

# Tipología y ciclo de vida de los datos: Práctica 2

Jorge Marchán Gutiérrez

Rafael Jiménez Sarmentero

mayo 2022

## Contents

<b>Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?</b>	<b>1</b>
<b>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.</b>	<b>3</b>
<b>Limpieza de los datos.</b>	<b>4</b>
¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos. . . . .	4
Identifica y gestiona los valores extremos. . . . .	6
<b>Análisis de los datos.</b>	<b>12</b>
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?) . . . . .	12
Comprobación de la normalidad y homogeneidad de la varianza. . . . .	12
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. . . . .	12
<b>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.</b>	<b>12</b>
<b>Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?</b>	<b>13</b>

---

**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 que pasajeros viajaban en el mismo y si han sobrevivido o no, por lo tanto podemos saber si el camarote o el tipo de camarote estan relacionados con una mayor supervivencia.
  - **SibSp y Parch:** De estas dos variables podemos extraer una sola, que hace referencia al numero de familiares que el pasajero tenia 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 caracter “.”.
3. La edad puede contener valores decimales al ser de tipo number y no entero.
4. El separador de columnas es el caracter “,”.

Para la limpieza de los datos resulta interesante conocer que 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

```
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 12 variables:
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...
## $ 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)"
## $ 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 extraer 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.

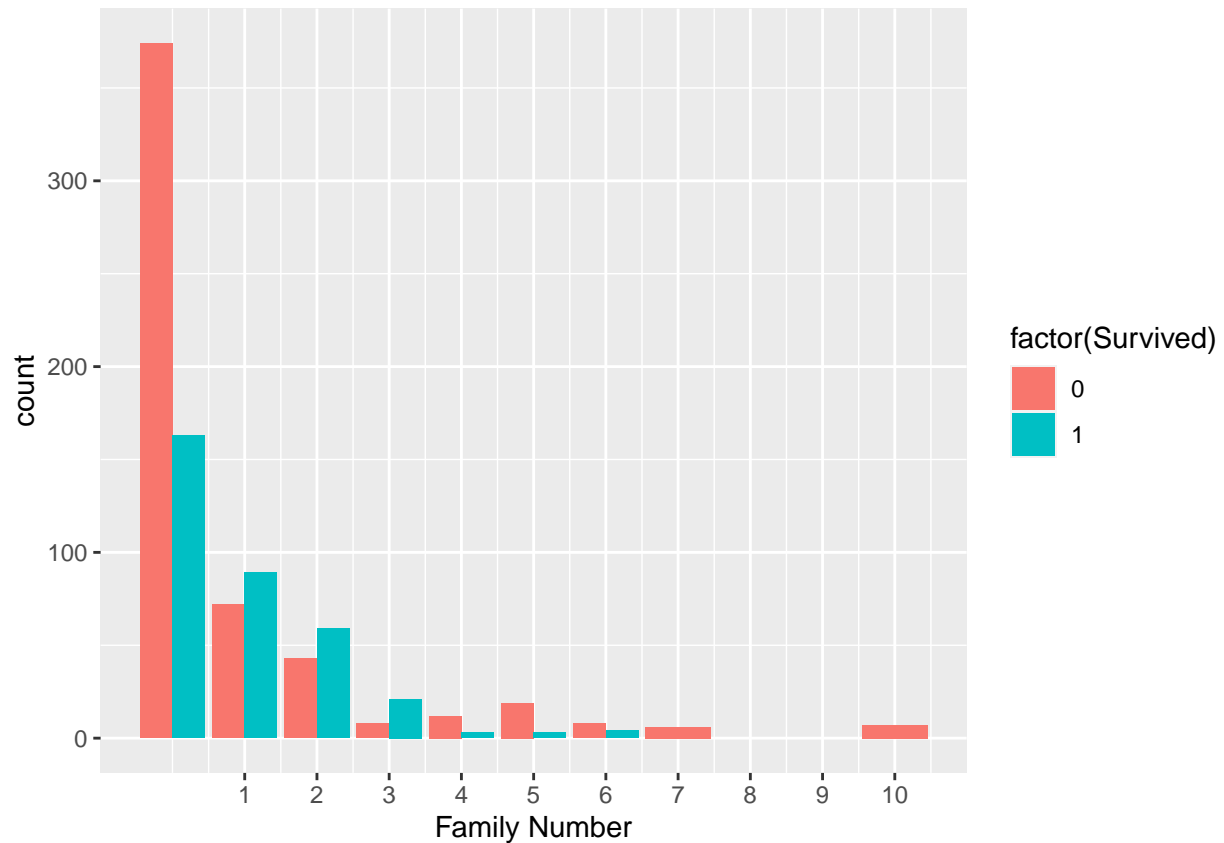
```
data$Fnumber <- data$SibSp + data$Parch
data$SibSp <- NULL
data$Parch <- NULL
```

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.

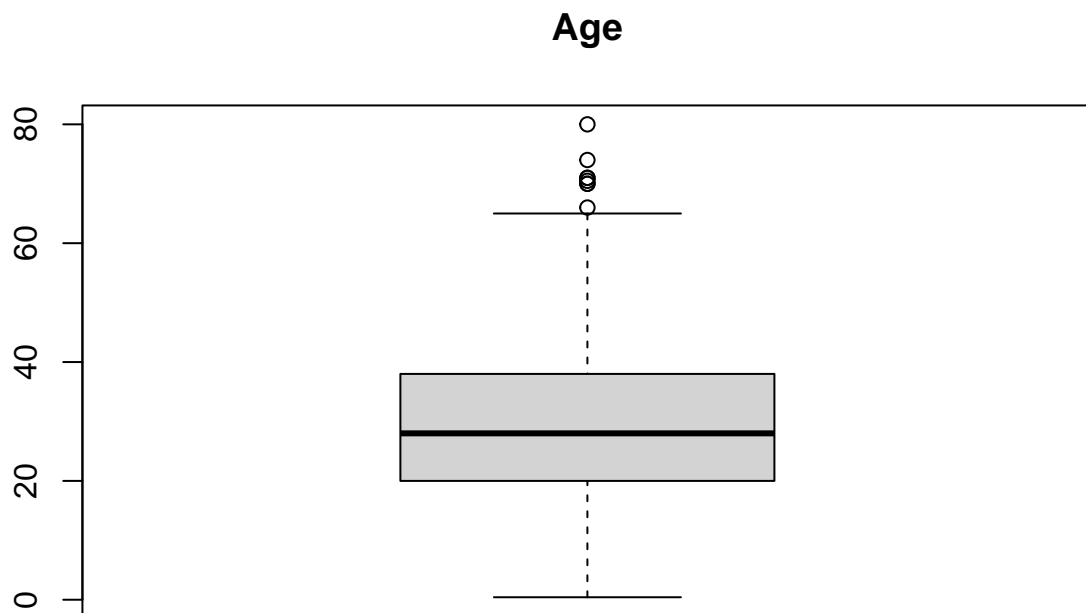
```
data$Ftype[data$Fnumber == 0] <- 'single'
data$Ftype[data$Fnumber < 4 & data$Fnumber > 0] <- 'small'
data$Ftype[data$Fnumber >= 4] <- 'large'

