

**DATA MINING WQD7005**

**INSTRUCTOR: PROFESOR MADYA DR. TEH YING WAH**

**Name: DENG QIAOYIN**

**No.: WQD170071**

**2018/2019 SEMESTER 2**

**Objective:**

Provide a user with recommendations around optimal actions to achieve investment objectives such as profits and return on investment.

**Table of content:**

Milestone 1: Acquisition of data

Milestone 2: Management of data

Milestone 3: Processing of data

Milestone 4: Interpretation of data

Milestone 5: Communication of insights of data

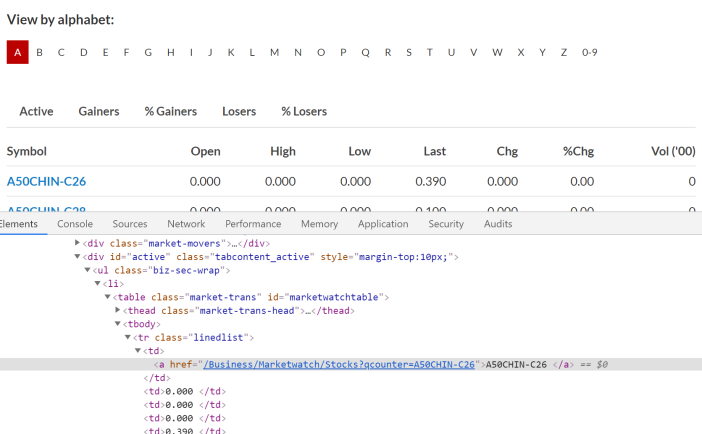
Milestone 6: Provide recommendations

**Milestone 1: Acquisition of data**

Data is crawled from this website:

<https://www.thestar.com.my/business/marketwatch/stock-list/?alphabet=A>

Using Python package selenium to crawl all the stock information based on the alphabetical order. The webpage is shown in Figure1.1.



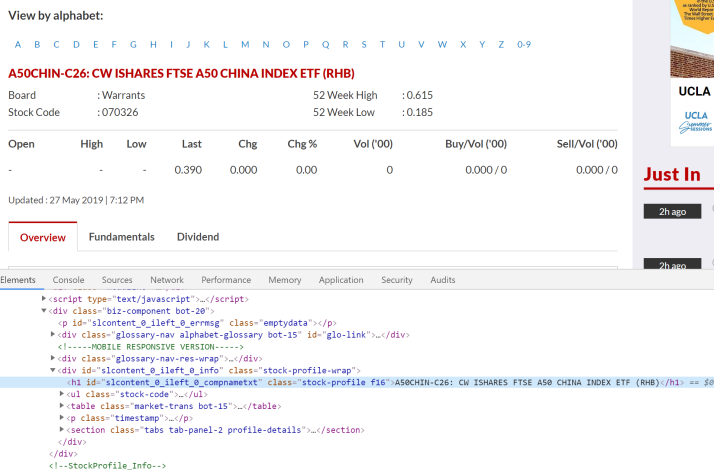


Figure 1.1 Webpage of stock data information

There are around 1800 stocks, and we acquire data by each day’s closing price. Information includes *Stock name, Stock code, Date, Time, Open price, Low price, High price, Closing price, Volumes, buy/volume, sell/volume*. Dataset is as shown in Figure 1.2 below.

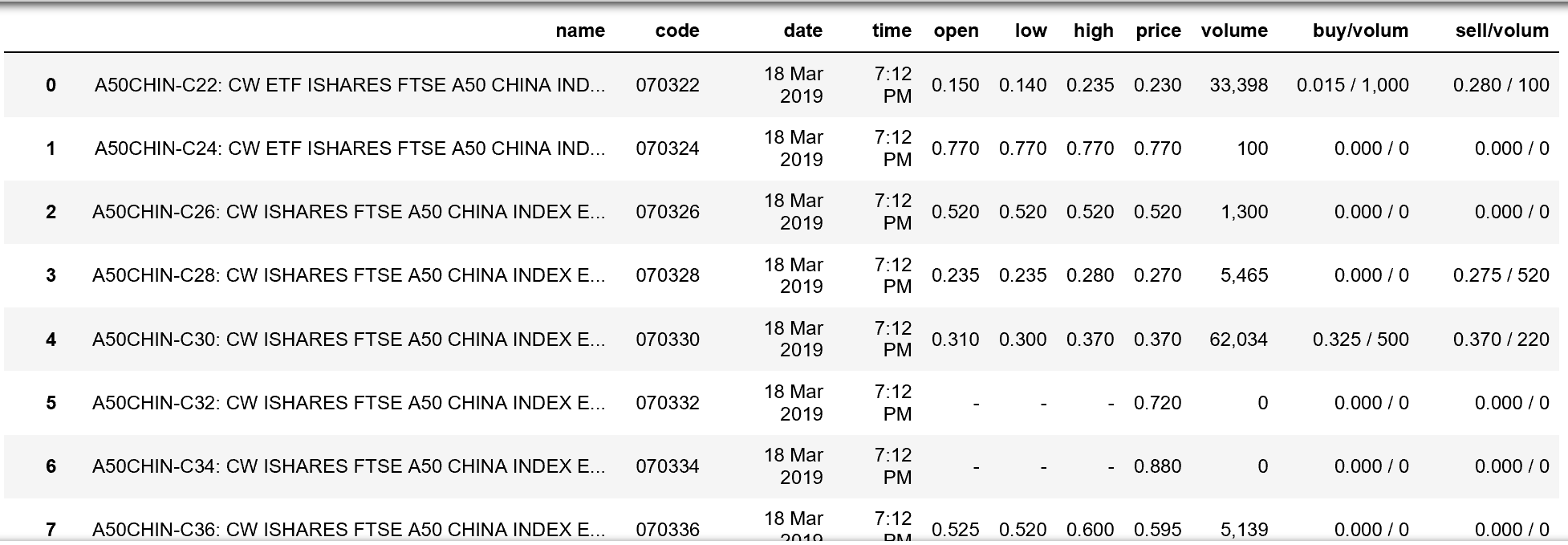


Figure 1.2 Stock dataset crawled from the webpage

Other than that, our group also crawl the currency data by each day.

<http://www.bnm.gov.my/?tpl=exchangerates>

This data includes *Currency unit, Date, Buy rate, Sell rate, Middle rate*, and the dataset is as shown in Figure 1.3.



Figure 1.3 Currency dataset variables

**Milestone 2: Management of data**

As we need to analysis the stock market and the analysis needs variables more than just open, close, low, high price data of a stock. Therefore, 3 more sources are add into the Data warehouse, and we add the *‘sector’* attribute to all data sources. The sector data in Figure 2.1 is crawl from the same source as the stock price data.

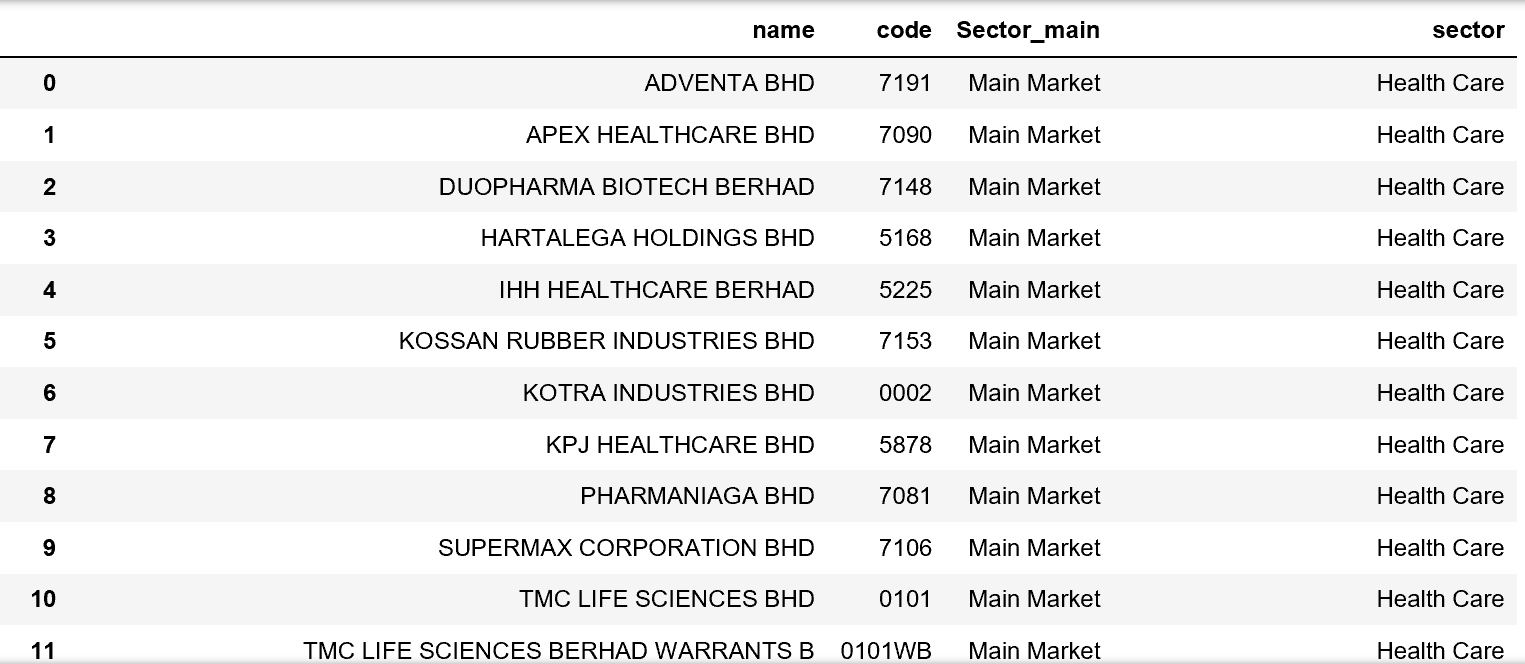


Figure 2.1 Sector dataset

* **Financial Reports**

There are two sources of Financial Reports, in Figure 2.2 and Figure 2.3, each one is list below with a web link and the webpage and result of the crawler.

