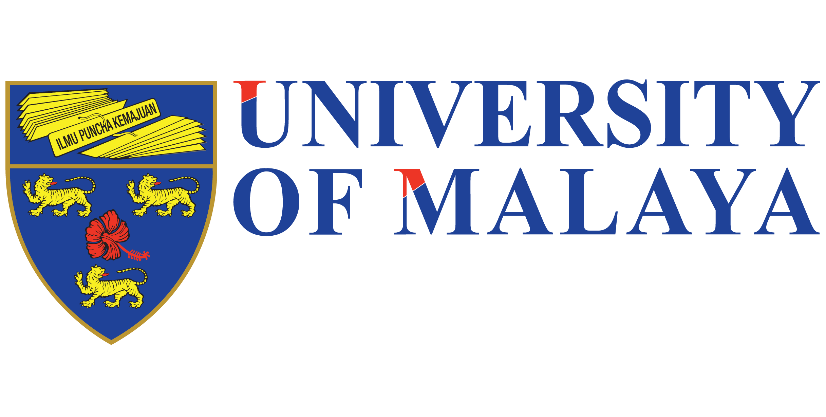
****

**MASTER OF DATA SCIENCE**

**2018/2019**

**WQD7005**

**DATA MINING**

**“MILESTONE 1 TO 6”**

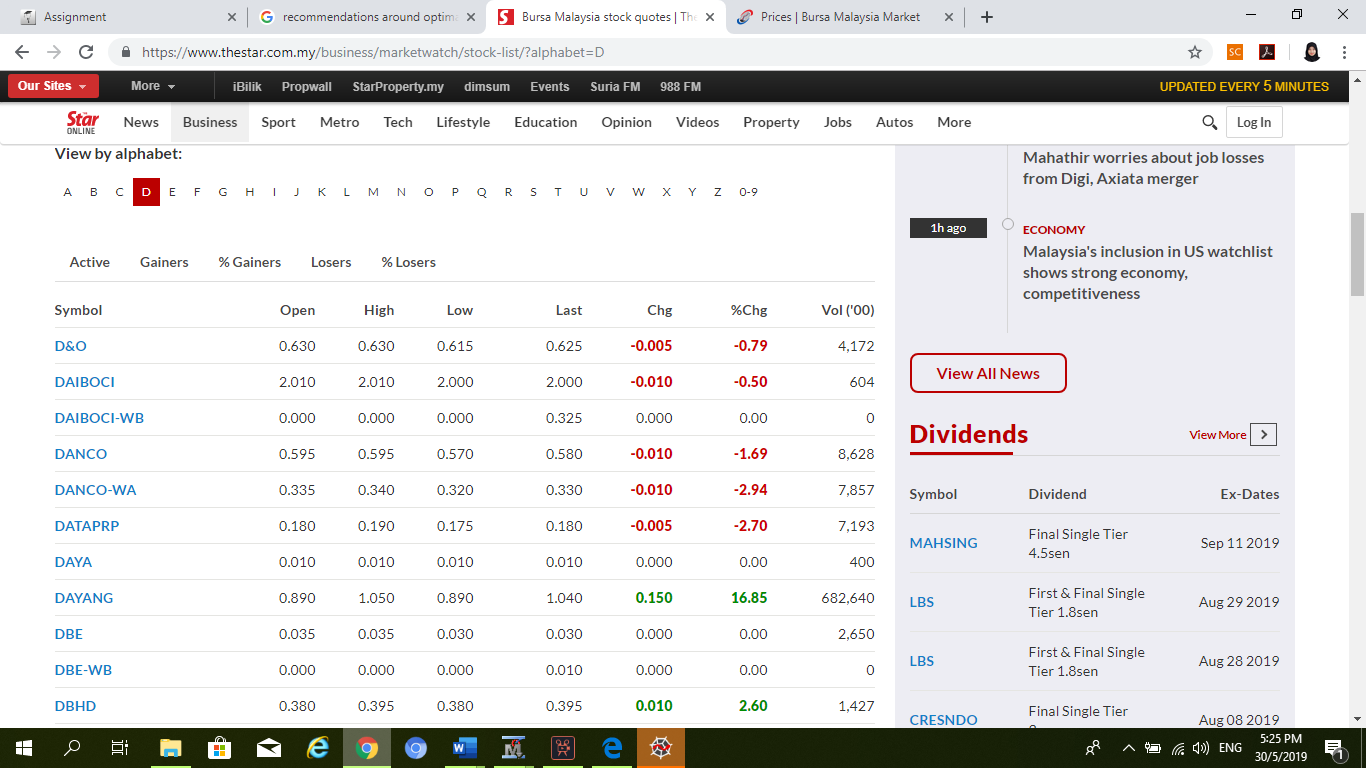
**NAME: WAN NOR ANISAH BINTI WAN SHAH RAHMAN**

**MATRIC NUMBER: WQD170092**

**MILESTONE 1**

**Acquisition of data (Group)**

In this milestone, we are required to crawl a data. The first thing to do is to identify which website need to crawl. For this project, we crawl data from “The Star.com” to get the real time stock market data. Figure 1 is the screenshot of The Star website.



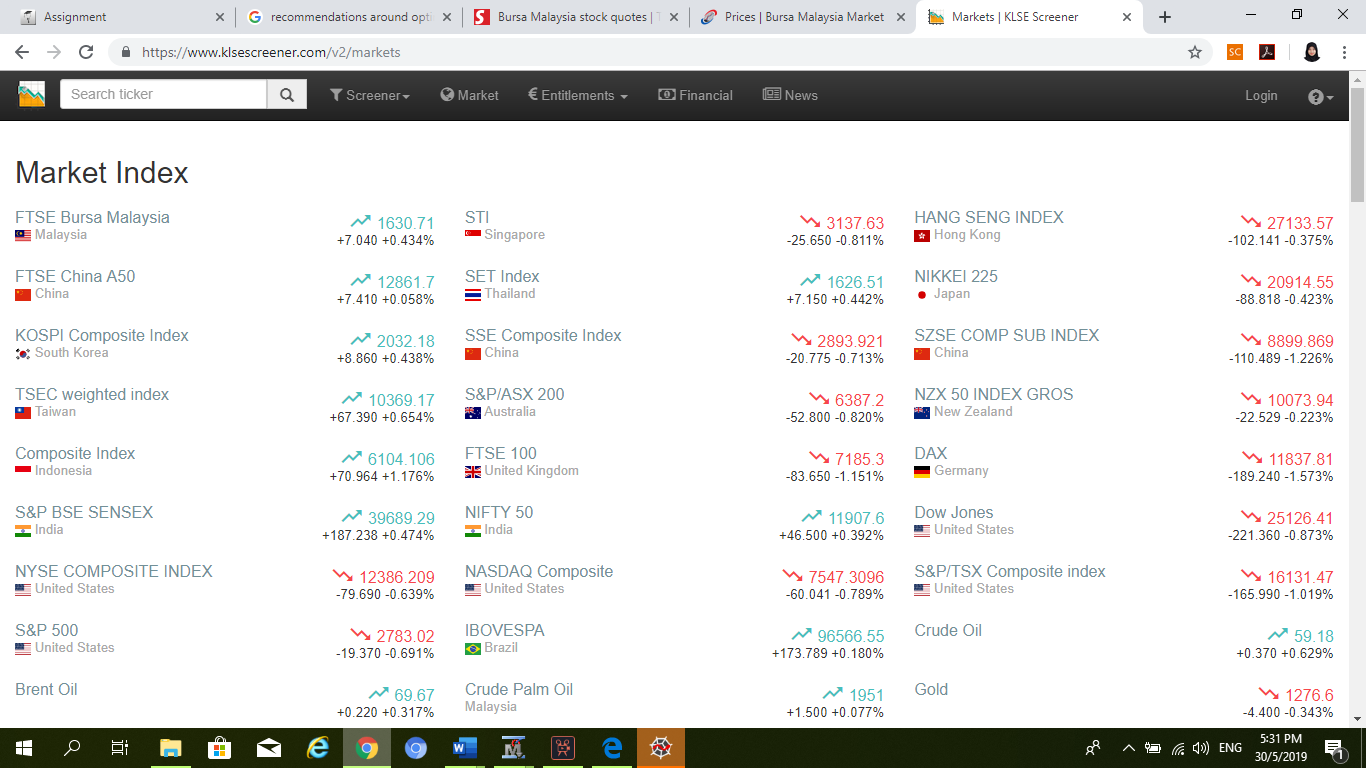
**Figure 1: Screenshot of The Star Website**

Since this website contains so many data from many companies, we try to limit the selected company based on the name on Bursa Malaysia website. Names of all company are download and saved in csv format. Then, based on these names, we crawl real time data daily for 2 weeks. Below is the screenshot for Bursa Malaysia website.