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")
```



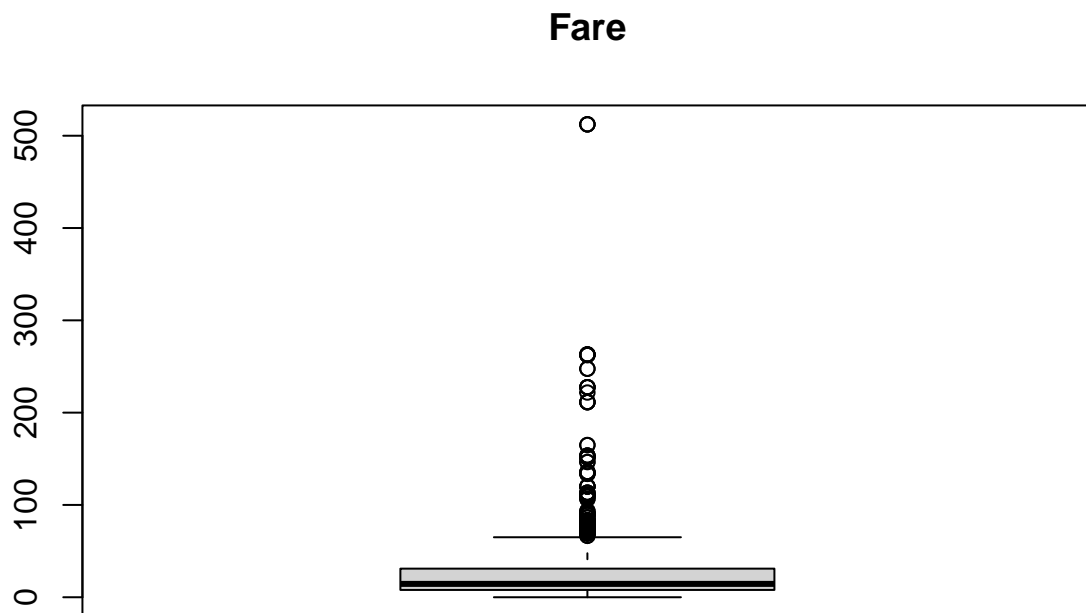
```
boxplot.stats(data$Age)$out
```

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

Del 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, por lo tanto estos datos no necesitan ser tratados.

Vamos a analizar ahora los valores extremos del atributo **Fare**

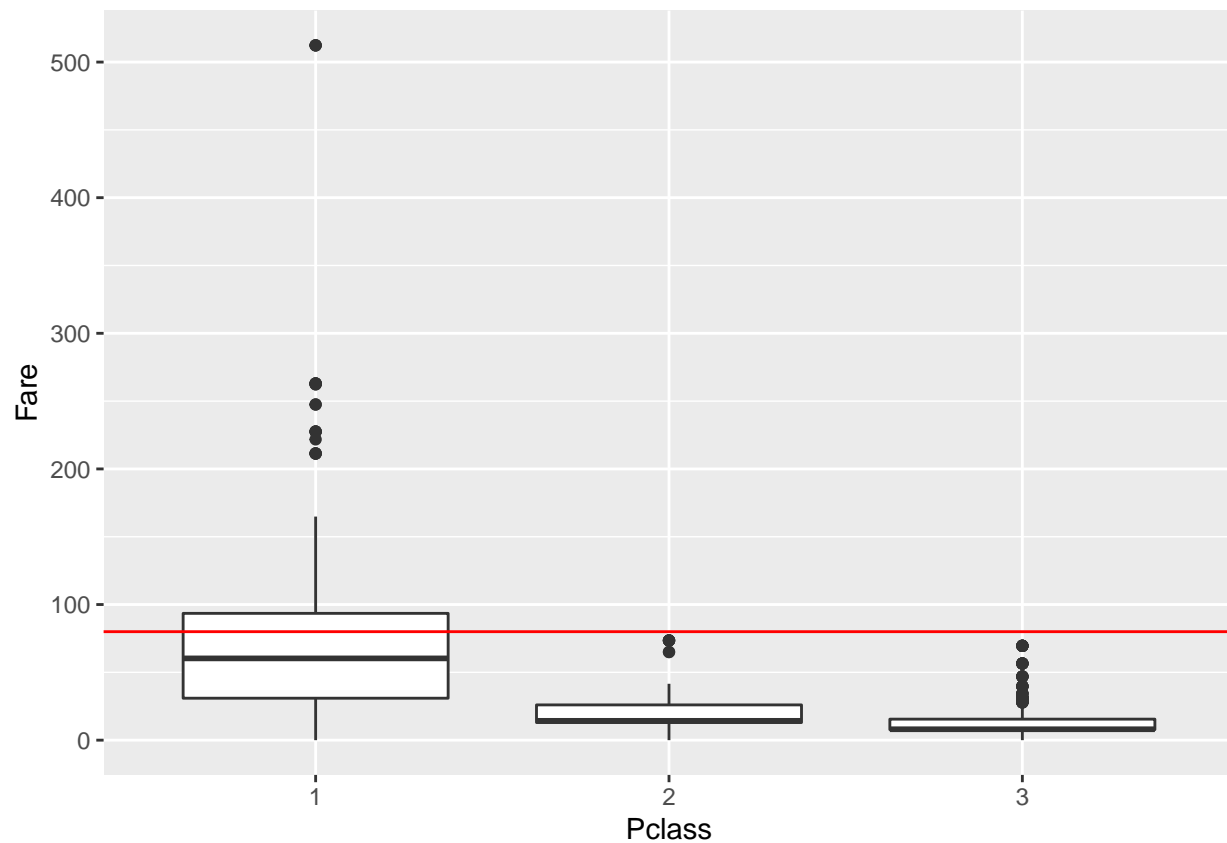
```
boxplot(data$Fare, main="Fare")
```



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 billete y hay menos de primera clase que del resto, por lo tanto es probable que los billetes de primera clase se consideren valores extremos, veamos los boxplot, por clase y marcando el límite para ser outliers en el atributo **Fare**