<https://www.malaysiastock.biz/Annual-Report.aspx>



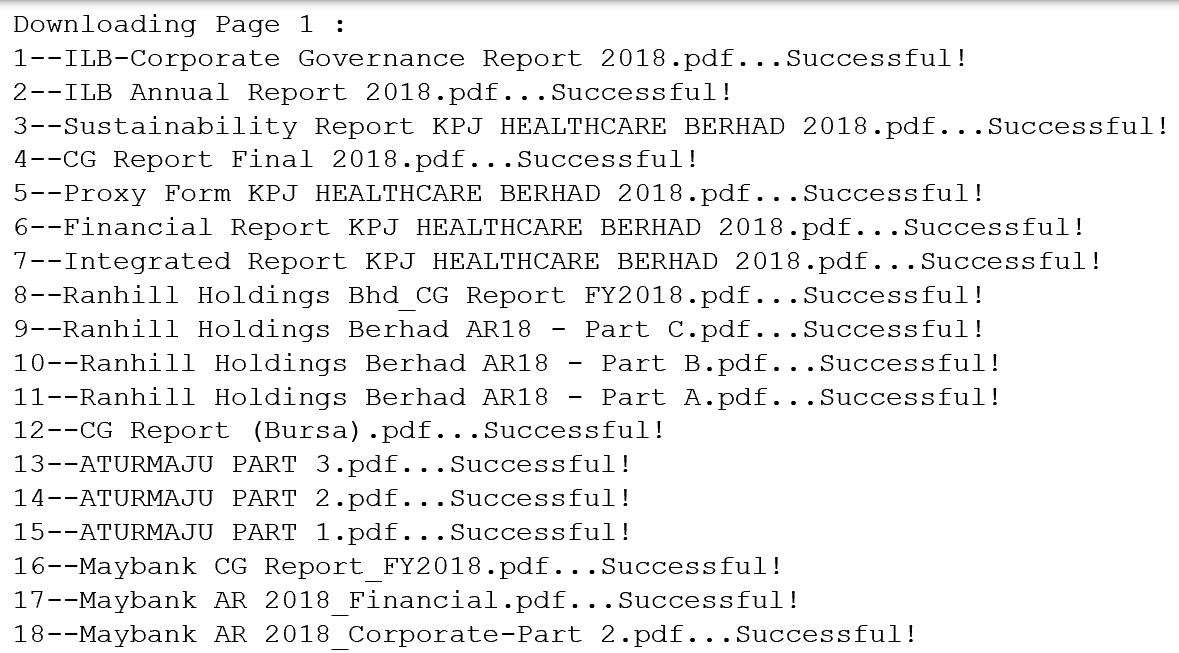


Figure 2.2 Financial Reports PDF webpage and the result of crawler

<https://www.klsescreener.com/v2/financial-reports>

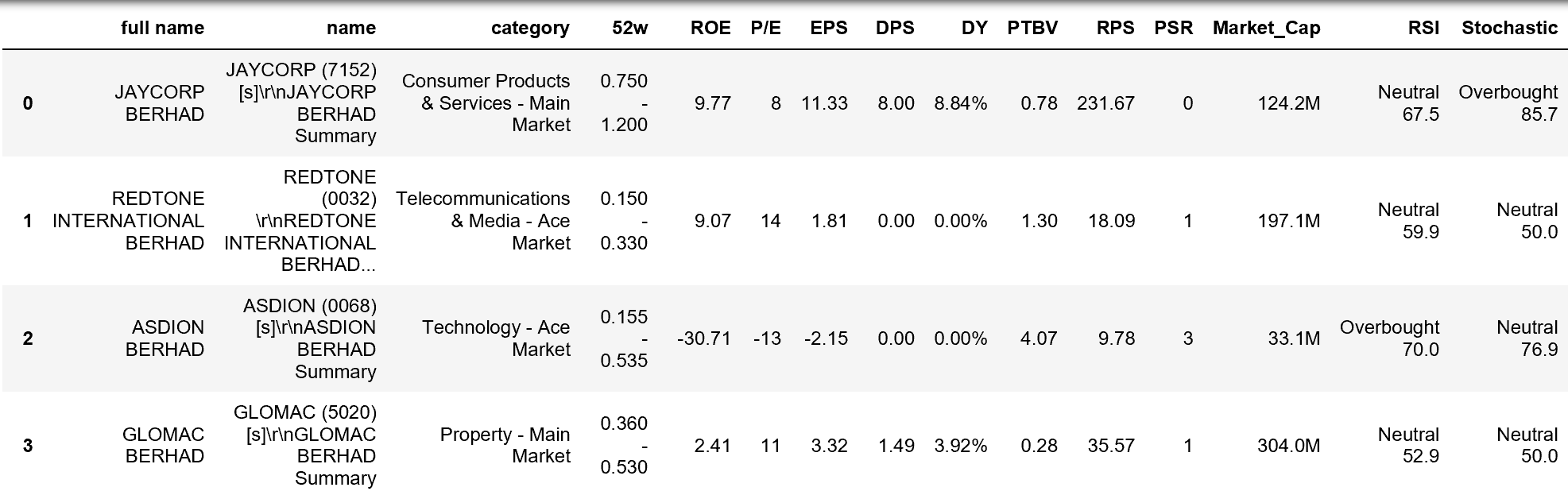
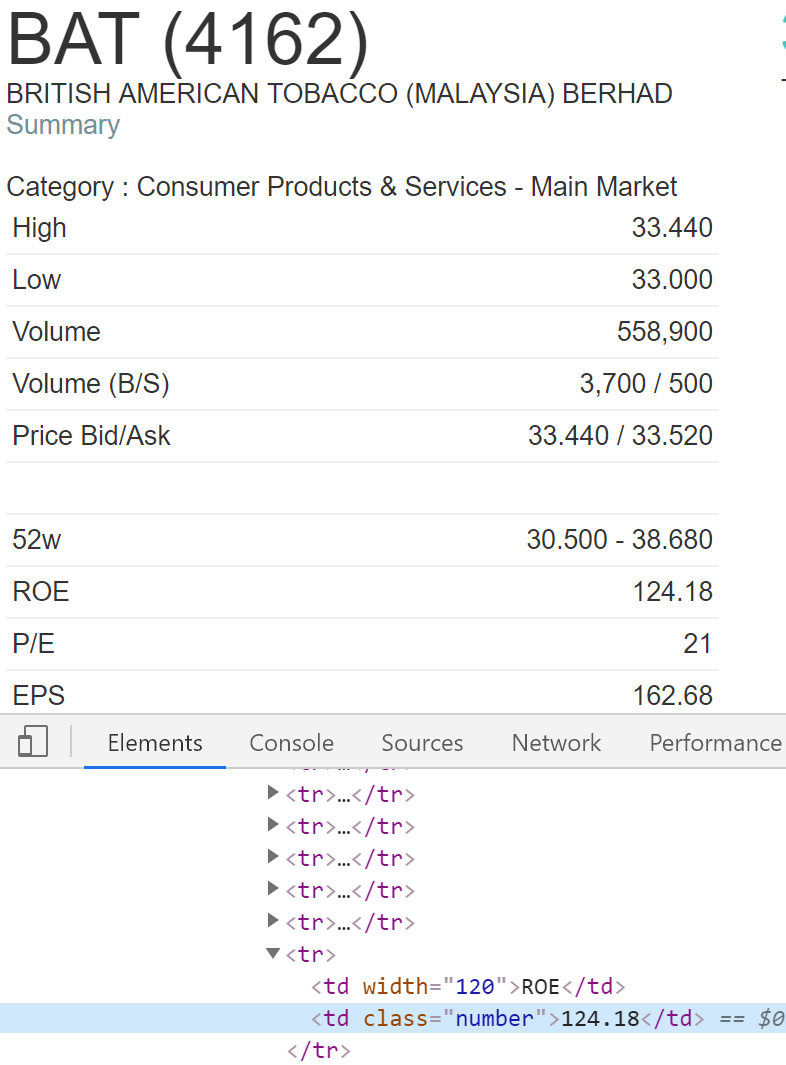


Figure 2.3 Financial data webpage and the result of crawler

* **News Data**

News data has 3 variables and all the information is include, Figure 2.4 shows the webpage and the result.

<https://www.klsescreener.com/v2/news>

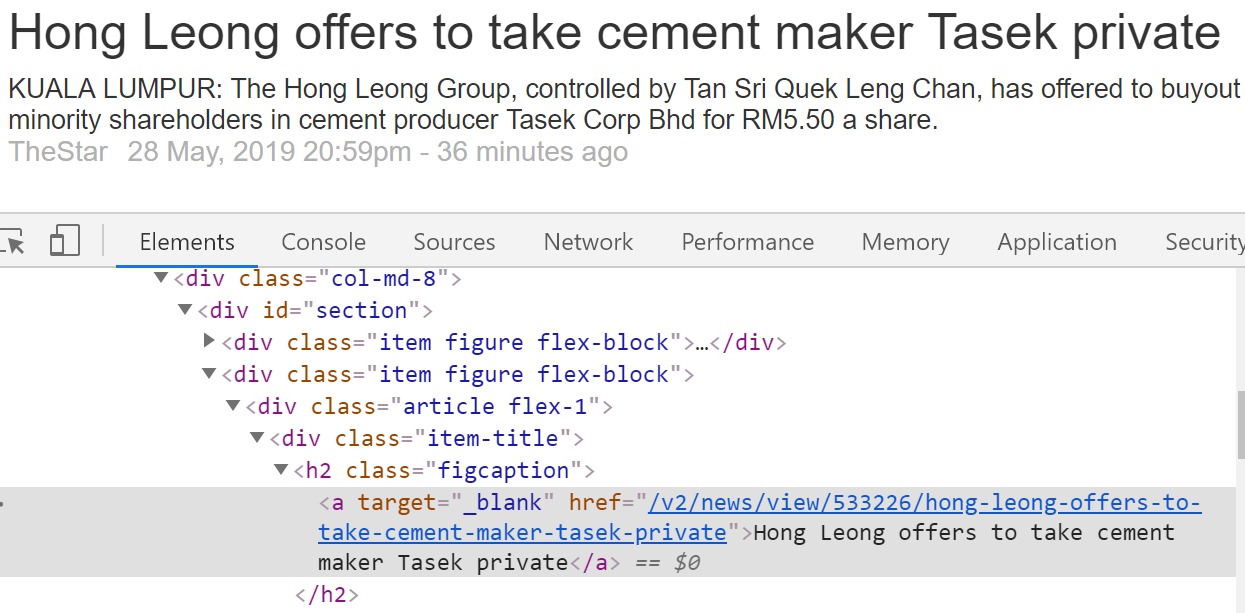


Figure 2.4 News data webpage and the result of crawler

* **Forum Data**

Forum data focus on each post title and the number of reply, Figure 2.5 shows the webpage and the result of crawler.

<http://www.investalks.com/forum/forum.php?mod=forumdisplay&fid=7&typeid=17&filter=typeid&typeid=17&page=1>

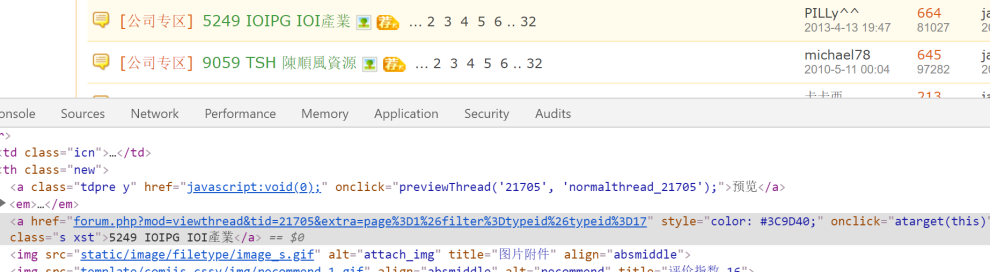




Figure 2.5 Forum data webpage and the result of crawler

Based on the data we had crawl down from the internet, we can draw a star schema in Figure 2.6, for the data warehouse.



Figure 2.6 Stock Data Star Schema

In the star schema, the Fact Table can be drill down to difference sources in difference dimension and each source (dimension) can roll up to the Fact Table using the ‘Stock Code’, and there are 5 dimension table.

**Milestone 3: Processing of data**

1. **Preprocessing of stock and follow by correlation analysis.**

First, all the stock information data is preprocessed and then merge them all together to create a one big stock dataset.

Second, calculate the percentage of change in price for each stock and create a new column to save it. Using the formula:

*Change = (todayPrice - yesterdayPrice) / yesterdayPrice*

Third, all the stock is filter by market and sector, and each of the sectors are saved to a separate file under the same market.

Fourth, based on each sector of the same market, the pivot table is creating to see the relationship of each stock.

Finally, the correlation of each stock is calculate and each level of correlation is highlight in excel file that shown below.

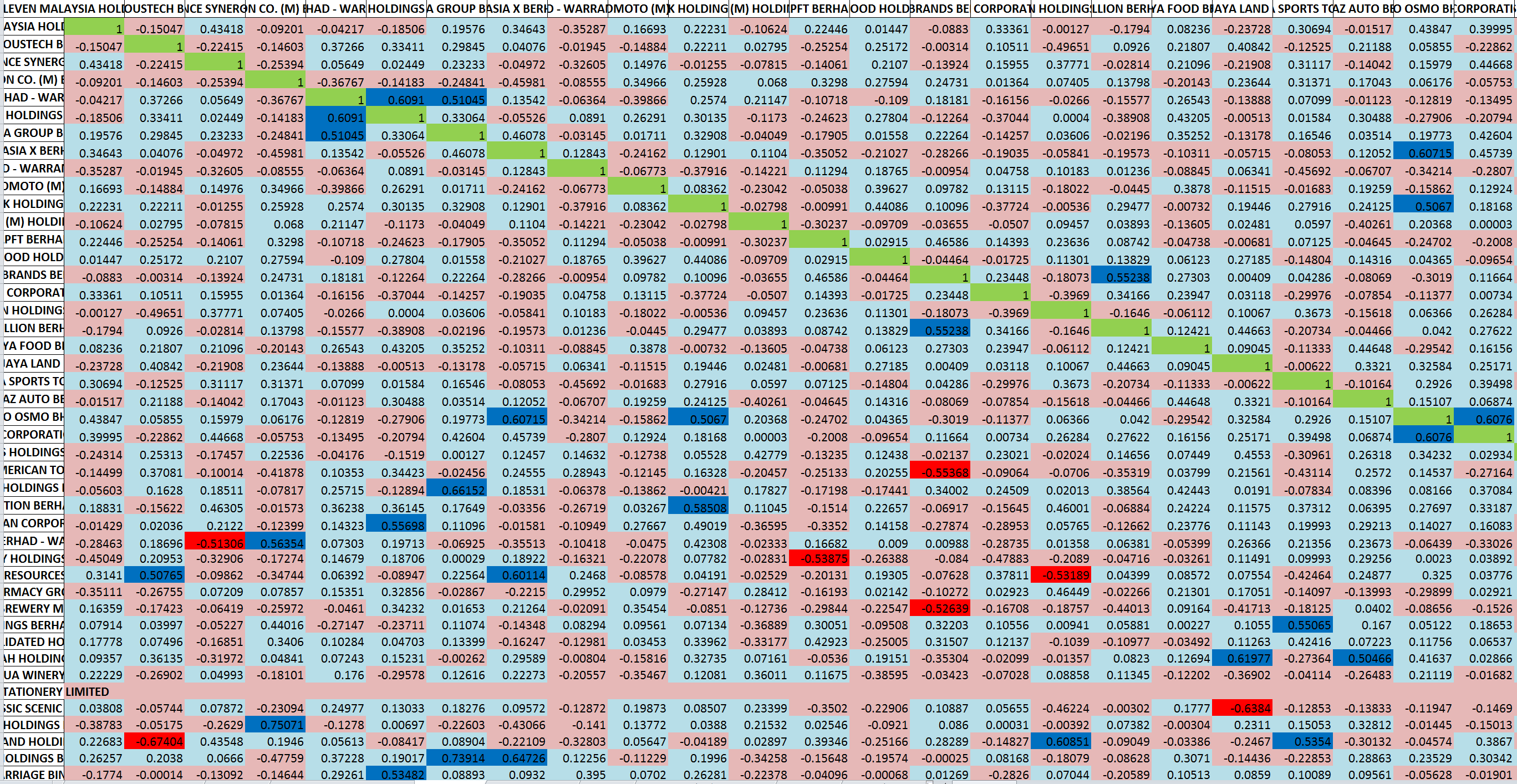


Figure 3.1 Correlation of stock

1. **Dimension reduction**

After the huge amount of stock data we obtain, we can easily be overwhelm by all the numbers in the dataset. The PAA and SAX technique can help us find the pattern in time series data quickly. For example, the company ‘A-RANK BHD’ stock price data is drawn four times with difference transformation.