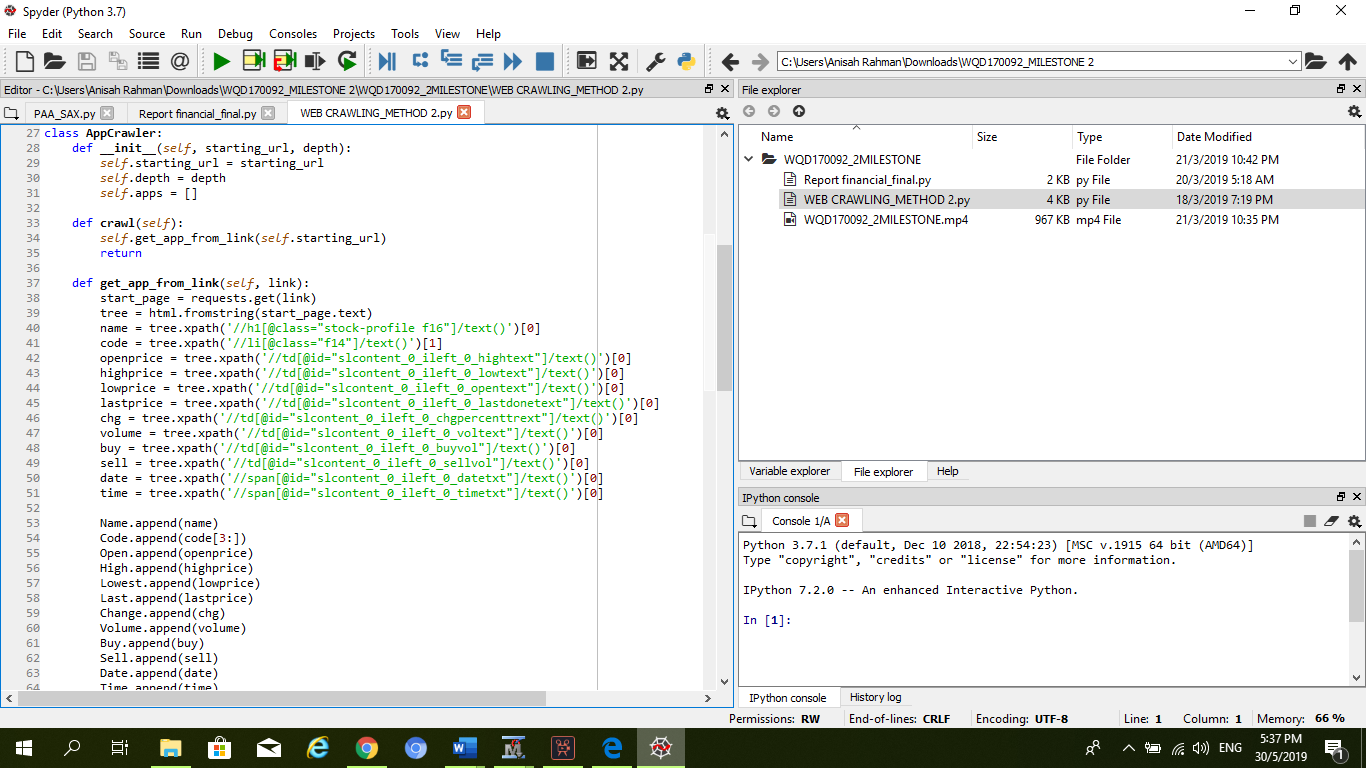
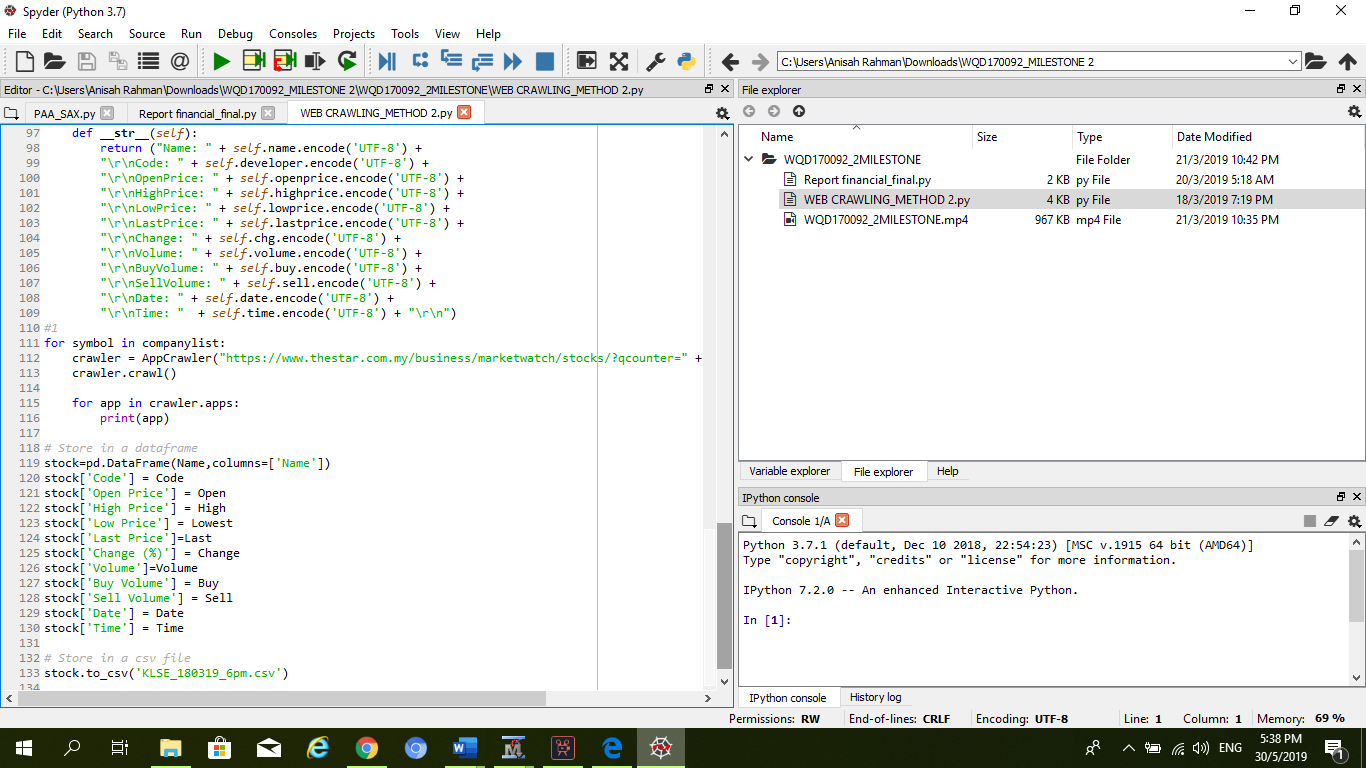
**Figure 2: Screenshot of Bursa Malaysia Website**

The third website we used to obtain our data is from KLSEscreener.com. This is the website that we used to get stock market data for Annual and Quarter. Annual is a dataset contains stock market data based on year while Quarter is based on month.

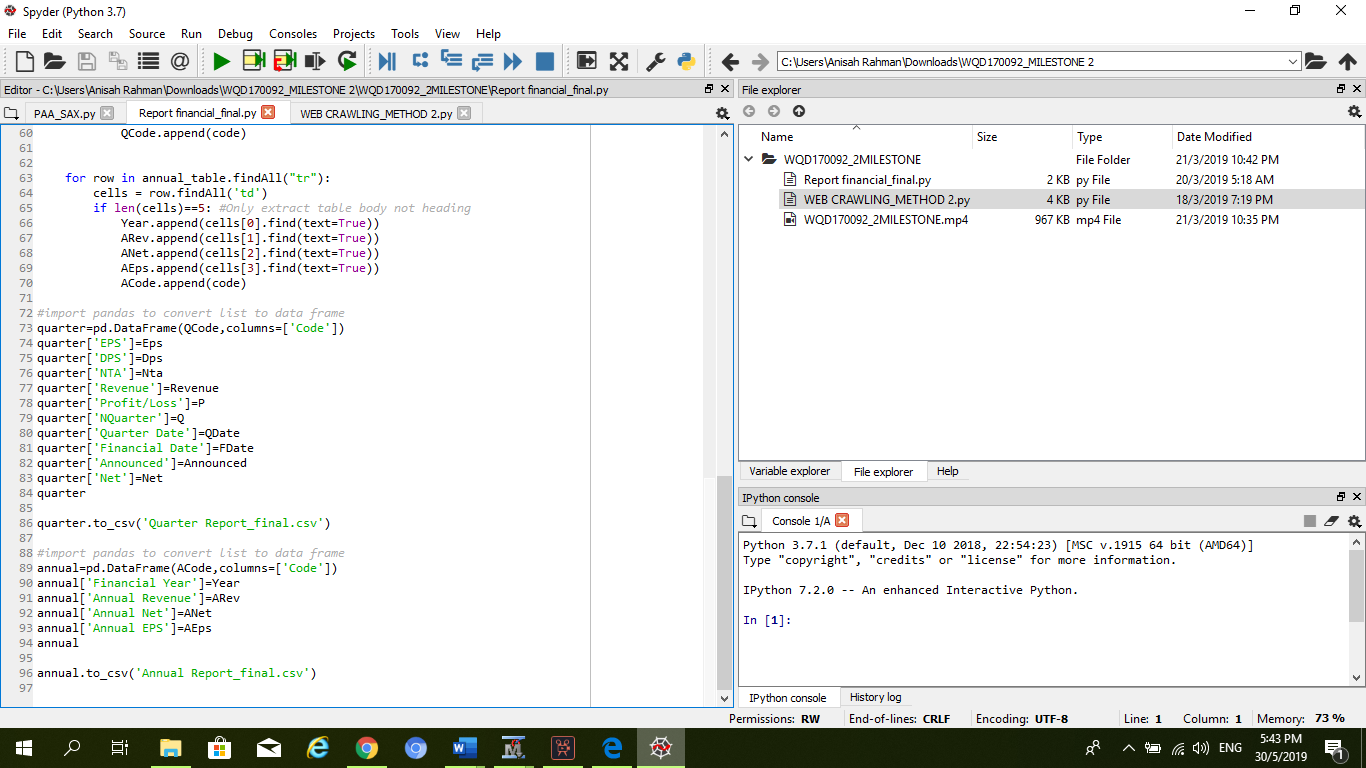
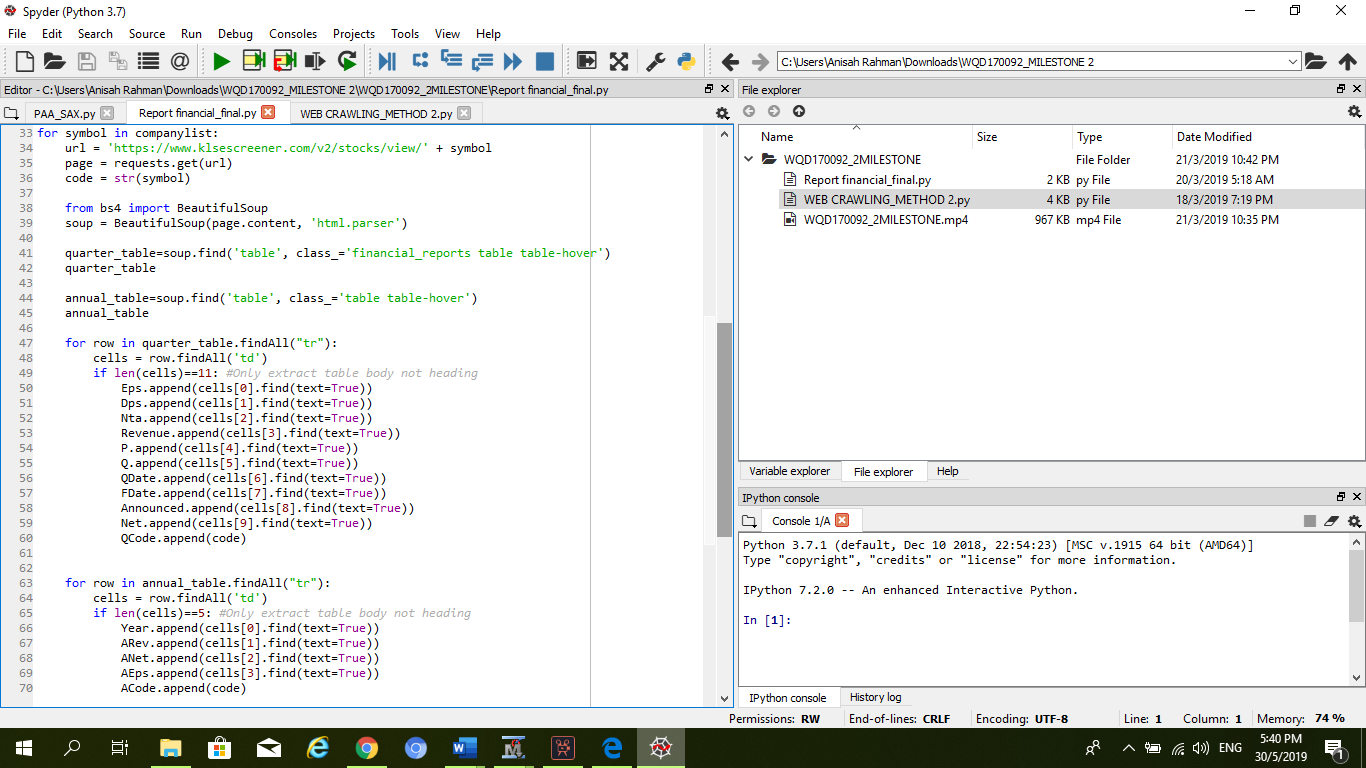


**Figure 3: Screenshot of KLSEscreener.**

In order to crawl all the data, we use python code run with spyder. Three different code need to be created since we are going to crawl three different datasets. For annual and quarter dataset, we just need to crawl them once while for daily dataset, we crawl it every day for two weeks. Below are the screenshots for all the codes.