```
ggplot(data, aes(x=Pclass, y=Fare)) +  
  geom_boxplot() +  
  geom_hline(aes(yintercept=80), colour='red')
```





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

```
##      PassengerId Survived Pclass
## 28             28         0       1
## 89             89         1       1
## 119            119         0       1
## 259            259         1       1
## 300            300         1       1
## 312            312         1       1
## 342            342         1       1
## 378            378         0       1
## 381            381         1       1
## 439            439         0       1
## 528            528         0       1
## 558            558         0       1
## 680            680         1       1
## 690            690         1       1
## 701            701         1       1
## 717            717         1       1
## 731            731         1       1
## 738            738         1       1
## 743            743         1       1
## 780            780         1       1
```

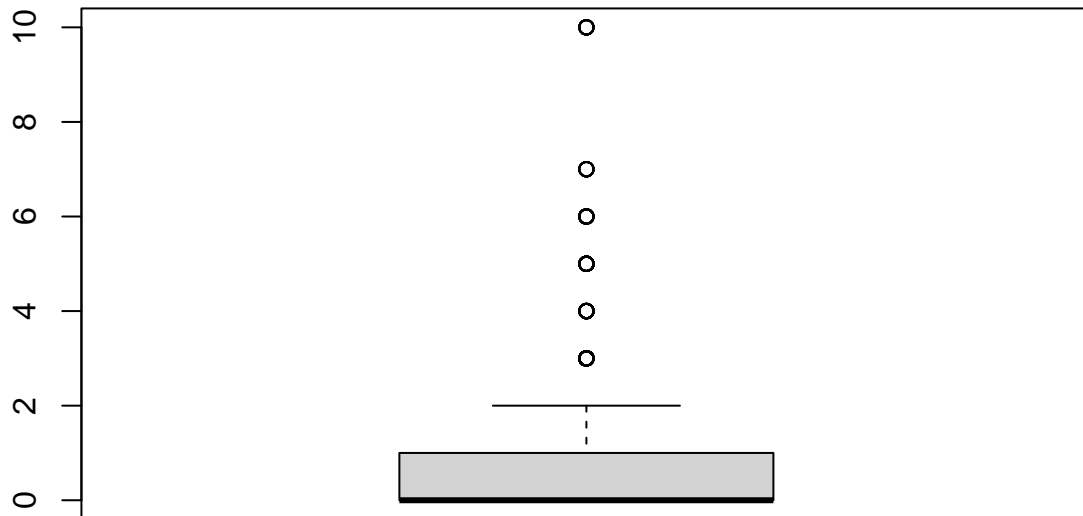
```
##                                     Name    Sex Age  Ticket
## 28                                Fortune, Mr. Charles Alexander  male  19   19950
```

## 89	Fortune, Miss. Mabel Helen	female	23	19950
## 119	Baxter, Mr. Quigg Edmond	male	24	PC 17558
## 259	Ward, Miss. Anna	female	35	PC 17755
## 300	Baxter, Mrs. James (Helene DeLaudeniére Chaput)	female	50	PC 17558
## 312	Ryerson, Miss. Emily Borie	female	18	PC 17608
## 342	Fortune, Miss. Alice Elizabeth	female	24	19950
## 378	Widener, Mr. Harry Elkins	male	27	113503
## 381	Bidois, Miss. Rosalie	female	42	PC 17757
## 439	Fortune, Mr. Mark	male	64	19950
## 528	Farthing, Mr. John	male	NA	PC 17483
## 558	Robbins, Mr. Victor	male	NA	PC 17757
## 680	Cardeza, Mr. Thomas Drake Martinez	male	36	PC 17755
## 690	Madill, Miss. Georgette Alexandra	female	15	24160
## 701	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18	PC 17757
## 717	Endres, Miss. Caroline Louise	female	38	PC 17757
## 731	Allen, Miss. Elisabeth Walton	female	29	24160
## 738	Lesurer, Mr. Gustave J	male	35	PC 17755
## 743	Ryerson, Miss. Susan Parker "Suzette"	female	21	PC 17608
## 780	Robert, Mrs. Edward Scott (Elisabeth Walton McMillan)	female	43	24160
##	Fare	Cabin Embarked Fnumber Ftype		
## 28	263.0000	