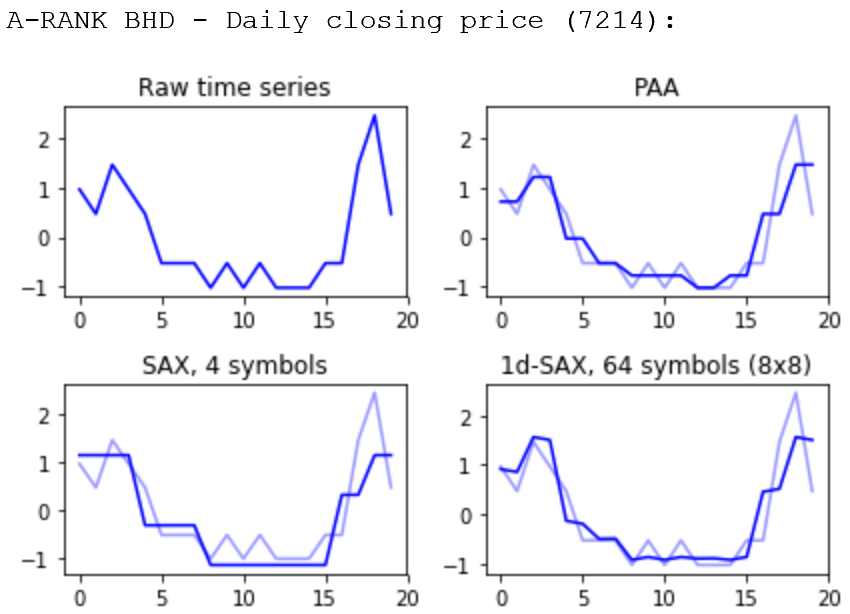


Figure 3.2 Stock data transformation with PAA and SAX

1. **Preprocessing of other sources datasets**

As we have difference sources of website, and we have obtained the raw data for each source, the next step is to preprocess the data.

* **Currency data**

In the currency data, the step is similar to that of the stock data. With all the days of the data, preprocess part is to combined them all together into a single dataset.

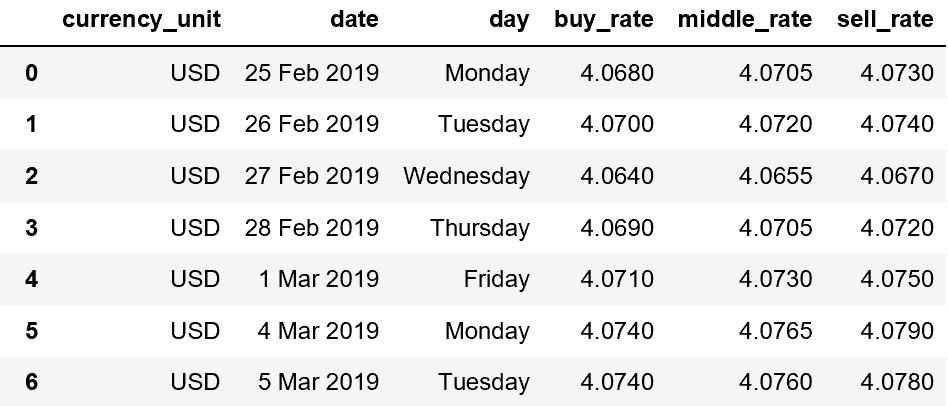


Figure 3.3 Preprocessed currency data

* **News data**

News data is crawled down once a week, and in the preprocess part, all the data is combined together, and the code and date is extracted to a separate column.



Figure 3.4 Preprocessed news data

* **KLSE data**

In KLSE data, the main information is about the financial report of each stock, and all the indicator can be crawl down from the website. And in this preprocess part, after combined all the dataset together, the new data is merge with the stock sector, and some changes is done, like add/drop variable, rearrange columns position, extract code from original column etc.

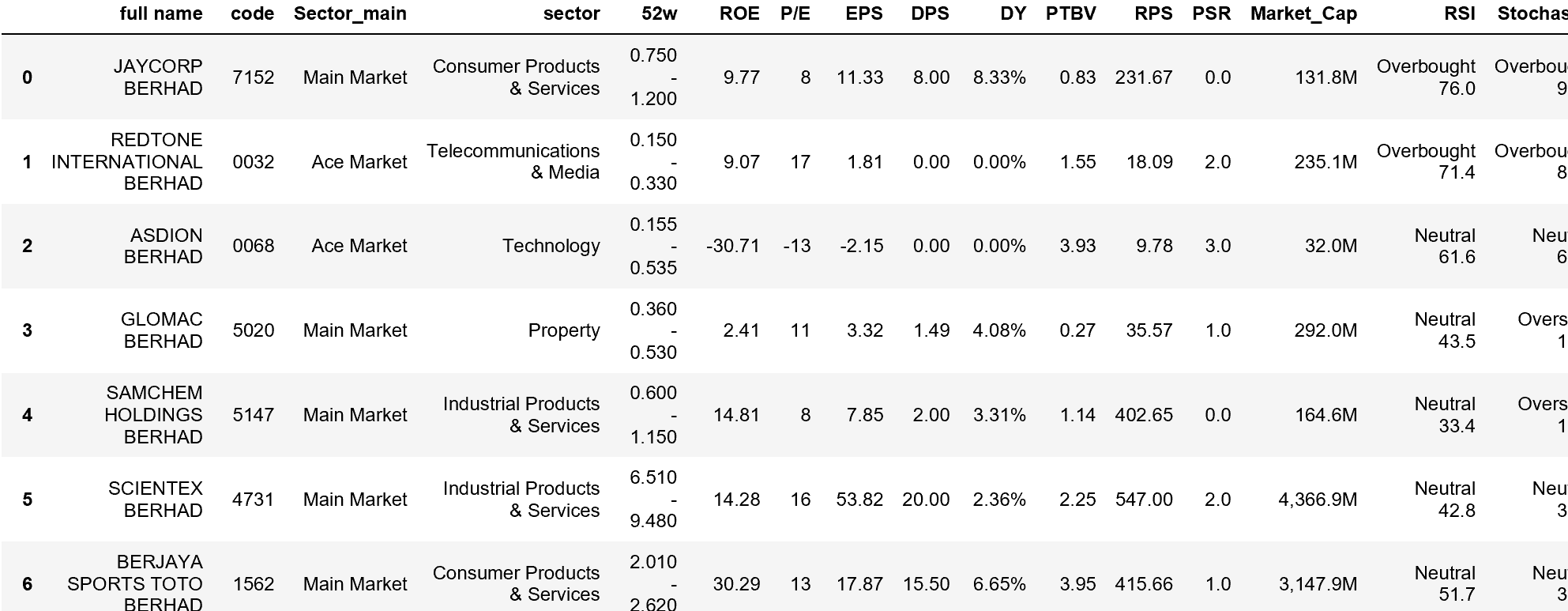


Figure 3.5 Preprocessed financial information data

* **Forum data**

The forum data needs to extract the stock code from the title column, and then merge the stock name from other dataset.

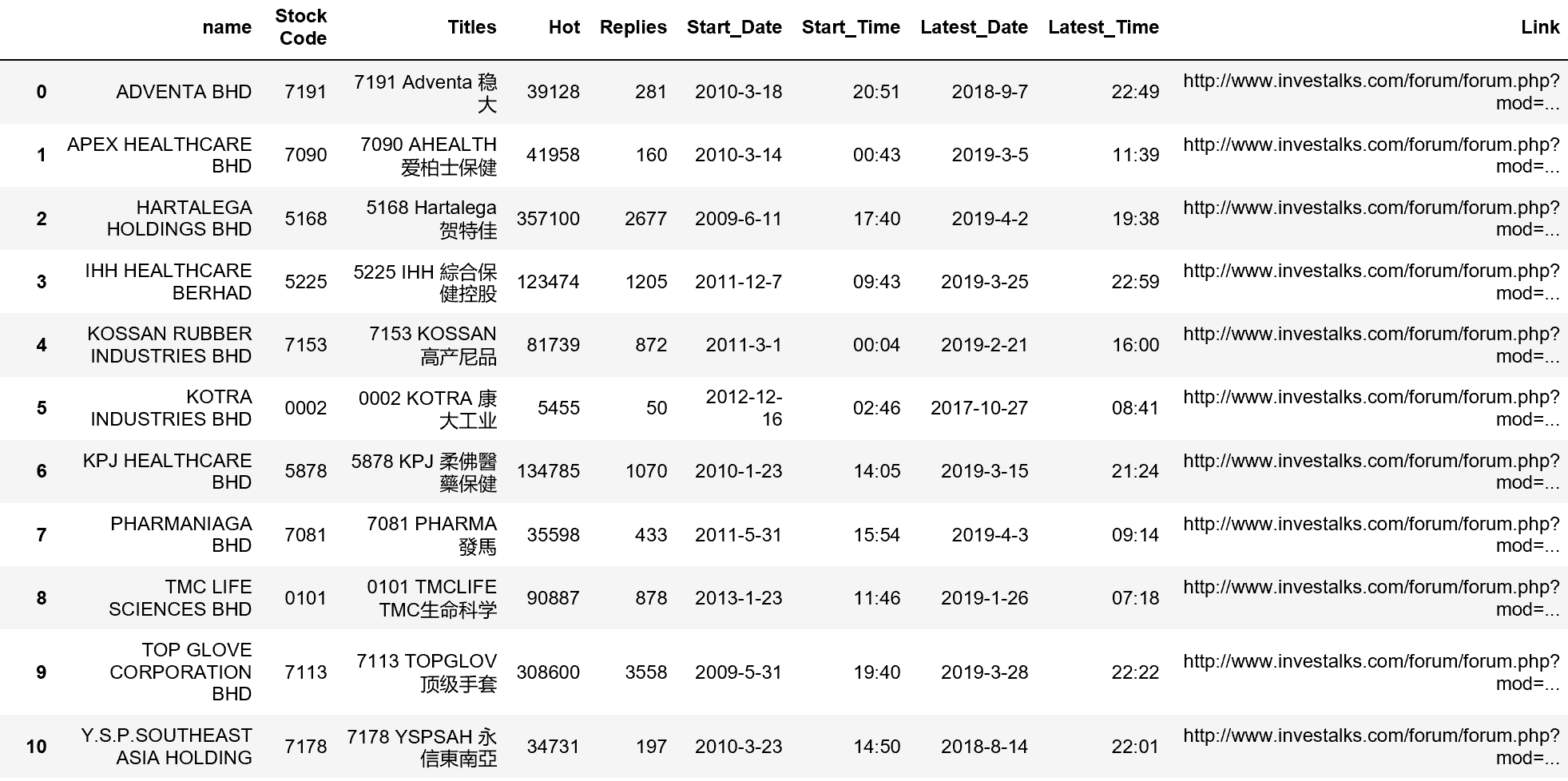


Figure 3.6 Preprocessed forum data

**Milestone 4: Interpretation of data**

With so many stocks out there in the market, this section is to start finding some of the good stock based on the dataset we have.

1. **Select stocks with good index**

To decide if a stock is good for investment or not, first is to use financial report to see how well is that particular stock preform last quarter. There are two sources relate to the financial report, one is the financial report of 2018 in PDF format, the other is already extracted the important information of a stock’s financial report that is about their financial satiation.

* **Financial report data**

In the PDF of financial report, all of the useful data is in a table and embedded in each PDF, and it is hard to use Python to retrieve a certain data that is written inside the table or picture. Therefore, the information we want is put inside a excel file manually done by each group member, and Figure 4.1 is the information of the PDF files.

