Homework 3- Masters Group 9

Group Members:

A) Oluwaseyitan Awojobi
B) Mohamed Elgendy
C) Ramya Koya
D) Karthik Sagar Tadi
E) Vijay Tulluri

SOLUTION

6. a. Yes, as seen in figure 2 above, the variable “Stretch” is not normally distributed.

b. Using the Data Mining Database (DMDB) and Graph Explore feature, we are able to confirm there are no missing values.

Interval Variable Summary Statistics

Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
FASHION		0	689	1	204	92.26	32.616	0.44484	0.7202
LEISURE		0	689	650	2929	1916.42	350.840	-0.16735	0.6126
ORIGINAL		0	689	823	2715	1849.38	291.608	-0.20216	0.1898
SALESTOT	total pairs of dungarees sold	0	689	2140	4594	4302.11	366.213	-3.77180	15.6242
STOREID		0	689	1	689	345.00	199.041	0.00000	-1.2000
STRETCH		0	689	2	1224	444.04	211.690	0.35207	0.6029

7. **Why should the variable SALESTOT be rejected?**

Being the total number of jeans sold (the sum of FASHION, LEISURE, STRETCH, and ORIGINAL), we should set SALESTOT as Rejected because it creates redundant information which can be obtained from the other variables (Fashion, Leisure, Original and Stretch).

9. Select the Cluster node and select Internal Standardization Standardization. What would happen if you did not standardize your inputs?

Answer:

When Internal Standardization which is the default setting is not done, the variables values are divided by the standard deviation before the cluster is performed but not subtracted by the mean. This allows for a correct number of clusters to be derived and less misclassifications.

With Standardization:

Root-Mean-Square Total-Sample Standard Deviation 242.9643

Root-Mean-Square Distance Between Observations 687.2069

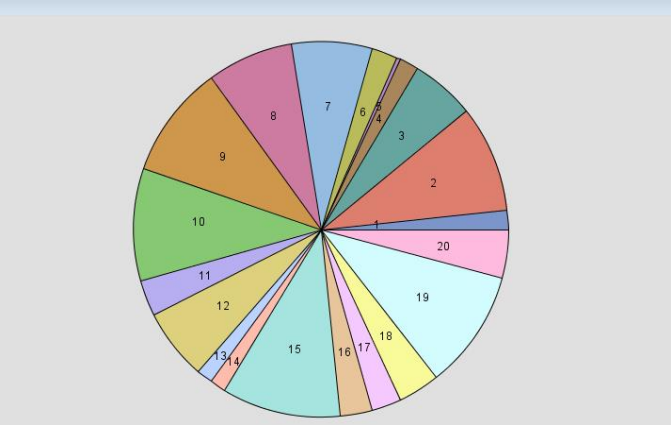


Figure 1-Without Standardization

Without Standardization:

