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 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.

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

Comprobamos qué columnas presentan 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
```

Y los sustituimos 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
```

Identifica y gestiona los valores extremos.

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?)

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?