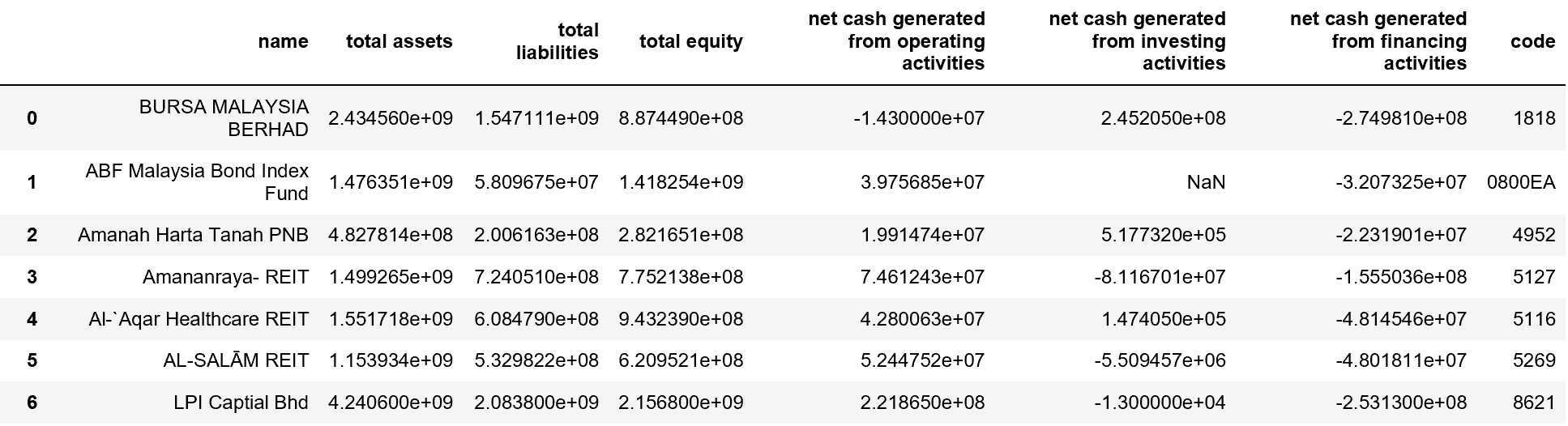


Figure 4.1 Financial data extracted from PDF files

Here we have data that describe the money flows in a stock, and the debt ratio is calculated by dividing total liabilities by total assets. The debt ratio is a financial ratio that measures the extent of a company's leverage. It can be interpreted as the proportion of a company's assets that are financed by debt. Generally, a ratio of 0.4 /40%, or lower is considered a good debt ratio. A ratio above 0.6/60% is generally considered to be a poor ratio, since there's a risk that the business will not generate enough cash flow to service its debt.

The second source has all the basic indicators of a stock that can directly crawl down from the website and merge with the sector information in the dataset.



Figure 4.2 Financial information data obtain from KLSE website

The ROE, P/E, EPS, DPS, DY, RSI, and Stochastic variables are used to assess the stock performance.

1. ROE stands for return on equity. It is a measure of how well a company uses investments to generate earnings growth, the higher the better.
2. P/E stands for price to earnings ratio. It is used for valuing companies and to find out whether they are overvalued or undervalued. The lower the better, means with little money can earn big money.
3. EPS stands for earning per share. The resulting number serves as an indicator of a company's profitability. The higher a company's EPS, the more profitable it is considered.
4. DPS stands for dividend per share. It is the sum of declared dividends issued by a company for every ordinary share outstanding, the higher the better.
5. DY stands for dividend yield. It shows what return expressed as dividends an investor will get per each dollar of his/her investment in company’s shares.
6. RSI stands for Relative Strength Index. It is a momentum oscillator that measures the speed and change of price movements.
7. Stochastic indicator shows us information about momentum and trend strength.

With all these indicator in the dataset, first the column name is renamed to standardize with other dataset, and reformat some variable, then with the RSI and Stochastic each having the value of “Oversold”, “Neutral”, “Overbought” are turn into number 1, 0, -1, and then get the subset of data that only left the stock with the result of RSI plus Stochastic that larger and equal to 0, which means that these stocks can buy or hold.

After that, sort the dataset in the order of higher ROE, lower PE, followed by higher EPS, DPS and DY to get the good stocks.

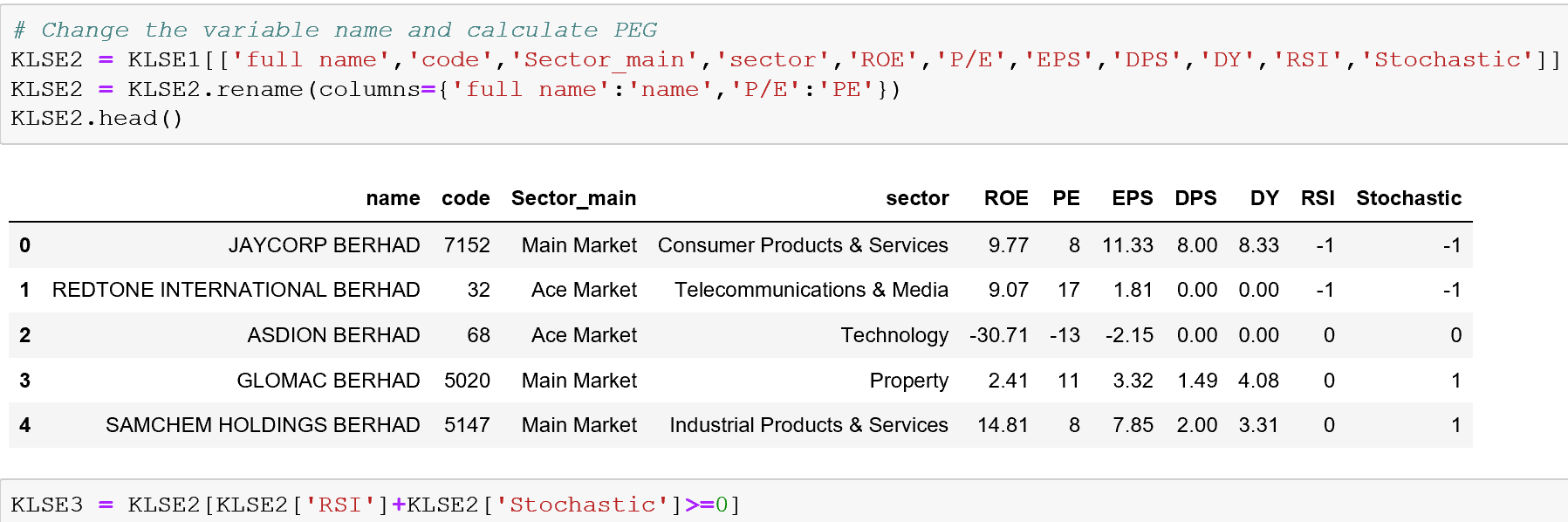
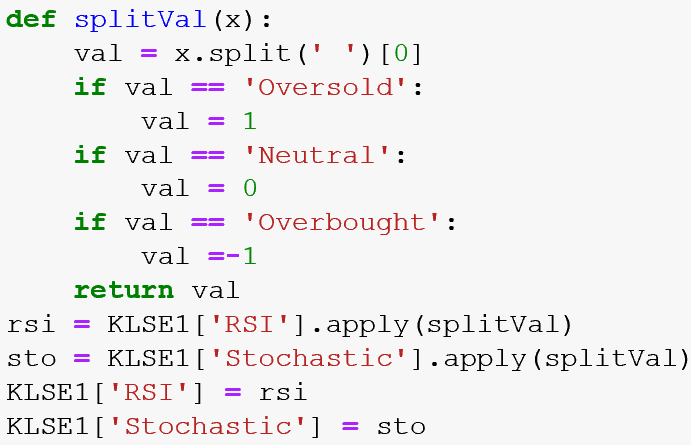
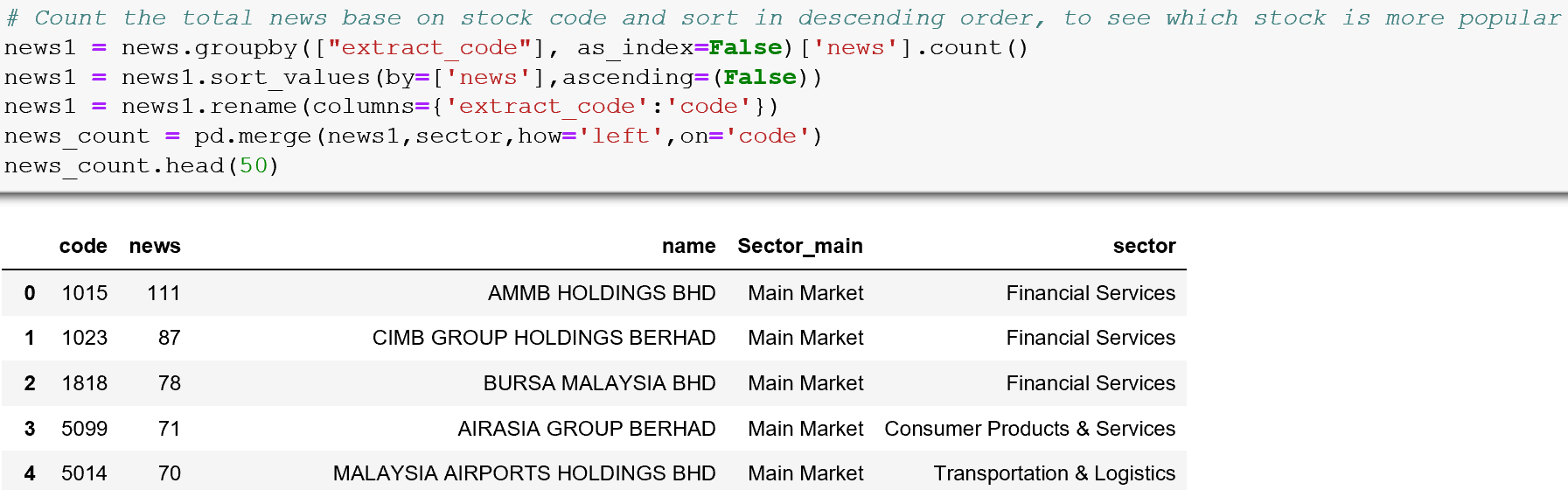




Figure 4.3 Python code to select stock and the results

* **News and Forum Data**

A stock’s performance is affected by the attention of investor. Generally, a stock with more news or comment means the stock is popular than others. Thus, the news dataset is group by the stock code and get the count of each stock to see the number of news within this period of time. And the forum dataset is group by code then sum all the replay to get the rank of the stocks.



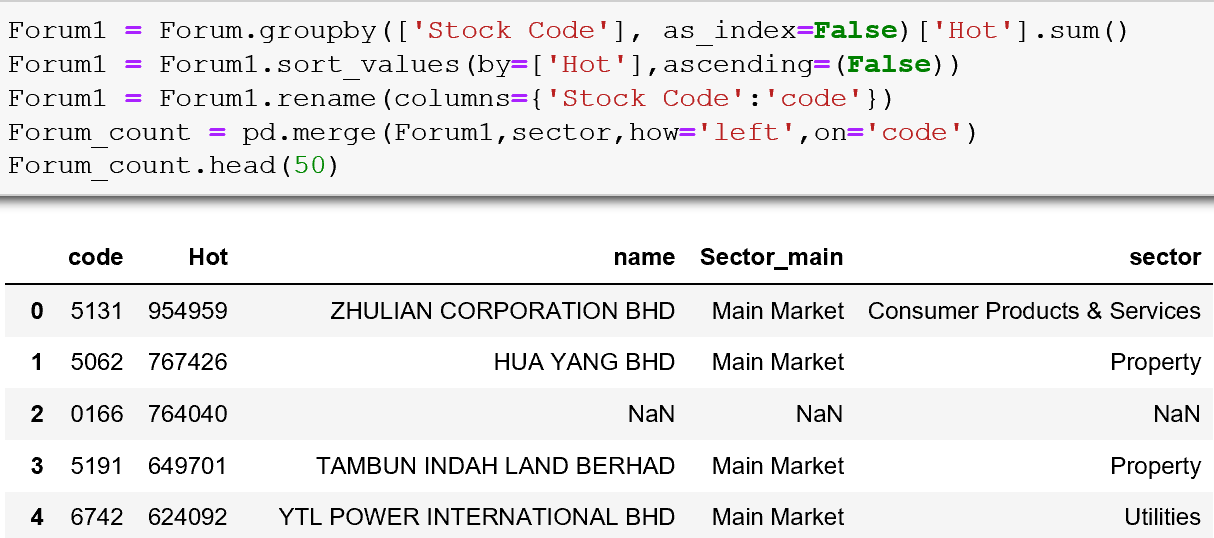


Figure 4.4 Python code to get the count of news and comments

Combining all those data set we can get the subset that choosing the top 50 stock from each dataset and then get the stock that is in two or more dataset as the final invest stock list and save as csv file.

