Milestone 6

Data Mining Goal:

Problem:  
To allow an investor to choose a stock to invest and try to get the maximum return and minimum lost for a short term investment.

To do:

Base on the daily stock market activity, identify some stocks which are able to be predicted whether the closing prices will raise or going down.

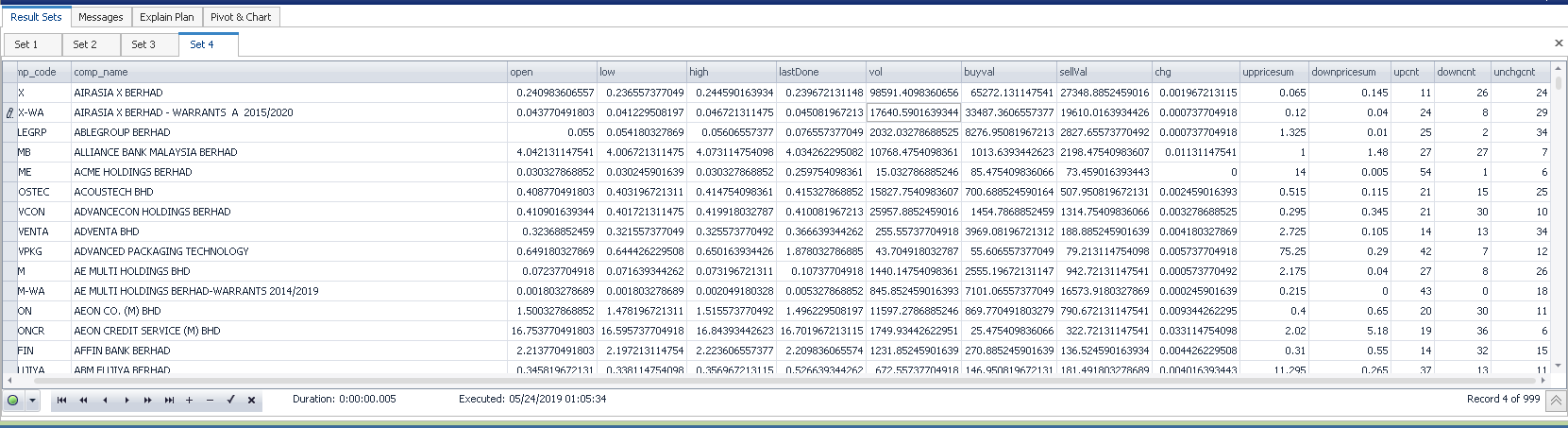
Prediction is made based on the probability to get raise, down or unchanged of the closing price change of the stocks activities for the period the data collected.

The process to achieve the goal:

1. Select the data from mysql

The data selected for each stock by calculating the mean of each columns for open, low, high, lastDone, vol, buyval, sellval, chg, v\_updownpercent. Only the stock category belong to ‘main’ is selected. Total number of stock is 999.

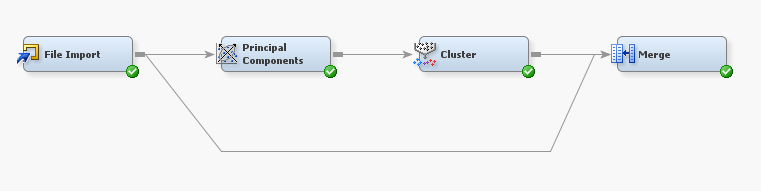
|  |
| --- |
| select  comp\_code, comp\_name, avg(open) as open, avg(low) as low, avg(high) as high,  avg(lastDone) as lastDone, avg(vol) as vol, avg(buyval) as buyval,  avg(sellval) as sellVal, avg(chg) as chg,  sum(  case when (lastDone - open) > **0** then lastDone - open  else **0**  end  ) as uppricesum,  sum(  case when (lastDone - open) < **0** then open - lastDone  else **0**  end  ) as downpricesum,  sum(  case when (lastDone > open) then **1**  else **0**  end  ) as upcnt,  sum(  case when (lastDone < open) then **1**  else **0**  end  ) as downcnt,  sum(  case when (lastDone = open) then **1**  else **0**  end  ) as unchgcnt  from (  select  distinct k.thedate, k.comp\_code, k.comp\_name, k.open, k.low,  k.high, k.lastDone, k.vol, k.buyval, k.sellval,  k.chg, k.v\_updownpercent  from upd\_klse k  inner join compcat c  on k.comp\_code = c.compCode  where c.category = 'main'  ) as t  group by comp\_code, comp\_name |

