**Solution Document**

**Assignment 2**

**Business Analytics with SAS – F16**

**(MIS 6324.502)**

Submitted by

Group 8

|  |  |
| --- | --- |
| Mayank Kothari | mkk160130 |
| Shamsundar Kulkarni | sxk162231 |
| Divya Jayaprakash | dxj160830 |
| Gautami Murugan | gxm161330 |
| Vijay Balaji | vxb161530 |

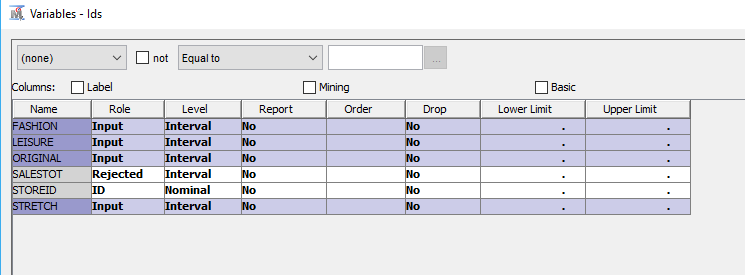
**Exercise 1: Clustering Stores**

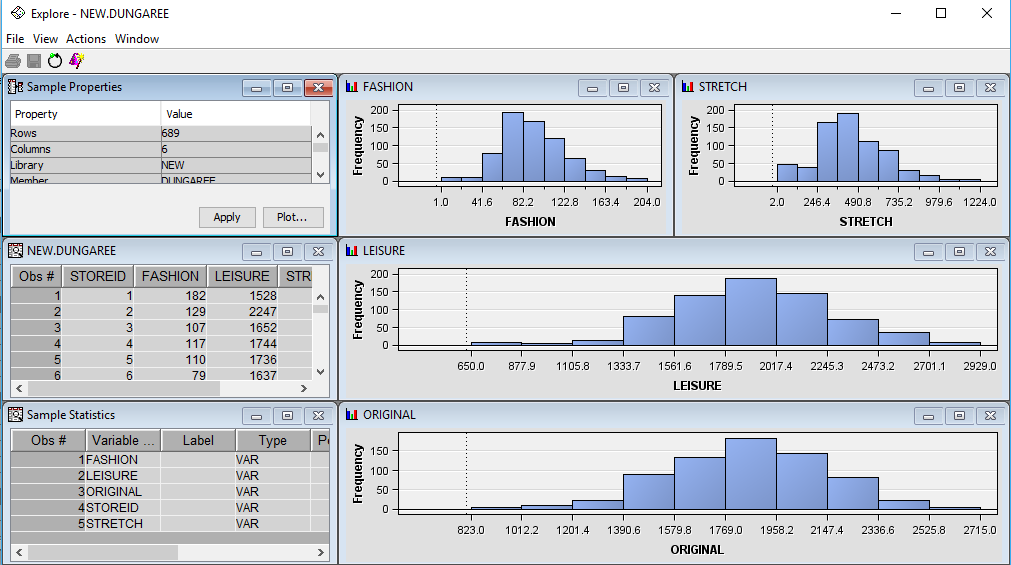
**3. Determine whether the model roles and measurement levels assigned to the variables are**

**appropriate. Examine the distribution of the variables. Answer the following with explanations.**

**a. Are there any unusual data values?**

**Result:**

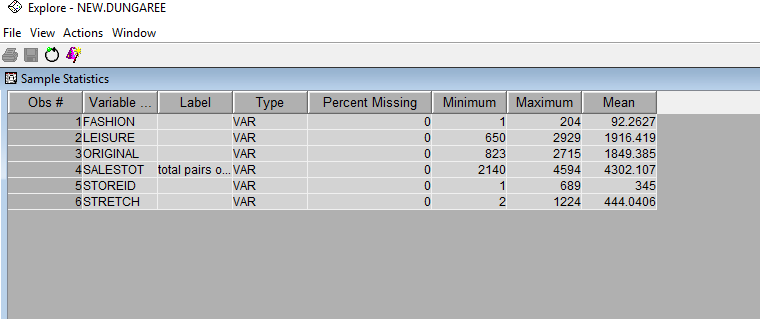
****

****

Unusual data values are not observed.

**b. Are there missing values that should be replaced?**

**Result:**



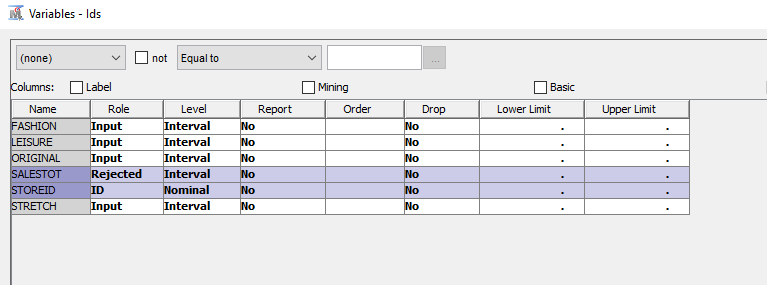
There are no missing values to be replaced.

**4. Assign the variable STOREID the model role ID and the variable SALESTOT the model role**

**Rejected. Make sure the remaining variables have the Input model role and the Interval**

**measurement level. Why should the variable SALESTOT be rejected?**

**Result:**



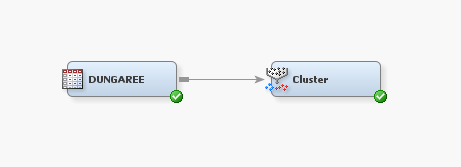
Variable SALESTOT gives the sum of the other input variables of the data set, so we don’t need SALESTOT separately as an input for our analysis.