**Figure 4: Part of python code for daily dataset.**

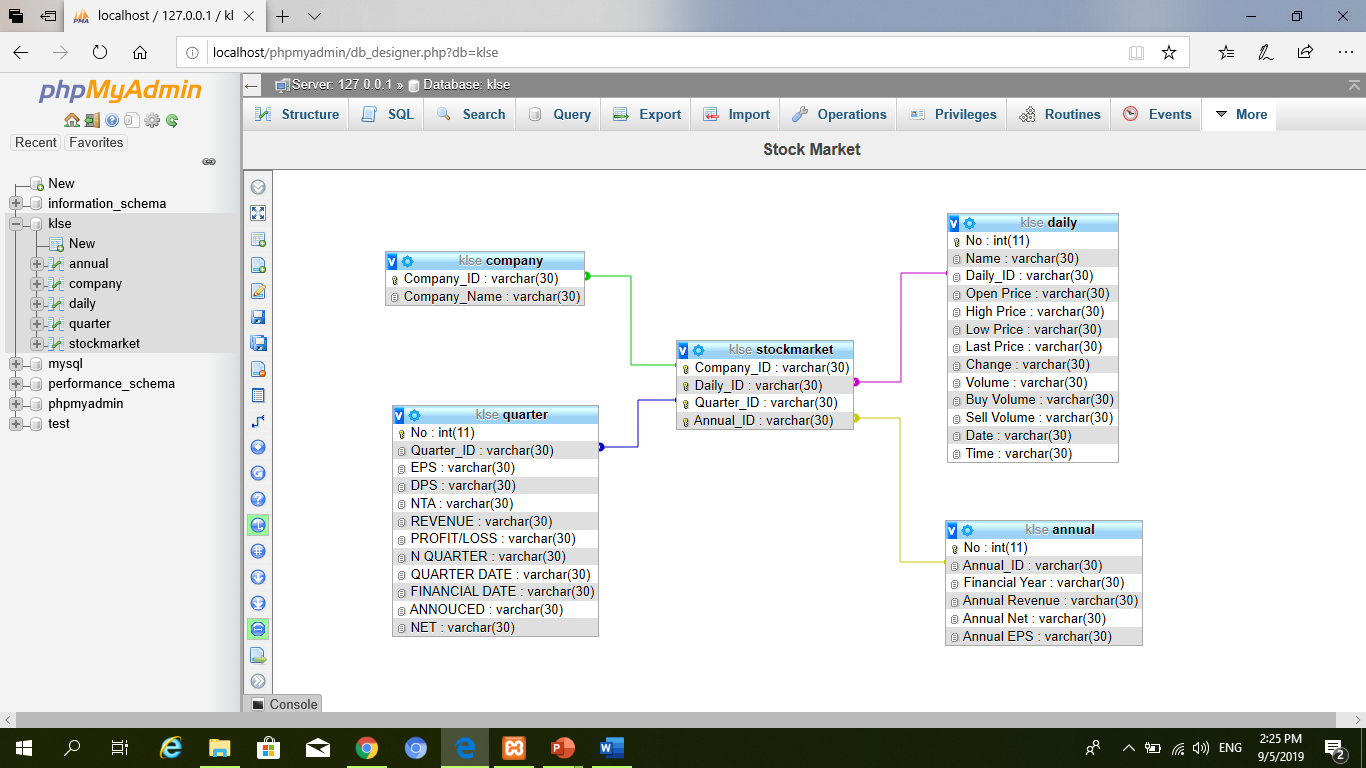


**Figure 5: Python code for Annual and Quarter Dataset**

**MILESTONE 2.**

**Management of data (Group).**

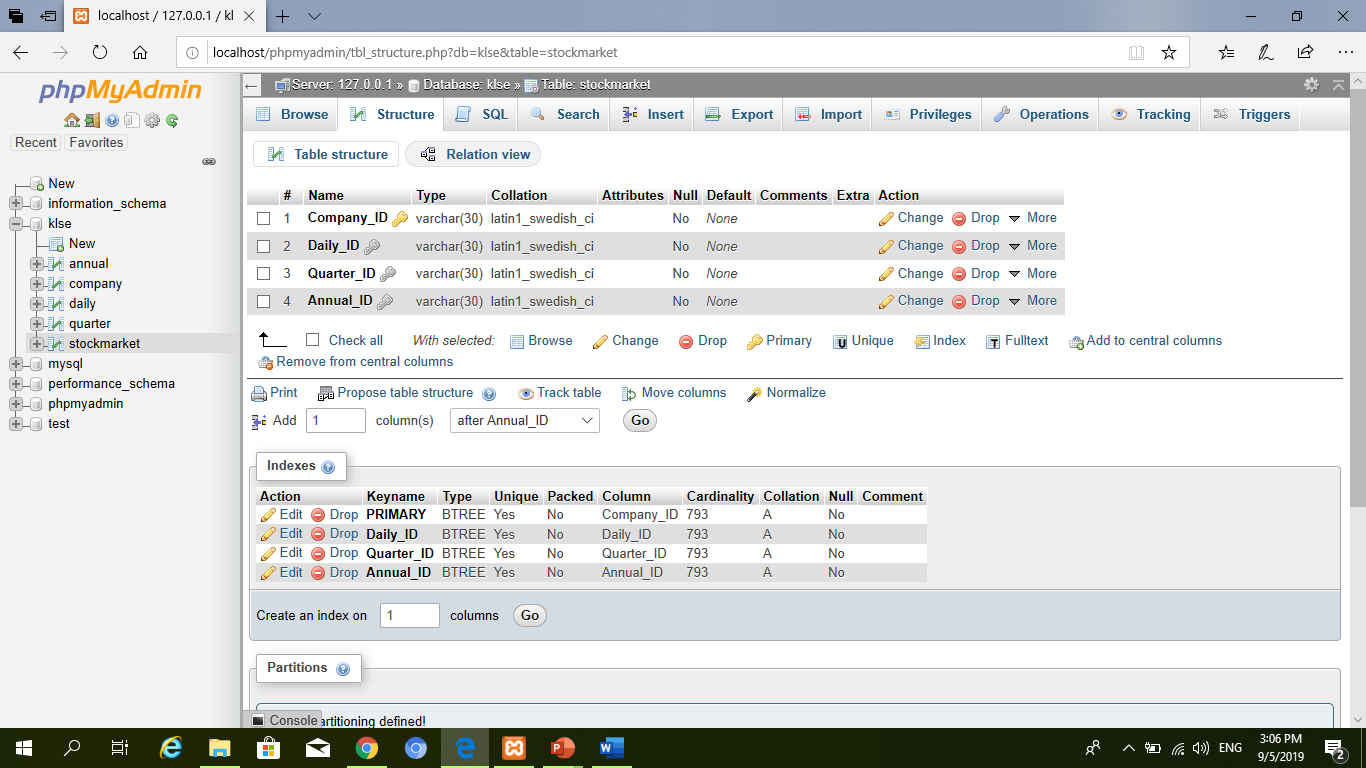
In this milestone, we are going to do star schema in phpMyAdmin and import data into hive.



**Figure 1: Star schema screenshot for stock market data.**

Figure 1 is the star schema for stock market datasets which is being saved in phpMyAdmin. When all the datasets are uploaded, the design of star schema is created in designer page. For star schema, there are two types of table which are fact table and the dimensional tables. Fact table is the table located in the middle that connected with the tables around it, called dimension tables. In this case, the fact table’s name is *stockmarket* and the dimensions table’s name are *company, daily, quarter, and annual.*

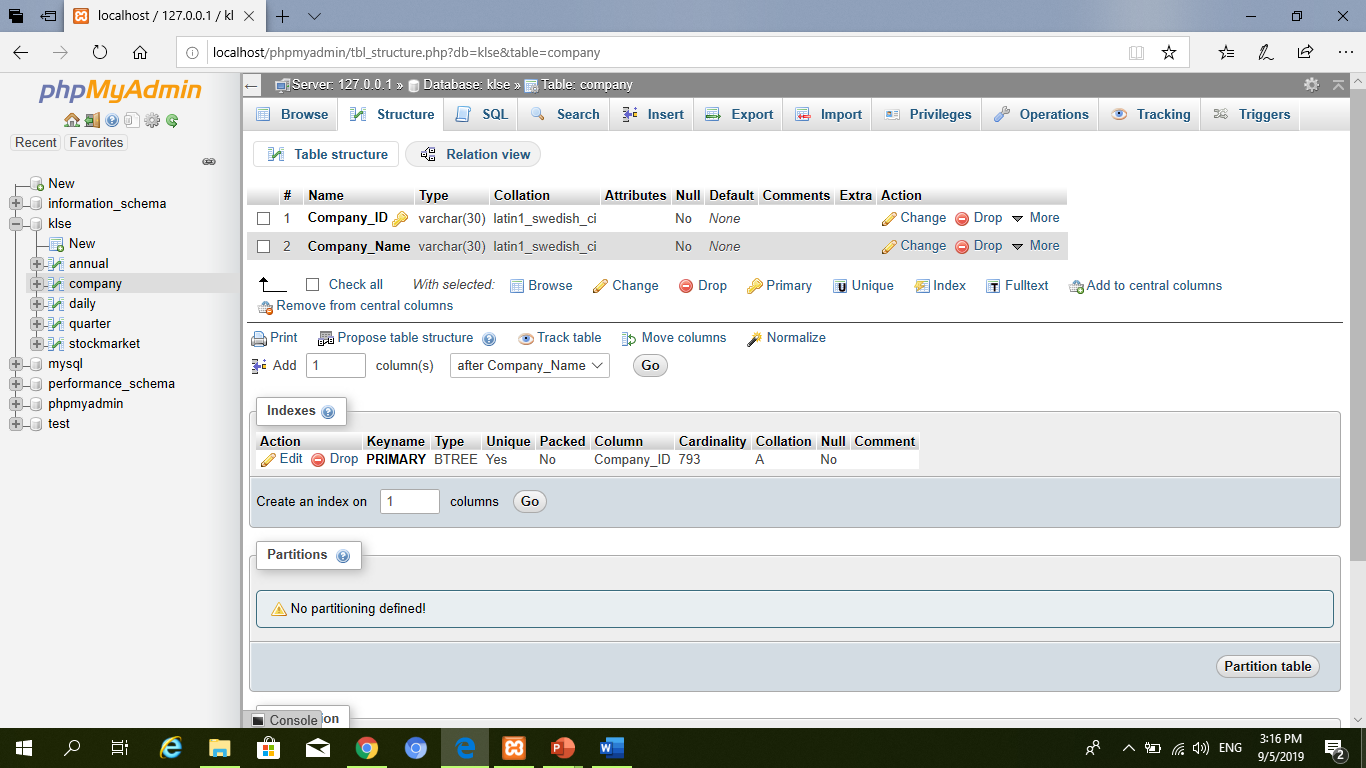
For fact table, there are four attributes which are *company\_ID, Daily\_ID, Quarter\_ID and Annual\_ID.* In order to connect those attributes to another attributes in dimension table, they need to be assigned to a key. In this case, all the attributes contain special character. Which means, all the character in those attributes are different from one another. Thus, primary key can be used. However, one table only can have one primary key. For that reason, *Company\_ID* hold primary key for this table. Other tables who have a special character too are assigned as a unique in order to allow them connected to other tables. Unique is just like a primary key which only can contains special character but, they can be many unique in one table. Figure 2 below is the screenshot of the stock market’s table structure.



**Figure 2: Table structure of stockmarket.**

Next, for dimension tables, four dimension tables are created. Same with the fact table, only one primary key can be assigned to one table. However, primary key only can contains the special or unique character which not have any duplications. Thus, primary key are used. Below are the details for every dimension tables.

