WQD 7004 Programming for Data Science

Assignment 3: Data Cleaning

Technical Analysis in Financial Data

|  |  |
| --- | --- |
| Ng Kang Wei | WQD170068 |
| Alireza Ghanbarzadeh | WQD170061 |
| Khashayar Namsehchi | WQD170034 |
| Khavish Punchoo | WQD170097 |

# The data

The dataset consisted of 20 companies stock listed on New York Stock Exchange (NYSE). The 20 companies chosen here are all financial institutions. The dataset consisted of the weekly fluctuation in prices for each of the company stock since 2012 to 2017. Each of the stock data is stored in a CSV file. There is 20 financial institutions stock chose thus there are 20 CSV files. Each of the file contains 18 columns and 1260 rows.

The stock dataset just contains the stock code or symbol used by the companies in NYSE but not the company name. In order to get the company name for each of the stock or company, there is another dataset that contains the company details and the stock code used by the companies. The companies’ details is stored in 1 CSV file and it has 8 columns and 506 rows.

# Purpose of dataset

The data can be used to analyze the past performance of the stocks and provide insights into the future performance of the stocks.

These are some of the analysis that can done on the dataset to achieve the said purpose:

* Portfolio allocation
* Calculate risk and Sharpe ratio
* Calculate return based on prices
* Price prediction
* Technical indicators computation and analysis
* Price chart visualization
* Visualizing technical indicators such as moving averages
* Forecasting future returns using past data

# Data Structure

This section explains the structure of the dataset. The stock data is a time series dataset that hold 5 years’ worth of data from 30/11/2012 to 29/11/2017 in the form of weekly stock prices fluctuations.

The structure of the stock dataset is as follows:

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Explanation |
| Date | String | Date in dd/mm/yyy format |
| Financial institution | String | Institution symbol |
| Location | String | Location (city) of institution |
| High | Numeric | Highest price in a day |
| Low | Numeric | Lower price in the day |
| Open | Numeric | The price of the share in the beginning of the trading day |
| Close | Numeric | The price of the share in the closing time of day |
| Volume | Numeric | Number of shares traded in a trading day |
| Aggregated Data 2 Days | Numeric | Mean of open/close/high/low of two trading days |
| Aggregated Data 3 Days | Numeric | Mean of open/close/high/low of two trading days |
| Aggregated Data for 5 Days | Numeric | Mean of open/close/high/low of two trading days |
| Number of employees | Numeric | Number of employees of an institution |
| Net changes 0 (numeric) | Numeric | Net price change of current day |
| Net change 0 (nominal) | String | Determine whether net change is positive or negative |
| Net change 5 (numeric) | Numeric | Net price change of current day of past 5 days |
| Net change 5 (nominal) | String | Determine whether net change is positive or negative |
| Net change 25 (numeric) | Numeric | Net price change of current day of past 25 days |
| Net change 25 (nominal) | String | Determine whether net change is positive or negative |

Besides the stock data, there is also a supplementary dataset. This supplementary dataset contains the company details. The stock dataset does not include the company name in the CSV file, just the stock code the company is listed on NYSE. The company dataset would provide more information on the company. The company dataset is stored in a single CSV file and contains 8 columns x 505 rows.

The structure of the company dataset is as follows:

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Explanation |
| Ticker symbol | String | The stock code in NYSe |
| Security | String | Company name |
| SEC Filings | String | The type of filings used by the company |
| GISC Sector | String | The sector the company is in |
| GISC Sub Industry | String | The sub industry the company is in |
| Address of Headquarters | String | The address of the company headquarters |
| Date first added | String | The date of the company is added to NYSE |
| CIK | String | A unique code for the company |

# Loading the dataset

Since the stock dataset is distributed on 20 CSV files. It would be time consuming and troublesome if it is loaded one by one. All the 20 stock CSV files are in the same directory and there are no other files within the directory. Therefore, the CSV files are loaded by checking the patterns of the files in the directory. The 20 files are loaded and merged into 1 single dataframe after loading.

For the loading dataset code to work, the 20 stocks CSV files must be in a directory call ‘nyse-financial-stocks’ and the company details (securities) CSV files would be at the same level as the nyse-financial-stocks directory.