**5. Add an Input Data Source node the diagram workspace and select the DUNGAREE data table as**

**the data source.**

**6. Add a Cluster node to the diagram workspace and connect it to the Input Data node.**

**Result:**



**7. Select the Cluster node and select Internal Standardization 🡺 Standardization. What would**

**happen if you did not standardize your inputs?**

**Result:**

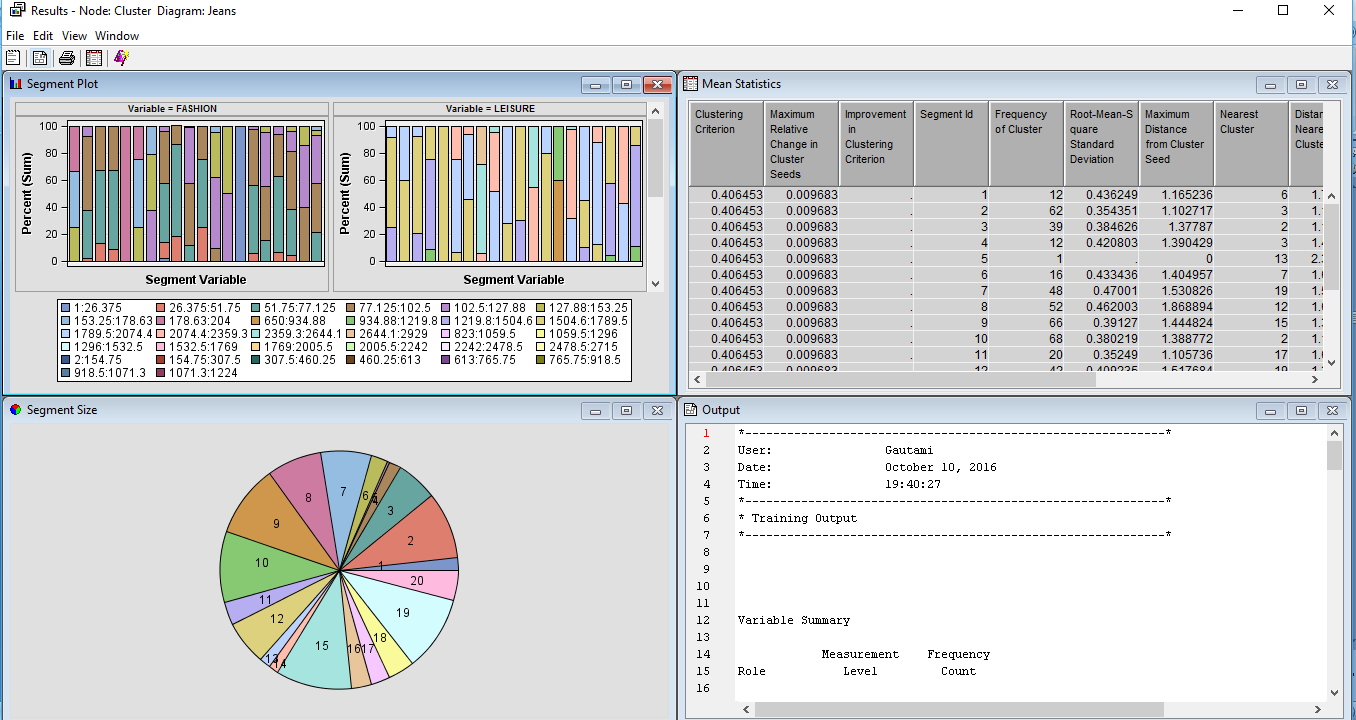
Clustering is based on similarity(measured in terms of distance). So the input variables should be scaled to the same level, hence we standardize our inputs.

In our case variables like ‘ORIGINAL’ and ‘LEISURE’ have a larger range compared to other inputs, so if we don’t standardize our input, these variables with larger range will influence the distance which in turn would influence clustering.

**8. Run the diagram from the Cluster node and examine the results. Does the number of clusters**

**created seem reasonable?**

**Result:**

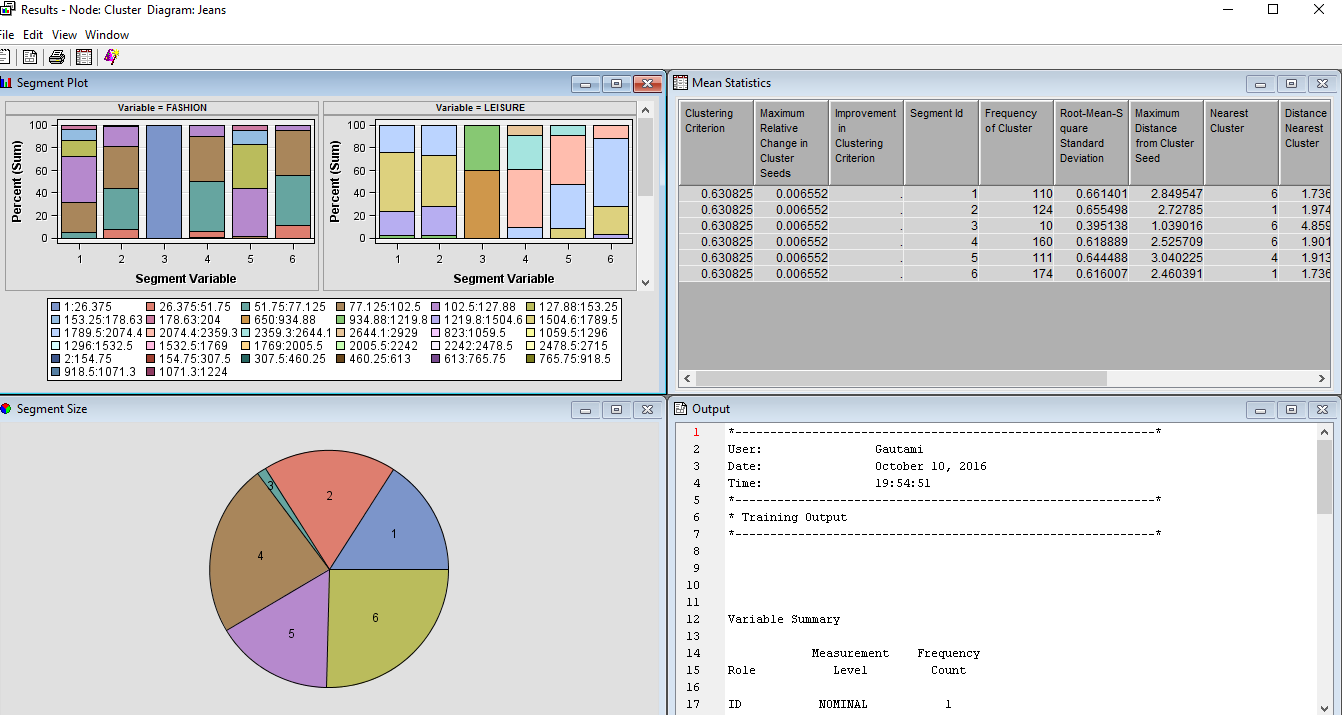
****

The number of clusters are not reasonable, excessive number of clusters are formed.

**9. Specify a maximum of six clusters and rerun the Cluster node. How does the number and**

**quality of clusters compare to that previously obtained?**

**Results:**

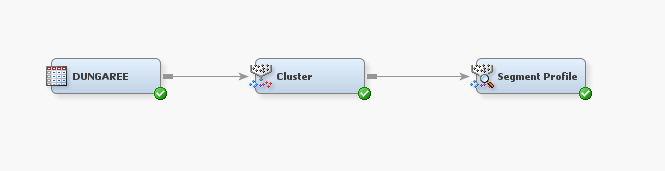


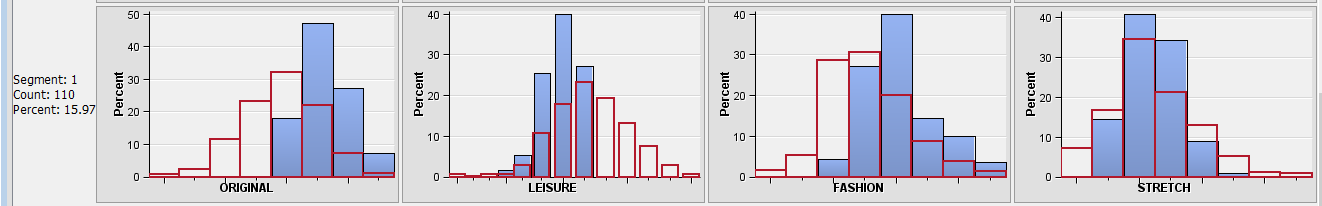
The previous cluster had 20 segments, compared to it we have lesser number of clusters this time i.e., 6 clusters.

