MASTER OF DATA SCIENCE

WQD7005 DATA MINING

LECTURER: PROFESOR DR. TEH YING WAH

MYEG STOCK MARKET ANALYSIS

NAME: ZULKANAIN BIN HASAN

MATRIC NO.: WQD180031

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

UNIVERSITY OF MALAYA

Semester I, 2019/2020

Assignment

This assignment should not only have learned in theory, but also have experience predictive analyzing and prescriptive analyzing data with the main objective of solving real-world problems. There are three key skills needed to succeed in data science, which we refer to as creating, connecting, and computing.

It consists of **6 parts with 6 milestones**:

1) Acquisition of data in your familiar domain
    a) Web crawling the real time data by using Python

2) Management of data
    a) Store data into hive data warehouse

3) Processing of data

4) Interpretation of data

5) Communication of insights of data

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

Table of Contents

1. Introduction

Savinderjit Kaur and Veenu Mangat (2012) say that “stock market is a place where buying and selling of stocks/ shares take place and to make maximum profit, right investment should be made at the right time”.

In their research, they also mentioned that data mining is being actively applied to stock market since the 1980s and the various aspects of stock market to which data mining has been applied include predicting stock indices, predicting stock prices, portfolio management, portfolio risk management, trend detection, designing recommender systems, etc.

1.1. Analysis goal:

1.1.1. Predict future trend of stock market by analyze selected counter for short term and long-term investment.

1.2. Analysis data:

1.2.1. Extract data from historical Bursa Malaysia Main Market.

1.2.2. Sample target based on high data trend.

1.2.3. Actual sample target approximately 10%.

2. Methodology

Umair Shafique and Haseeb Qaiser (2014) say that “KDD or Knowledge Discovery Databases is the process of extracting the hidden knowledge according from databases. KDD requires relevant prior knowledge and brief understanding of application domain and goals. KDD process model is iterative and interactive in nature. SEMMA stand for (Sample, Explore, Modify, Model, and Access) is data mining method developed by SAS institute. It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically a logical organization of the functional tools for them”.

Figure 2.1 shows the comparison between KDD and SEMMA. In details, most of the process is quite similar thus for this project, both methodology will be used.

KDD	SEMMA
Pre KDD	----------
Selection	Sample
Pre processing	Explore
Transformation	Modify
Data mining	Model
Interpretation/Evaluation	Assessment
Post KDD	----------

Figure 2.1: KDD vs SEMMA

Figure 2.2 show details process in SEMMA methodology. In this assignment, SEMMA is widely used and this report structure also is based on this methodology starting from Sample, Explore, Modify, Model and Assess.

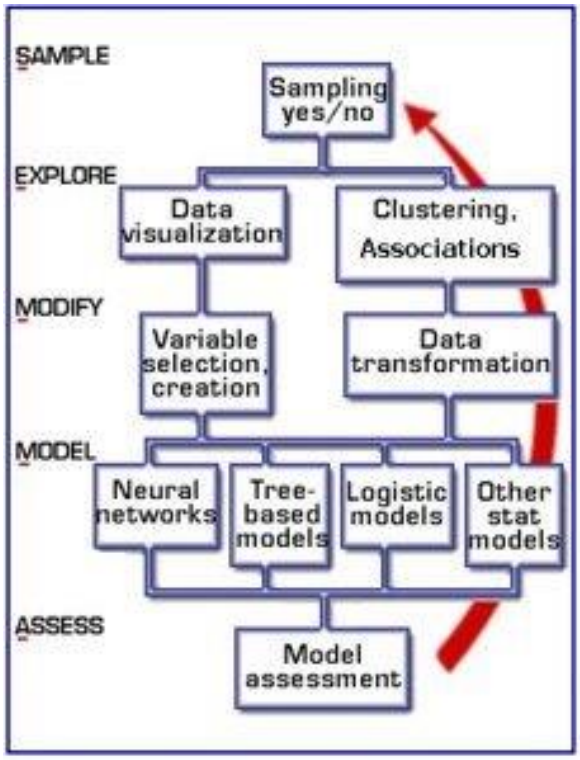

Figure 2.2: SEMMA methodology

2.1. Sample

Acquisition of data is one of the challenge in this assignment. Stock market data is selected as a domain. For this assignment, MYEG counter will be a focus. To get this data, few step need to follow as shown in Figure 2.3.

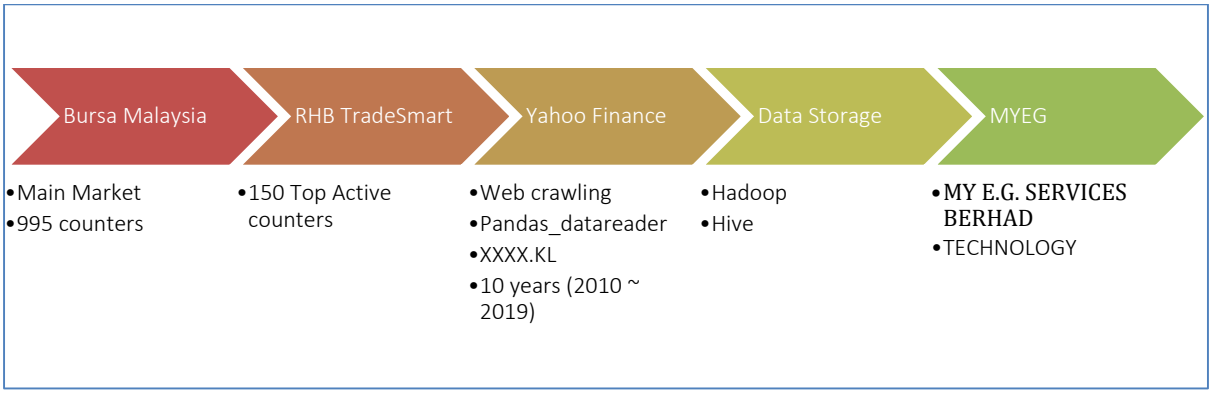

Figure 2.3

RHB TradeSmart platform is used to access historical data of Bursa Malaysia as shown in Figure 2.4. In this platform, 150 Top Active counters is identified.

