The Titanic project

## The different steps :

1. Data description
2. Preprocess
3. …

# Data description

Train df: len of 891

Test df: len of 418

Variables :

* PassengerId (not explanatory)
* Pclass (possible expl) ordinal
* Name (probably not expl)
* Sex (possible expl) nominal
* Age (possible expl)
* Sibsp (possible expl) nominal
* Parch (possible expl) nominal
* Ticket (possible expl)
* Fare (possible expl)
* Cabin (possible expl) nominal
* Embarked (possible expl) nominal
* Survived: dependent variable

# Preprocess

First: how to handle missing values:

117 for Age

687 for cabin

2 for Embarked

The **first theory** is:

* 687 Nan represents people with no cabin 🡪 lower class. Should compare the social class with the amount of nan value
* 2 Nan value for Embarked is almost 0, maybe remove it?
* 117 are too much to remove, should it be calculated as the mean, or should a model be used to compute the missing values?

Now let’s plot no cabin vs class label

Maybe add a column: have cabin: 1 or 0 because the name of the cabin doesn’t seem to show the Pclass (Pclass of 1 in B.., C.. and E..).

Then are there any unnecessary columns?

* Ticket, probably not necessary, just a reference
* Name, not necessary? Maybe by keeping the last name+ looking amount sisters and parents can give an idea of the family? Or to know the family just look at the one with the same cabin?
  + Approach for later!
  + For now, the label name is removed