Except segment 3, all other segments are highly populated, so the number of observations are more and the quality is better than the previous cluster.

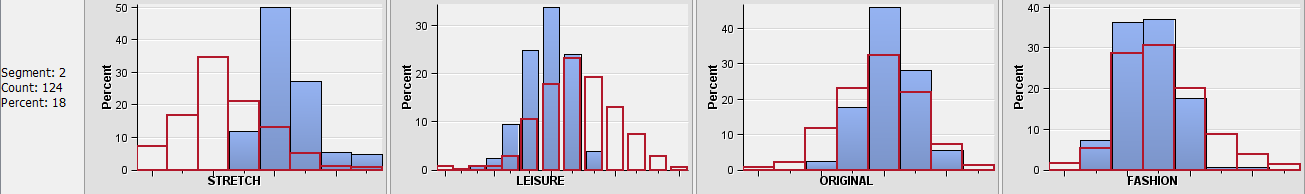
**10. Use the Segment Profile node to summarize the nature of the clusters.**

**Result:**

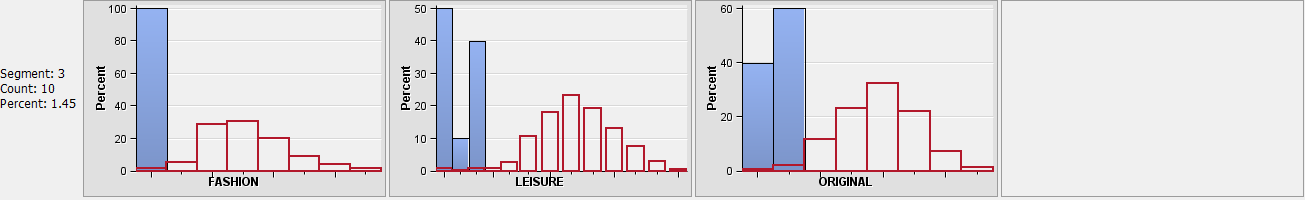
****



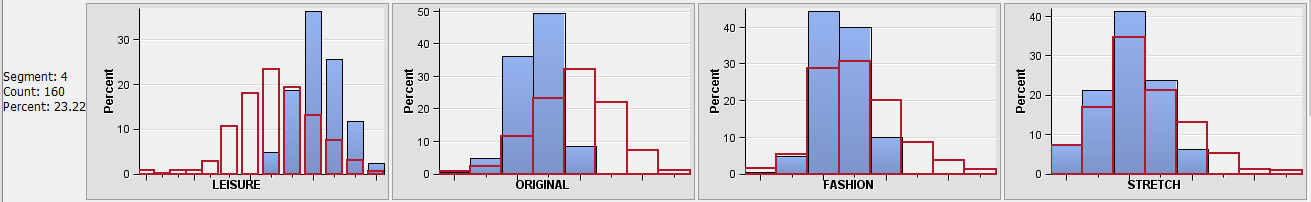
Segment 1 contains stores selling higher than average number of ORIGINAL jeans.



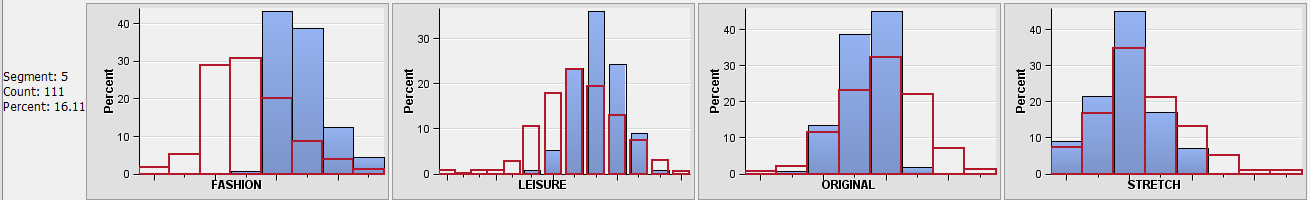
Segment 2 contains stores selling higher than average number of STRETCH jeans.



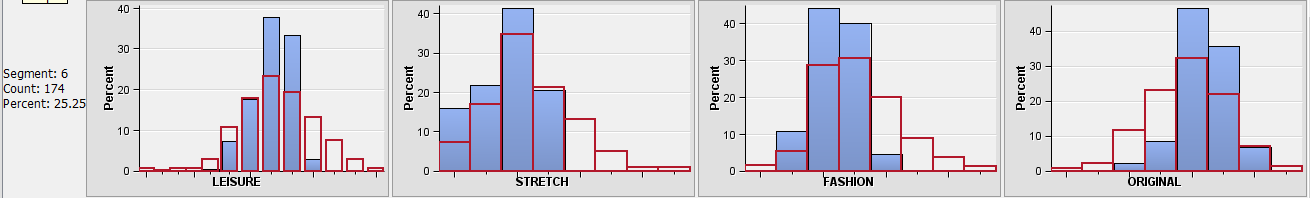
Segment 3 contains stores selling small number of all jean types.



Segment 4 contains stores selling higher than average number of LEISURE jeans.



Segment 5 contains stores selling higher than average number of FASHION jeans

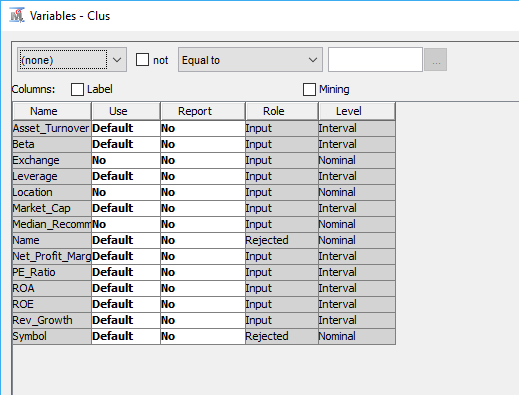


Segment 6 has stores selling higher than average number of ORIGINAL jeans, but lower than average number of STRETCH and FASHION jeans.