Root-Mean-Square Total-Sample Standard Deviation 251.9923

Root-Mean-Square Distance Between Observations 712.7418

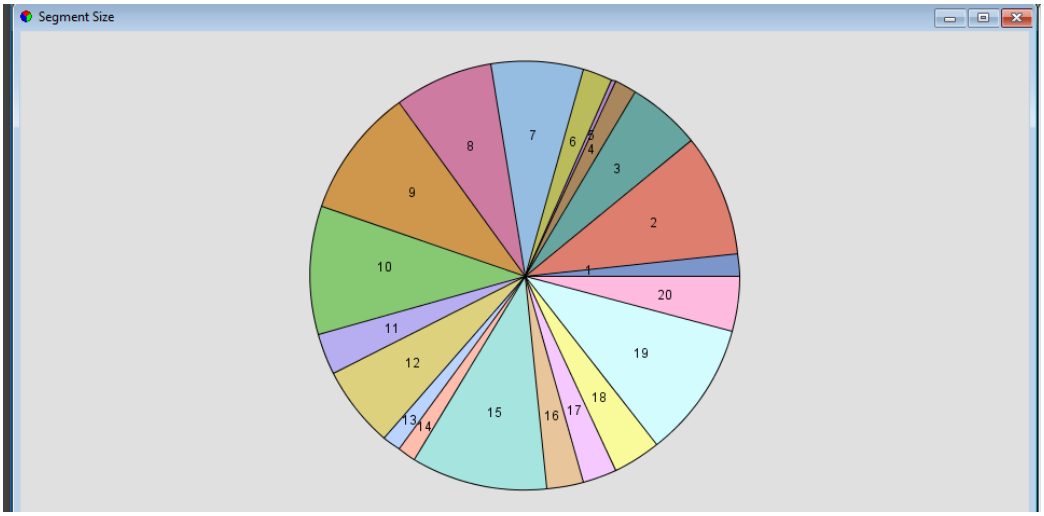


Figure 2- Standardized

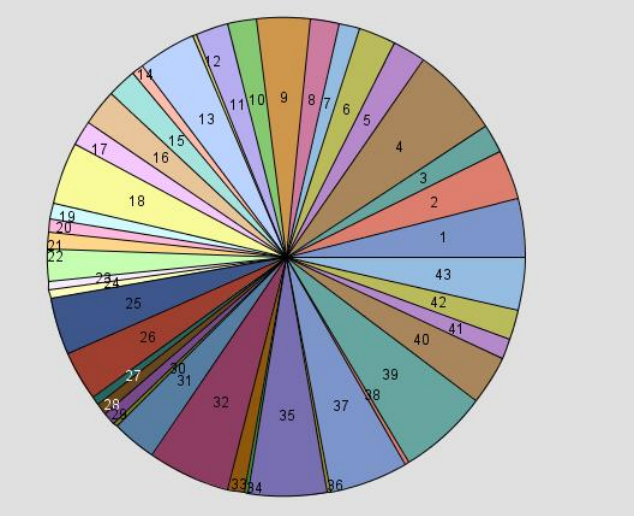


Figure 3- cluster output

10. **Does the number of clusters created seem reasonable?**