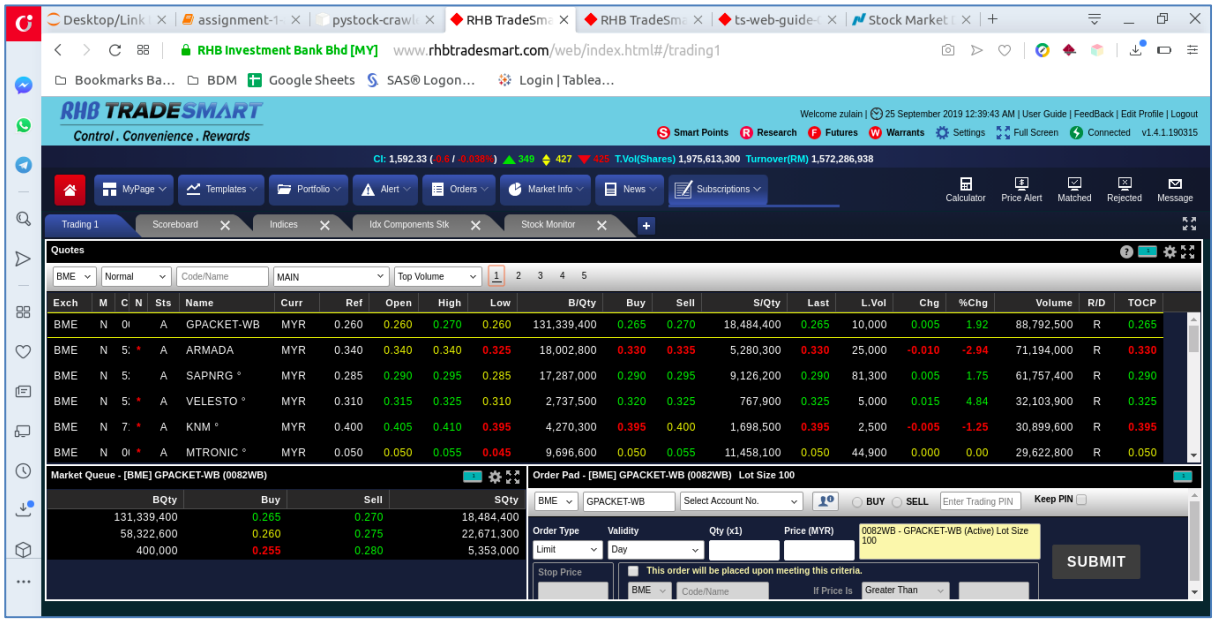

Figure 2.4

Stock data is unable to crawl at RHB TradeSmart platform due to website limitation. Due to this issue, all stock code and stock name for all 150 counters then is refer to Yahoo Finance website as shown in Figure 2.5. A bit different here, stock code that are using in Yahoo Finance is written as xxxx.KL. Next, all stock name in Yahoo Finance is saved for crawl activity.

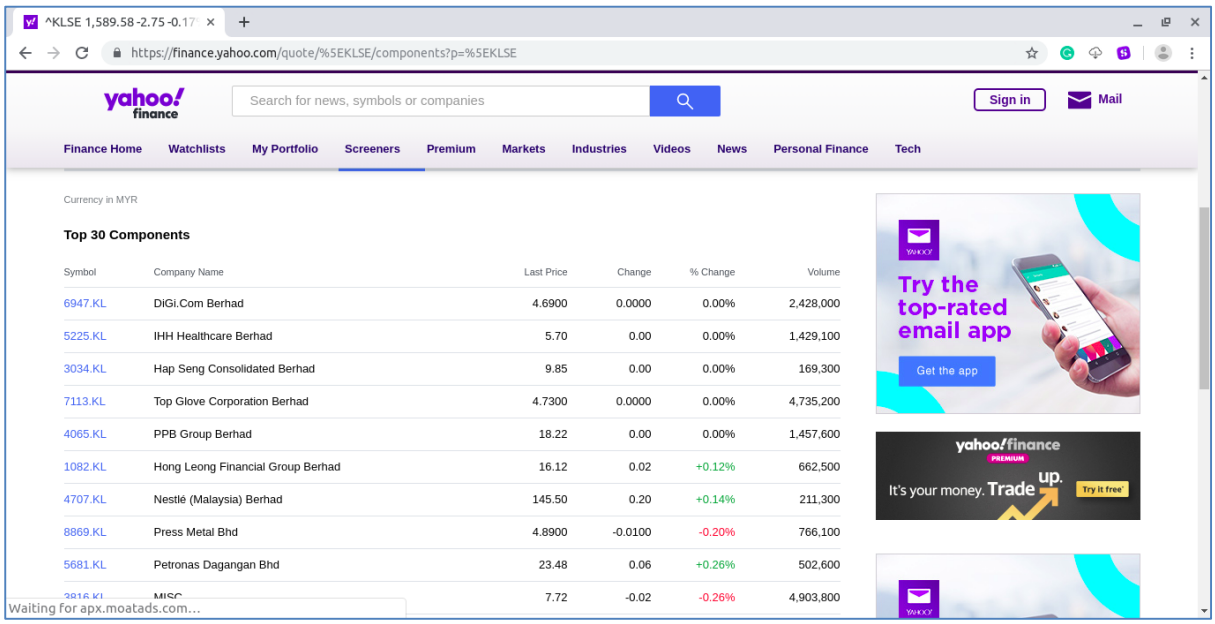

Figure 2.5

Dataset from Yahoo Finance is being crawl and downloaded using Phyton with pandas-datareader. Historical data is downloaded from the year 2010 till 2019. Figure 2.6 shows head of the data that be crawl with pandas-datareader.

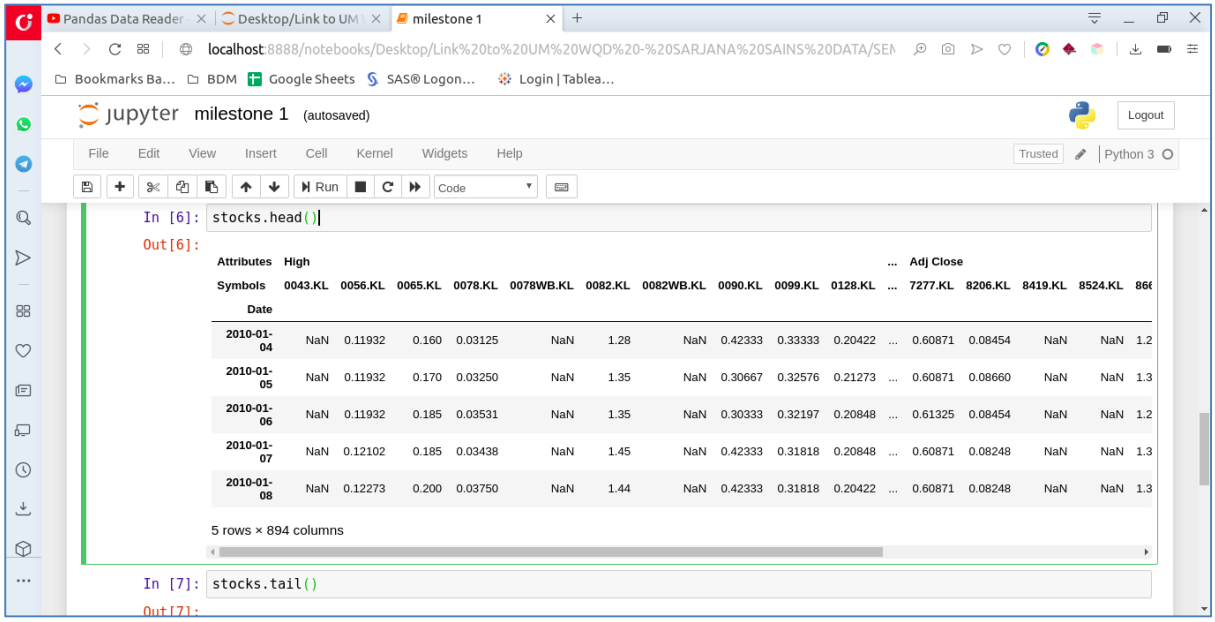

Figure 2.6

Downloaded data then stored in Hive. Before stored in Hive, firstly Hive need to be installed in Hadoop environment. Figure 2.7 shows step by step on how to install Hadoop and Hive.

