**Goal**: build a model to predict the **“survival”** of a passenger

Hints:

1. Load the dataset as a dataframe, replace all missing values “?” with NA (in R) or NaN (in Python)
2. Subset the dataset: keep only columns 2,3,5,6,7,8,10,12
3. For missing values in column “Age”: substitute with the average value of Age
4. For missing values in other columns: remove rows with NA
5. Split the dataset into Training Data (80%) and Testing Data (20%)
6. Build a classification model using Training Data
7. Test & assess the model
8. Briefly describe your results

### Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = upper, 2 = middle, 3 = lower |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard |  |
| parch | # of parents / children aboard |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |