# Pràctica 2: Neteja i anàlisi de les dades

Adrián Alonso Gonzalo i Alexandre Vidal De Palol

Maig/Juny 2022

## Contents



# 1 Descripció i càrrega del dataset.

El conjunt de dades train.csv s'ha obtingut del web https://www.kaggle.com/c/titanic.

Aquest conjunt de dades conté informació sobre la tripulació del Titanic amb 12 tipus de variables i un total de 891 registres ( passatgers).

Les variables d'aquesta mostra son:

- PassengerId: Numero de passatger.
- Survived: Supervivència (0=No, 1=Si).
- Pclass: Classe de tiquet (1=Primera, 2=Segona, 3=Tercera) .
- Name: Nom.
- Sex: Sexe.
- Age: Edat.

- SibSp: Germans / Cónjugues a bord del Titanic.
- Parch: Pares / nens a bord del Titanic.
- Ticket: Número de ticket.
- Fare: Preu del ticket.
- Cabin: Numero de cabina.
- Embarked: Port de embarcament.

Lectura del fitxer.

```
dataset <- read.csv("../data/input/train.csv")
str(dataset)
```

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

Observem que que el dataset conté 3 tipus de variables del quals caràcter, numèric i enter.

# 2 Integració i selecció de les dades d'interès a analitzar.

El procés d'integració i selecció de les dades es realitzarà al llarg del procés de neteja i anàlisi de les diverses variables del conjunt de dades d'entrenament del dataset.

En aquest procés es pretén anar analitzant les diferents variables en el procés de neteja i anàlisi, i en funció de les característiques que es vagin observant de les diverses variables es prendrà la decisió d'utilitzar un conjunt seleccionat el qual pugui ser útil per a la predicció del model i la comprovació amb el conjunt de test o validació.

El resultat del projecte pot respondre a possibles causes de mort dels tripulants que no van sobreviure a la tragèdia del Titanic, permetent establir models d'inferència sobre les causes relatives a la mortalitat entre diversos tipus de passatgers. Per altra banda la implementació de un model d'interès sobre quines han sigut variables que han influït més o menys en la supervivència del naufragi.

# 3 Neteja de dades.

## 3.1 Valors buits (missing values).

## 3.2 Valors extrems (outliers).

## 3.3 Altres accions per a la neteja del joc de dades.

Primerament, observem que la primera variable "PassengerId" no és res més que un identificador, per tant procedim a eliminar-la del conjunt de dades ja que no ens interessa per l'estudi.

```r
# eliminació de la primera columna PassengerId
dataset <- dataset[,-1]
```

Observem que les variables 'Survived' i 'Pclass' són de tipus enter, però la seva funció es indicar una categoria. Per tant, procedim a convertir-les en tipu factor.

```r
# transformació de les variables

dataset$Survived <- as.factor(dataset$Survived)
class(dataset$Survived)
```

```
## [1] "factor"
```

```r
dataset$Pclass <- as.factor(dataset$Pclass)
class(dataset$Pclass)
```

```
## [1] "factor"
```

# 4 Anàlisi de les dades.

## 4.1 Selecció de grups a analitzar/comparar.

## 4.2 Normalitat i homogeneïtat de la variància.

## 4.3 Aplicació de proves estadístiques per a la comparació de grups.

# 5 Representació de resultats (taules i gràfiques).

# 6 Conclusions.