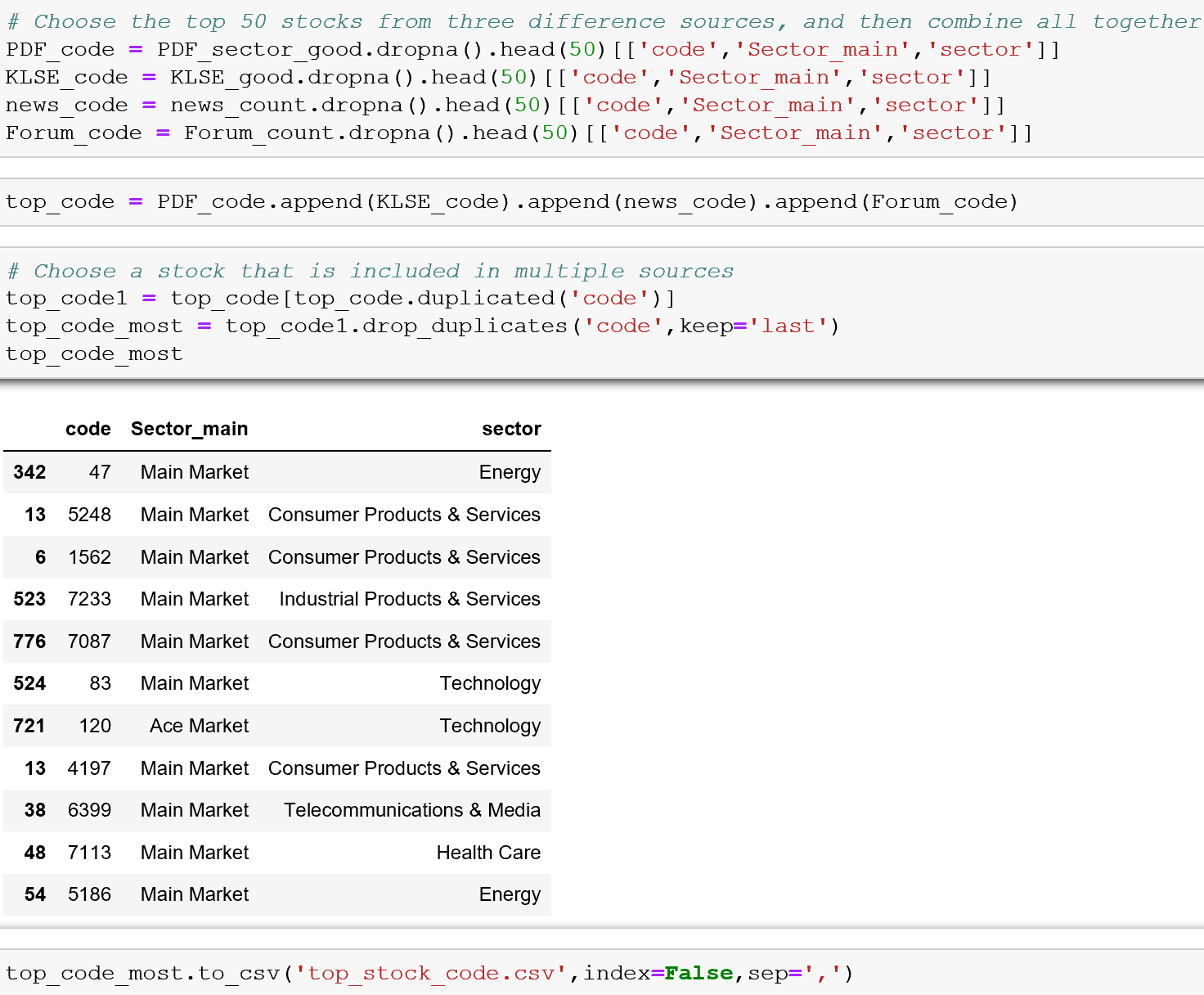


Figure 4.5 Find the stock list and save as csv file

Then, to get the subset of stocks, which we consider as good stocks, based on the indicator from the stock price dataset. Figure 4.6 is the stock that we’ve got, there are total 28 stocks in the list, and they are all from the ‘Main Market’ and ‘Ace Market’, and distribute in 8 difference sectors.

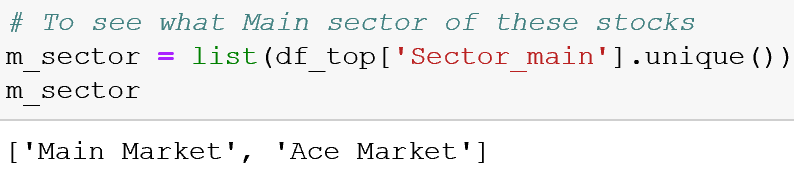
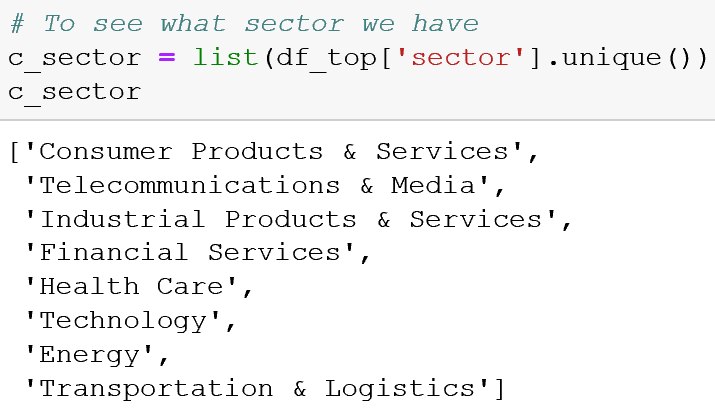
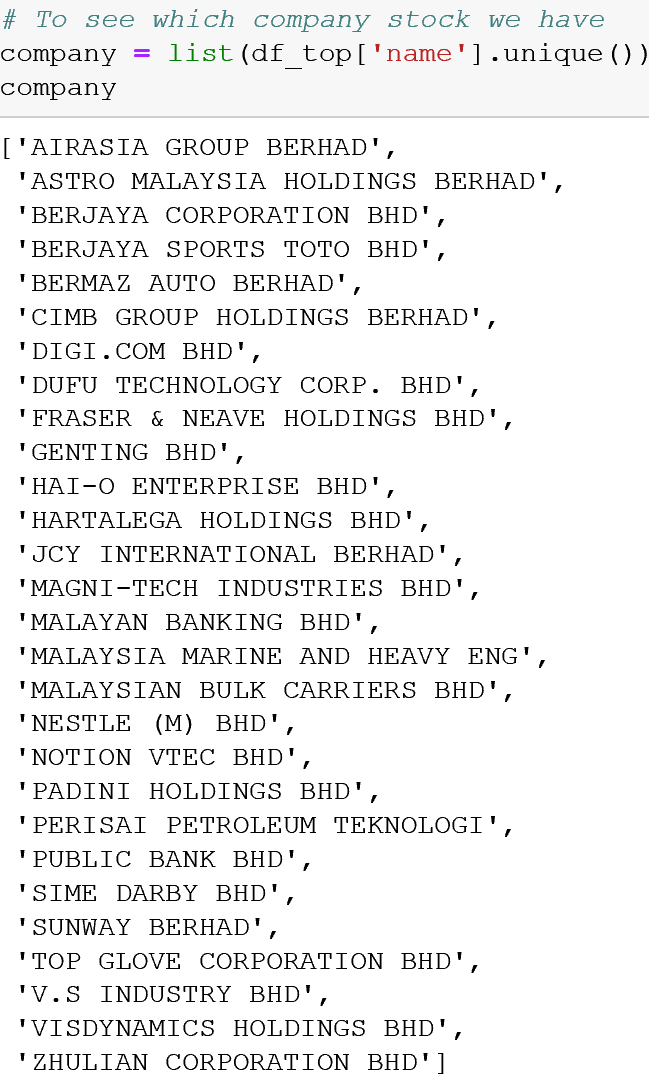
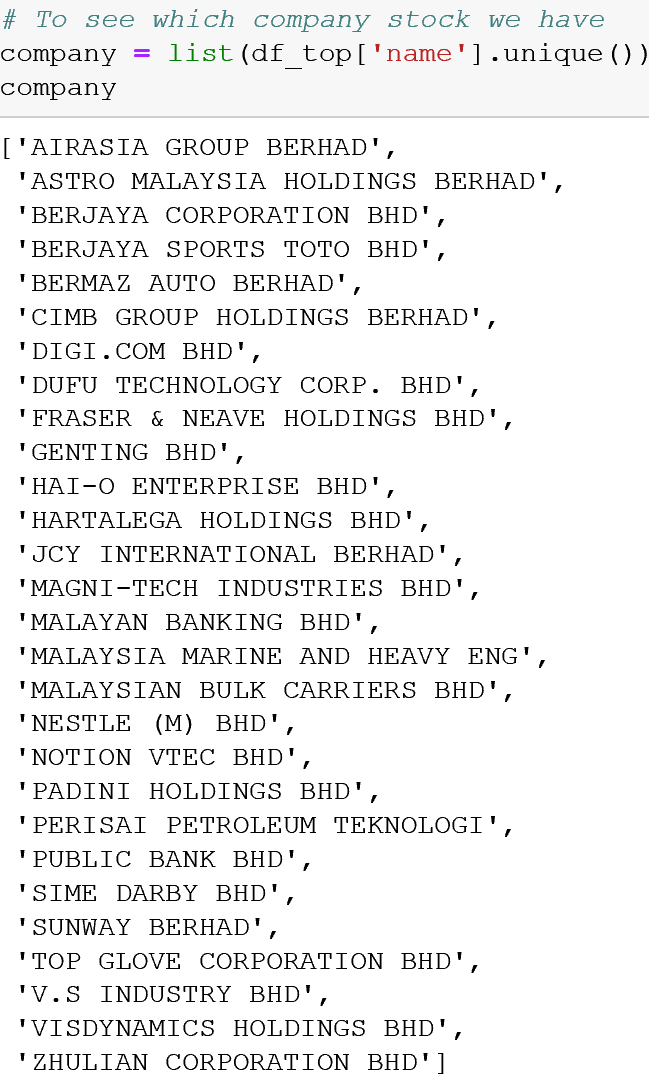


Figure 4.6 Final stock lists

* **Sentiment analysis**

After select the stock, the next step is to visualize the stock’s performance. First is to get the sentiment of news data for each observation. As the news data contain both English and Chinese, it is best to separate the news dataset, and choose the ‘SnowNLP’ package to calculate the Chinese news sentiment, and ‘TextBlob’ package to calculate the English news sentiment.

Since in the Chinese sentiment, the score range from 0 to 1, and the English sentiment score range from -1 to 1, we needs to standardize the two sentiment score and combined into one dataset, and the score becomes -1 indicate negative emotions, 0 indicate neutral emotions, and 1 indicate positive emotions.

1. **Visualization of a stock’s performance**

After the selection of some good stock, the Consumer Products & Services sector has the most stocks, 11 stocks of them, in the list. Therefore, this report will focus on the sector of Consumer Products & Services and those 11 stocks. For the visualization, the steps are univariate analysis, bivariate analysis, some technical graph, and correlation within those 11 stocks.

First, all the stocks are extract to a different DataFrame, and prepare for the analysis.