The number of clusters did not seem reasonable as it is difficult to read and interpret to its users.

11. **How does the number and quality of clusters compare to that previously obtained?**

Limiting the cluster to 6 makes it easily interpretable for the users.

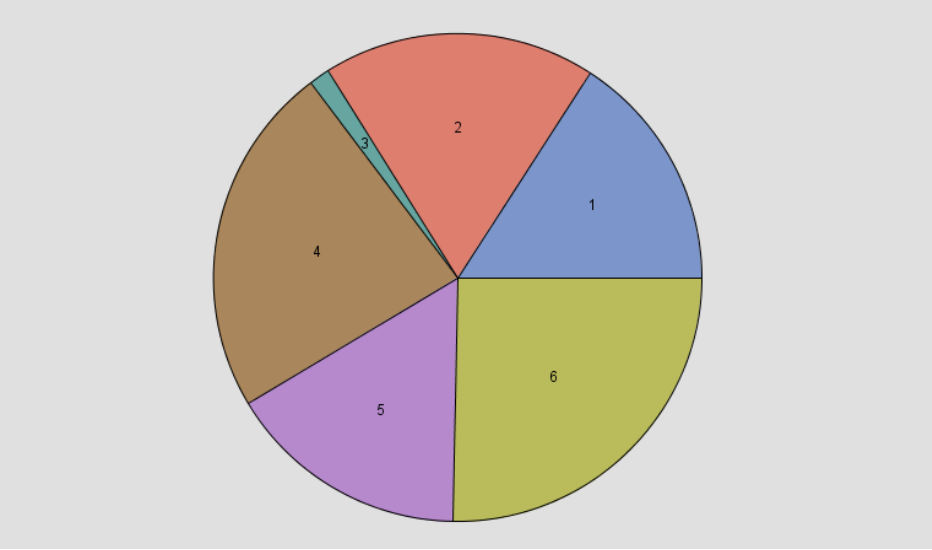
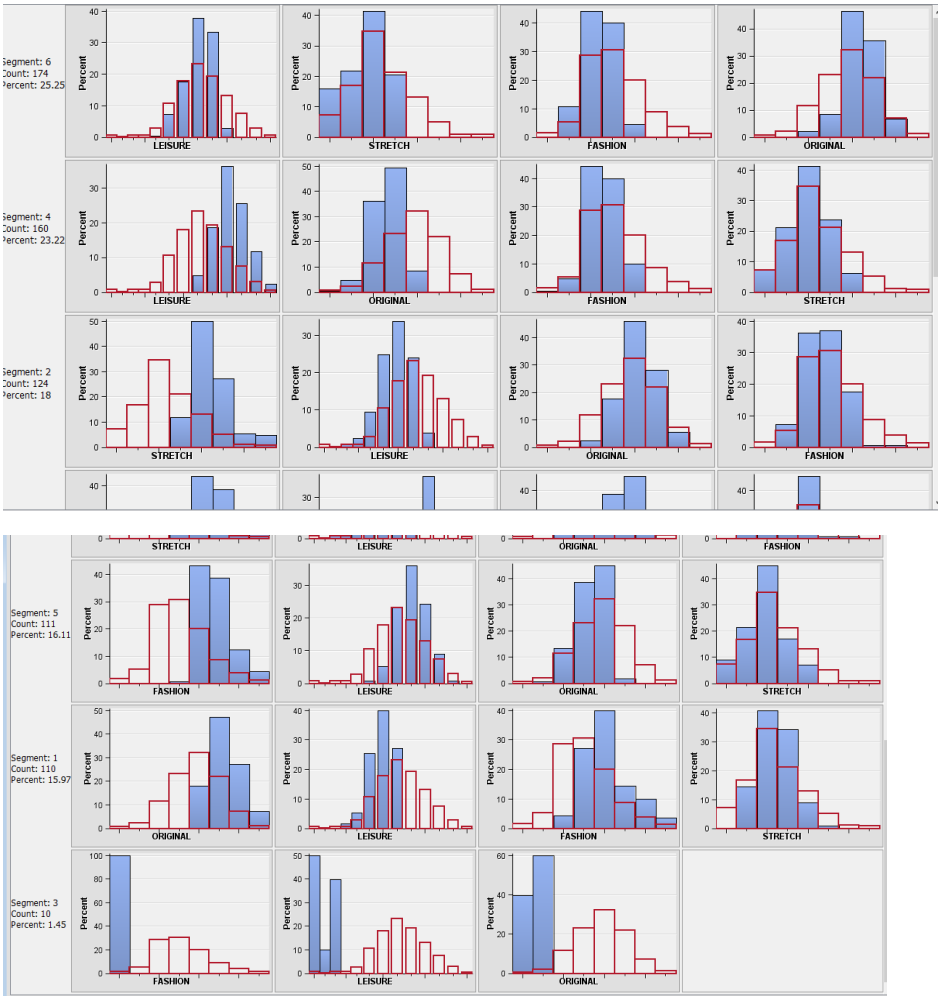


