Basics in ML with Python

For the two folling exercises, one may rely on the notebook `Introduction to the Pandas library`

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...

The Titanic dataset may be downloaded on
`https://sites.google.com/site/marianneclausel/home/enseignements-20-21/cirm?authuser=0`

Exercise 1 : 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

Exercise 2 : 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'.

For the two next sections, one may rely on the notebook `Introduction to Principal Component Analysis`

Exercise 3 : Analysis of Financial Time series

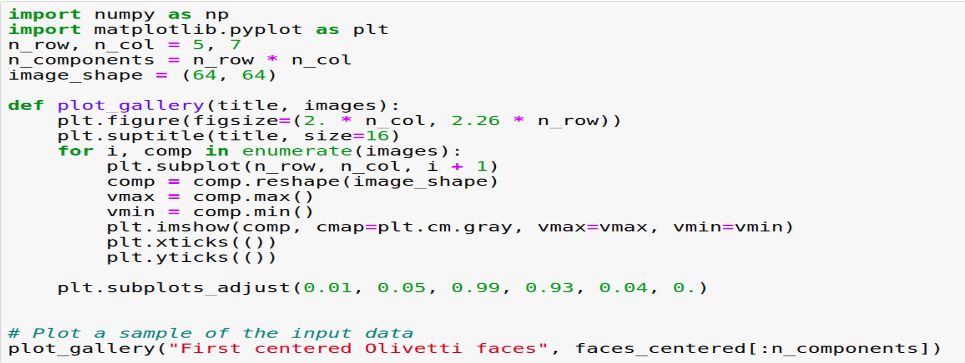
We want to understand the information contained in financial time series. The dataset is `rs.csv` and can be downloaded on Arche.

1. Import the data into a dataframe `rs`. this dataframe may contain missing values. Imputation of missing values can be done using the function `fillna` : `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.fillna.html`
2. To process the data in order to perform PCA, convert this dataframe into a numpy array using `to_numpy()`
3. Perform a PCA with two components. Transform the data projecting on the two first principal components
4. Plot the two first components
5. Visualize the data in the first two-component space
6. Compare with t-SNE

Exercise 4 : the Olivetti faces dataset

We want to perform PCA on a classical dataset : the Olivetti faces dataset.

1. Import the dataset using the function `fetch_olivetti_faces`
2. Center the faces : `faces_centered = faces - faces.mean(axis=0)`
3. We now plot the faces



```
import numpy as np
import matplotlib.pyplot as plt
n_row, n_col = 5, 7
n_components = n_row * n_col
image_shape = (64, 64)

def plot_gallery(title, images):
    plt.figure(figsize=(2. * n_col, 2.26 * n_row))
    plt.suptitle(title, size=16)
    for i, comp in enumerate(images):
        plt.subplot(n_row, n_col, i + 1)
        comp = comp.reshape(image_shape)
        vmax = comp.max()
        vmin = comp.min()
        plt.imshow(comp, cmap=plt.cm.gray, vmax=vmax, vmin=vmin)
        plt.xticks(())
        plt.yticks(())

    plt.subplots_adjust(0.01, 0.05, 0.99, 0.93, 0.04, 0.)


# Plot a sample of the input data
plot_gallery("First centered Olivetti faces", faces_centered[:n_components])
```

4. Perform PCA with 20 components. Plot the 20 first components