**Exercise 2: Clustering Pharmaceutical Firms**

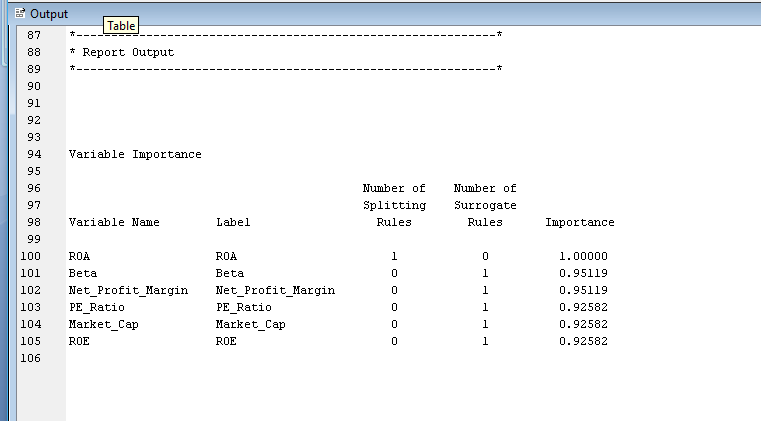
**1. Use only the quantitative variables (1‐9) to cluster the 21 firms. Use the default settings in SAS**

**Enterprise Miner.**

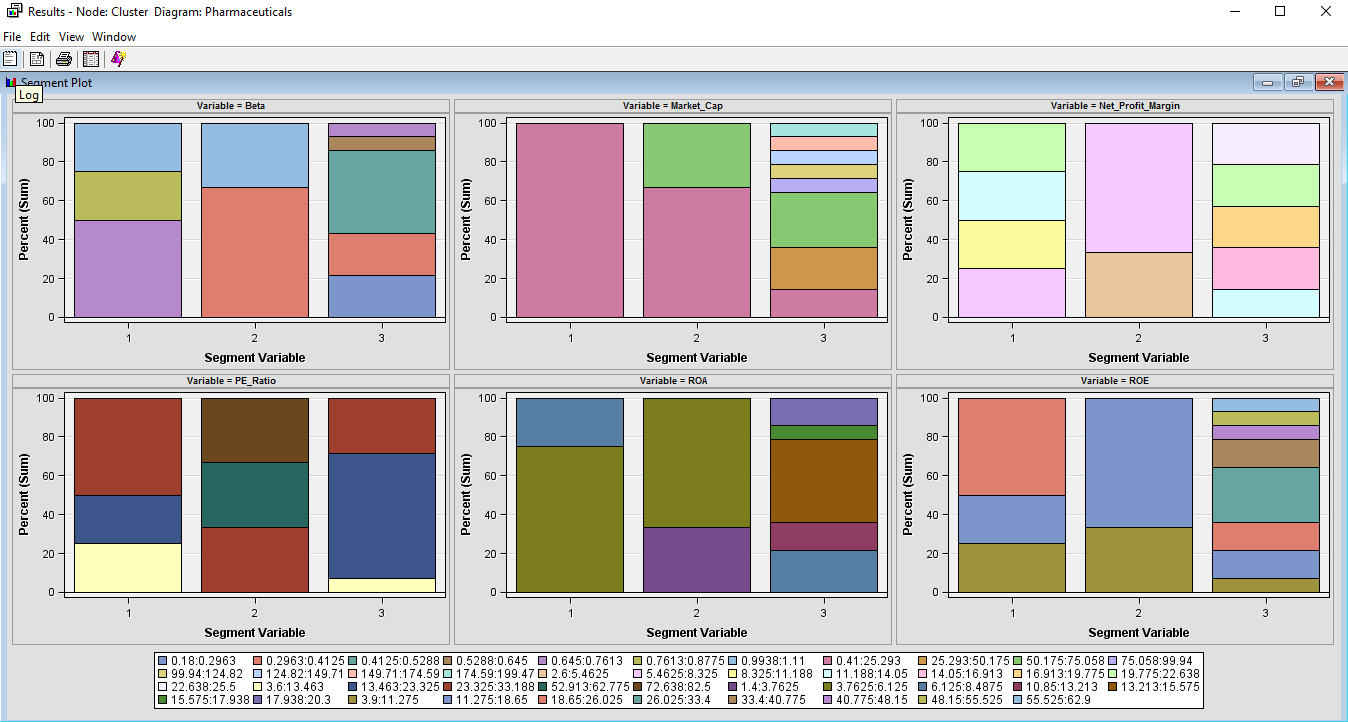


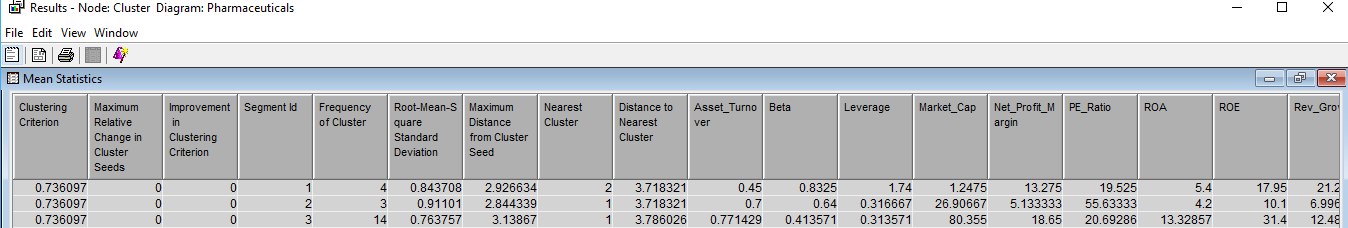
**2. Interpret the clusters with respect to the quantitative variables that were used in forming the**

**clusters.**



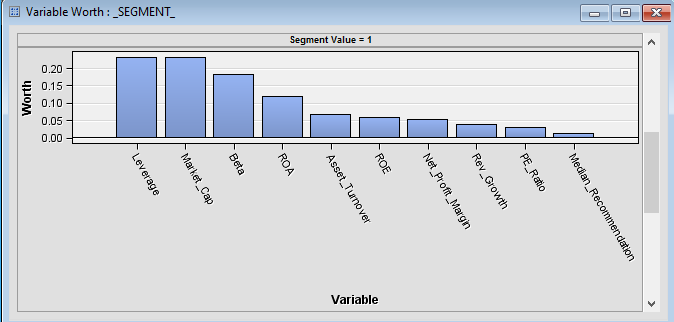
The variable of importance used for clustering are: ROA, Beta, Net\_Profit\_Margin, PE\_Ratio, Market\_Cap and ROE.