Figure 4- Readable Cluster



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

Segment 1: Stores in this segment sold mostly Original jeans as it is skewed to the right compared to the whole sample. Stores in this segment sold less Leisure jeans.

Segment 2: Stores in this segment sold mostly Stretch jeans followed by the Original Jeans as confirmed in the image above. In addition, stores in this segment sold less leisure jeans as it is skewed to the left.

Segment 3: Stores in this segment sold Leisure, Fashion and Original jeans, none of which generated high sales. They are all highly skewed to the left compared to the whole sample.

Segment 4: Stores in this segment sold more Leisure jeans as confirmed above as it is highly skewed to the right compared to the whole sample. Original, Fashion and Stretch Jeans are skewed to the left hence generated less sales.

Segment 5: Stores in this segment sold more Fashion jeans and Leisure jeans as they are both skewed to the right compared to the whole sample. Stores in this segment sold less original and Stretch jeans.

Segment 6: Stores in this segment sold more Original jeans, and less stretch and Fashion Jeans.

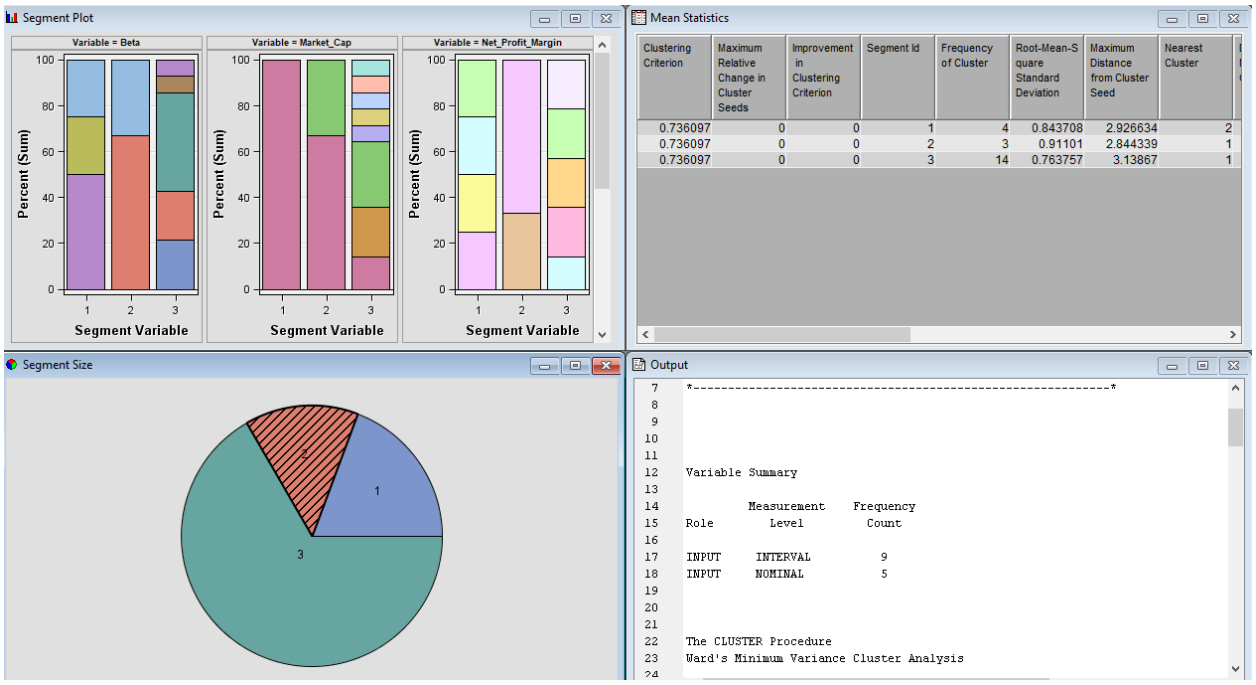
CASE 2:
Pharmaceuticals

1.

(none) not Equal to

Columns: Label Mining

Name	Use	Report	Role	Level
Asset_Turnover	Yes	No	Input	Interval
Beta	Yes	No	Input	Interval
Exchange	No	No	Input	Nominal
Leverage	Yes	No	Input	Interval
Location	No	No	Input	Nominal
Market_Cap	Yes	No	Input	Interval
Median_Recomm	No	No	Input	Nominal
Name	No	No	Input	Nominal
Net_Profit_Marg	Yes	No	Input	Interval
PE_Ratio	Yes	No	Input	Interval
ROA	Yes	No	Input	Interval
ROE	Yes	No	Input	Interval
Rev_Growth	Yes	No	Input	Interval
Symbol	No	No	Input	Nominal



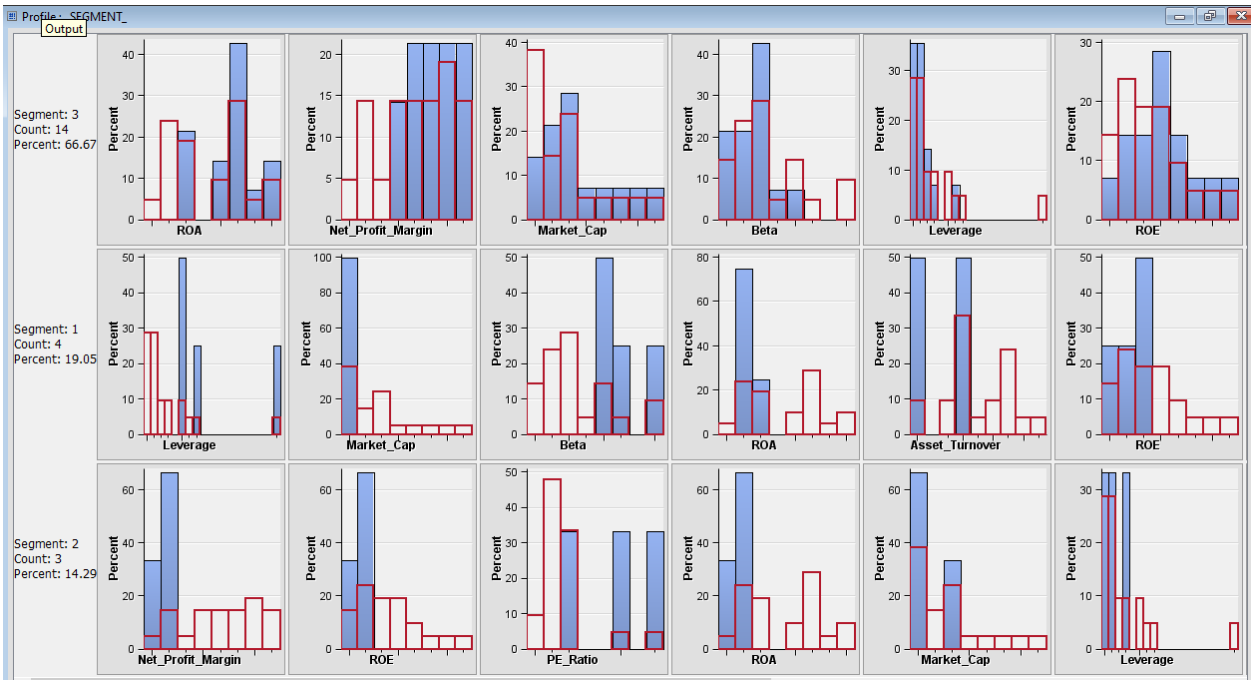
2. With the figure below, we can tell using the quantitative variables chosen that:

