**Practice #7: The final dataframe**

**General:**

You know how to:

* Gather csv file from yahoo finance with request or directly with read\_csv() function
* Create one dataframe from several dataframes

**Objectives:**

* Use the two files containing a list of companies’ symbols (ex “AAPL” for apple) to create a big list with symbols of all companies.
* Use this list to download all of this company csv data (directly with the read\_csv() function of pandas) into dataframes.
* Add a column to each dataframe with the symbol of the downloaded company.
* Merge all dataframe into one huge dataframe.
* Save it using “parquet”.
* Clean the final dataframe from incorrect values

**Steps:**

1. **Create a big list of symbols**

*You have two files which contains a list of compagnies’ symbols in Europe and in US. You will merge them to create a complete list of all symbols.*

* Use the “with open()” function to open both csv files.
* The csv file is only on row with a high number of items
* Store their content (list of symbols) in two different variables
* Create the final list with all symbols by merging both lists

1. **Downloading**

*You have now a big list of compagnies’ symbols. You will use them and the url of yahoo finance to download csv file of all these compagnies.*

* Create a loop on the 10th first symbols (making a loop on all symbols is extremely long).
* For every iteration, download the csv file by using the read\_csv() function of pandas and the following URL: “https://query1.finance.yahoo.com/v7/finance/download/”SYM”?period1=0&period2=1661904000&interval=1d&events=history&includeAdjustedClose=true”
  + Where “SYM” is the symbol of the company (change at every iteration)
  + The following function: pd.read\_csv(url) allow you to directly download data from the url into a dataframe
* Add a column with the symbol of the company to the created dataframe
* Add the dataframe to a list of dataframes

1. **Creating the final dataframe**

*You have now a list of dataframes, you will merge them into one big dataframe and save it using parquet.*

* Merge the list of dataframes into one big dataframe (use the concat() function of pandas)
* Save the dataframe into a file using the df.to\_parquet() function (install pyarrow and fastparquet package first)
* Compare the saved file with the file saved with the df.to\_csv() function

1. **Cleaning your dataframe**

*You have dataframe with a huge amount of data, clean it to remove incorrect entries.*

* Load the final dataframe named ‘df\_final\_US\_EUR.parquet’ with the df.read\_parquet() function
* Look into your dataframe and remove:
  + Empty rows (Nan)
  + Rows with 0 prices
  + Rows with incorrect type
  + Others incorrect rows…