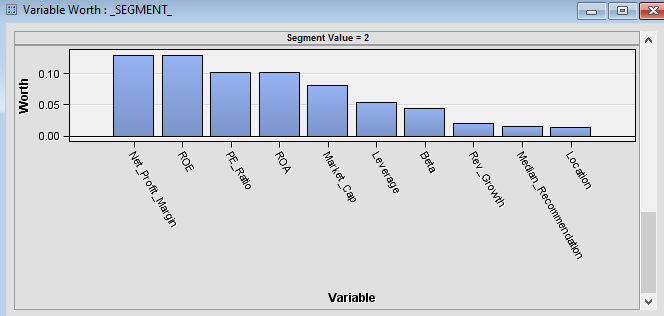


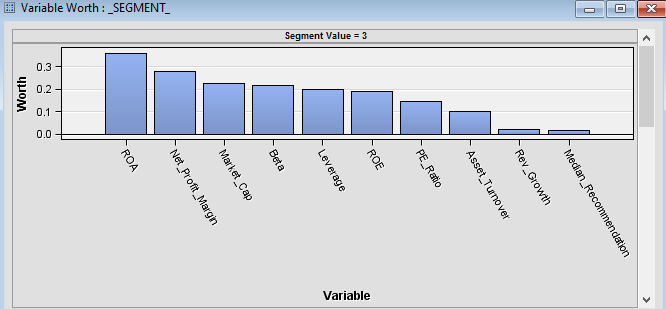


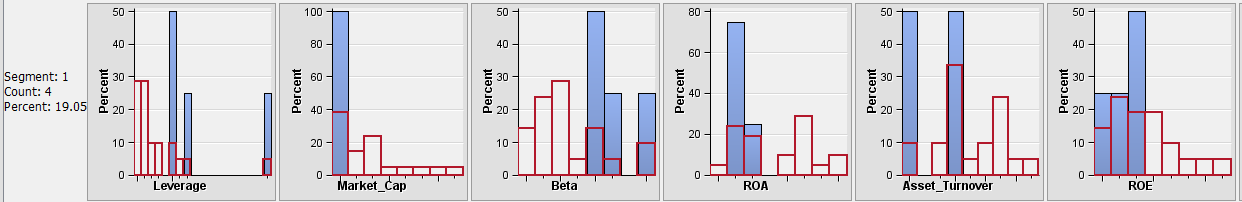
* From the mean statistics we can see frequency of clusters in each segment i.e. no of observations in each segment. We can check how populated each cluster is.
* It gives the mean value of each quantitative variable in each cluster.
* It also tells us details about the nearest cluster and the distance to the nearest cluster.

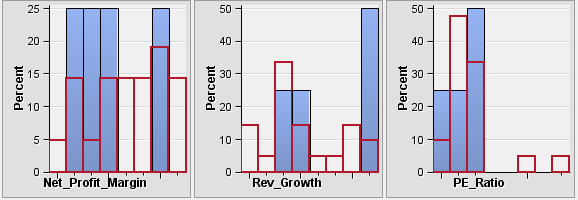
Using Segment Profile Node:

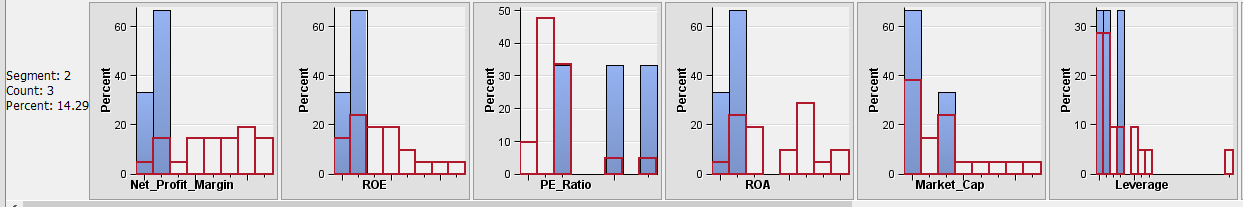


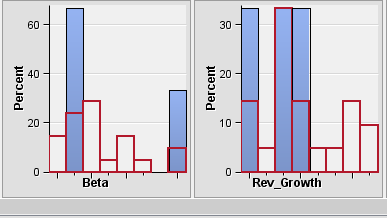


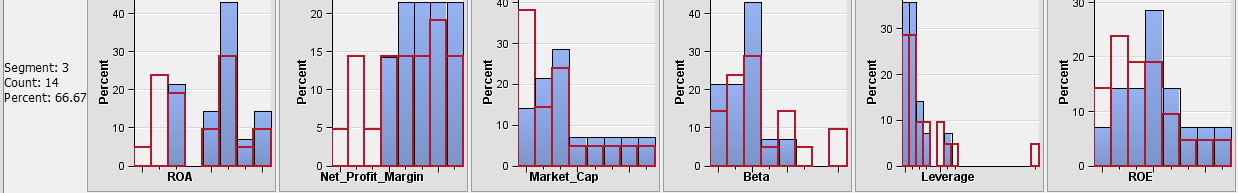


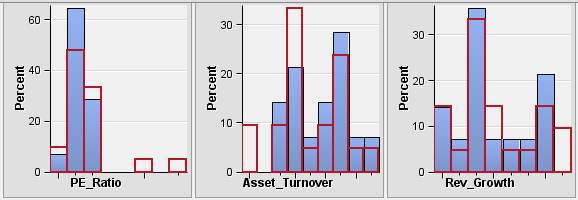












**3. Is there a pattern in the clusters with respect to the qualitative variables (10‐12) (those not used**

**in forming the clusters)?**

**Results:**

No we don’t see any specific pattern with respect to qualitative variables.

**4. Provide an appropriate name for each cluster using any or all of the variables in the dataset.**

**Don’t describe the cluster, name it.**

**Results:**

Cluster 1: High Leverage

Cluster 2: High Net\_Profit\_Margin

Cluster 3: High ROA

**5. Do the clusters formed seem reasonable? Try different numbers of clusters and examine the**

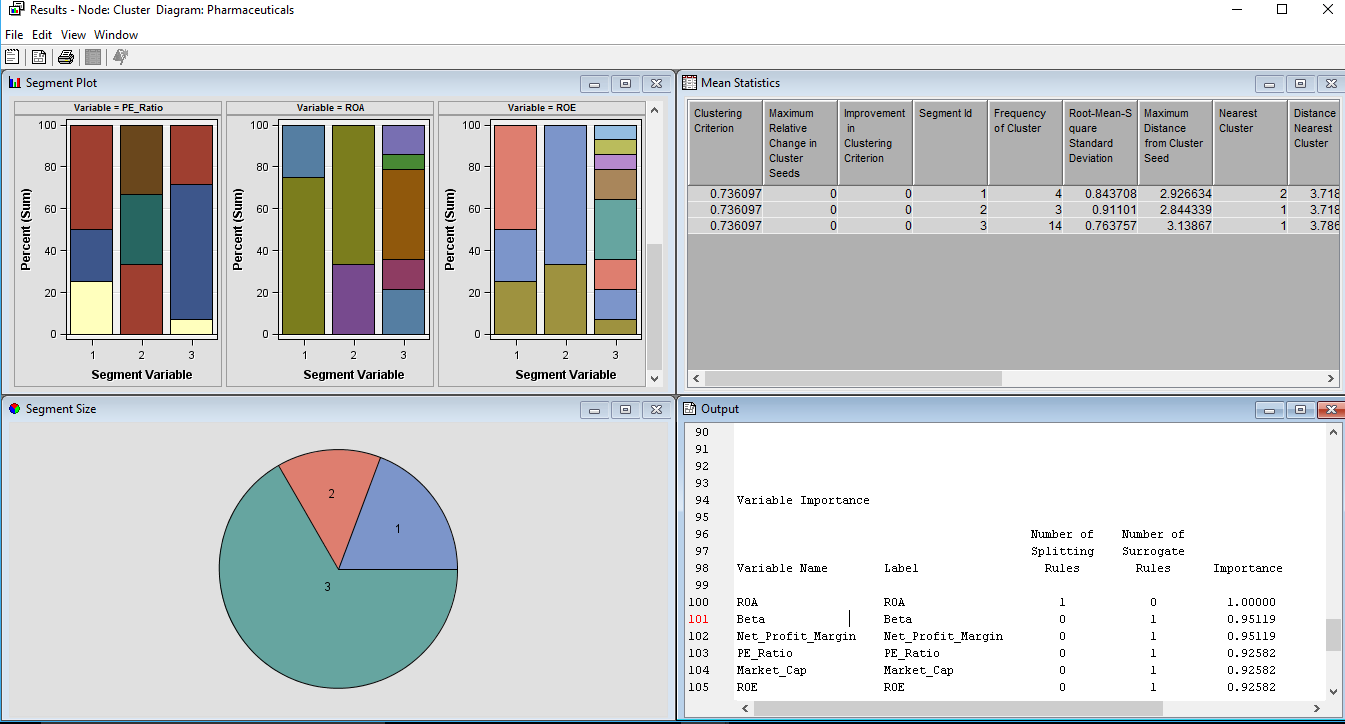
**results. Feel free to experiment with other criteria as needed. Explain the reasons for your**