Clusters 1: This has the highest leverage, beta and estimated revenue growth relative to the whole sample, and it contains 4 pharmaceutical firms. Relative to the other clusters, the other variables are low

Cluster 2: This has the highest price/earnings ratio relative to the overall sample and contains 3 pharmaceutical firms. Relative to the overall sample, the other variables ae low. These firms appear to be growing pharmaceutical firms that are steadily securing their place in the market and are expected to be well-established in the future.

Cluster 3: has the highest asset turnover, market capitalization, return on assets, net profit margin and return on equity relative to the overall sample. It contains 14 pharmaceutical firms. Relative to the overall sample, It has low beta, leverage and price/earnings ratio These firms appear to be well-established pharmaceutical firms that have high capital, profit and market share.

Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Asset_Turnover	Beta	Leverage	Market_Cap	Net_Profit_Margin	PE_Ratio	ROA	ROE	Rev_Growth
1	4	0.843708	2.926634	2	3.718321	0.45	0.8325	1.74	1.2475	13.275	19.525	5.4	17.95	21.2375
2	3	0.91101	2.844339	1	3.718321	0.7	0.64	0.316667	26.90667	5.133333	55.63333	4.2	10.1	6.996667
3	14	0.763757	3.13867	1	3.786026	0.771429	0.413571	0.313571	80.355	18.65	20.69286	13.32857	31.4	12.48929



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

Yes, there is a pattern with the clusters with respect to the qualitative variables. The stock exchange didn't have any influence in the clustering.

Location

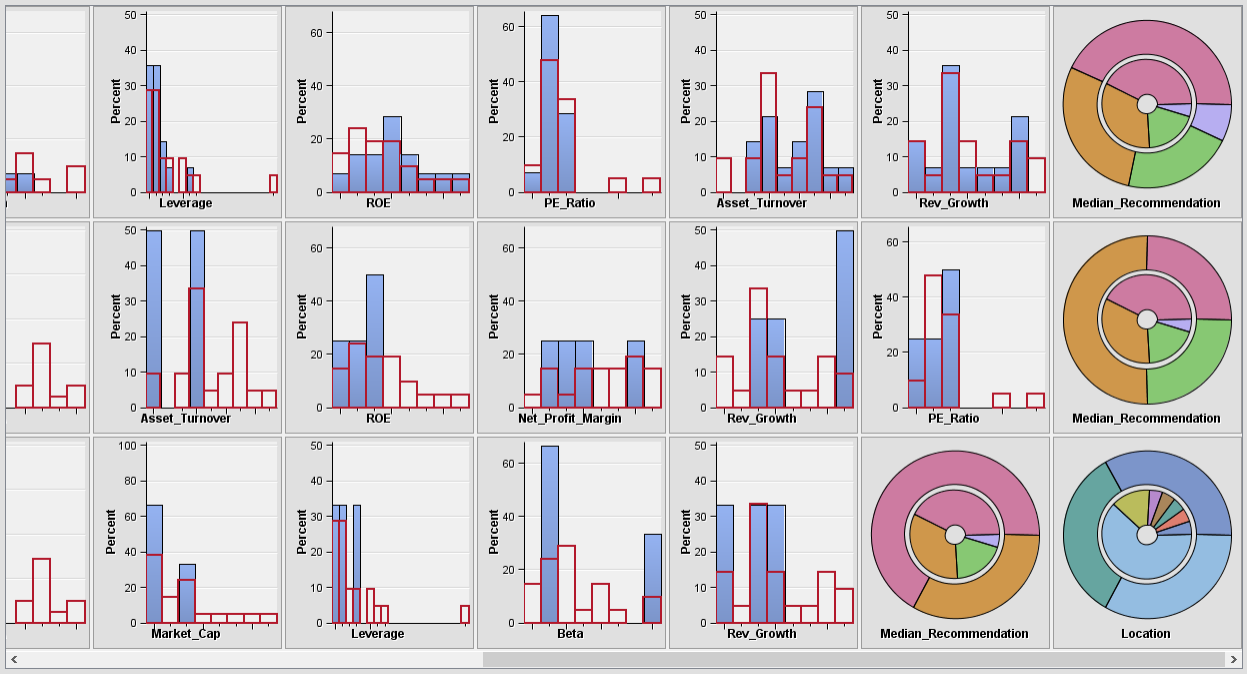
Cluster 2 firms are in Canada, U.S. and Germany. A high percentage of the pharmaceutical firms are located in the U.S. (61.9%), with UK at (14.3%) and other Countries at (4.8%).