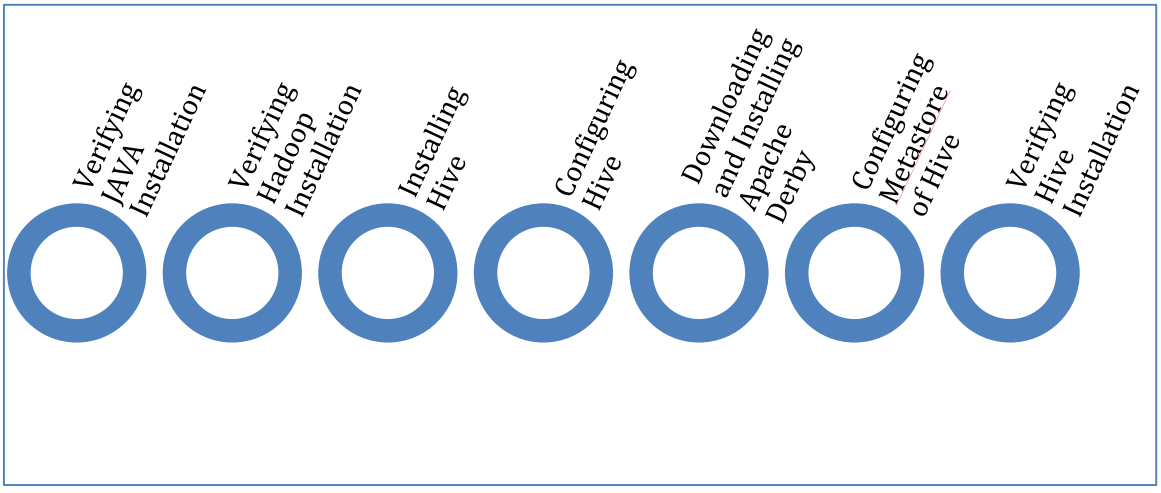

Figure 2.7

Cleanse data is a step of preprocessing in Knowledge Discovery in Databases (KDD). MYEG dataset is selected from Hive since this is a target data to analyse. Figure 2.8 shows example on how to select data from Hive

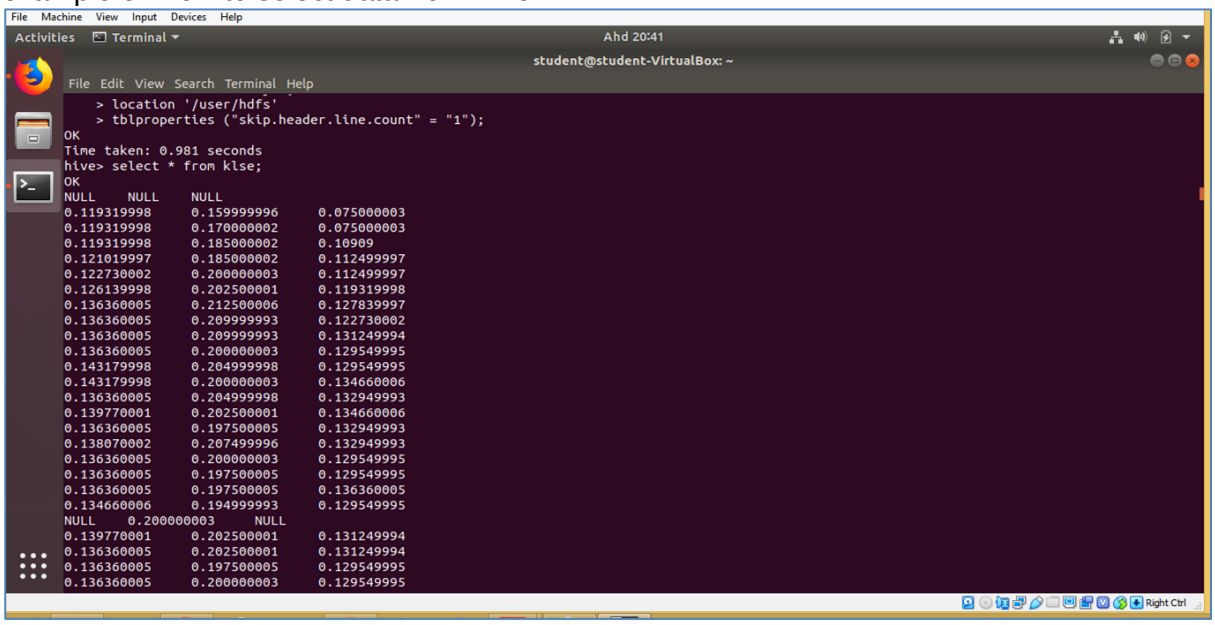

Figure 2.8

2.2. Explore

Exploring MYEG data using SAS Enterprise Miner found it have 7 attributes with 2405 instances as shown in Figure 2.9. Type of variable is Date, Volume, Open, High, Low, Close and Adj_Close.

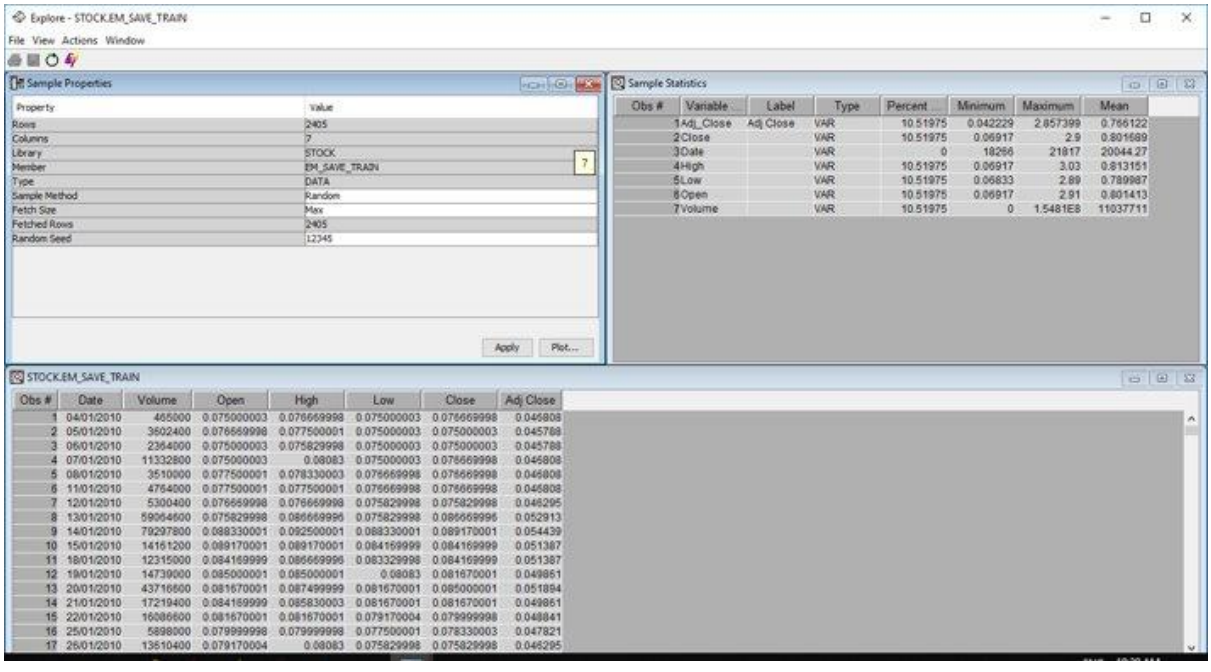

Figure 2.9

Observed highest point for MYEG always in Q1 financial year as shows in Figure 2.10. Also observed, there is missing value for MYEG data from mid-2018 till mid-2019.