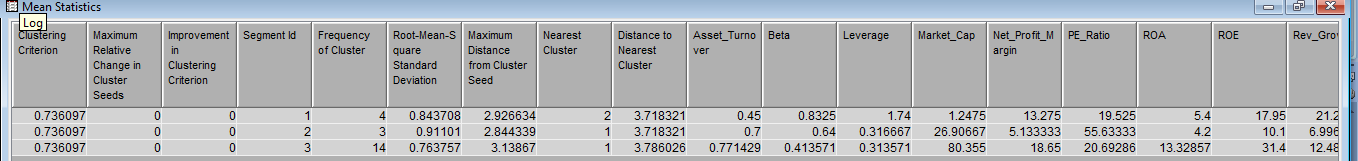
**selections, and identify the best clustering in your opinion (justify).**

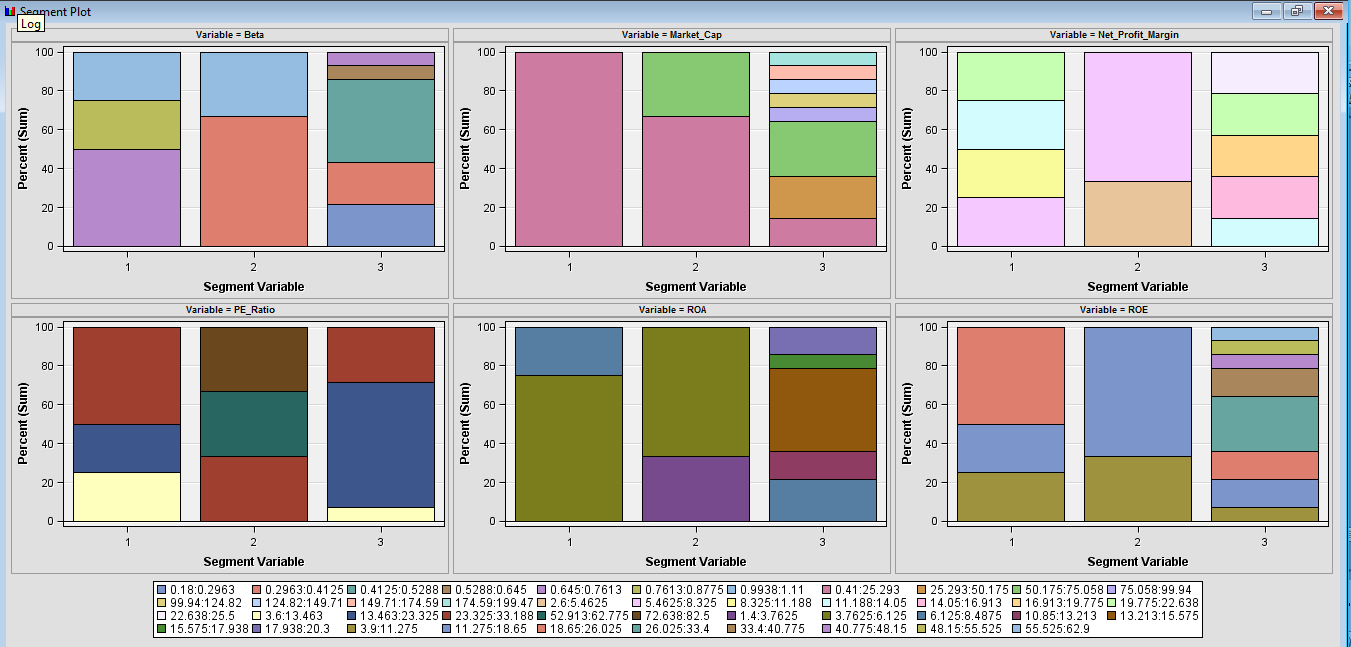
**Results:**

Number of clusters: 3 (Automatic specification in SAS enterprise miner)

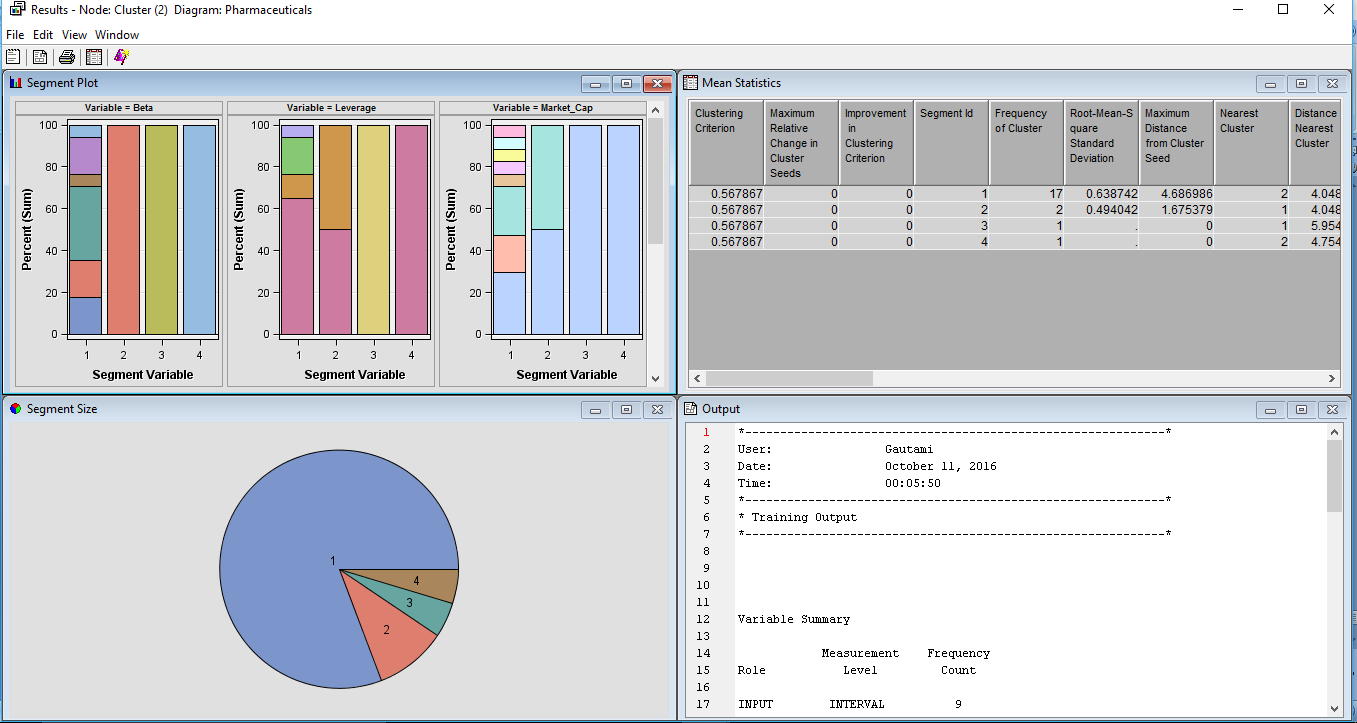
Yes, the number of clusters formed is reasonable.