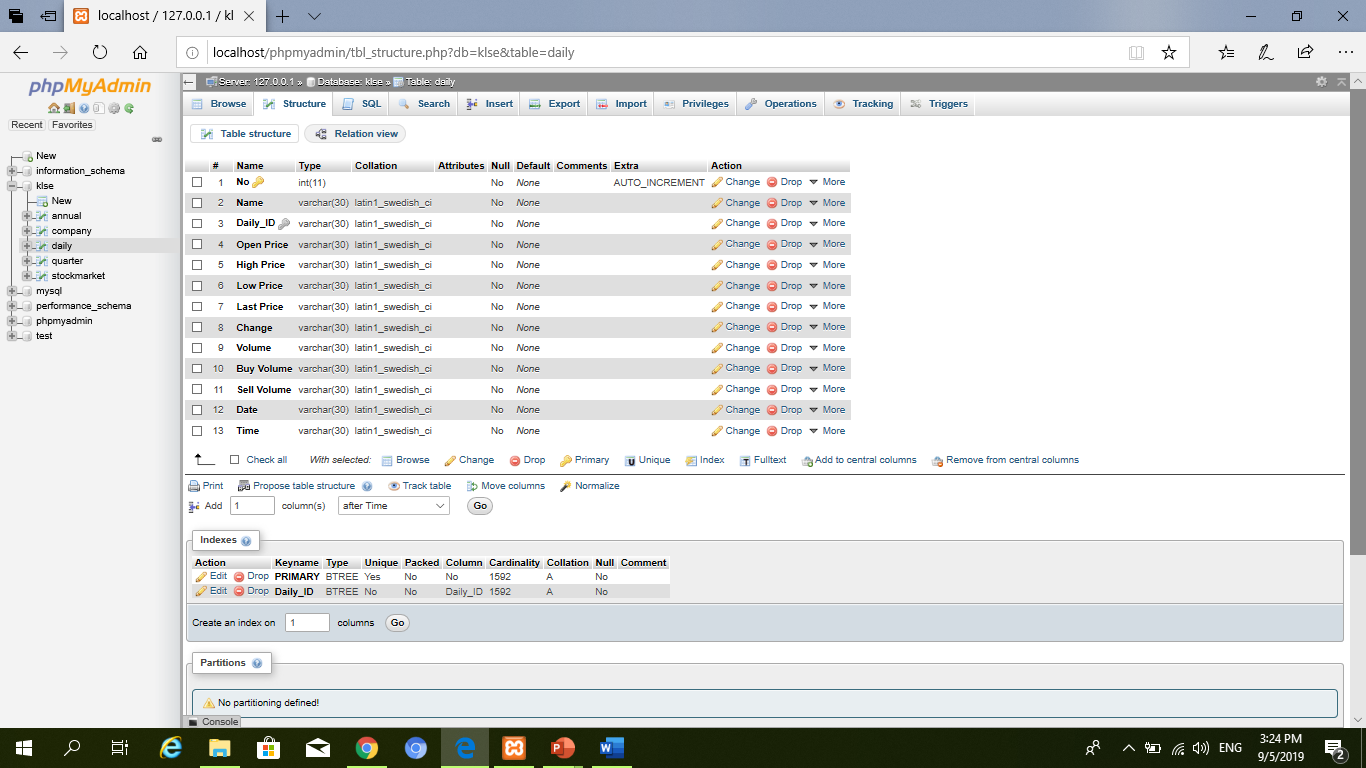
1. **Company**



**Figure 3: Table structure of Company.**

For company table, Company\_ID can be assigned as a primary key since it is contains only unique character. In this attribute, all of company ID is stated. Because of this attribute is the selected attribute to be connected with fact table, foreign key is not appropriate to be assigned

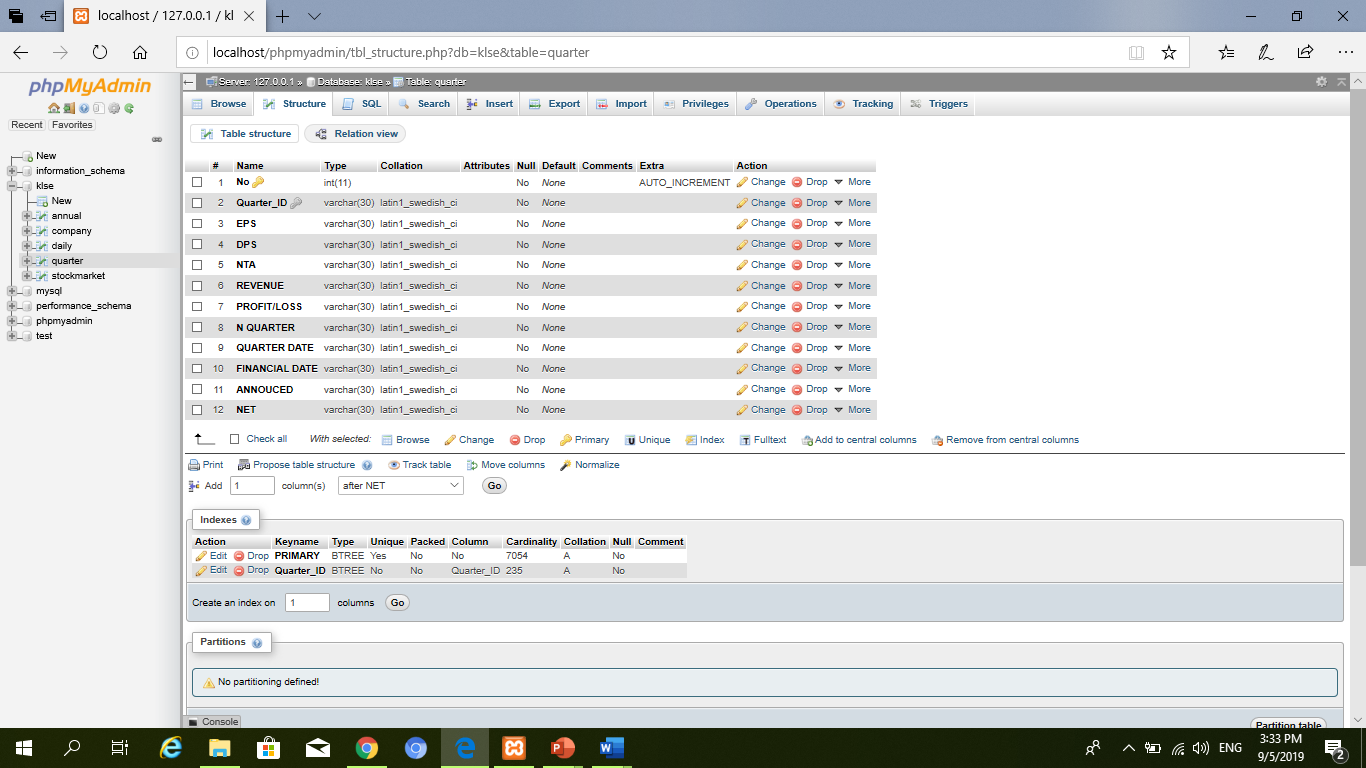
1. **Daily.**



**Figure 4: Table structure for daily.**

Daily table is a bit different from company table. For this table, two types of key are assigned. Primary key is assigned to the *No* which is auto increment by the database itself. Everytime the new data import, the continuous number will be created. Thus, there are repeated number for this attribute. However, in order to connect with the fact table, foreign key is important. Daily\_ID is a foreign key for this table. It contains all the daily ID of the company. But the ID is repeated for the same company.

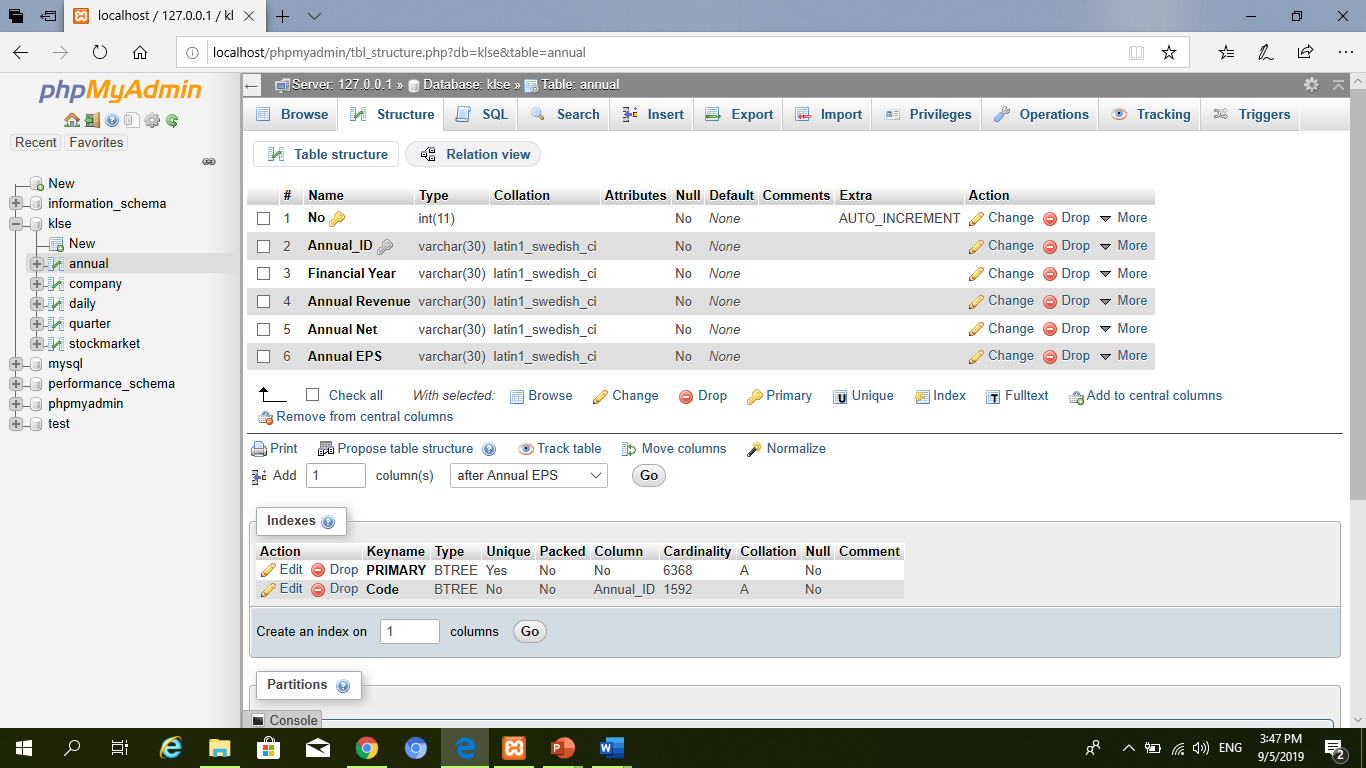
1. **Quarter.**



**Figure 4: Table structure for Quarter.**

Quarter table is same as daily table. It contains primary key and foreign key. Primary key is assigned to the no which is auto increment to and the foreign key is assign to Quarter\_ID attribute which contains the repeated ID.

1. **Annual.**



**Figure 5: Table structure for Annual.**

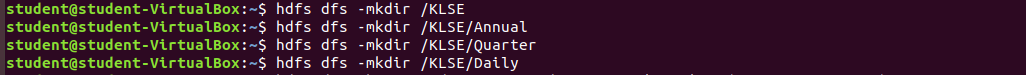
Same as quarter table, annual table contains primary key and foreign key. Primary key is assigned to the no which is auto increment to and the foreign key is assign to Annual\_ID attribute which contains the repeated ID.