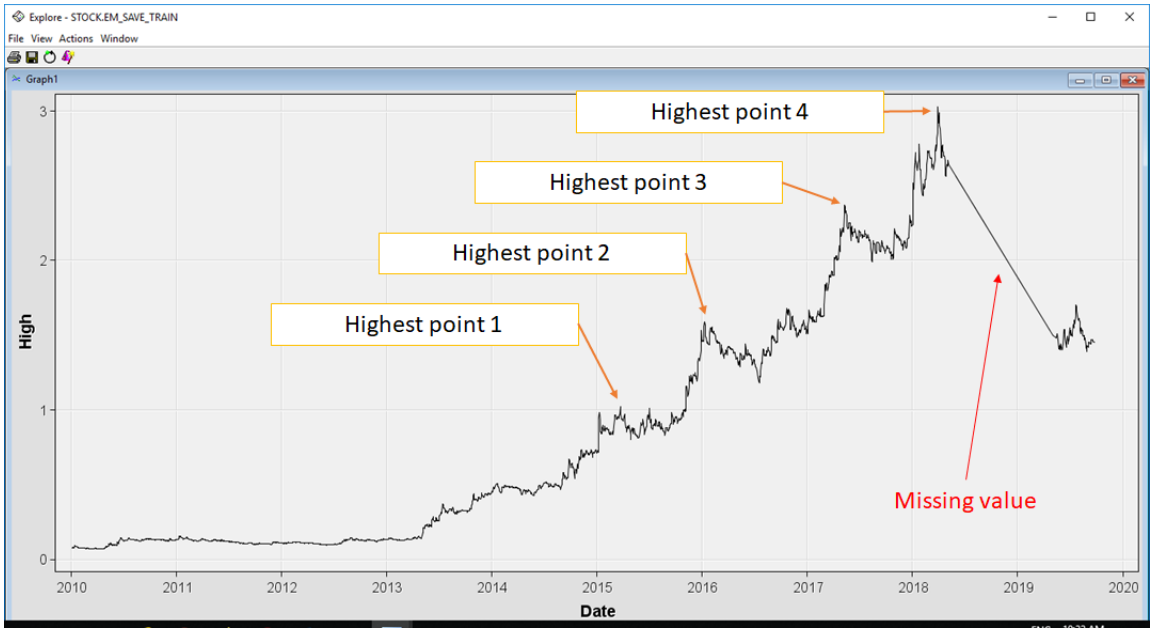

Figure 2.10

Curious with that observation then further analysis have been done. Google search shown MYEG released it Financial Report in Q1 every year as per Figure 2.11. Thus it make sense why at this period the MYEG stock is always achieve highest point.

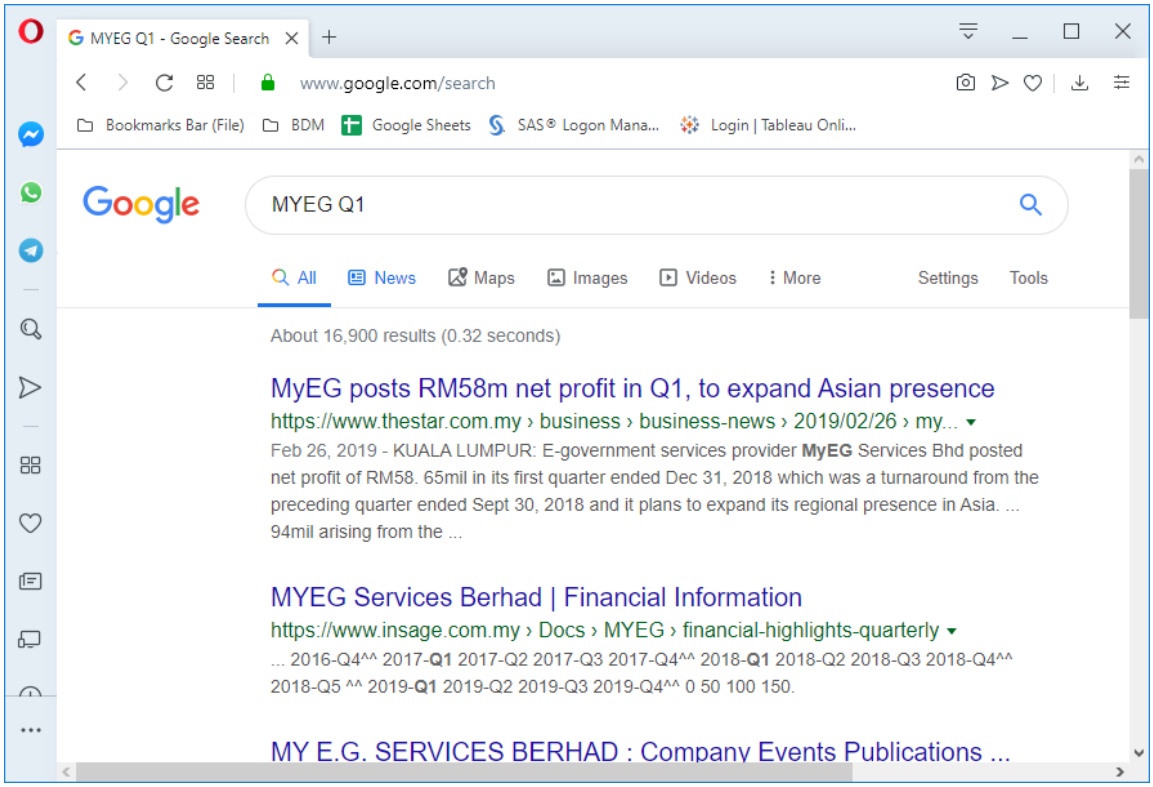

Figure 2.11

Detail observations tabulated from year 2010 to 2019 as shown in Figure 2.12. Variable Adj_Close, Close, High, Low and Open shows similar pattern. MYEG stock is at mid-point which is 1.25 to 1.55 in year 2019.