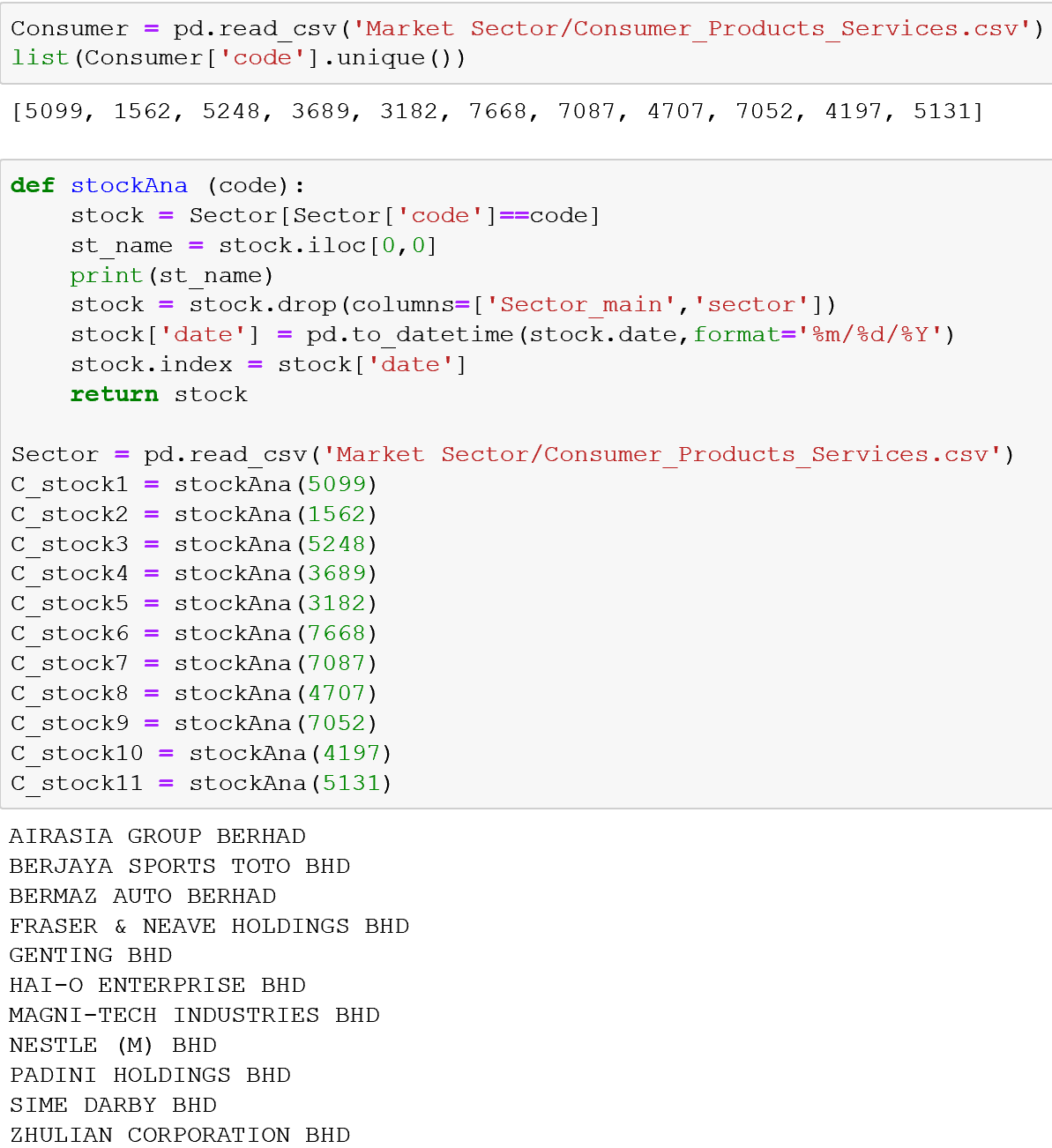
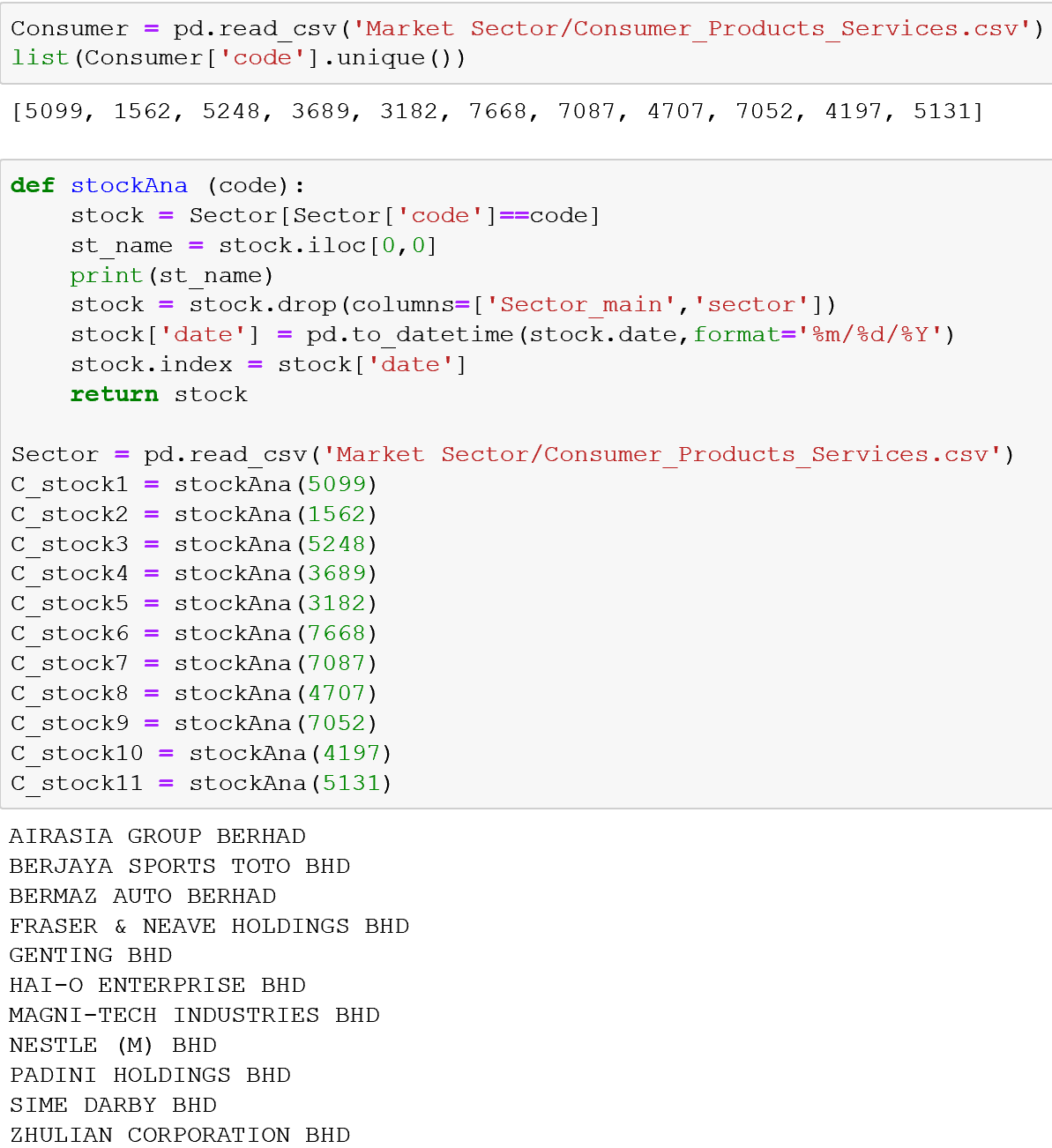
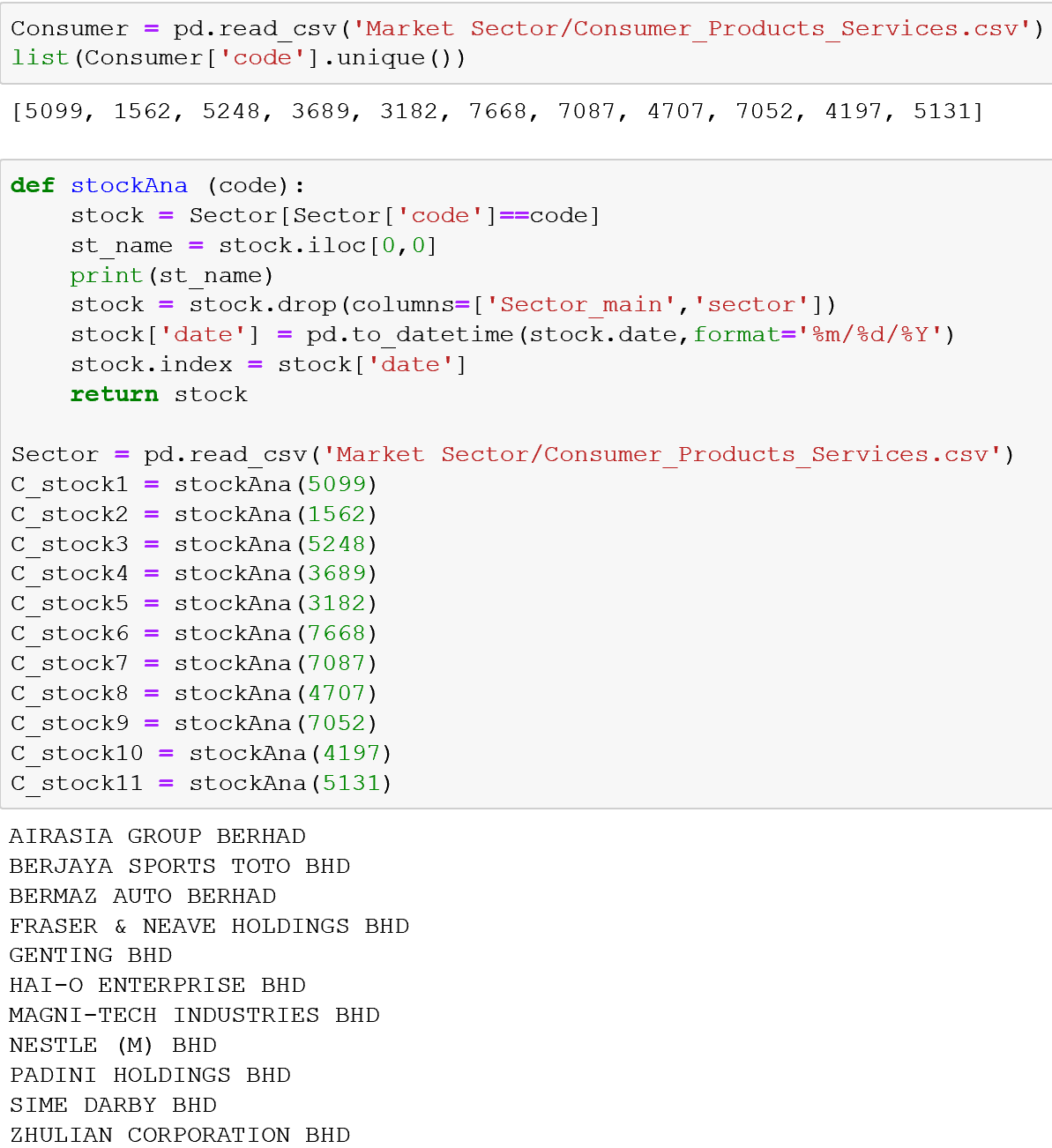


Figure 4.7 The 11 stocks in Consumer Products & Services sector

* **Univariate analysis**

In this part, the first stock in the list, AIRASIA GROUP BERHAD, is chosen for the analysis. Since the stock data are all time series data, we can plot the graph that shows the data trend from day to day using the ‘price’ variable.

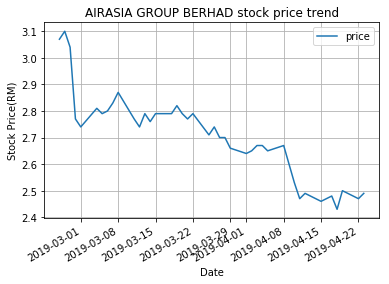


Figure 4.8 Stock price trend of AIRASIA GROUP BERHAD

As we can see from Figure 4.8, the stock price trend is going downward. Then let’s plot the other variables.

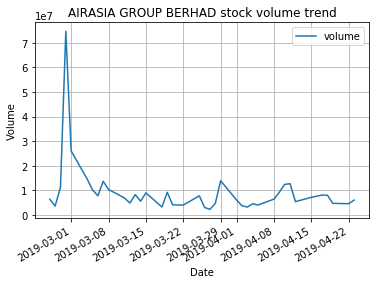
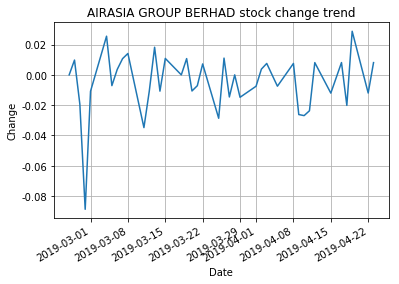
 

Figure 4.9 Stock trends in volume and change of the stock

In fact, price of a stock is affect by the volume of a stock, and the change is based on the closing price of each day.

* **Bivariate analysis**

In this part, we can visual how one variable affecting another variable. Figure 4.10 shows the moving average and the price data in the same graph, it shows the smoother trend of price trend in 5 days basic.

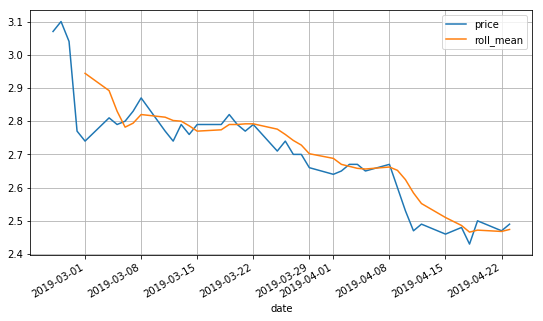


Figure 4.10 Price vs Moving average

In univariate analysis, the pattern of price and volume is related to each other. From Figure 4.11 we can see the volume go down and the price also go down, and vice versa, most of the time.

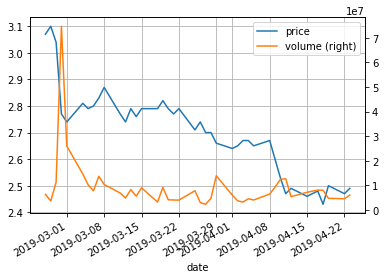


Figure 4.11 Price vs Volume

Knowing that this stock price is going downward, let’s take a look at the Market index trend and the stock price trend. As shown in Figure 4.12, the two trends are similar, and that means the stock is one of the stock that has great effects on the whole Market.



Figure 4.12 Stock price vs Market index

News also affects the price of a stock. There are three types of sentiment, positive, neutral, and negative. Because no one stock can get news every day, so the day without the news data is label as neutral in the dataset, and Figure 4.13 shows the price and sentiment of the stock, we can see that positive news reflects on increasing price, and vice versa.

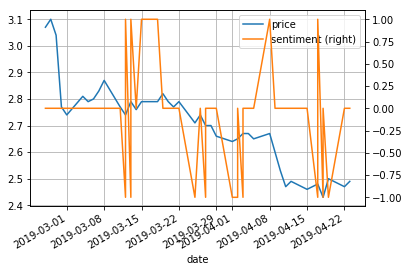


Figure 4.13 Stock price vs Sentiment analysis

In short, the rise and fall of stocks is affected by many factors, which may be due to policy reasons, environmental reasons, internal reasons, and the correlation between different companies.

* **Candlestick chart**

Line segment chart is feasible, but the data for each day has at least four variables (opening, stock price, stock price and closing), we hope to find a visualization that does not require us to draw four different lines to see the trend of these four variables. In general, we use the candlestick chart (also known as the Japanese candlestick chart) to visualize financial data. As shown in Figure 4.14. The candlestick chart was first used by Japanese rice merchants in the 18th century.