**Roll Up (Drill-up) and Roll Down (Drill-down).**

Roll up is the process of summarize data. It works by climbing up hierarchy or by dimension reduction. We can filter our data to show only the selected data. Thus, we will have a less amount of data to be shown.

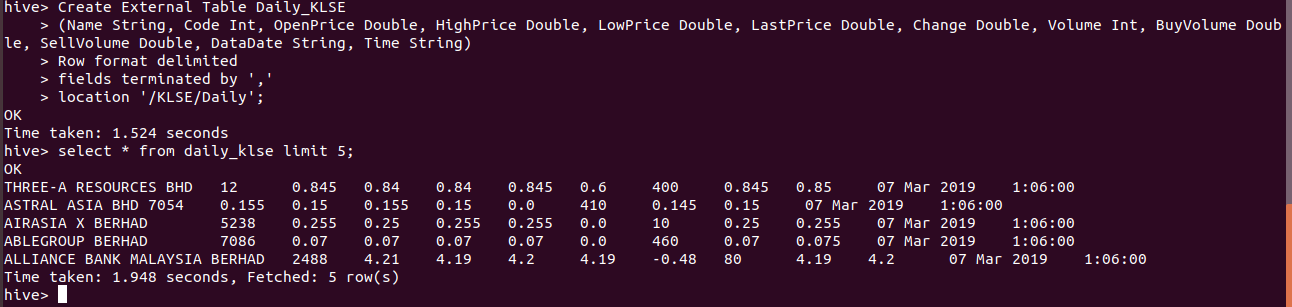
Drill down is the reverse process of Roll up. While roll up is from more detailed data to less detailed data, drill down is the process from less detailed data to more detailed data. By means, we can have more information that we need from the certain data we have by doing drill down process.

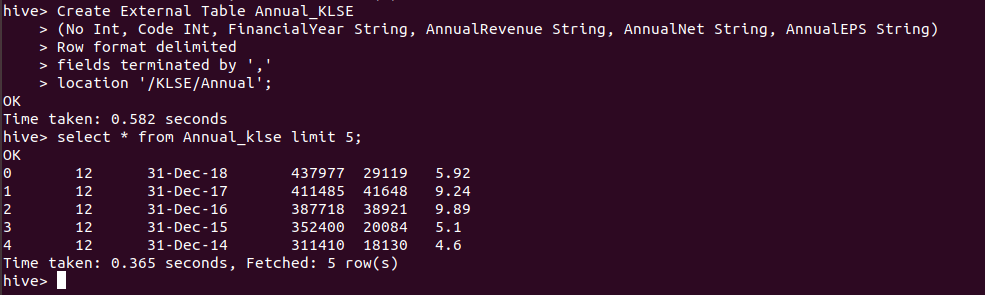
For hive, since we have three types of data, we create one main directory consist of another three sub directories names Annual, Quarter and Daily. When the directory is already created, the next step is to import all those three datasets from local host to the hdfs before they could be imported into hive table. Figure 6 and 7 is the screenshot of creating directories and importing datasets.

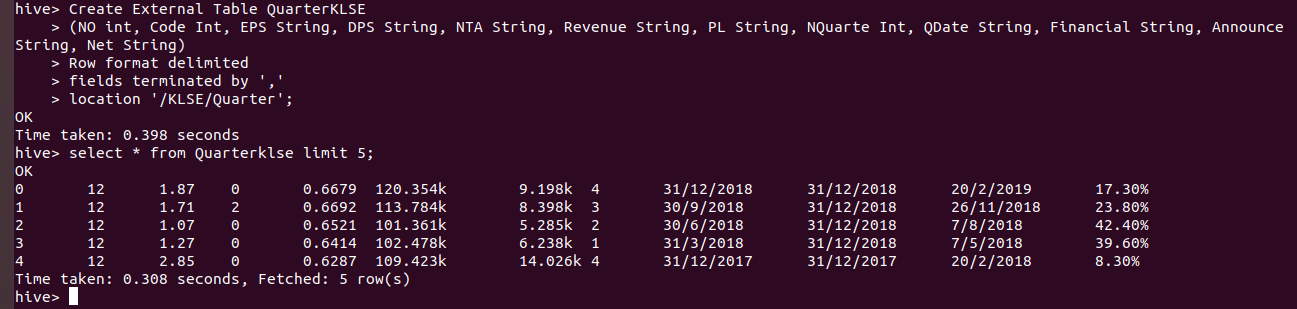
**Figure 6: Create Directories.**



**Figure 7: Import Datasets.**

Then, we need to create table in hive before import datasets. Hive command is similar with mysql, thus, we can show our imported dataset directly in hive. Below are the screenshots of all the tables and their results.

**Figure 8: Create and show Table for Daily Dataset**

**Figure 9: Create and show Table for Annual Dataset**

**Figure 10: Create and show Table for Quarter Dataset**

**Milestone 3.**

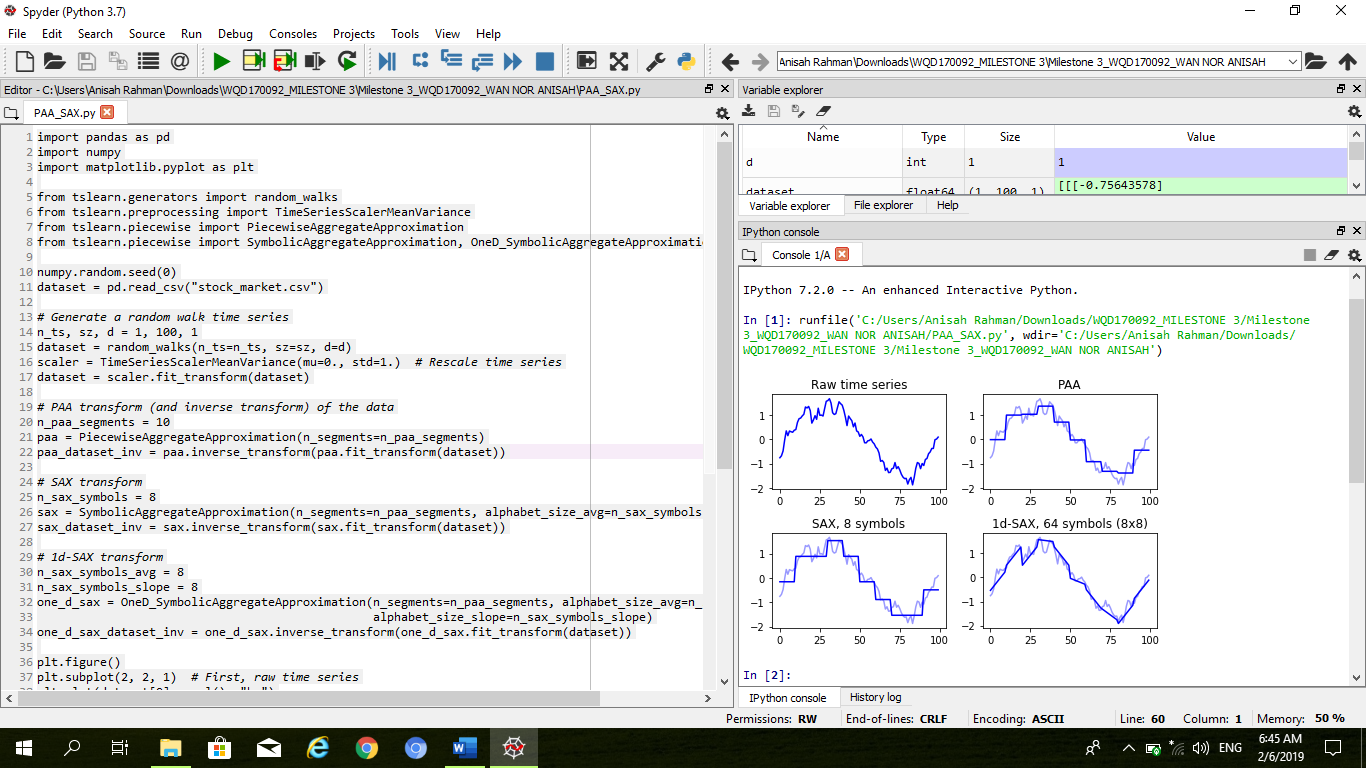
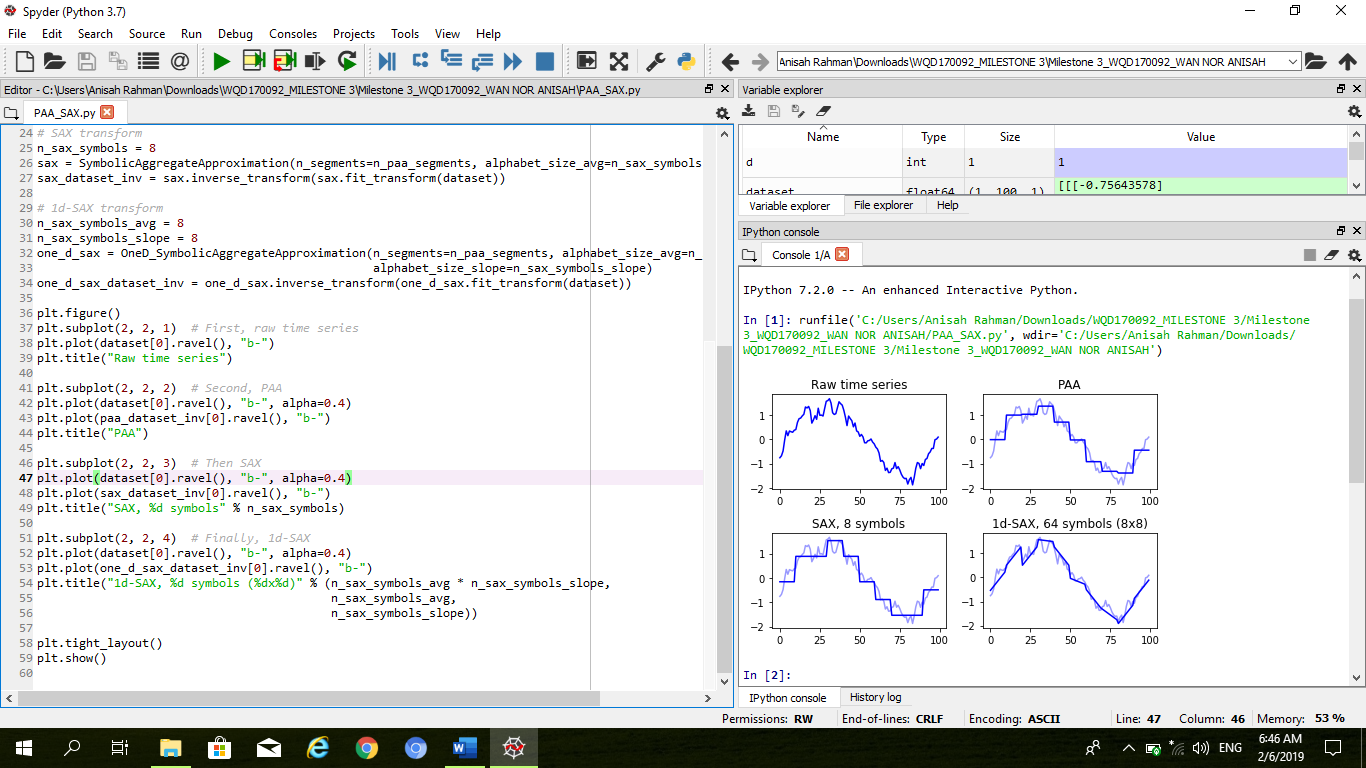
**Processing of data (Group).**

For this milestone, we are requested to do PAA SAX. PAA stands for Piecewise Aggregate Approximation and SAX stands for Symbolic Aggregate Approximation.