C23 C25 C27 S 5	large	
## 89	263.0000	C23 C25 C27 S 5	large	
## 119	247.5208	B58 B60 C 1	small	
## 259	512.3292	<NA> C 0	single	
## 300	247.5208	B58 B60 C 1	small	
## 312	262.3750	B57 B59 B63 B66 C 4	large	
## 342	263.0000	C23 C25 C27 S 5	large	
## 378	211.5000	C82 C 2	small	
## 381	227.5250	<NA> C 0	single	
## 439	263.0000	C23 C25 C27 S 5	large	
## 528	221.7792	C95 S 0	single	
## 558	227.5250	<NA> C 0	single	
## 680	512.3292	B51 B53 B55 C 1	small	
## 690	211.3375	B5 S 1	small	
## 701	227.5250	C62 C64 C 1	small	
## 717	227.5250	C45 C 0	single	
## 731	211.3375	B5 S 0	single	
## 738	512.3292	B101 C 0	single	
## 743	262.3750	B57 B59 B63 B66 C 4	large	
## 780	211.3375	B3 S 1	small	

Observamos como los valores considerados outliers para **Fare** están asociados con la clase en la que viajan, mientras más alta es la clase y mayor numero de pasajeros por 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")
```

## 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 como 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, veamos los pasajeros con Fnumber = 10 existentes en el dataset

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

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

```
## 202 CA. 2343 69.55 <NA>      S      10 large
## 325 CA. 2343 69.55 <NA>      S      10 large
## 793 CA. 2343 69.55 <NA>      S      10 large
## 847 CA. 2343 69.55 <NA>      S      10 large
## 864 CA. 2343 69.55 <NA>      S      10 large
```

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

---

## 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 que hay con la supervivencia, de los atributos **Pclass**, **Sex**, **Embarked** y **Ftype**.

### Relación entre **Survived** y **Pclass**

```
#if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
#if (!require('corrplot')) install.packages('corrplot'); library('corrplot')
#if (!require('ggcorrplot')) install.packages('ggcorrplot'); library('ggcorrplot')
#cor(as.numeric(data$Survived), as.numeric(c(data$Pclass, data$Sex))) %>% corrplot(method = "color", tl
```

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

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.

---

**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.**

---



---

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

---