# M2.851 - Tipología y ciclo de vida de los datos

Práctica 2

Alumnos:

Michaelle Estefanía Valenzuela Sangoquiza

Juan Manuel Penalta Rodríguez.

**Índice**

[M2.851 - Tipología y ciclo de vida de los datos 1](#__RefHeading___Toc1001_781917998)

[​ 1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? 3](#__RefHeading___Toc2709_2179713084)

[​ 2. Integración y selección de los datos de interés a analizar. 4](#__RefHeading___Toc2711_2179713084)

[​ 3. Limpieza de los datos. 4](#__RefHeading___Toc2713_2179713084)

[​ 3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 4](#__RefHeading___Toc2715_2179713084)

[​ 3.2. Identificación y tratamiento de valores extremos. 5](#__RefHeading___Toc2717_2179713084)

[​ 4. Análisis de los datos. 6](#__RefHeading___Toc2719_2179713084)

[​ 4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 6](#__RefHeading___Toc2721_2179713084)

[​ 4.2. Comprobación de la normalidad y homogeneidad de la varianza. 6](#__RefHeading___Toc2723_2179713084)

[​ 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. 6](#__RefHeading___Toc2725_2179713084)

[​ 5. Representación de los resultados a partir de tablas y gráficas. 7](#__RefHeading___Toc2727_2179713084)

[​ 6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 7](#__RefHeading___Toc2729_2179713084)

[​ 7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. 7](#__RefHeading___Toc370_2578945720)

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

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

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

[https://github.com/mishuvale91/titanic-dataset/tree/main/data]

En total el dataset dispone de **891 entradas**, con **1**2 **campos** para cada entrada. Los campos disponibles se describen a continuación:

* **Survival**: Variable numérica que indica si el pasajero supervivió o murió en el hundimiento.
  + 1 = Superviviente
  + 0 = No superviviente
* **Pclass**: Variable numérica que indica el tipo de ticket que tenía el pasajero.
  + 1 = Primera Clase
  + 2 = Segunda Clase
  + 3 = Tercera Clase
* **Sex**: Variable alfanúmerica que indica el género del pasajero: female o male
* **Age**: Variable numérica que indica la edad en años del pasajero.
* **Sibsp**: Variable numérica que indica el número de hermanos o esposos del pasajero que viajan con él.
* **Parch**: Variable numérica que indica el número de padres o hijos del pasajero que viajan con él.
* **Ticket**: Variable alfanumérica con el número de ticket del pasajero.
* **Fare**: Variable numérica que indica la tarifa pagada por el pasajero.
* **Cabin**: Variable alfanumérica que indica la cabina o cabinas que ocupaba el pasajero y sus parientes.
* **Embarked**: Variable alfanumérica que indica el puerto de embarque del pasajero.
  + C = Cherbourg
  + Q = Queenstown
  + S = Southampton

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

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

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

Se procede a la lectura del conjunto de datos “titanic-data”, que contiene 891 registros con 12 columnas, donde se realiza la selección de los siguientes campos:

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

Teniendo como resultado de la selección 10 campos prioritarios para realizar el análisis de predicción de supervivencia excluyendo los campos “Name” y “Cabin”.

## 3. Limpieza de los datos.

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

Veamos que campos tienen columnas con valores no definidos:

colSums(is.na(titanic\_data))

PassengerId Survived Pclass Sex Age SibSp Parch

0 0 0 0 177 0 0

Ticket Fare Embarked

0 0 0

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

titanic\_data$Age[is.na(titanic\_data$Age)] <- mean(titanic\_data$Age,na.rm=T)

Veamos ahora que campos tienen valores vacios:

colSums(titanic\_data=="")

PassengerId Survived Pclass Sex Age SibSp Parch

0 0 0 0 0 0 0

Ticket Fare Embarked

0 0 2

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

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

titanic\_data$Embarked[c(62, 830)] <- "C"

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

apply(titanic\_data,2, function(x) length(unique(x)))

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

cols<-c("Survived","Pclass","Sex","Embarked")

for (i in cols){

titanic\_data[,i] <- as.factor(titanic\_data[,i])

}

### 3.2. Identificación y tratamiento de valores extremos.

Los valores extremos o outliers son aquellos que parecen no ser congruentes si los comparamos con el resto de los datos. Para identificarlos se utilizar la función boxplots.stats() de R, la cual se emplea a continuación.

Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

boxplot.stats(titanic\_data$Survived)$out

# Levels: 0 1

boxplot.stats(titanic\_data$Pclass)$out

# Levels: 1 2 3

boxplot.stats(titanic\_data$Sex)$out

# Levels: female male

boxplot.stats(titanic\_data$Embarked)$out

# Levels: C Q S

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

boxplot.stats(titanic\_data$Age)$out

# 2 58 55 2 66 65 0 59 71 70 2 55 1 61 1 56 1 58 2 59 62 58 63 65 2 0 61 2 60 1 1 64 65 56 0 2 63 58

# 55 71 2 64 62 62 60 61 57 80 2 0 56 58 70 60 60 70 0 57 1 0 2 1 62 0 74 56

**Con estos valores, para el campo Age, se puede observar que hay muchos pasajeros cuya edad excede los valores más comunes, es decir, los valores superiores a 64, por lo que se puede deducir que hay personas mayores a bordo del barco.** Pero ninguno parece que tenga un valor que no coincida con una edad valida.

boxplot.stats(titanic\_data$Fare)$out

# 71.2833 263.0000 146.5208 82.1708 76.7292 80.0000 83.4750 73.5000 263.0000 77.2875 247.5208 73.5000

# 77.2875 79.2000 66.6000 69.5500 69.5500 146.5208 69.5500 113.2750 76.2917 90.0000 83.4750 90.0000

# 79.2000 86.5000 512.3292 79.6500 153.4625 135.6333 77.9583 78.8500 91.0792 151.5500 247.5208 151.5500

# 110.8833 108.9000 83.1583 262.3750 164.8667 134.5000 69.5500 135.6333 153.4625 133.6500 66.6000 134.5000

# 263.0000 75.2500 69.3000 135.6333 82.1708 211.5000 227.5250 73.5000 120.0000 113.2750 90.0000 120.0000

# 263.0000 81.8583 89.1042 91.0792 90.0000 78.2667 151.5500 86.5000 108.9000 93.5000 221.7792 106.4250

# 71.0000 106.4250 110.8833 227.5250 79.6500 110.8833 79.6500 79.2000 78.2667 153.4625 77.9583 69.3000

# 76.7292 73.5000 113.2750 133.6500 73.5000 512.3292 76.7292 211.3375 110.8833 227.5250 151.5500 227.5250

# 211.3375 512.3292 78.8500 262.3750 71.0000 86.5000 120.0000 77.9583 211.3375 79.2000 69.5500 120.0000

# 93.5000 80.0000 83.1583 69.5500 89.1042 164.8667 69.5500 83.1583

La variable Fare (precio del billete) tiene muchos valores que podrían ser posibles valores extremos. La primera idea es que podrían ser debido a la existencia de tres clases distintas de pasajeros, que logicamente pagarían valores muy distintos. Aún así filtrando por el campo Pclass siguen saliendo bastantes valores para cada uno de los valores de Pclass:

boxplot.stats(filter(titanic\_data,Pclass==1)$Fare)$out

boxplot.stats(filter(titanic\_data,Pclass==2)$Fare)$out

boxplot.stats(filter(titanic\_data,Pclass==3)$Fare)$out

## 4. Análisis de los datos.

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

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

### 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

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

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

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

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa |  |
| Redacción de las respuestas |  |
| Desarrollo código |  |