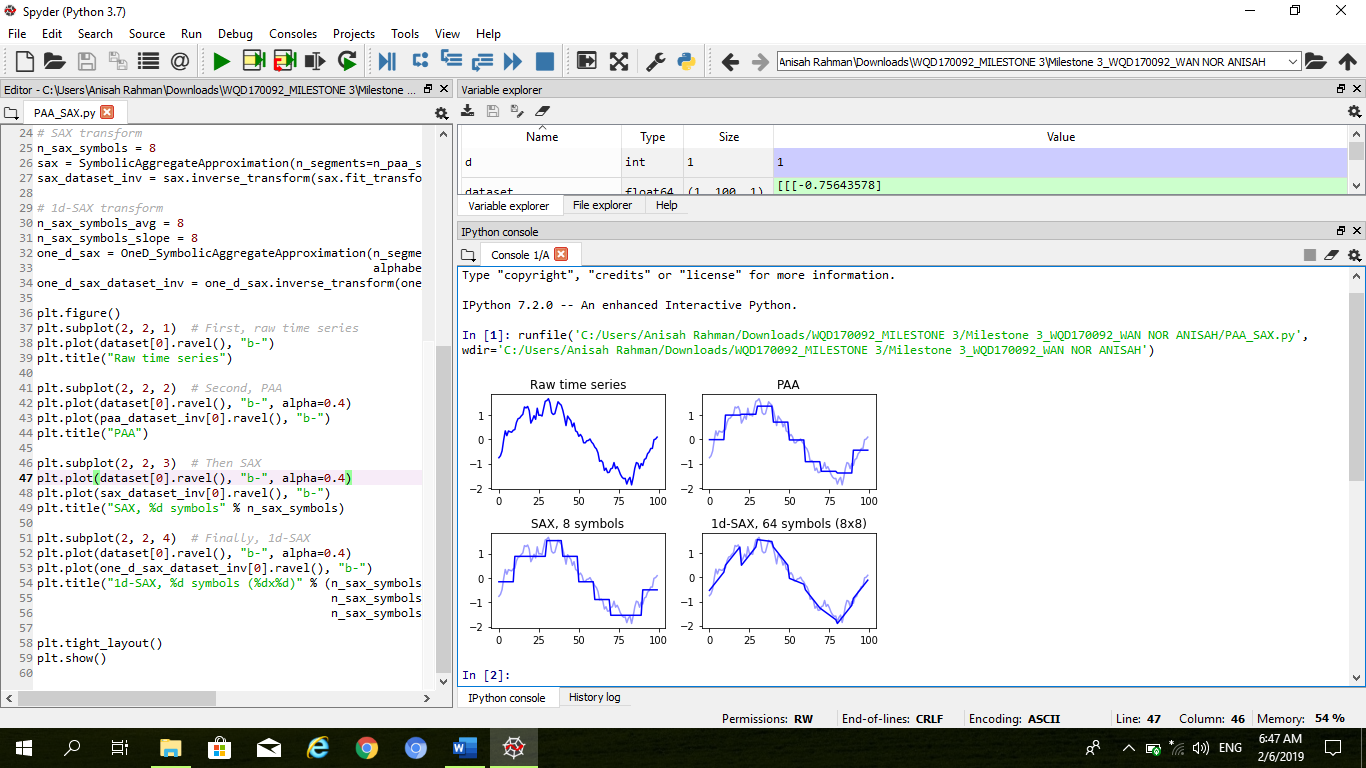
Piecewise Aggregate Approximation (PAA) is a very simple dimensionality reduction method for time series mining. It minimizes dimensionality by the mean values of equal sized frames, which misses some important information and sometimes causes inaccurate results in time series mining. While SAX is a method of discretizing time series, to better understand patterns and motifs. It involves binning the time series and then translating these discrete bins into words, which are discrete objects made of discrete letters.

We used our daily dataset as an input dataset. This dataset first needs to be processed through PAA in order to reduce the dimension of it. Then, the next step is to do SAX where the data are converted into symbol. This can help in well understanding the pattern of our data. Thus, by using this pattern, we can predict the future pattern.

Figure 1 is the code for PAA and SAX using python. Figure 2 is a comparison between PAA, SAX and 1d-SAX features.



**Figure 1: PAA and SAX code using python**



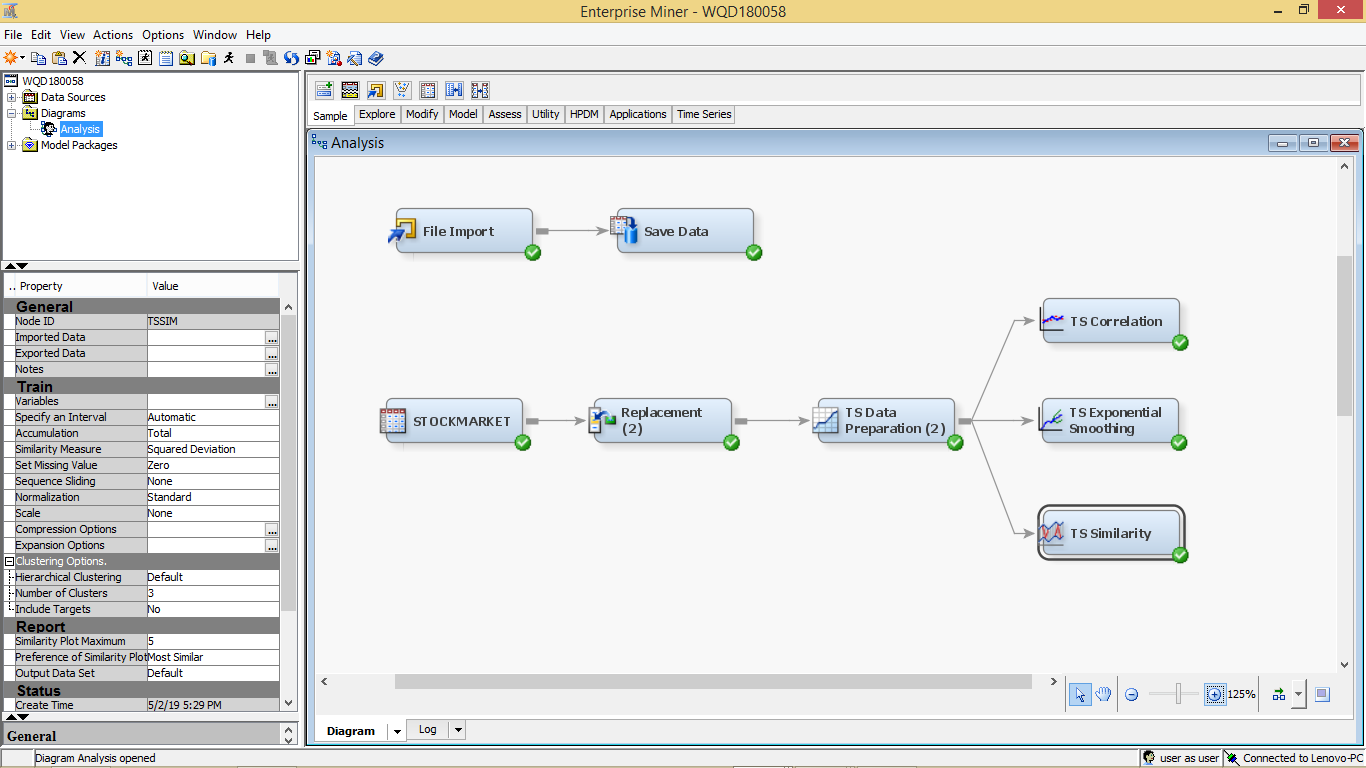
**Figure 2: comparison between PAA, SAX and 1d-SAX features.**