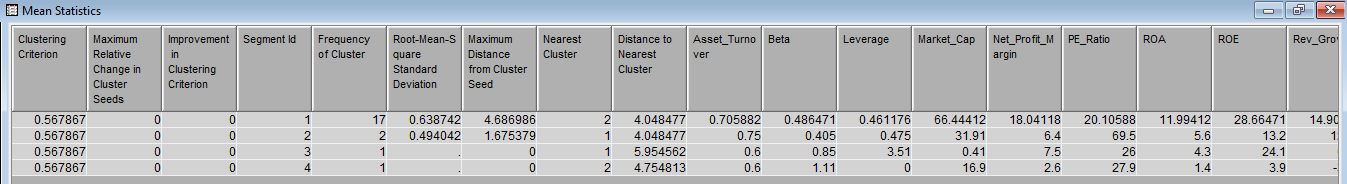


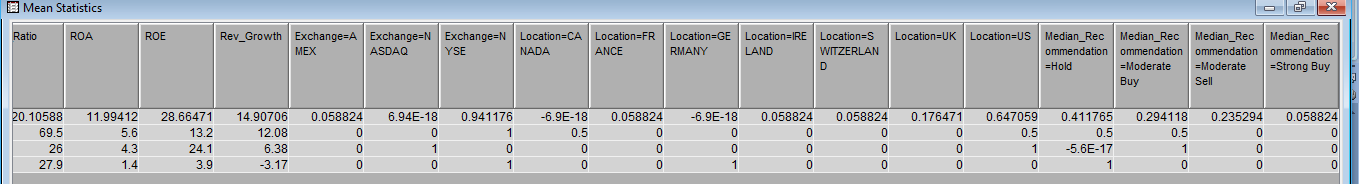
****

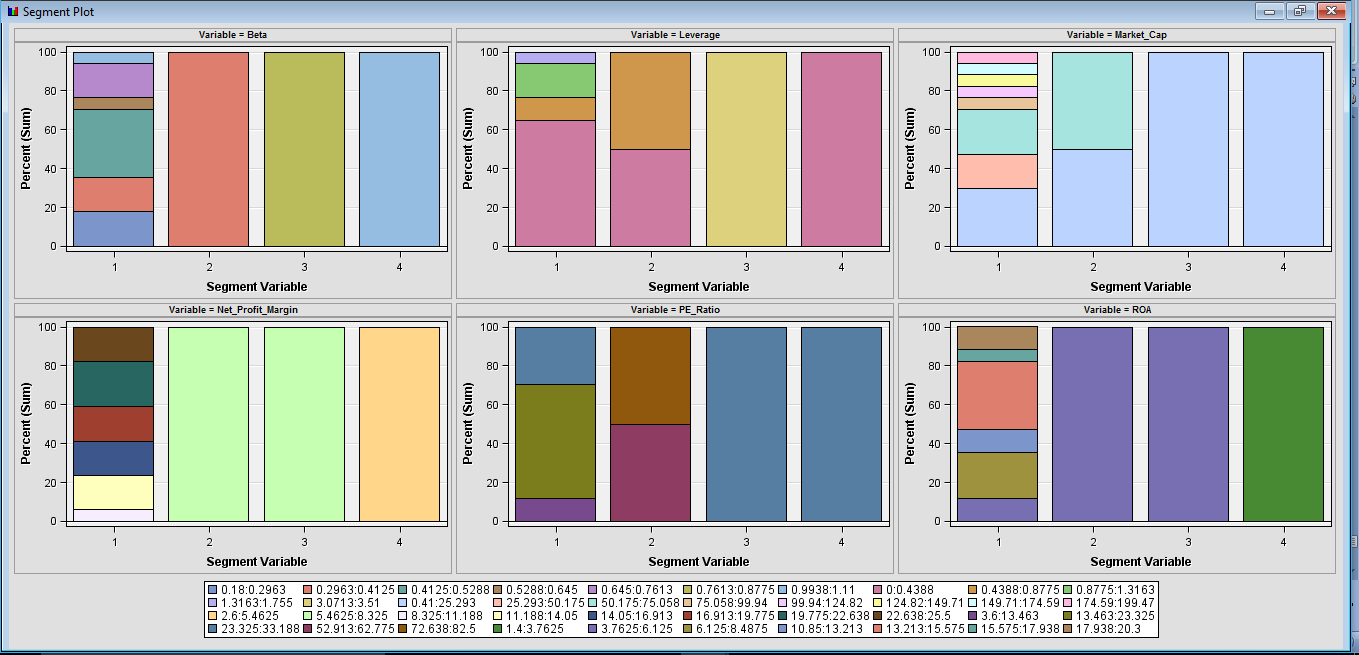
****

Number of clusters: 4(User Specified)