The result:  


1. Load the data into SAAS Enterprise miner

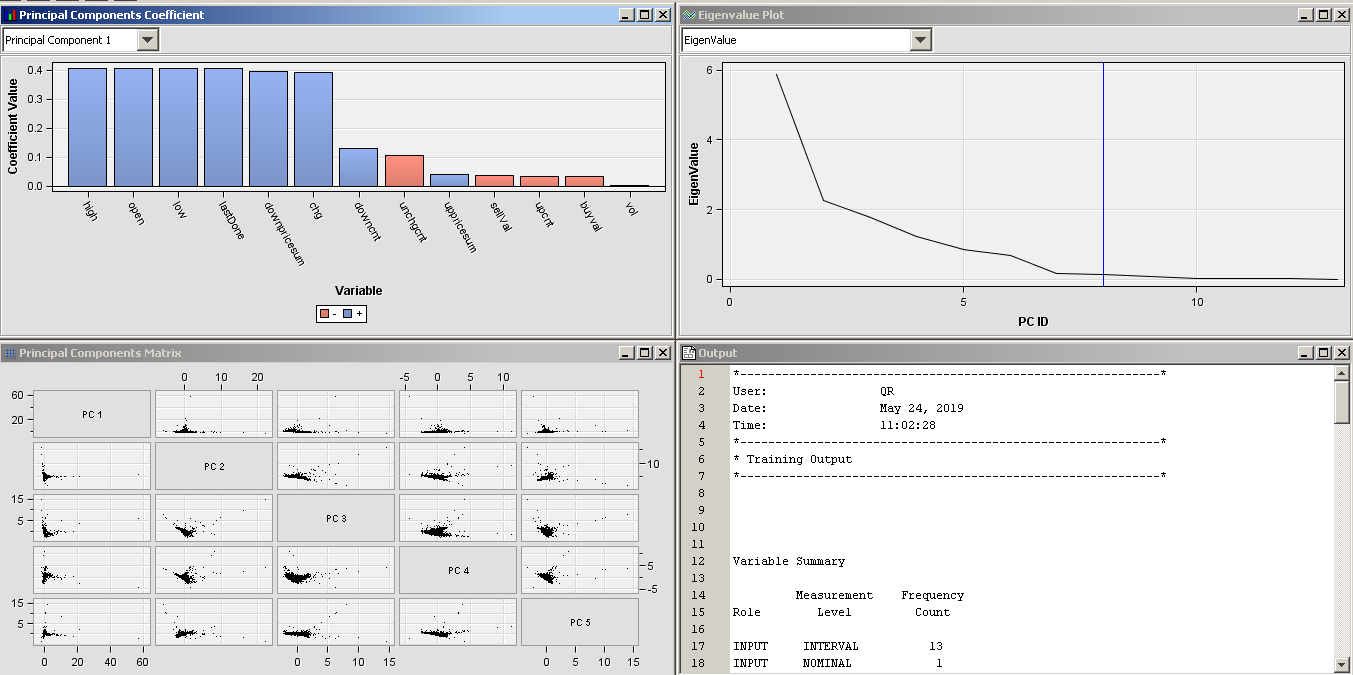
The selected data was exported into CSV file, percentageChg\_main\_sumupdown.csv, and loaded into SAAS Enterprise miner.

1. Performance Clustering observation

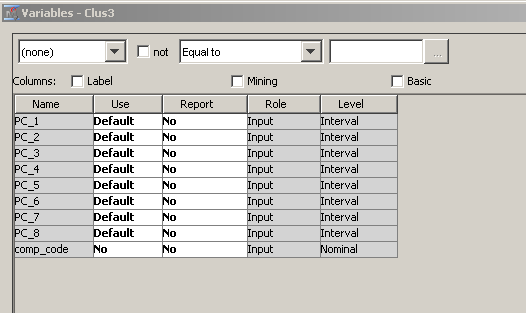


1. The data is loaded into Enterprise miner
2. A Principal components added for analysis the principal components of the dataset.