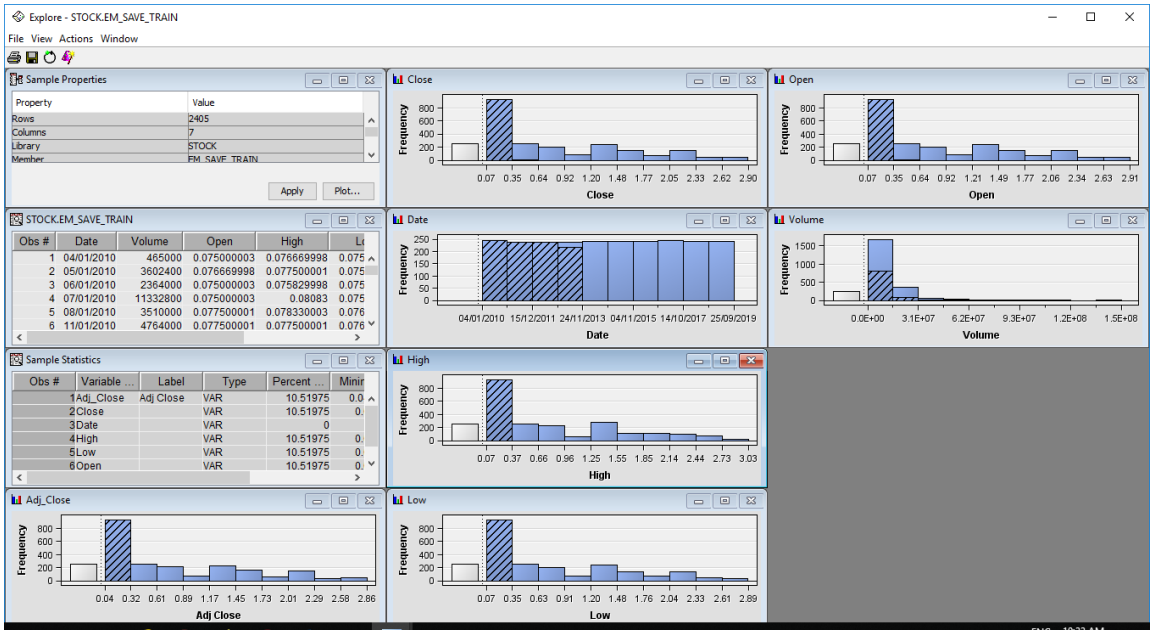

Figure 2.12

Histogram proved that the highest point (High variable) for MYEG is in range of 0.07 to 0.37 with 935 times as shown in Figure 2.13.

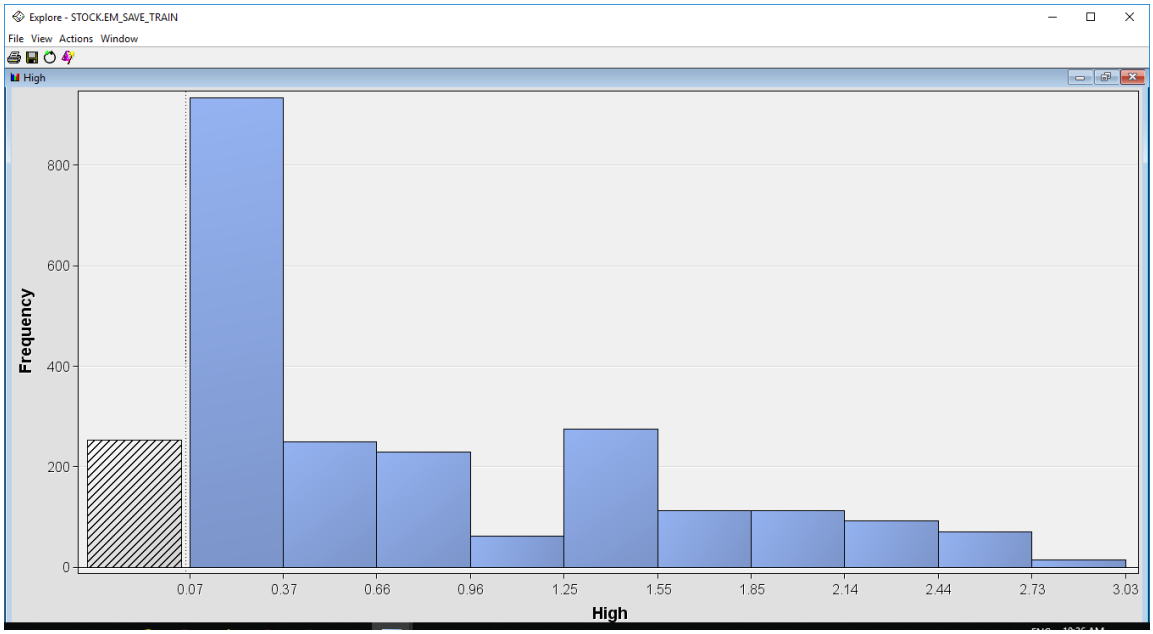

Figure 2.13

Prescriptive analysis is carried out with clustering by using Cluster node in SAS Enterprise Miner. Finding, cluster 10 is the biggest cluster with the lowest Volume (4302217) while cluster 9 is the smallest cluster however it consist of the biggest Volume (80175967) as shown in Figure 2.14. Adj_Close is the importance variable with 1.0000 points followed by Open, Low, Close and Volume.