Median Recommendation:

Cluster 1 firms' stocks are majorly moderate buy (50%) with moderate sell and hold having 25% each.

Cluster 2 firms' stocks are majorly hold (66.7%) and moderate buy at (33.3%).

Cluster 3 stocks are mostly "Hold" at (42.9%), moderate buy at (28.6%), moderate sell at 21.4% and strong buy is the least at 7.14%.



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.

Cluster 1- The Average

Cluster 2- The Minutest

Cluster 3 -The Prime

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).**

a. Yes, the clusters formed seem reasonable.

b. 5 clusters

Mean Statistics

	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Asset_Turnover	Beta	Leverage	Market_Cap	Net_Profit_Margin	PE_Ratio	ROA	ROE	Rev_Growth
0	0	1	1		0	3	4.106133	0.6	1.11	0	16.9	2.6	27.9	1.4	3.9	-3.17
0	0	2	2	0.558122	1.183956	5	4.210877	0.75	0.405	0.475	31.91	6.4	69.5	5.6	13.2	12.08
0	0	3	2	0.695678	1.475755	5	3.353698	0.6	0.75	2.48	1.505	9.25	22.95	5.55	22.75	10.185
0	0	4	12	0.71708	2.912824	5	3.621509	0.808333	0.435833	0.320833	89.54583	19.425	20.93333	14.35833	33.96667	9.905
0	0	5	4	0.688295	2.295848	3	3.353698	0.425	0.5975	0.635	13.1	15.65	17.675	6.2	14.575	30.1425

Figure 5- 5 clusters

10 Clusters

Mean Statistics

	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Asset_Turnover	Beta	Leverage	Market_Cap	Net_Profit_Margin	PE_Ratio	ROA	ROE	Rev_Growth
0	0	1	1	.	0	5	2.447177	1	0.35	0.34	122.11	21.1	18	20.3	62.9	21.87
0	0	2	2	0.558122	1.183956	7	3.262776	0.75	0.405	0.475	31.91	6.4	69.5	5.6	13.2	12.08
0	0	3	2	0.3056	0.648275	7	2.327313	0.55	0.28	0.27	25.21	14	19.25	7.15	16	27.995
0	0	4	8	0.556538	2.153929	7	2.558471	0.8	0.3925	0.245	87.31	18.7625	21.975	13.95	28.6625	8.005
0	0	5	1	.	0	1	2.447177	0.8	0.65	0.16	199.47	25.2	23.6	19.2	45.6	25.54
0	0	6	1	.	0	7	3.458389	0.6	1.11	0	16.9	2.6	27.9	1.4	3.9	-3.17
0	0	7	2	0.546016	1.158274	3	2.327313	0.75	0.555	0.86	4.45	11.1	20.3	7.3	18.15	10.52
0	0	8	1	.	0	7	3.787418	0.6	0.85	3.51	0.41	7.5	26	4.3	24.1	6.38
0	0	9	2	0.563263	1.194862	3	3.015849	0.3	0.915	1	0.99	17.3	16.1	5.25	13.15	32.29
0	0	10	1	.	0	4	2.876038	0.6	0.63	1.12	48.19	25.5	13.1	13.4	54.9	0.36