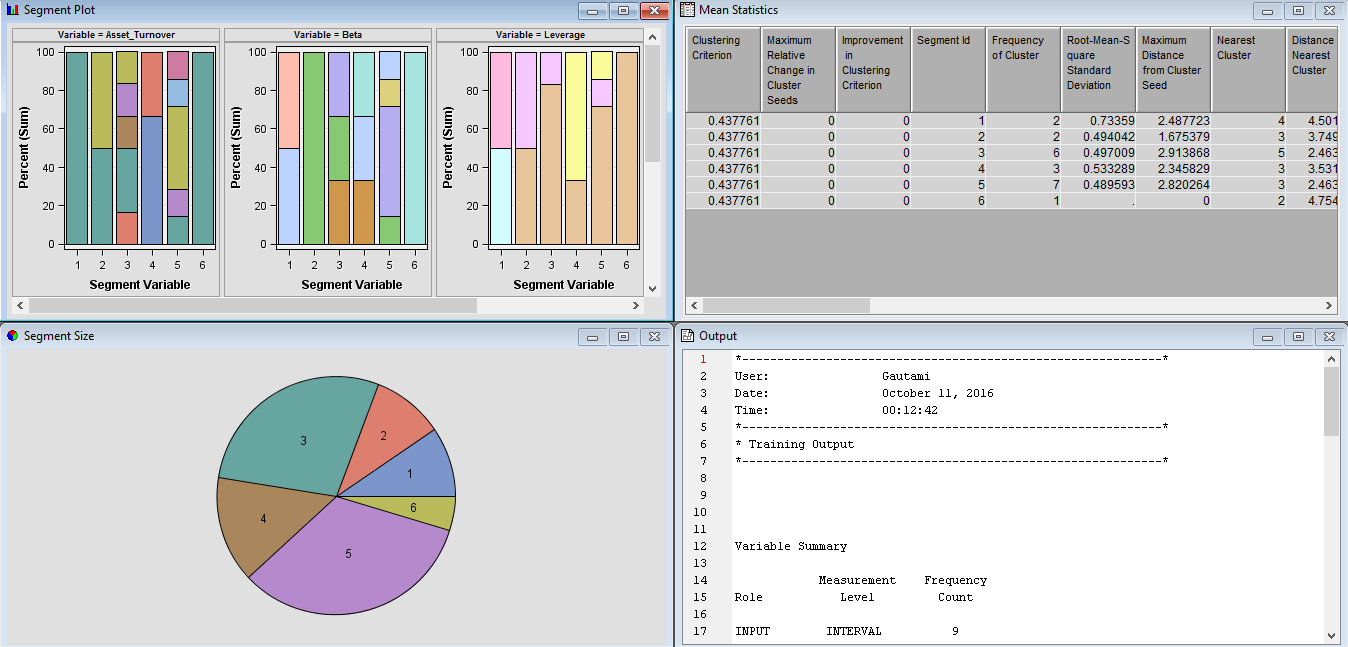
****

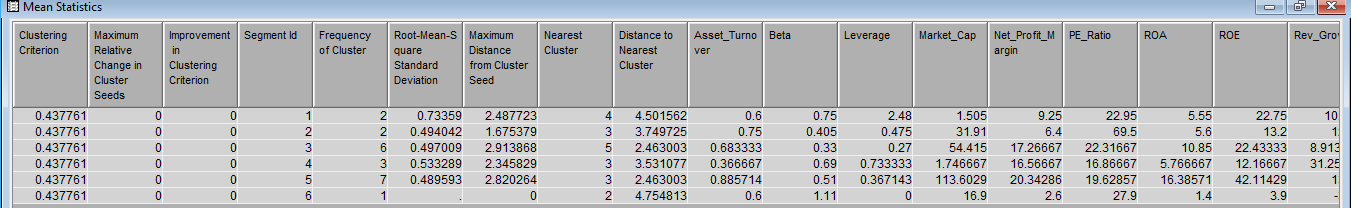
****

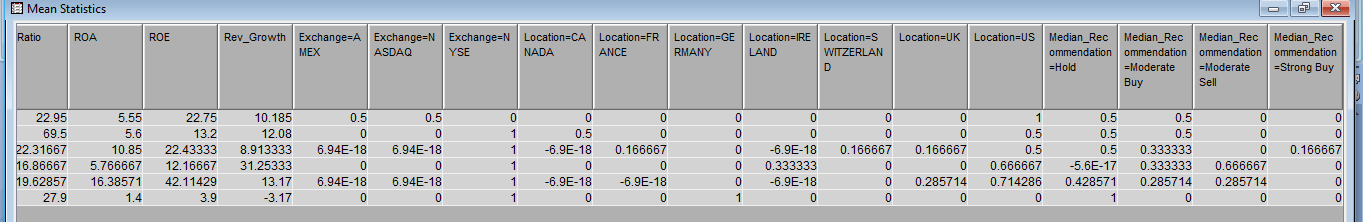
****

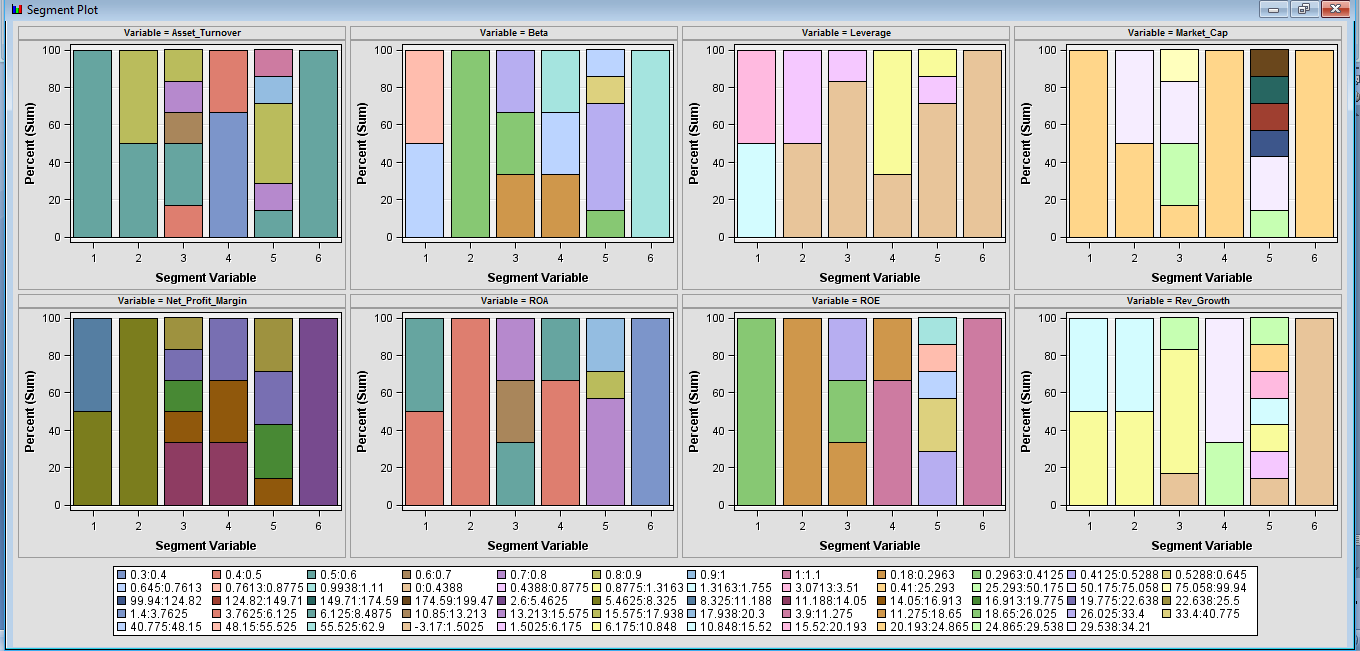
****

Number of clusters: 6(user specified)









From the above clusters formed we can see the following observation:

* When number of clusters formed are 4, we can observe we have only 1 observation in segment 3 and segment 4. Hence those segments are not populated and analysis of data is not appropriate.
* When number of clusters formed are 6, we can observe we have only 1 observation in segment 6. Hence that segment is not populated and analysis of data is not appropriate.

The frequency of clusters in each segment is less in both the cases.

Hence, according to our analysis, the 3 clusters formed by the automatic specification in SAS enterprise miner are the reasonable number clusters to analyze the Data.