**Content** :

This Data repo contains the following datasets (in .csv format): **(refer to ‘archive’ folder)**

* 2014\_Financial\_Data.csv
* 2015\_Financial\_Data.csv
* 2016\_Financial\_Data.csv

This datasets contains 200+ financial indicators, that are commonly found in the 10-K filings each publicly traded company releases yearly, for a plethora of US stocks (on average, 4k stocks are listed in each dataset).

Important remarks regarding the datasets:

1. The third-to-last column, Sector, lists the sector of each stock. Indeed, in the US stock market each company is part of a sector that classifies it in a macro-area. Since all the sectors have been collected (Basic Materials, Communication Services, Consumer Cyclical, Consumer Defensive, Energy, Financial Services, Healthcare, Industrial, Real Estate, Technology and Utilities), the user has the option to perform per-sector analyses and comparisons.
2. The second-to-last column, PRICE VAR [%], lists the percent price variation of each stock for the year. For example, if we consider the dataset 2015\_Financial\_Data.csv, we will have:
   * 200+ financial indicators for the year 2015;
   * percent price variation for the year 2016 (meaning from the first trading day on Jan 2016 to the last trading day on Dec 2016).
3. The last column, class, lists a binary classification for each stock, where
   * for each stock, if the PRICE VAR [%] value is positive, class = 1. From a trading perspective, the 1 identifies those stocks that an hypothetical trader should BUY at the start of the year and sell at the end of the year for a profit.
   * for each stock, if the PRICE VAR [%] value is negative, class = 0. From a trading perspective, the 0 identifies those stocks that an hypothetical trader should NOT BUY, since their value will decrease, meaning a loss of capital.

The columns PRICE VAR [%] and class make possible to use the datasets for both classification and regression tasks:

* If the user wishes to train a machine learning model so that it learns to *classify* those stocks that in buy-worthy and not buy-worthy, it is possible to get the targets from the class column;
* If the user wishes to train a machine learning model so that it learns to *predict* the future value of a stock, it is possible to get the targets from the PRICE VAR [%] column.

**Task :**

1. **Get the general information and look of the dataset [1 point]**
2. **Data cleaning and outlier detection [2 points]**
3. **Carry out Basic EDA and draw conclusions/observations for potential key metrics affecting the performance of Stocks [4 points]**
4. **Visual and/or statistical relation(s) between input and output variables [3 points]**