Figure 6-10 Clusters

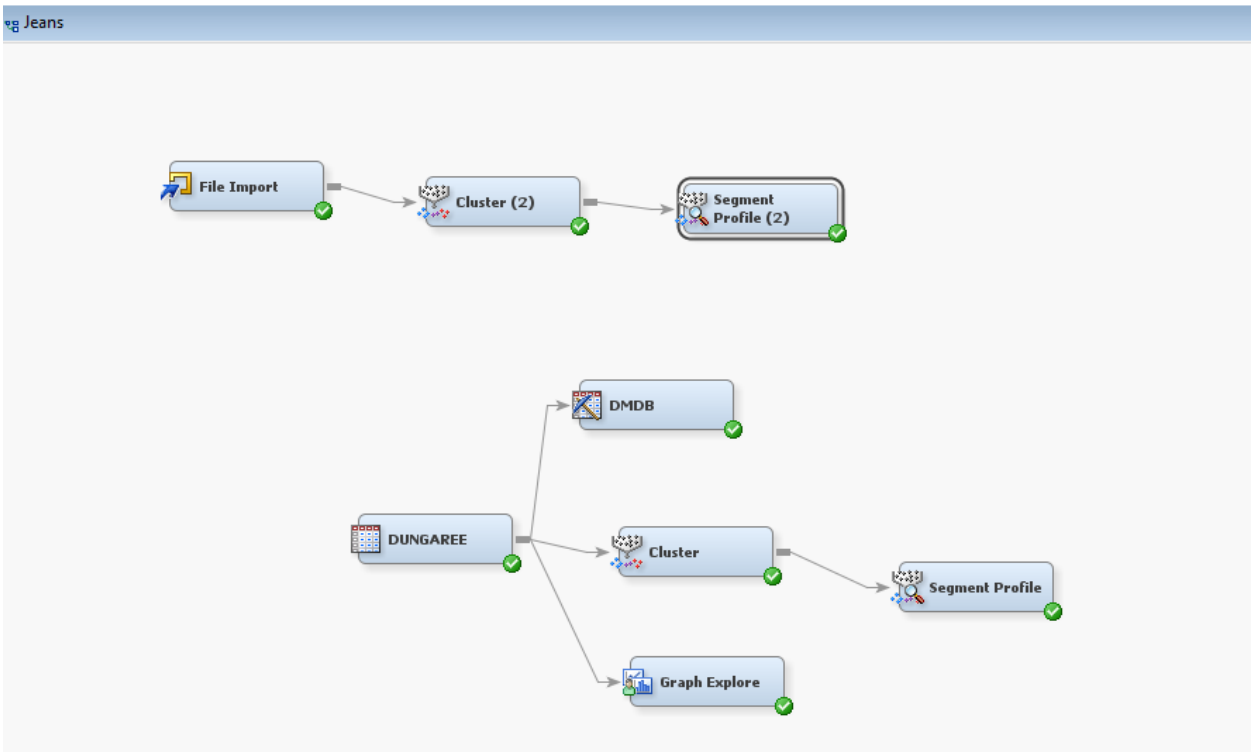
2 Clusters

Mean Statistics

	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Asset_Turnover	Beta	Leverage	Market_Cap	Net_Profit_Margin	PE_Ratio	ROA	ROE	Rev_Growth
0	0	1	18	0.929064	4.648845	2	3.582239	0.7	0.506667	0.630556	62.77556	17.45556	20.43333	11.56667	28.41111	14.43333
0	0	2	3	0.91101	2.844339	1	3.582239	0.7	0.64	0.316667	26.90667	5.133333	55.63333	4.2	10.1	6.996667

Figure 7- 2 clusters

c. The purpose of using 2, 5 and 10 clusters is to compare how much different they are from the default cluster choice. Choosing 2 clusters would be ideal as Cluster 2 in the original cluster accounted for Location; However, choosing 2 clusters, this qualitative variable does not appear to affect the clusters. Also, the distance within cluster sum of squared error is significant in 2 clusters hence we chose 2 clusters to be the best option. For Cluster 2 also, there is good classification of the clusters.


Jeans
File Import
Cluster (2)
Segment Profile (2)
DMDB
DUNGAREE
Cluster
Segment Profile
Graph Explore