S_surviving_prob <- (t["1", "S"] * 100) / t["Sum", "S"]
sprintf("S surviving probability = %f",S_surviving_prob)
```

```
## [1] "S surviving probability = 36.280684"
```

En este gráfico podemos observar cómo la mayoría de pasajeros embarcaron en el puerto `S`. Sin embargo, también observamos que sobrevivió un mayor porcentaje (55%). Esto puede ser debido a múltiples factores, como la clase de *ticket* que tenía la gente que embarcaba en ese puerto o el tipo de familias.

Comprobación de la normalidad y homogeneidad de la varianza.

Vamos a comprobar la normalidad de los atributos numéricos aplicando el Test Shapiro-Wilk, que sirve para contrastar si los datos siguen una distribución normal. Vamos a utilizarlo en la variable `Age`:

- Hipótesis nula (H_o): Los datos de la muestra *no son diferentes* a una distribución normal.
- Hipótesis alternativa (H_1): Los datos de la muestra *son diferentes* a una distribución normal.

Aceptaremos la hipótesis nula cuando el p-value que nos devuelva el test sea mayor a 0.05 (p_value > 0.05).

Rechazaremos la hipótesis nula (por lo tanto, aceptamos la alternativa) en caso contrario (p_value < 0.05).

```
print(shapiro.test(data$Age))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Age
## W = 0.98146, p-value = 7.337e-08
```

```
print(shapiro.test(data$Fnumber))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Fnumber
## W = 0.67924, p-value < 2.2e-16
```

```
print(shapiro.test(data$Fare))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Fare
## W = 0.52809, p-value < 2.2e-16
```

Dados los resultados de los distintos tests podemos afirmar que ninguno de los atributos numéricos del *dataset* sigue una distribución normal.

A continuación vamos a realizar pruebas para comprobar la homogeneidad de la varianza, también conocida como homocedasticidad. Existen diferentes tests para esto; nosotros vamos a utilizar el test de Fligner-Killeen (que se trata de un test que compara varianzas basándose en la mediana) para determinar si las dos poblaciones tienen una varianza similar en función de la supervivencia. Además, este tipo de test es aplicable cuando no se cumple la normalidad en la muestra, como hemos demostrado anteriormente que es nuestro caso para la variable `Age`.

- Hipótesis nula (H_0): Las varianzas de todas las poblaciones son iguales.
- Hipótesis alternativa (H_1): Alguna varianza difiere del resto.

```
a <- data[data$Survived == "0", "Age"]
b <- data[data$Survived == "1", "Age"]
fligner.test(x = list(a,b))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, b)
## Fligner-Killeen:med chi-squared = 1.0436, df = 1, p-value = 0.307
```

En este caso, p-value > 0.05 por lo tanto aceptamos la hipótesis nula, lo que significa que las varianzas de las poblaciones son iguales; por lo tanto existe homocedasticidad.

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.

A continuación vamos a aplicar contrastes de hipótesis sobre las muestras para determinar si un factor influye en la supervivencia. Las hipótesis nula y alternativa de nuestros contrastes son:

- Hipótesis nula (H_0): Los dos factores son independientes.
- Hipótesis alternativá (H_1): Los factores son dependientes.

```
frec <- table(data$Survived, data$Sex)
chisq.test(frec)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  frec
## X-squared = 205.03, df = 1, p-value < 2.2e-16
```

Obtenemos un p-value < 0.05, por lo tanto rechazamos la hipótesis nula en favor de la alternativa y llegamos a la conclusión de que los factores `Age` y `Survived` son dependientes.

Ahora vamos a comprobar la correlación entre todas las variables numéricas para observar si el precio del billete estaba relacionado con el número de familiares a bordo:

```
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

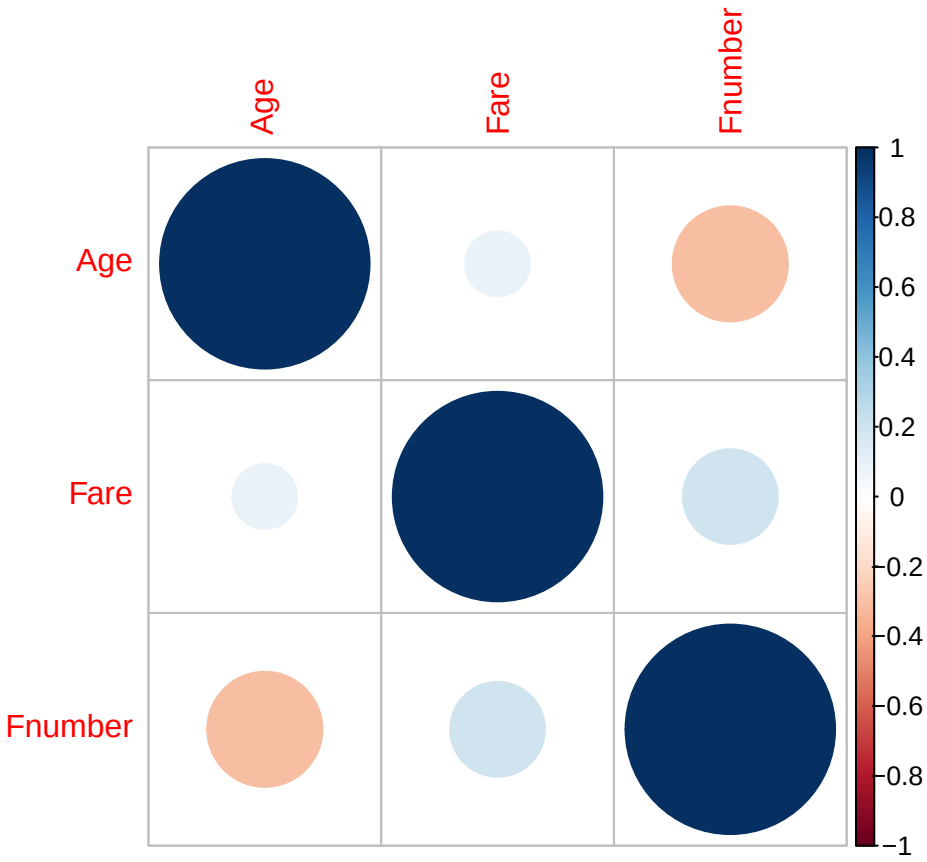
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
if (!require('corrplot')) install.packages('corrplot'); library('corrplot')
```

