ACTU PS5841 Data Science in Finance & Insurance - Spring 2022 (Y. Wang)
Assignment - 9
Assigned 5/12/22, Due 5/12/22 (Thur)

Problem 1. Data Wrangling
The spreadsheet workbook, **data_raw.xls**, contains information, which is stored in 4 worksheets, about a number of public companies and their stocks. Please code to produce a single pandas dataframe, to the specifications below, and export it into a csv file, **data_out.csv**.

· Retain distinct fields only. For example, there are two fields, labeled ”Security Price”, that contain the same information. Only one should be kept.

· Assume stocks with null values for “Dividend Yield” are non-dividend-paying stocks. Replace those null values with 0.00

· Convert the strings in the “Market Capitalization” field to the correct numerical values. For example, $123.45M should be converted to 123,450,000.00, $123.45B to 123,450,000,000.00, $123.45T to 123,450,000,000,000.00

· The “Equity Summary Score” field provides a numerical indication of sentiment of independent research firms on each stock. Please translate the Equity Summary Scores into sentiment categories as follows, and record them in the ”Analyst Sentiment” field.

```
[0.1,  1.0] = very bearish
[1.1,  3.0] = bearish
[3.1,  7.0] = neutral
[7.1,  9.0] = bullish
[9.1,  10.0] = very bullish
```

The resulting dataframe contains the following.

```
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Symbol                                     3061 non-null   object
 1   Company Name                               3061 non-null   object
 2   Security Type                              3061 non-null   object
 3   Security Price                             3061 non-null   float64
 4   Equity Summary Score                       3061 non-null   float64
 5   Volume (90 Day Avg)                        3057 non-null   float64
 6   Market Capitalization                      3061 non-null   float64
 7   Dividend Yield                             3061 non-null   float64
 8   Company Headquarters Location              3061 non-null   object
 9   Sector                                     3060 non-null   object
 10  Industry                                   3060 non-null   object
 11  Optionable                                 3061 non-null   object
 12  Price Performance (52 Weeks)               2988 non-null   float64
 13  Total Return (1 Yr Annualized)             2988 non-null   float64
 14  Beta (1 Year Annualized)                   2988 non-null   float64
 15  Standard Deviation (1 Yr Annualized)       2990 non-null   float64
 16  S&P Global Market Intelligence Valuation   3047 non-null   float64
 17  S&P Global Market Intelligence Quality     3044 non-null   float64
 18  S&P Global Market Intelligence Growth Stability  3046 non-null   float64
 19  S&P Global Market Intelligence Financial Health  2989 non-null   float64
 20  P/E (Price/TTM Earnings)                   2145 non-null   float64
 21  PEG Ratio                                      836 non-null    float64
 22  EPS Growth (Proj This Yr vs. Last Yr)          2763 non-null   float64
 23  Institutional Ownership                        2981 non-null   float64
 24  Institutional Ownership (Last vs. Prior Qtr)   3060 non-null   float64
 25  Analyst Sentiment                              3061 non-null   object
```

You may find the following resources useful.

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_excel.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.concat.html
https://stackoverflow.com/questions/14984119/python-pandas-remove-duplicate-columns
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.fillna.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.replace.html
https://stackoverflow.com/questions/43096522/remove-dollar-sign-from-entire-python-pandas-dataframe
https://www.skytowner.com/explore/converting_k_and_m_to_numerical_form_in_pandas_dataframe

The file, **data_prepared.csv**, represents the ”correct” output from Problem 1. You can load this file as the starting point for Problem 2.

Problem 2. Classification
Let’s explore if we can predict “Analyst Sentiment” with information at hand.
Let’s focus on non-REIT common stocks only, and exclude records for “Common Stock (REIT)”, “Depository Receipt”, “Unit Trust Fund” and “Unit Trust Fund (REIT)”.
Consider a multiclass logistic regression to predict ”Analyst Sentiment”, using the following features

```
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Security Price                                2598 non-null   float64
 1   Volume (90 Day Avg)                           2598 non-null   float64
 2   Market Capitalization                         2598 non-null   float64
 3   Dividend Yield                                2598 non-null   float64
 4   Total Return (1 Yr Annualized)                2598 non-null   float64
 5   Beta (1 Year Annualized)                      2598 non-null   float64
 6   Standard Deviation (1 Yr Annualized)          2598 non-null   float64
 7   S&P Global Market Intelligence Valuation      2598 non-null   float64
 8   S&P Global Market Intelligence Quality        2598 non-null   float64
 9   S&P Global Market Intelligence Growth Stability  2598 non-null   float64
 10  S&P Global Market Intelligence Financial Health  2598 non-null   float64
 11  Institutional Ownership                       2598 non-null   float64
 12  Institutional Ownership (Last vs. Prior Qtr)  2598 non-null   float64
```

For this exercise, please drop any record(row) if it contains a null value for any field.

Please report the estimated error rate for a random prediction, using 10-fold cross-validation.

You may find the following resources useful.
`https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

The file, **data_prepared.csv**, represents the ”correct” output from Problem 1. You can load this file as the starting point for Problem 3.

Problem 3. Regression, Dimension Reduction
Let’s explore if we can predict a “Equity Summary Score” with information at hand.
Let’s focus on non-REIT common stocks only, and exclude records for “Common Stock (REIT)”, “Depository Receipt”, “Unit Trust Fund” and “Unit Trust Fund (REIT)”.
Consider a principal components regression (PCR) to predict Equity Summary Score. Please use the following raw features as input, but standardize them when performing principal component analysis.

```
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Security Price                             2598 non-null   float64
 1   Volume (90 Day Avg)                        2598 non-null   float64
 2   Market Capitalization                      2598 non-null   float64
 3   Dividend Yield                             2598 non-null   float64
 4   Total Return (1 Yr Annualized)             2598 non-null   float64
 5   Beta (1 Year Annualized)                   2598 non-null   float64
 6   Standard Deviation (1 Yr Annualized)       2598 non-null   float64
 7   S&P Global Market Intelligence Valuation   2598 non-null   float64
 8   S&P Global Market Intelligence Quality     2598 non-null   float64
 9   S&P Global Market Intelligence Growth Stability  2598 non-null   float64
 10  S&P Global Market Intelligence Financial Health  2598 non-null   float64
 11  Institutional Ownership                    2598 non-null   float64
 12  Institutional Ownership (Last vs. Prior Qtr)     2598 non-null   float64
```

For this exercise, please drop any record(row) if it contains a null value for any field.

[*a*] Estimate and plot the cumulative % of variance explained vs the number of PCs included. Note that 100% of variance is explained when all 13 PCs are included.

[*b*] Using 10-fold cross-validation, explore the potential for dimension reduction.
[*b*1] Plot the estimated root mean squared error(RMSE) of a random prediction vs the number of PCs used in the PCR.
[*b*2] If you do not use all 13 PCs, what is the optimal number of PCs to include in your PCR?
[*b*3] Fit the model using the optimal number of PCs and report the coefficient of each included PC score.

You may find the following resources useful.
`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`