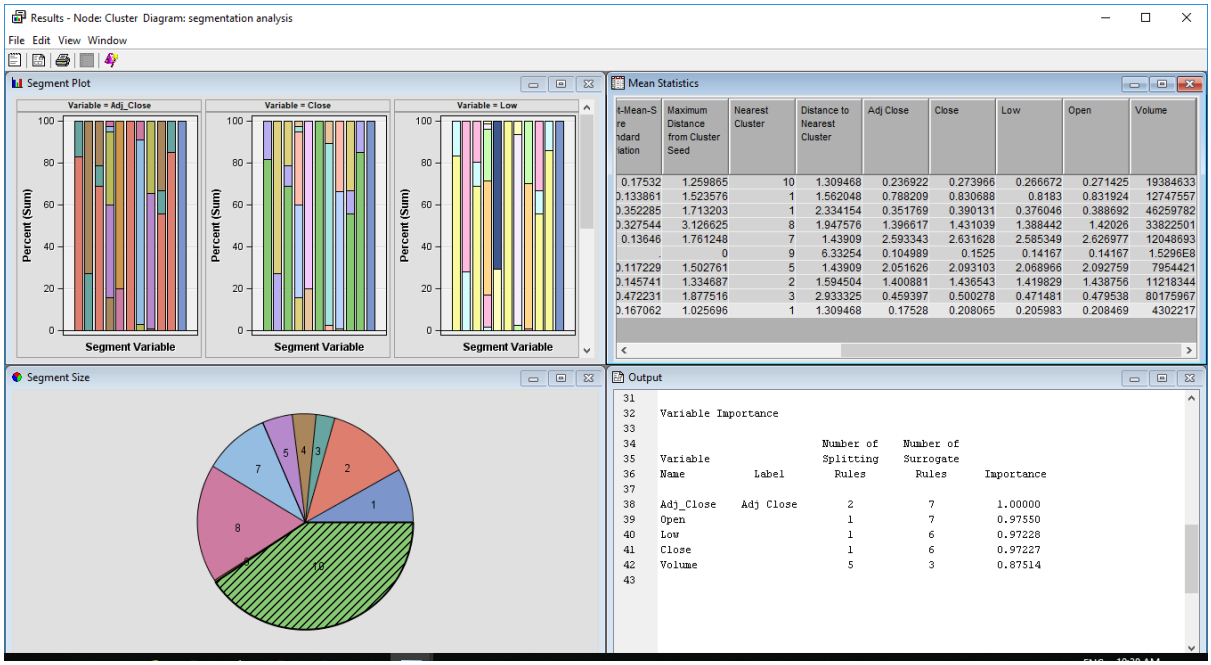

Figure 2.14

2.3. Modify

Before move to Model step, setting of High variable have been Modify by using Replacement node. Limit Method is set to User Specified and Replacement Lower Limit is set to 1. Refer to Figure 2.15, setting done to High variable Data Set Allocation by using Data Partition node with Training = 50 and Validation = 50.

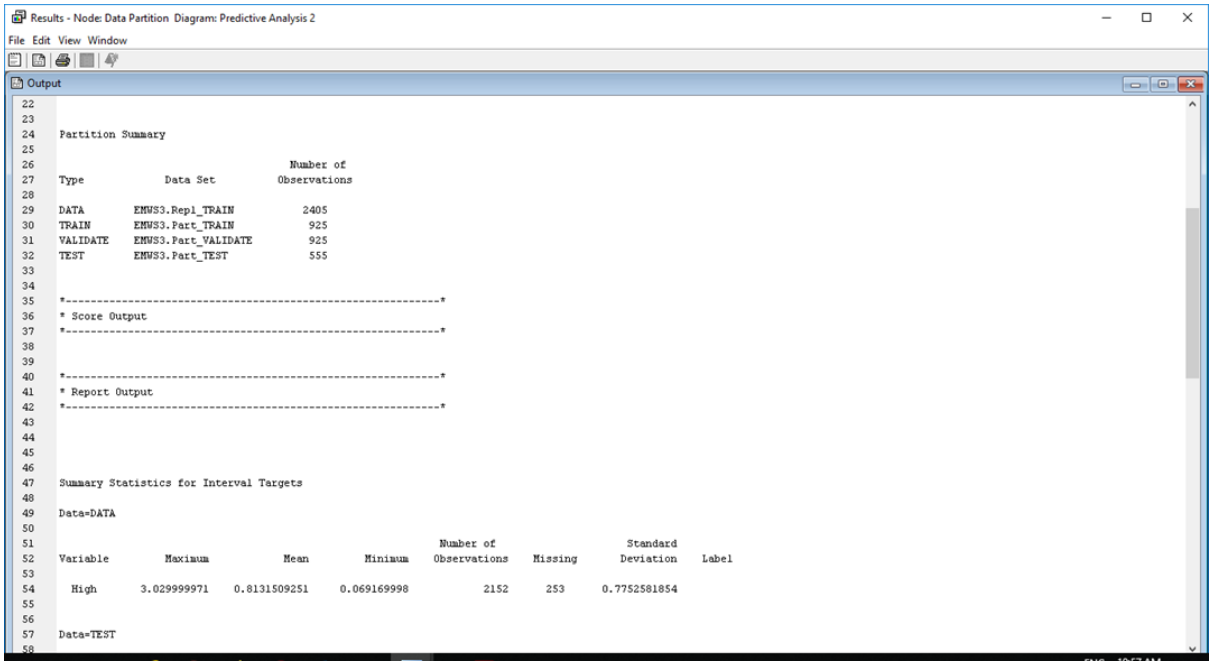

Figure 2.15

2.4. Model

Decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences. Decision tree is used in this assignment to perform predictive analysis. Initial decision tree result shown as per Figure 2.16. <1.0035 or Missing shows higher concentration in Split 1.

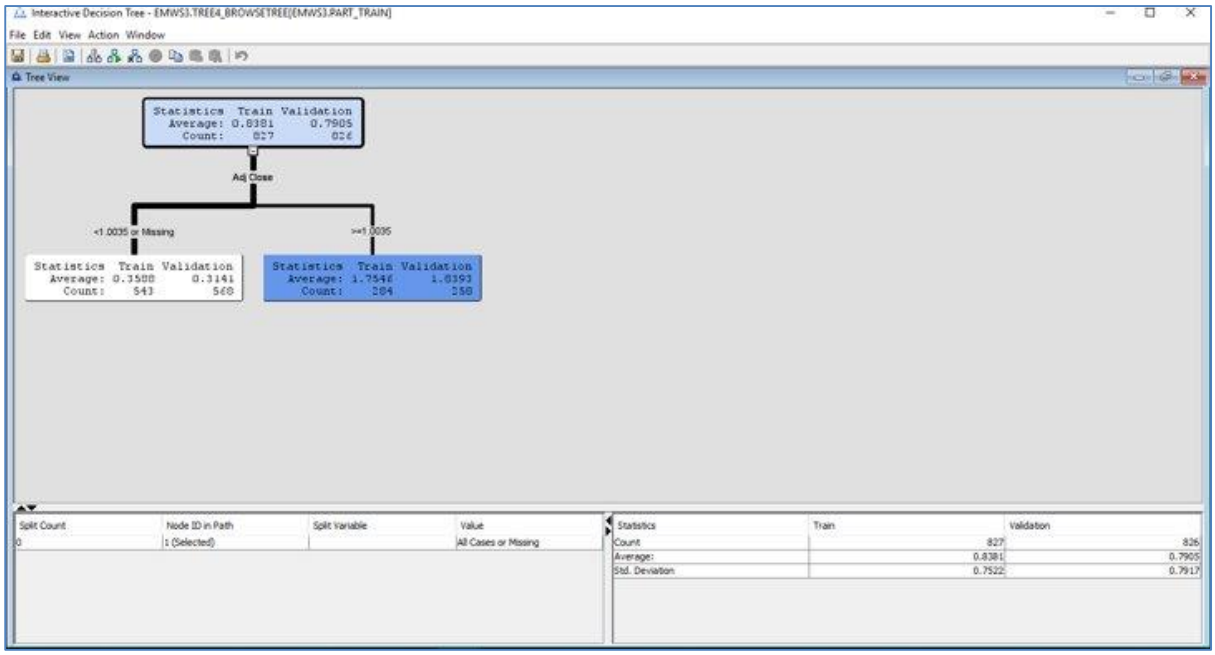

Figure 2.16

Figure 2.17 show maximal tree. In total there is 44 leaf-trees generated.