```
## Loading required package: corrplot
```

```
## corrplot 0.92 loaded
```

```
d <- select_if(data, is.numeric)
M <- cor(d)
corrplot(M)
```



De la matriz de correlación podemos deducir que los datos no están muy correlacionados. Existe una ligera relación entre `Fnumber` y `Fare` y entre `Fare` y `Age`, pero no es significativa.

Por último, vamos a aplicar la regresión logística debido a que queremos predecir una variable categórica (`Survived`), que puede tomar solo dos valores. Por lo tanto, la regresión lineal queda descartada.

De los datos que hemos visto y analizado a lo largo de toda la práctica, podemos deducir que la edad (`Age`), el sexo (`Sex`), la clase (`Pclass`), el puerto de embarque (`Embarked`) y el tamaño de familia a bordo (`Ftype`) influyen en la supervivencia. Por lo tanto, van a ser nuestras variables independientes, y el atributo `Survived` es nuestra variable dependiente.

```
model <- glm(
  Survived ~ Age + Sex + Pclass + Embarked + Ftype,
  data = data,
  family = 'binomial'
)

model2 <- glm(
  Survived ~ Age + Sex + Pclass + Embarked,
  data = data,
  family = 'binomial'
)

summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Embarked + Ftype,
```

```
##     family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8221  -0.6411  -0.3902   0.6002   2.5121
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.375729   0.578102   4.110 3.96e-05 ***
## Age         -0.041586   0.008426  -4.936 7.99e-07 ***
## Sexmale     -2.672213   0.225932 -11.828  < 2e-16 ***
## Pclass2     -1.294928   0.301809  -4.291 1.78e-05 ***
## Pclass3     -2.354290   0.300714  -7.829 4.92e-15 ***
## EmbarkedQ   -0.723203   0.612761  -1.180    0.238
## EmbarkedS   -0.336145   0.275028  -1.222    0.222
## Ftypesingle  1.905359   0.477967   3.986 6.71e-05 ***
## Ftypesmall   2.006791   0.474574   4.229 2.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 621.35  on 703  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 639.35
##
## Number of Fisher Scoring iterations: 5
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Embarked, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6464  -0.6808  -0.3979   0.6367   2.4715
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.036825   0.430758   9.371  < 2e-16 ***
## Age         -0.036082   0.007715  -4.677 2.92e-06 ***
## Sexmale     -2.515793   0.209293 -12.020  < 2e-16 ***
## Pclass2     -1.144614   0.290678  -3.938 8.23e-05 ***
## Pclass3     -2.409565   0.291179  -8.275  < 2e-16 ***
## EmbarkedQ   -0.814190   0.567903  -1.434   0.1517
## EmbarkedS   -0.493651   0.266886  -1.850   0.0644 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
```

```
## Residual deviance: 642.68  on 705  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 656.68
##
## Number of Fisher Scoring iterations: 5
```

Si comparamos la bondad del ajuste de estos dos modelos (AIC) vemos que el primer modelo, el que tenía más variables, tiene un AIC menor. Por lo tanto, la bondad de ajuste es mejor; deberíamos optar por el primer modelo.

Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

Los diferentes gráficos y tablas los hemos ido incluyendo a lo largo de toda la práctica.

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

Al principio de la práctica nos hemos planteado responder a la pregunta acerca de si había características de los pasajeros que impactaban sobre la tasa de supervivencia. A lo largo del trabajo, las hemos ido respondiendo y, finalmente, con la aplicación de un modelo de regresión logística, hemos visto que sí y hemos a la conclusión de que factores como el tamaño de familia, el sexo o la clase en la que se viajó afectaron a la tasa de supervivencia: los pasajeros tenían más probabilidades de sobrevivir si viajaban en primera clase, eran mujer y viajaban en una familia pequeña, mientras que el peor caso de supervivencia eran hombres de familias grandes que viajaban en tercera clase.