Figure 4.14 Candlestick chart of the stock

The green line in the candlestick chart represents the closing price of the trading day is higher than the opening price (profit), and the red line represents the opening price of the trading day is higher than the closing price (loss). The tick marks represent the highest and lowest prices of the day's trade.

* **Correlation of stocks**

First, plot the entire stock price in a graph, as shown in Figure 4.15.

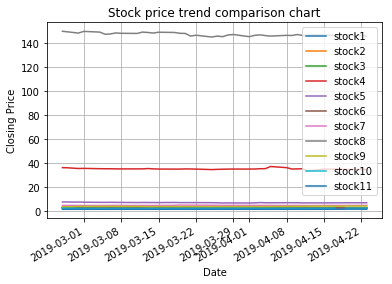


Figure 4.15 Stocks price trend

The problem of this graph: Although the absolute value of the price is important (expensive stocks are difficult to buy, this will not only affect their volatility but also affect the ease with which you trade them). But in the transaction, we pay more attention to the change in the price of each stock rather than its price.

A "better" solution is to visualize the information we actually care about: the return of stocks, which requires us to make the necessary data transformations. There are many ways to convert data. One of the conversion methods is to compare the stock price of each trading day with the stock price starting from the time period we care about. That is:

*Return = Price(t) / Price(0)*

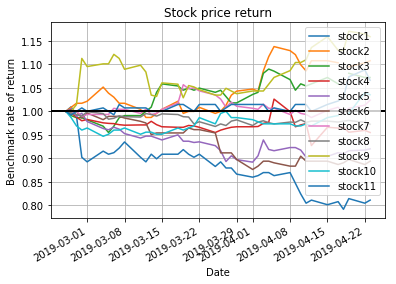


Figure 4.16 Stocks price return

Now we can see how high the return of each stock is from the date we care about. And we can see that the correlation between these stocks is slightly positive. They basically move in the same direction, which is difficult to observe in other types of charts.

We can use logarithmic differences to represent stock price changes:

*Change(t) = log(price(t)) - log(price(t-1))*

The log here is the natural logarithm. The advantage of using a logarithmic difference is that the difference value can be interpreted as a percentage difference in the stock and is not affected by the denominator.

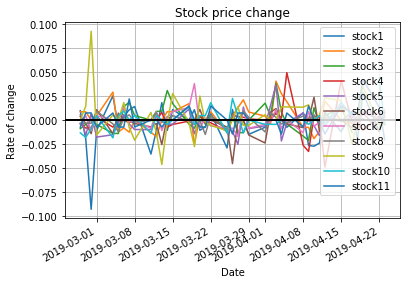


Figure 4.17 Stocks price change

Next is to plot the scatter matrix to visually see the correlation within these 11 stocks.

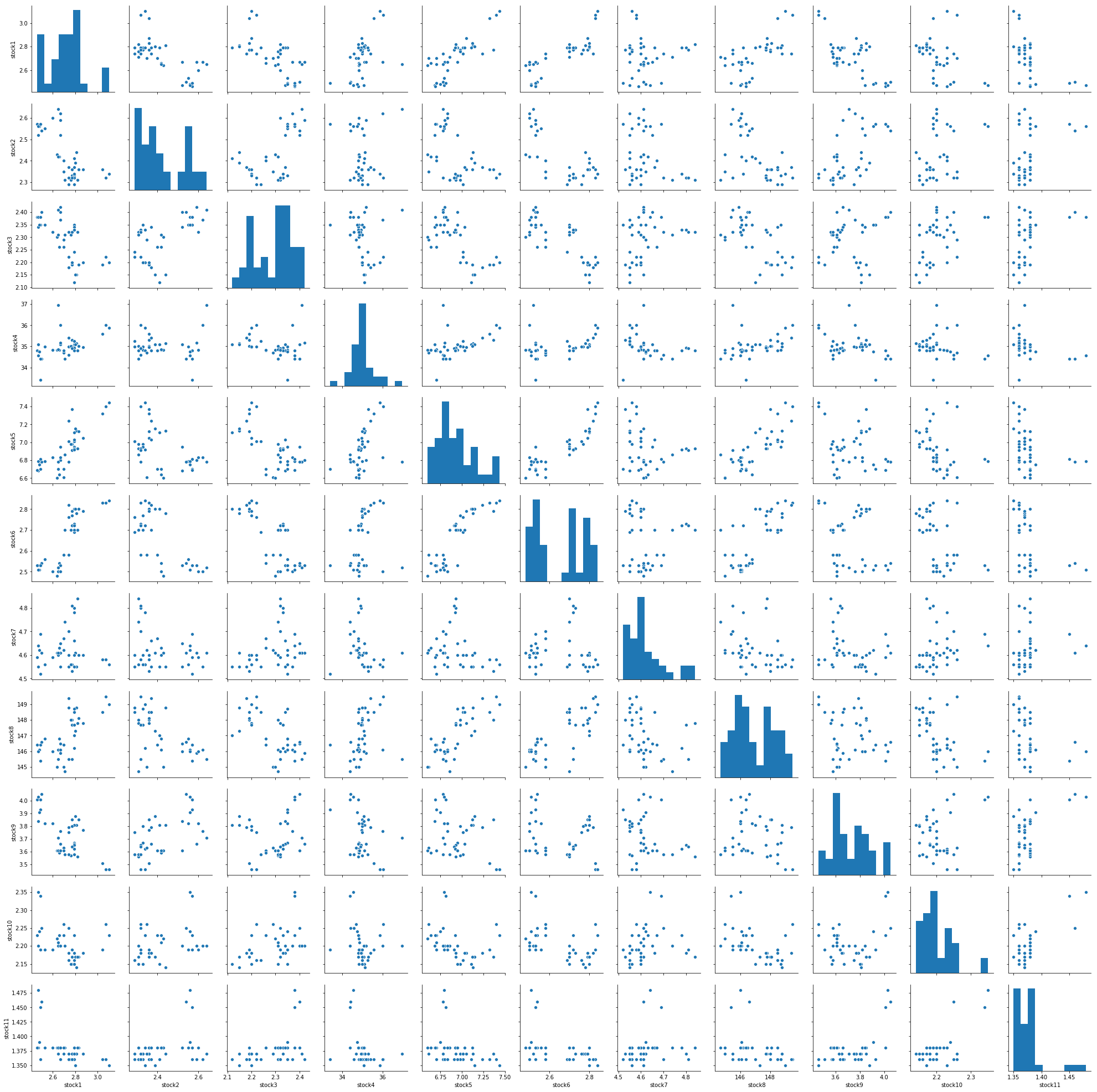


Figure 4.18 Stocks correlation comparison

We can see that stock 1, stock 5, stock 6 and stock 8 have a positive correlation, while stock 11 has no correlation among these stocks.

To better verify this correlation, the SAX technique can help to find pattern between those stocks. Using the SAX to convert the time series data into different symbol, because of the dataset is small and only have around 40 days of price data, only 4 symbol, a, b, c and d is used to compare the stocks.

Table 4.1 Stock data transformation

|  |  |  |  |
| --- | --- | --- | --- |
| **Stock1** | ddccbbaaa | **Stock7** | abbddbabd |
| **Stock2** | abaaaacddd | **Stock8** | dcdcaabba |
| **Stock3** | aaaccccddd | **Stock9** | acbaabcdd |
| **Stock4** | dccbbdbba | **Stock10** | caaabdcbbd |
| **Stock5** | ddccaaaab | **Stock11** | bbbbbbbdd |
| **Stock6** | ddccbaaaa |  |  |

With the result of transformation, stock 1, stock 5, stock 6 and stock 8 have same pattern. These 4 stocks are positive correlate and when one stock’s price goes up, other stocks will likely to go up as well. Thus the 4 stocks are selected as the stock portfolio and label them to prepare for modeling.

Use the stock code of the 4 stocks, and extract from the stock price dataset, and based on the percentage of price change and sentiment score in the dataset to label the day of the stock that is buy, hold, or sell. Dataset is shown below.



Figure 4.19 Stock dataset with labels

**Milestone 5: Communication of insights of data**

Stock investment can be divided into two parts, one is the selection of stocks, and the other is to decide whether to buy or hold or sell of those stocks.

1. **Selection model**

The tool to build the model is SAS Enterprise Miner, and there are two main diagrams in the SAS project. The first one is the selection of stock, the dataset used is from all the sources we have collected and through preprocess of the dataset, the variables is shown in Figure 5.1. And the diagram of this dataset is shown in Figure 5.2.

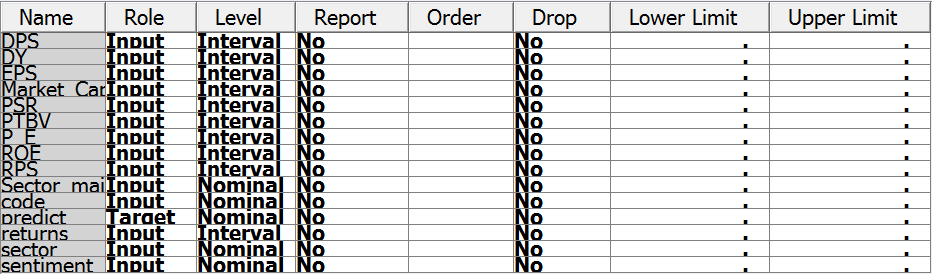


Figure 5.1 Metadata of the selection model

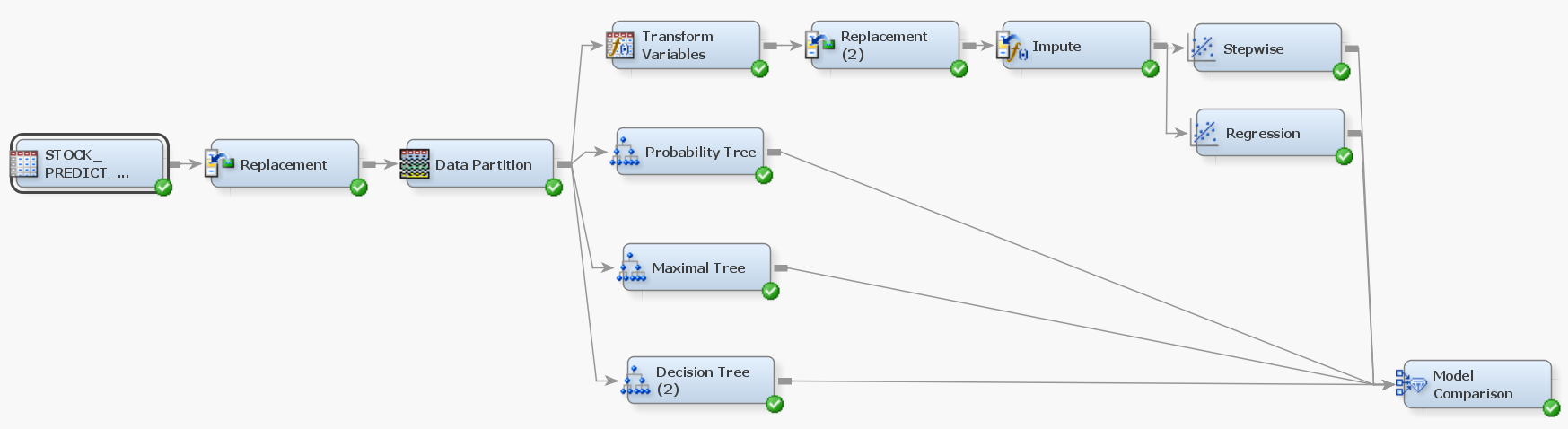


Figure 5.2 SAS diagram of the selection model

With the Decision Tree model result shows, the ‘returns’ variable has the highest information gain, follow by ‘sentiment’ variable. And the output shows 96% accuracy.