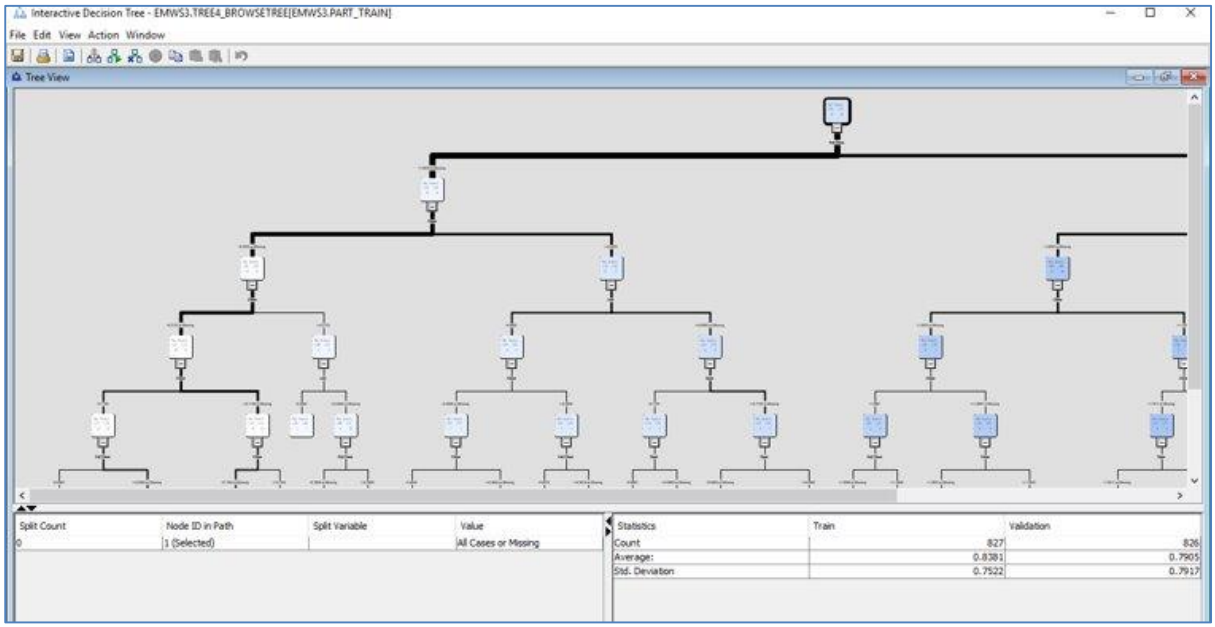

Figure 2.17

Based on 44 leaf-trees generated, 40 leaf-trees is a lower misclassification rate as shown in Subtree Assessment in Figure 2.18.

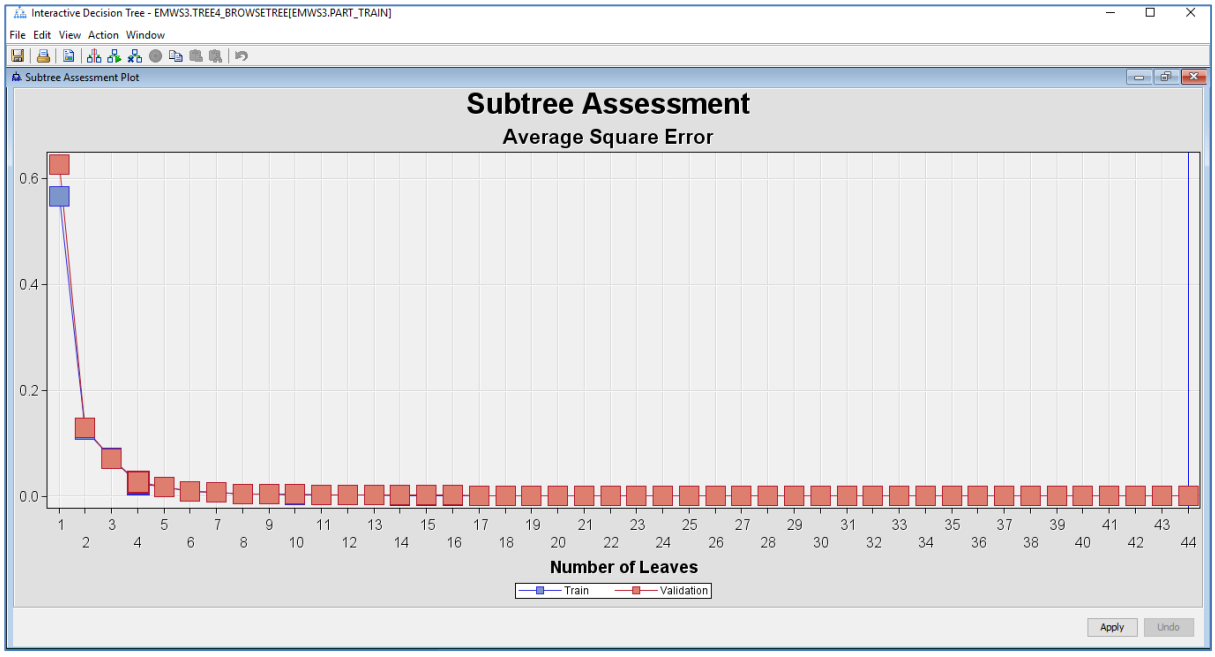

Figure 2.18

Full result of decision tree is shown in Figure 2.19.

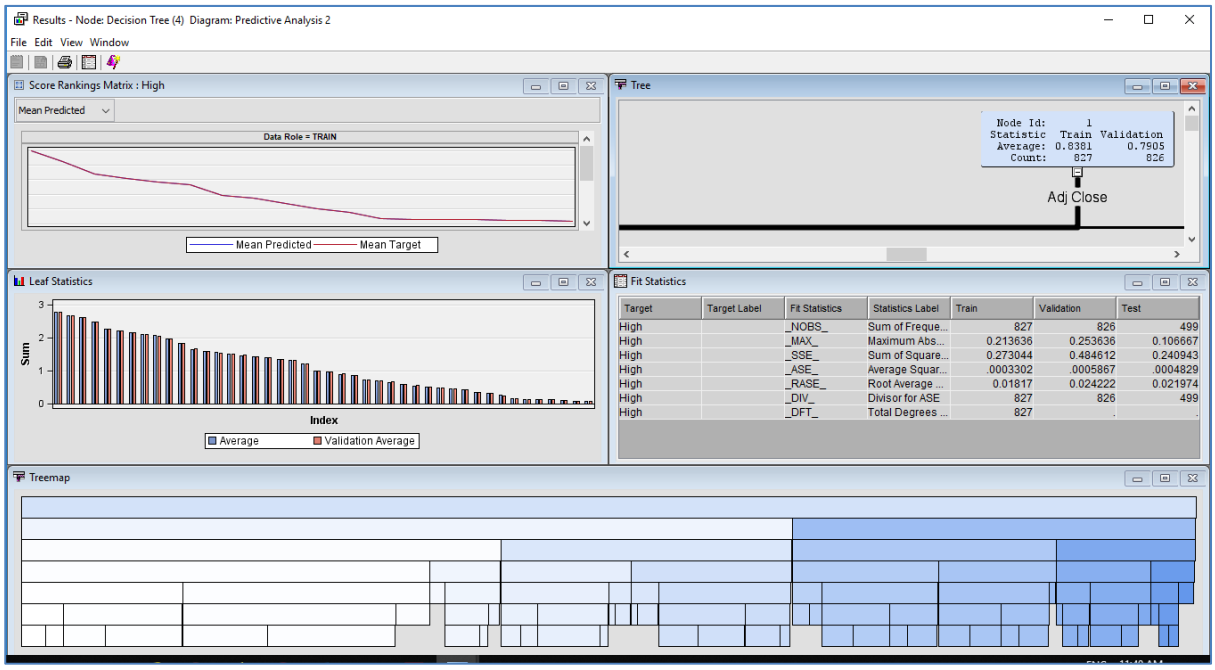

Figure 2.19

2.5. Assess

Using Segment Profile node, hidden knowledge inside data can be assess. Figure 2.20 shows full result of Segment Profile. There is 10 segment generated with segment 10 has the highest frequency which is 812 followed by segment 8 (351 frequency).