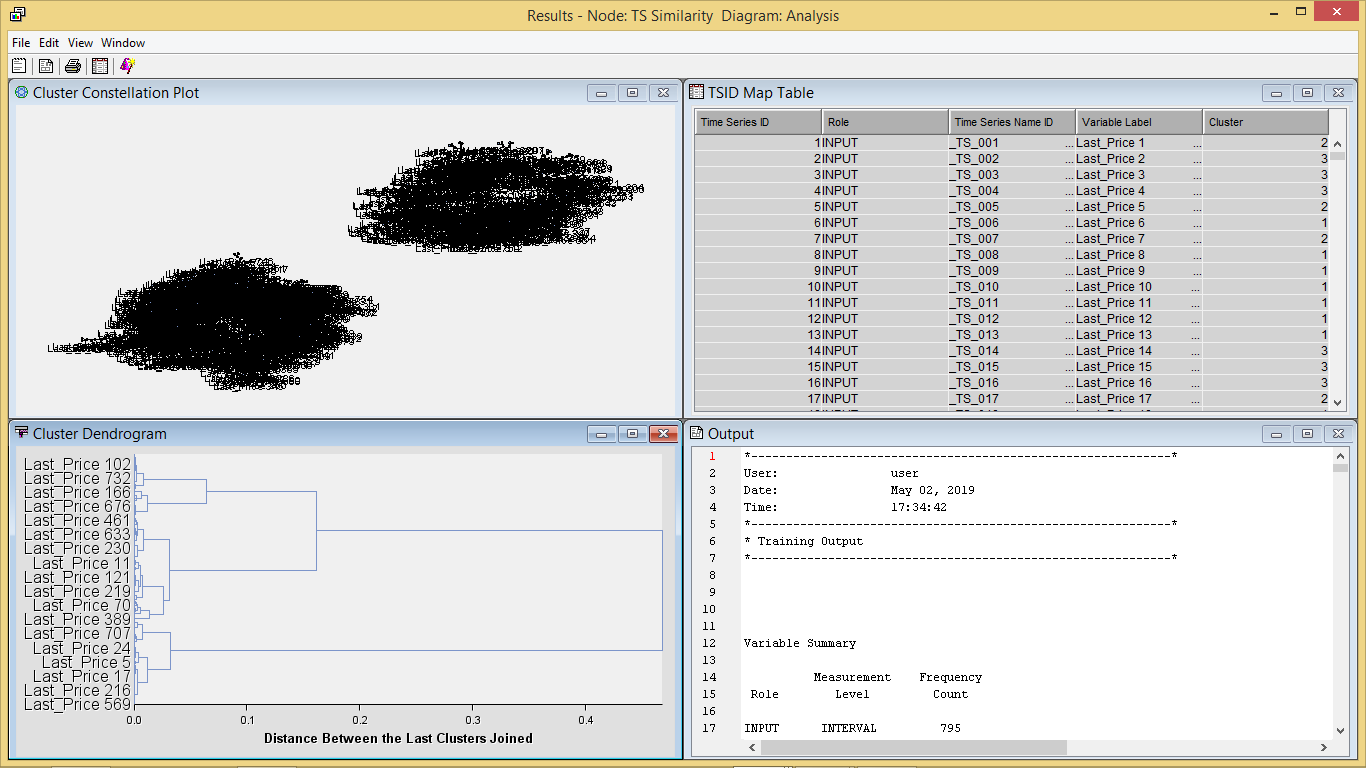
**Milestone 4.**

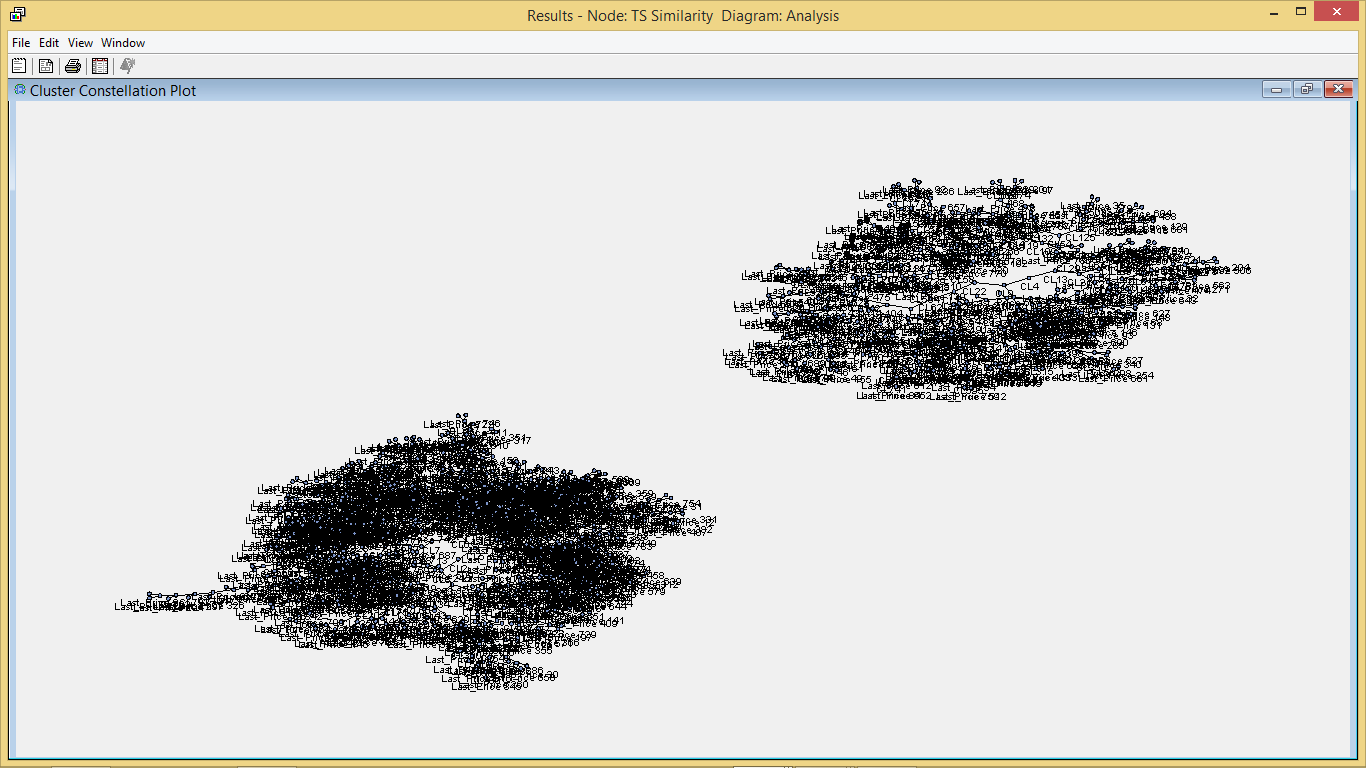
**Interpretation of data (Individual).**

For interpretation of data, this data are clustered based on their last price similarity. From this, we can see how many clusters the data can have to ease us identify which company is the good choice to make investment.

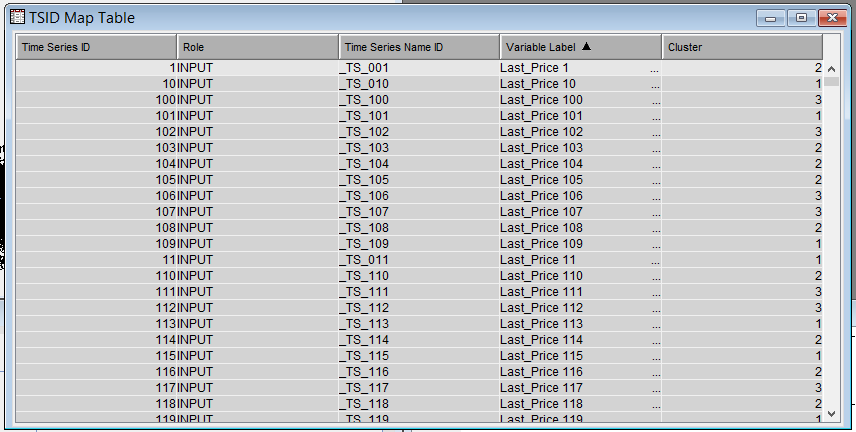
Figure 1 is how the nodes arranged in order to get cluster for the last price data. Figure 2 is the results from TS similarity nodes. It shows the cluster plot, dendogram, map table and the output. In figure 3, shows only cluster plot.

****

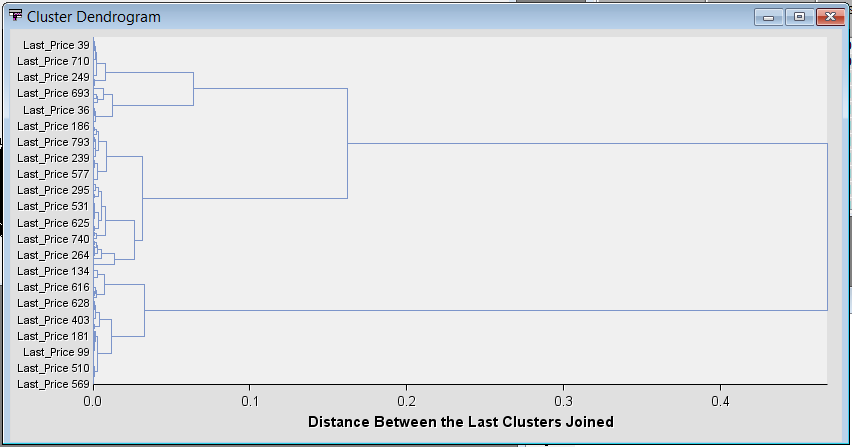
**Figure 1: SAS Nodes.  
Figure 2: TS Similarity Diagram Analysis.**

**Figure 3: Cluster Result.**

Based on the Figure 3 above, there are 2 distinct clusters that we could see. However, it is actually divided into 3 clusters as per table below.



**Figure 4: TSID Map Table.**

The cluster dendrogram below explain how the clusters are divided into three.

