Importation and preprocessing in Python

Exercise 1 : survival on Titanic dataset

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history.

Women and children first? The aim is to understand how survivors of Titanic were selected...

Importation of the data and description of the dataset

1. In this first practical session, we shall work on the dataset `titanic.csv` on the survival of the passengers of Titanic. Download this dataset as a data frame
2. Describe the dataset `titanic` : features, nature of the features, number of observations
3. Basic statistics : mean of each variable, quartiles
4. Percentage of missing values for each column. Sort by descending values

Basic graphic analysis

We want to understand what features could contribute to a high survival rate. It would make sense if everything except 'PassengerId', 'Ticket' and 'Name' would be correlated with a high survival rate.

1. Get rid off the features 'PassengerId', 'Ticket' and 'Name' which seem irrelevant to analyse the data
2. We focus on the features 'Age' and 'Sex'.
   (a) Separate the dataset into men and women
   (b) Display the distribution of the age survivors and non survivors according to the sex. Comment
3. At first glance is there some link between 'Embarked' and 'Survival'.
4. At first glance is there some link between 'Pclass' and 'Survival'.