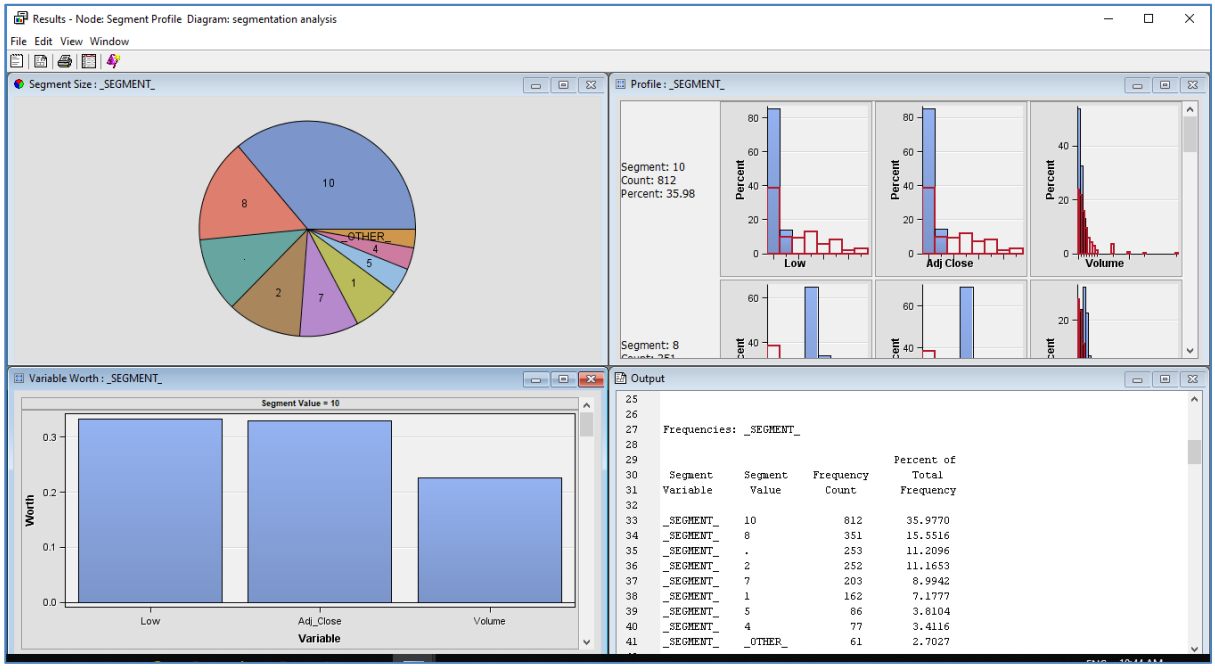

Figure 2.20

Figure 2.21 shows Profile for each segment. For segment 10, value for Low, Adj_Close and Volume is at minimum point. Difference with the 2nd higher segment which is segment 8, Adj_Close and Low value is at center point but Volume is at minimum point compared to the average.

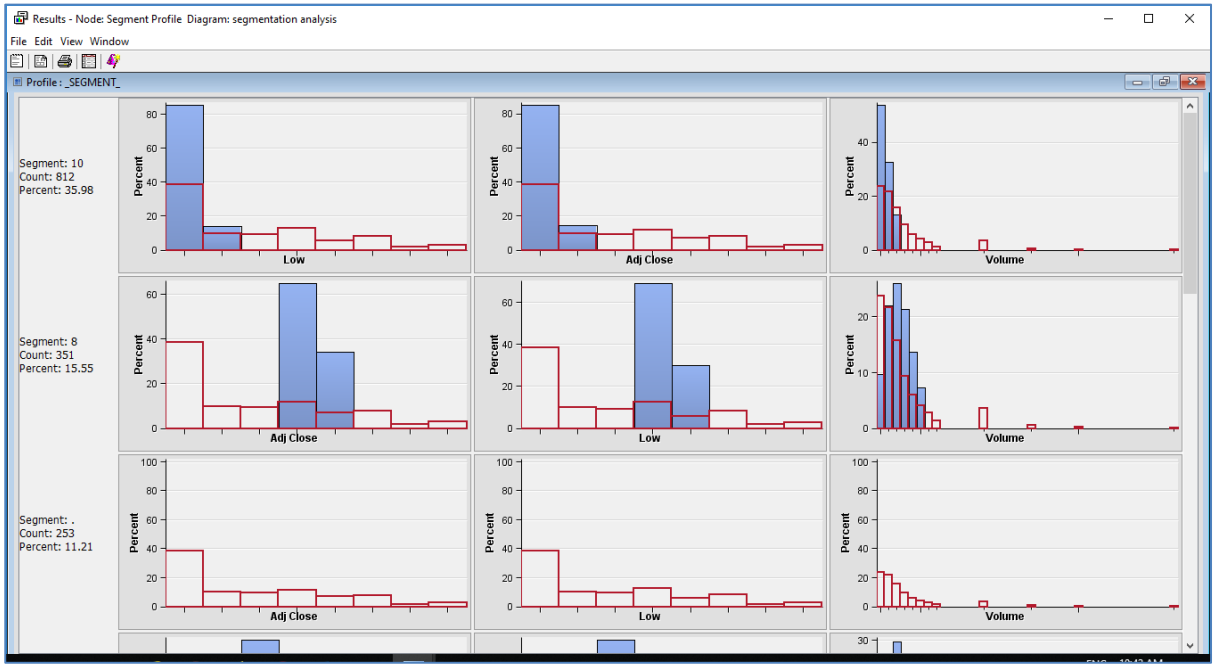

Figure 2.21

3. Conclusion and Recommendation

For this assignment, here is the results:

- MYEG stock shows good uptrend from year 2013 to 2018.
- Q1 every year is a good period to gain higher profit in investment (positive sentiment due to company Financial Report).
- The best time to invest in MYEG stock is in Q3 (lowest point every year).
- Even though current value of MYEG stock is at mid-point which is 1.25 to 1.55 it still shows a good sign to improved based similar position as per mid-2017.

Overall in this assignment, all technique in data mining have been discover by using KLSE stock market dataset as a case study. Data mining does not only help to make an analysis however it is more than that.

As summary, data mining can help the investor to face risks in order to make a profit however more study needs to be conducted to gain a confidence level and lowering the risk.