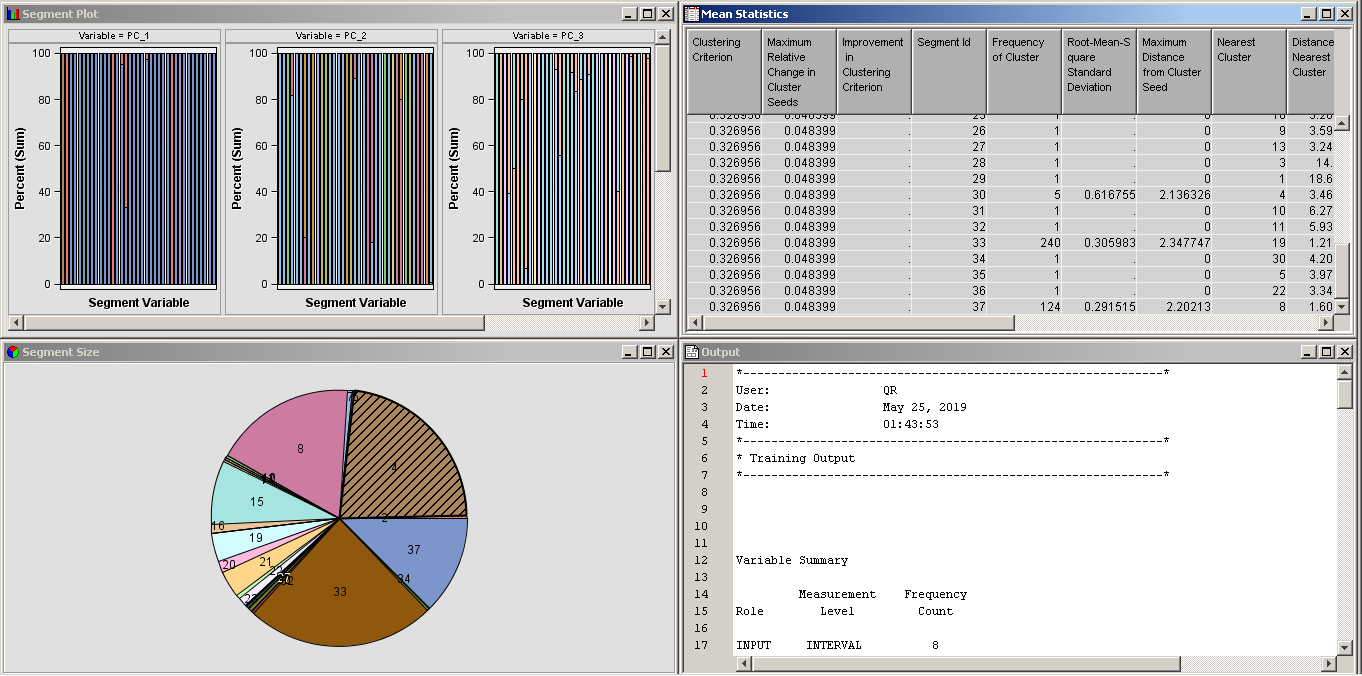
There are 8 sets of principal components and the result of the principal components are added to the dataset for clustering.



1. The clustering node is applied with default setting and only the principal components are used for clustering.

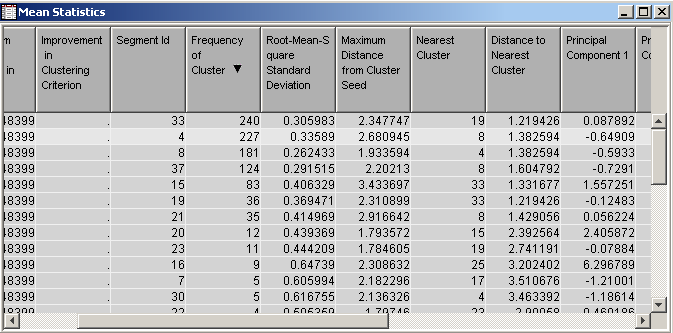


The result of the clustering for all the attributes and new added principal component sets is shown as below:



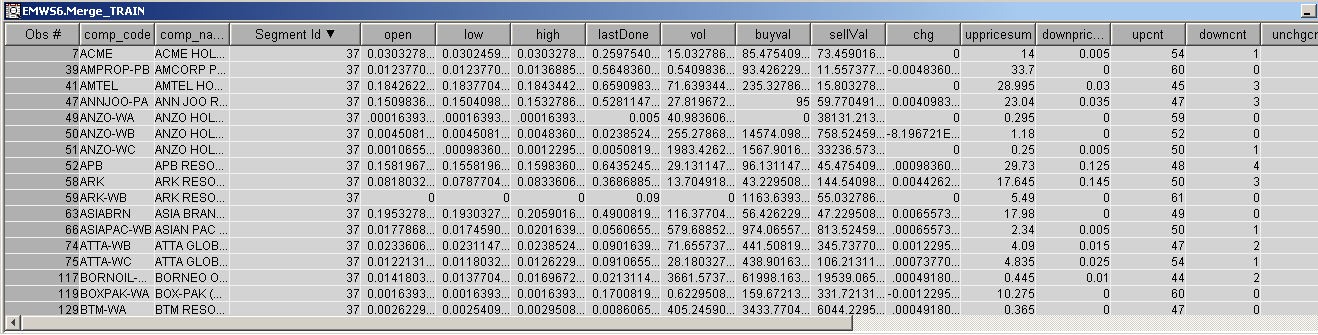
There are 37 clusters created, however, just the clusters with high frequency