**3**

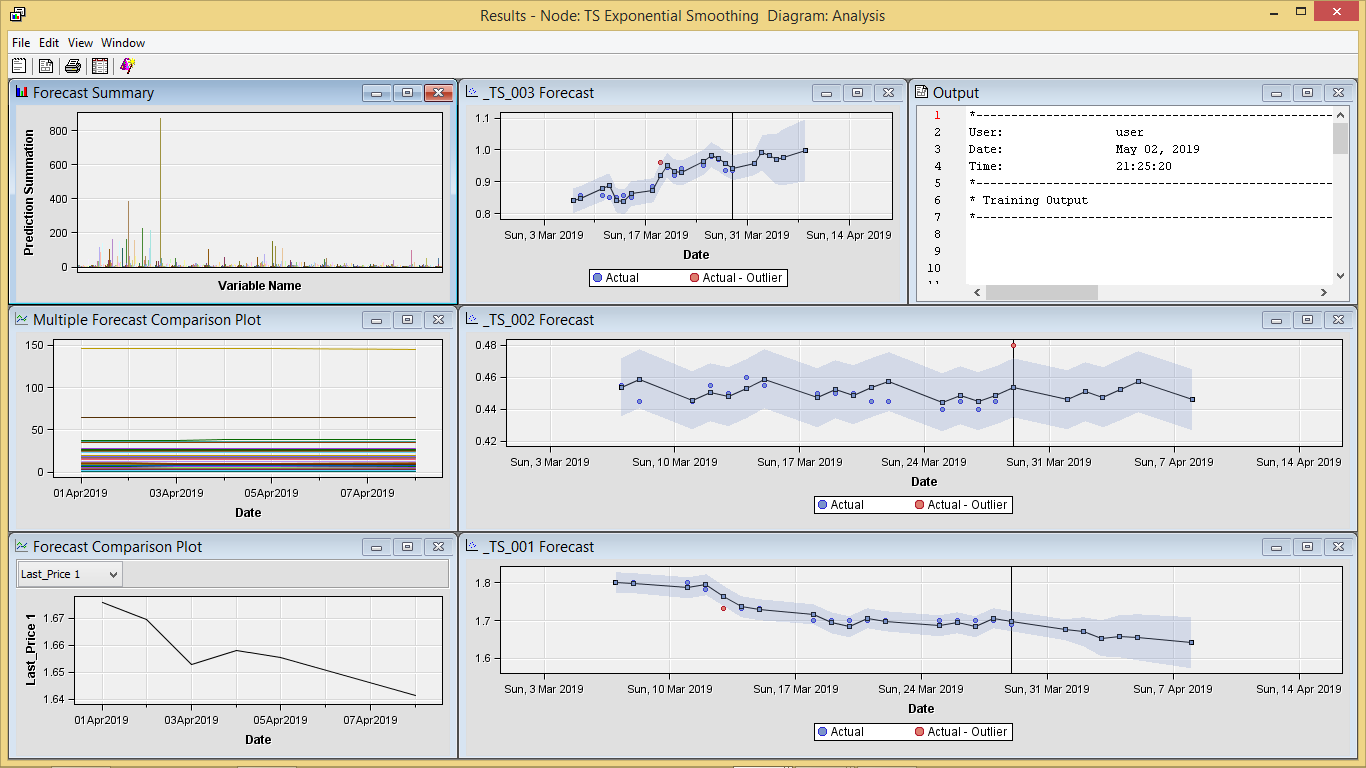
**2**

**1**

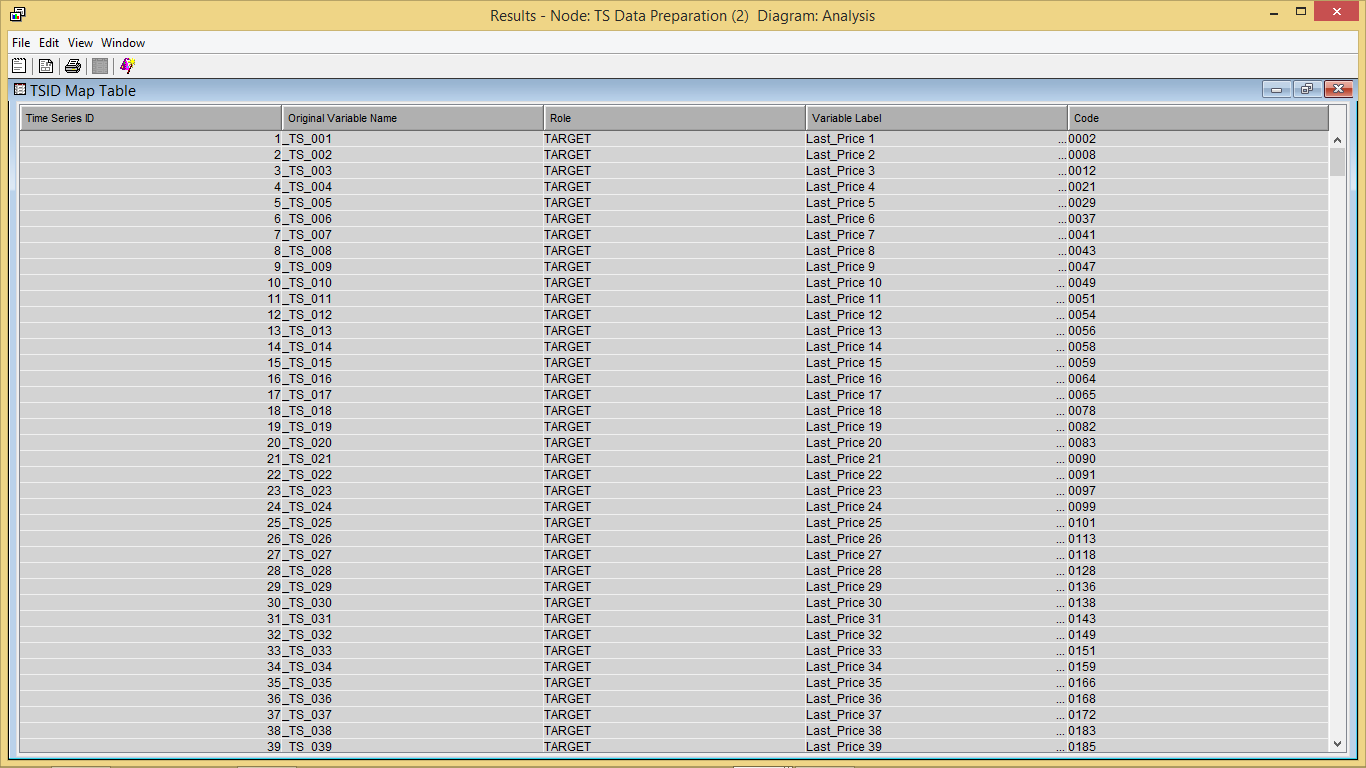
**Figure 5: Cluster Dendrogram.**

**Milestone 5.**

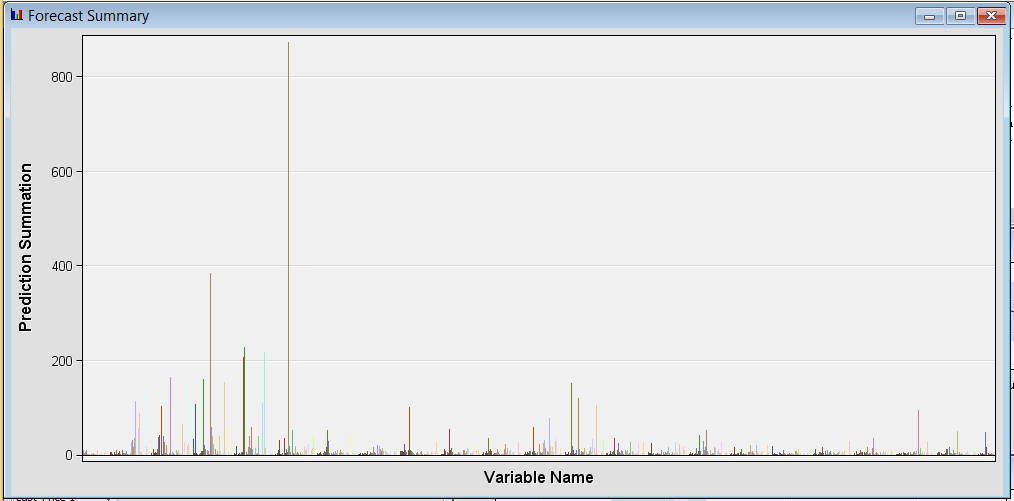
**Communication of insights of data (Individual)**

 For this milestone, prediction is made. Thus, we can choose which company can be a good selection in order to make investment. Below is the result from TS Exponential Smoothing. The nodes are arranged just like what have been showed in Milestone 4 figure 1.

**Figure 1: TS Exponential Smoothing Diagram Analysis.**

Figure 2 below shows the description for each company (CODE) which is assign to a new TS ID.

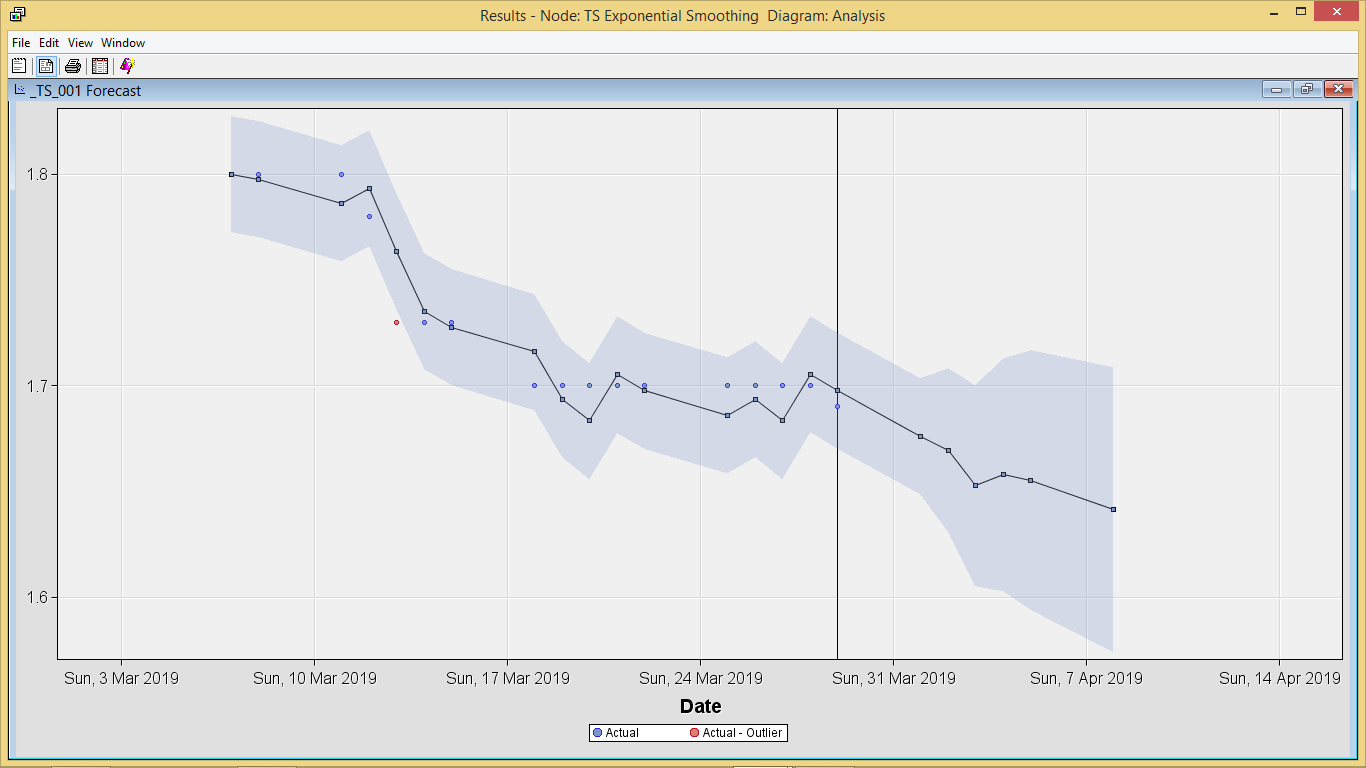
**Figure 2: TSID Map Table**

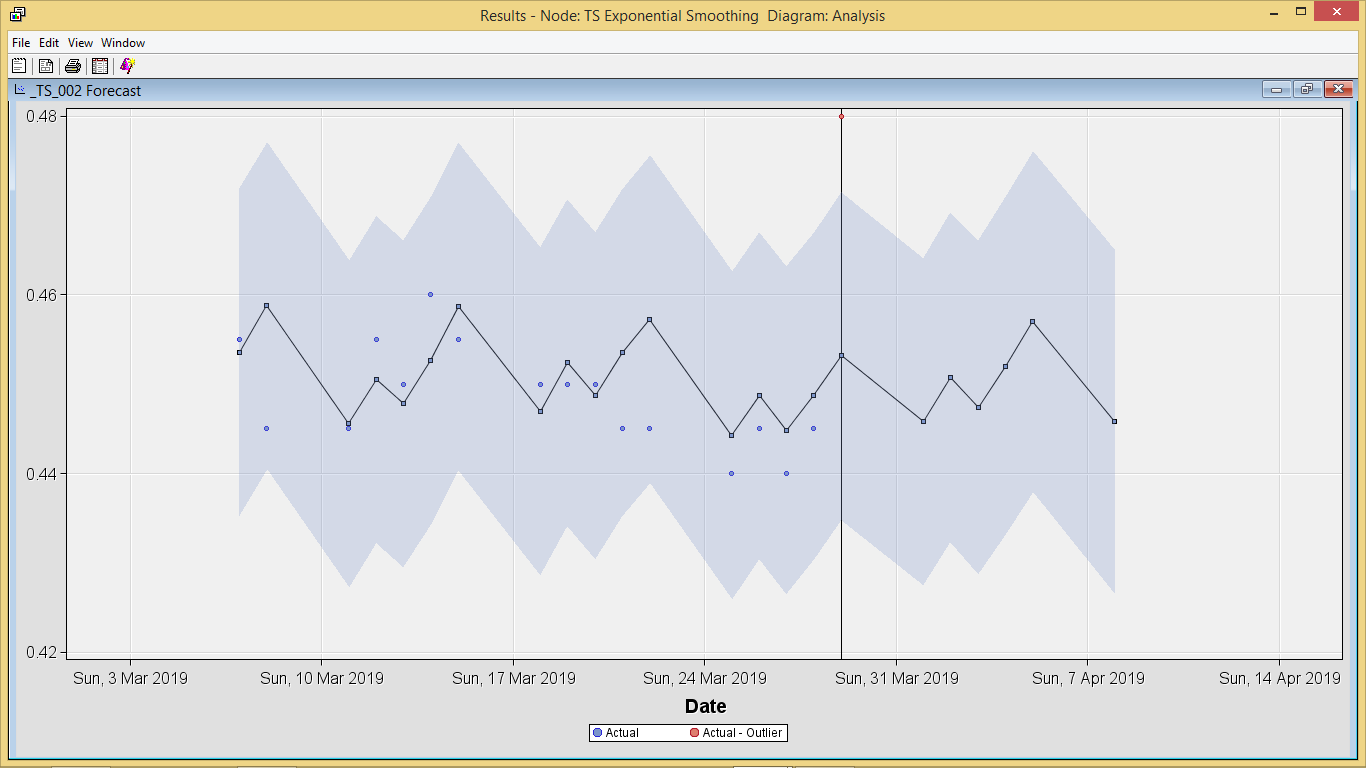


**Figure 3: Forecast Summary.**

**Milestone 6.**

**Recommendation.**

This milestone is a continuous result from previous milestone. We can see, the result of the prediction can be divided into three types. Figure 1 shows that the company is forecasted to decline. Thus, it is not a good choice to invest in such of this company. While figure 2 shows that the company only in stable situation. It is not going up nor down. This can be a choice for investor. However, the investor needs to reconsider if they intend to invest in this type of company for a long time period. Figure 3 is a very good type of investment because the company are predicted to go up.

**Figure 1: Type 1 company.**

**Figure 2: Type 2 Company.**

**Figure 3: Type 3 Company.**