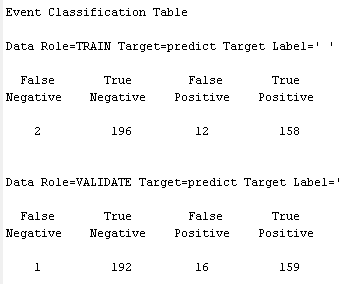
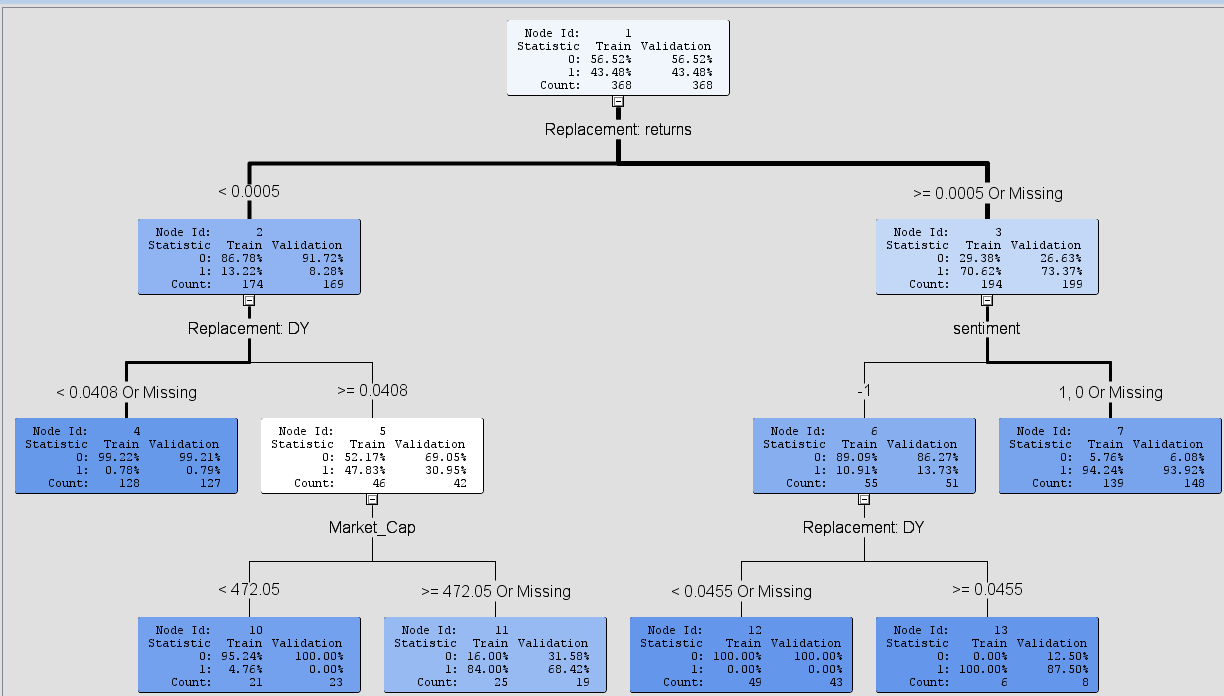


Figure 5.3 Decision Tree plot and Classification Table

In the model comparison, the output shows that the decision tree model is better than the regression model.

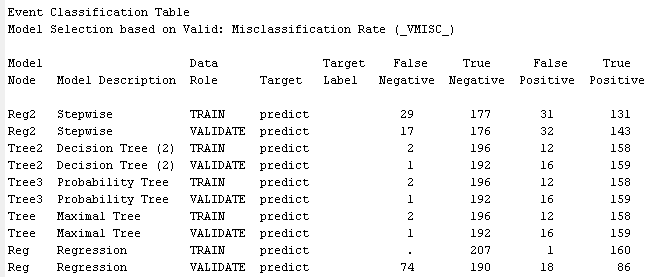


Figure 5.4 Selection model comparisons output

1. **Prediction model**

The dataset contains 11 columns and 4 stocks extracted from Milestone 4.

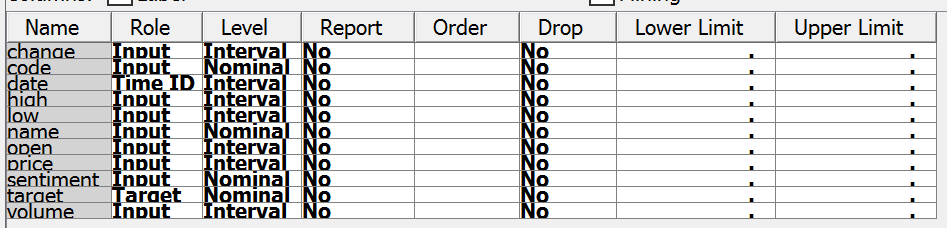


Figure 5.5 Metadata of prediction model

And dataset is put into the Enterprise Miner to build the Decision Tree Model and Regression Model, the SAS diagram is as shown below.

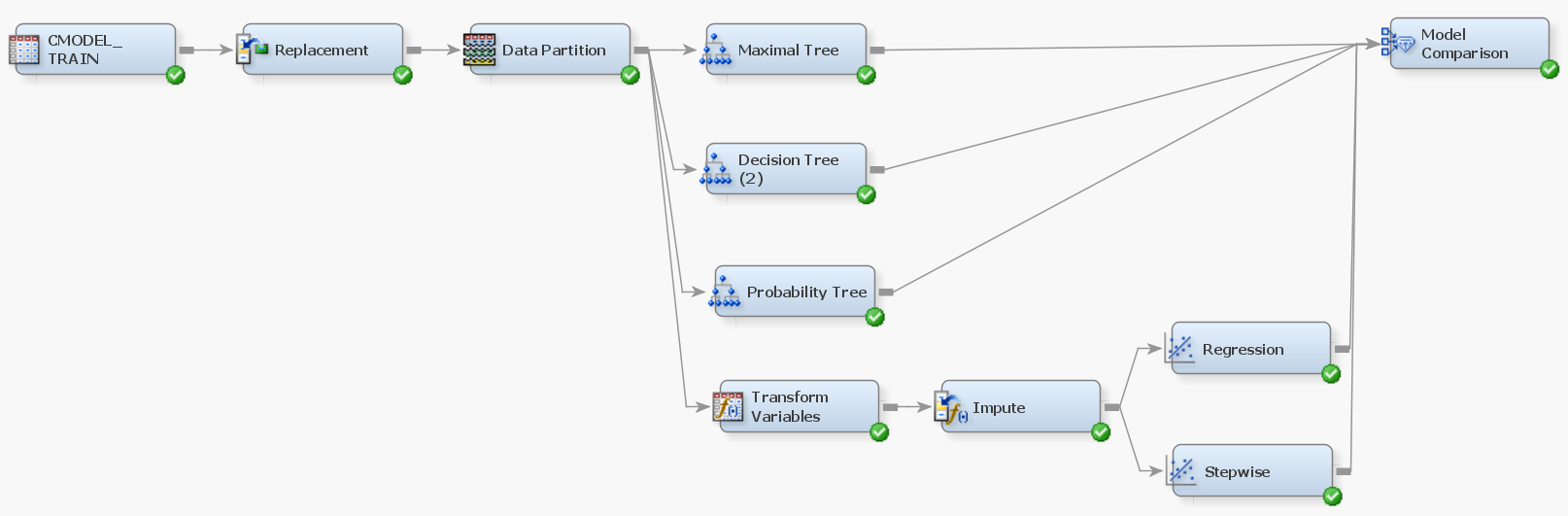


Figure 5.6 SAS diagram of prediction model

Decision tree model shows the ‘change’ variable has the highest information gain, and that decide the split of the tree, and sentiment is less important in the dataset. The output of the decision tree model has 93% accuracy.

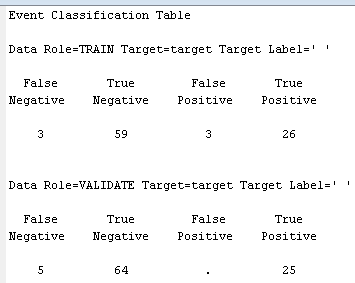
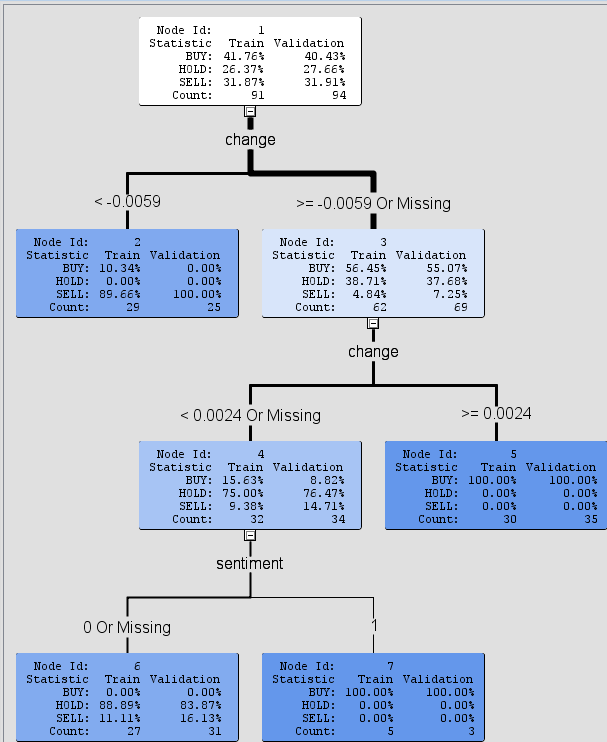


Figure 5.7 Decision Tree plot and Classification Table

The comparison of decision tree model and regression model shows that decision tree model is robust and has better accuracy than the regression model.

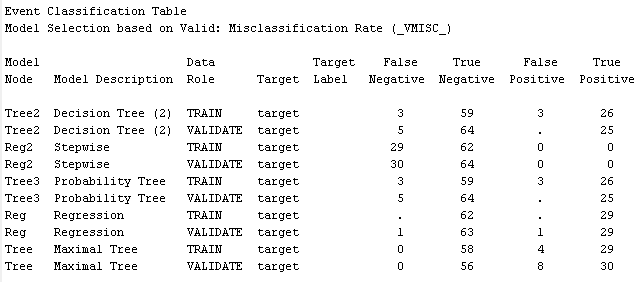


Figure 5.8 Prediction model comparisons output

**Milestone 6: Provide recommendations**

Financial quantitative investment refers to the process of using computer technology and adopting certain mathematical models to implement investment ideas and realize investment strategies.

The advantages of quantitative investment are:

- Avoid subjective emotions, human weaknesses and cognitive biases to make choices more objective.

- At the same time, we can observe multiple angles and multi-level models, track market changes in a timely manner, and constantly discover new statistical models to find trading opportunities.

- After deciding on the investment strategy, the effect can be verified by back inspection.

Figure 6.1 Quantitative strategy

The steps of investment in this project consist mainly 3 parts.

* 1. Understand industry trends

Use news data and forum data to find the hottest stocks, and to sort and find the most popular industries amount all the sectors.

* 1. Selecting the leading stocks in the industry

Use stock indicators such as ROE, PE, EPS, DPS, DY, RSI, Sentiment, etc. to screen out better stocks for analysis.

* 1. Specific analysis of stocks

By looking at the timing diagram of the variables and the candlestick chart, the overall trend of the stock is obtained, the profit and loss indicator is defined, and the data is labeled, respectively, buy, hold, and sell.

There are around 40 days of record of each stock in the dataset, assume that we have brought the stock AIRASIA GROUP BERHAD in the first day, we can calculate the change from day one till day 40.

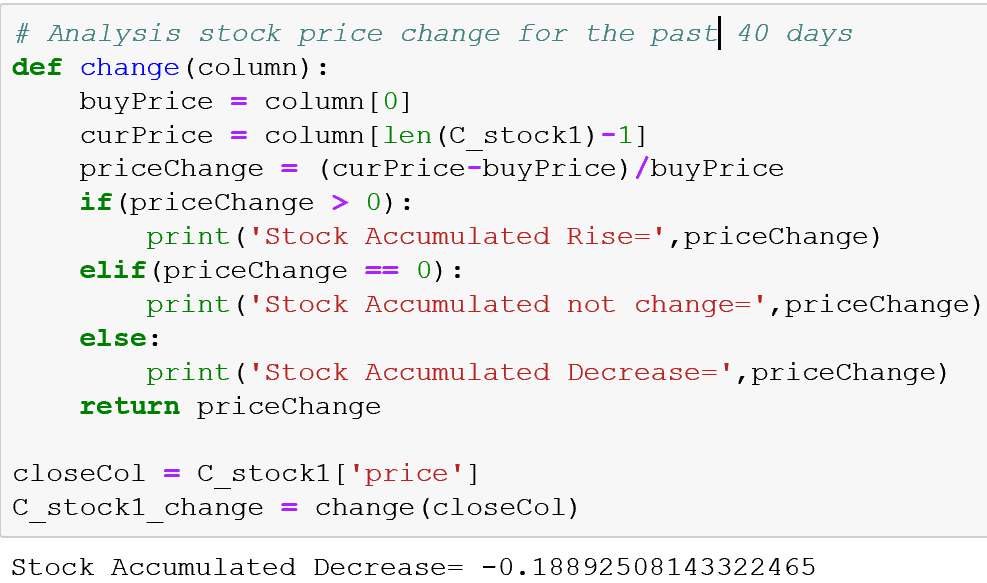


Figure 6.2 AIRASIA GROUP BERHAD stock price change for the past 40 days

As Figure 6.2 shows, the stock, AIRASIA GROUP BERHAD has decrease 19% compare to the first day and the visualization also indicate the same trend of the stock.

By analyzing such a stock in this way, it can be extended to multiple stocks for analysis and comparison. First, the user can understand the real situation of the market, select a stock of a sector to invest, and compare the same sector to select a stock that is more worthy of investment. With those indicators, a user can choose their own stock portfolio and then decide whether to buy or hold or sell of the stock.

**Conclusion**

In stock market, there are lots of information needs to consider before taking action. That means if we are about to analysis the stock market, the thing we need first is the data that is complete and useful.

However, as the group members are lack of the knowledge of stock market, the data sources in this report are simple and not big enough, and the results cannot be use in real investment.

There are still large spaces to enhance this project, the data, the model, the strategy and so on. But we’ve learn a lot from this process, we learn to crawl data from different website, we learn to use Python to manage our data, we learn to evaluate the data, we learn to use Tableau and SAS to perform visualization and modeling of data, and so on, and that’s the value of this entire project.