The frequency of the cluster is listed as shown as below:

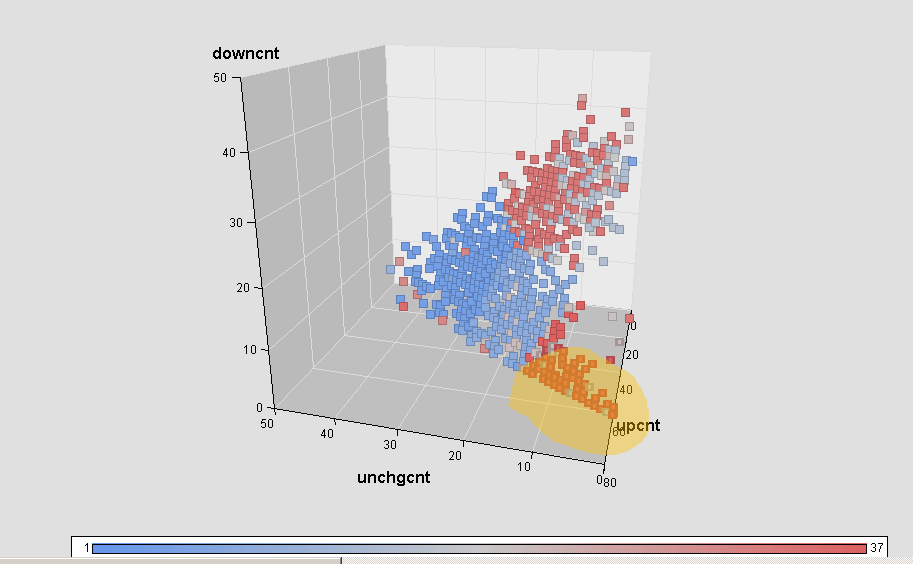


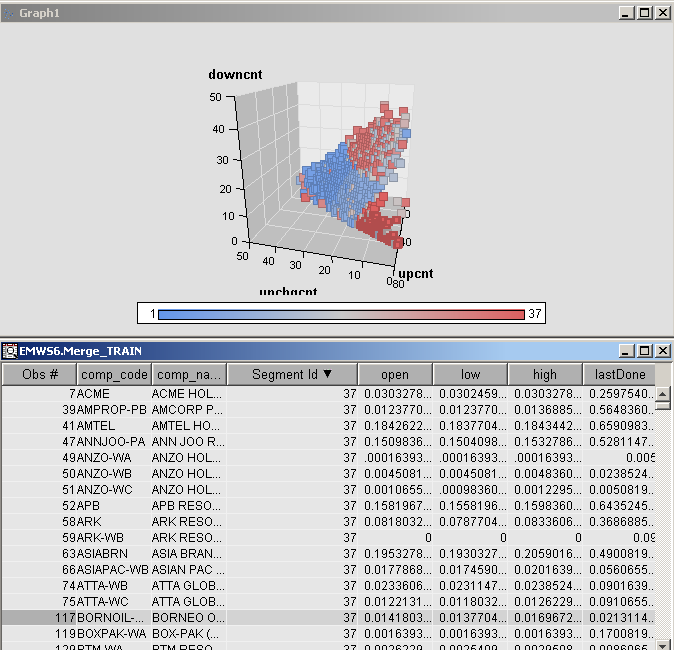
The cluster id 33, 4, 8 and 37 have more meaningful with higher frequency.

1. Clustering result is merged to the dataset, so that each of the stocks will be assigned the segment Id.



1. Plot a graph to describe the occurrence of stock price getting up or down or remain unchanged during the period of data collection. The color is representing the cluster ID.



Only the stock with max occurrence of stock price is up and minimum occurrence of stock is decreased and minimum occurrence of price remained unchanged. These stocks have potential increase price and active in the market and suitable for the investor want to have shorten investment.   
  


Mostly the stocks which from segment Id 37 meets the criteria of the short term and fast investment.