SAS Assignment 4 – Clustering

Clustering Stores & Pharmaceutical Firms

JOO YONG YOON

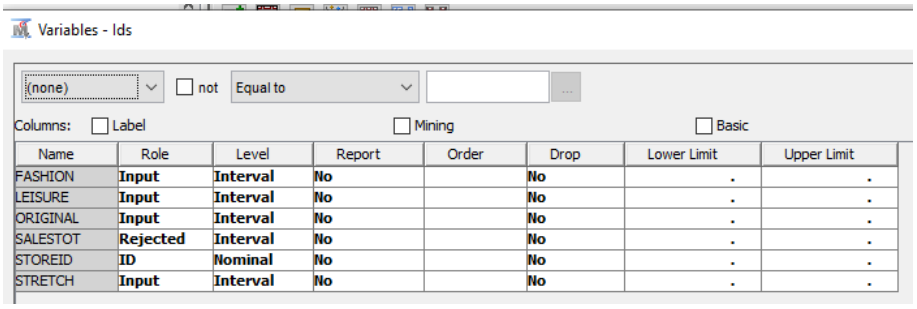
MI 353

Dr. CEVIKPARMAK

23 April 2023

Exercise 1: Clustering Stores

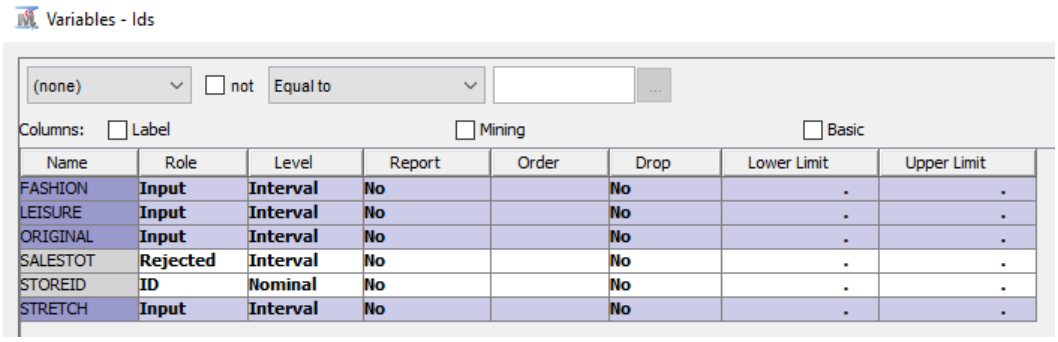
***** Please refer to the figure(s) provided for the question.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
FASHION	Input	Interval	No		No	.	.
LEISURE	Input	Interval	No		No	.	.
ORIGINAL	Input	Interval	No		No	.	.
SALESTOT	Rejected	Interval	No		No	.	.
STOREID	ID	Nominal	No		No	.	.
STRETCH	Input	Interval	No		No	.	.

Figure 1. Variable Settings

6. Explore the data. Determine whether the model roles and measurement levels assigned to the variables are appropriate. Examine the distribution of the variables.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
FASHION	Input	Interval	No		No	.	.
LEISURE	Input	Interval	No		No	.	.
ORIGINAL	Input	Interval	No		No	.	.
SALESTOT	Rejected	Interval	No		No	.	.
STOREID	ID	Nominal	No		No	.	.
STRETCH	Input	Interval	No		No	.	.

Figure 2. Explore Data Input (I)

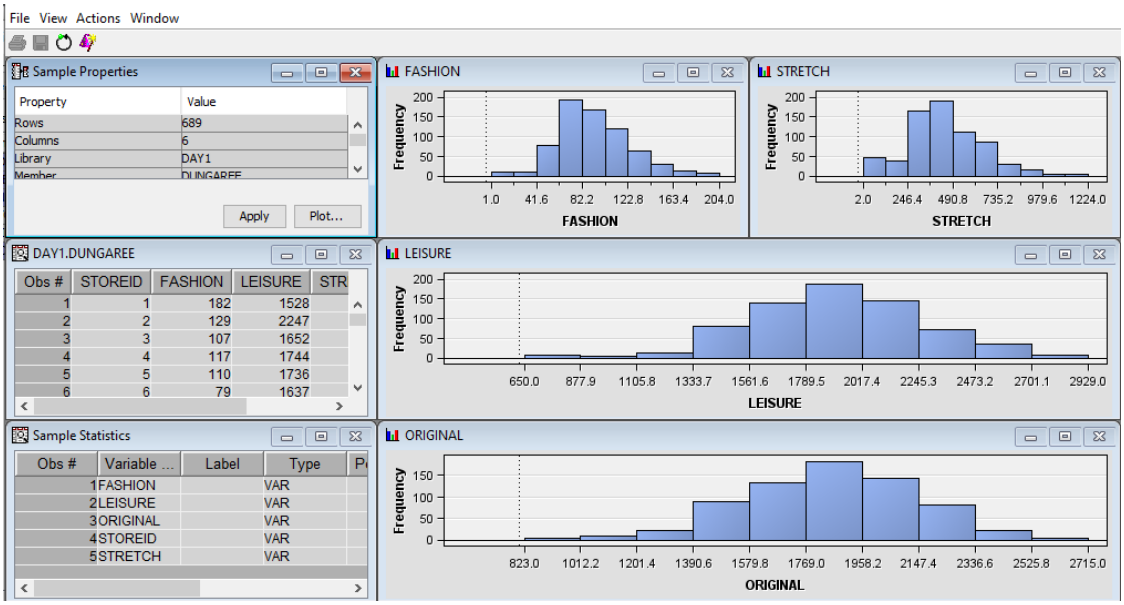


Figure 3. Explore Data Input (II)

a. Are there any unusual data values?

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

Responses for both a & b: To examine the variable distribution, I selected the relative variables only and chose explore option. It was not necessary to explore the ID and Rejected value, yet I figured SALESTOT and STOREID are the only distribution that do not have a normal or slightly skewed distribution. Their histograms do not follow any particular patterns. Other than that it seems there is no presence of any unusual data and all values seem to exist, hence there are no missing values that can be found.

7. 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?

Please refer to Figure 1. SALESTOT is to be rejected because it shouldn't be considered as a value independent to others. The SALESTOT is the sum of the all the other variables that are input into the data. In specific, it is highly likely that there are different levels of sales for each type of style. This explains why the histogram does not follow a normal distribution because each style has a different number of sales associated with it.

8. Add a Cluster node to the diagram workspace and connect it to the data source node.

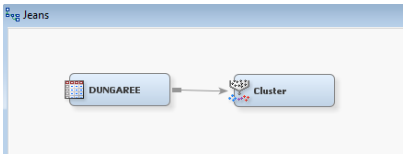


Figure 4. Cluster Node

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

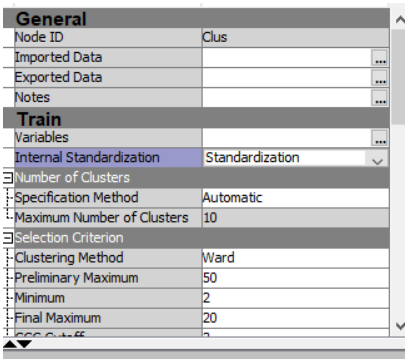


Figure 5. Internal Standardization Setting

It is important to chose standardization option because otherwise the clustering will take place with respect to ranges solely, which means the clustering will occur strictly on the inputs with the largest range.

10. Run the diagram from the Cluster node and examine the results. Does the number of clusters created seem reasonable?

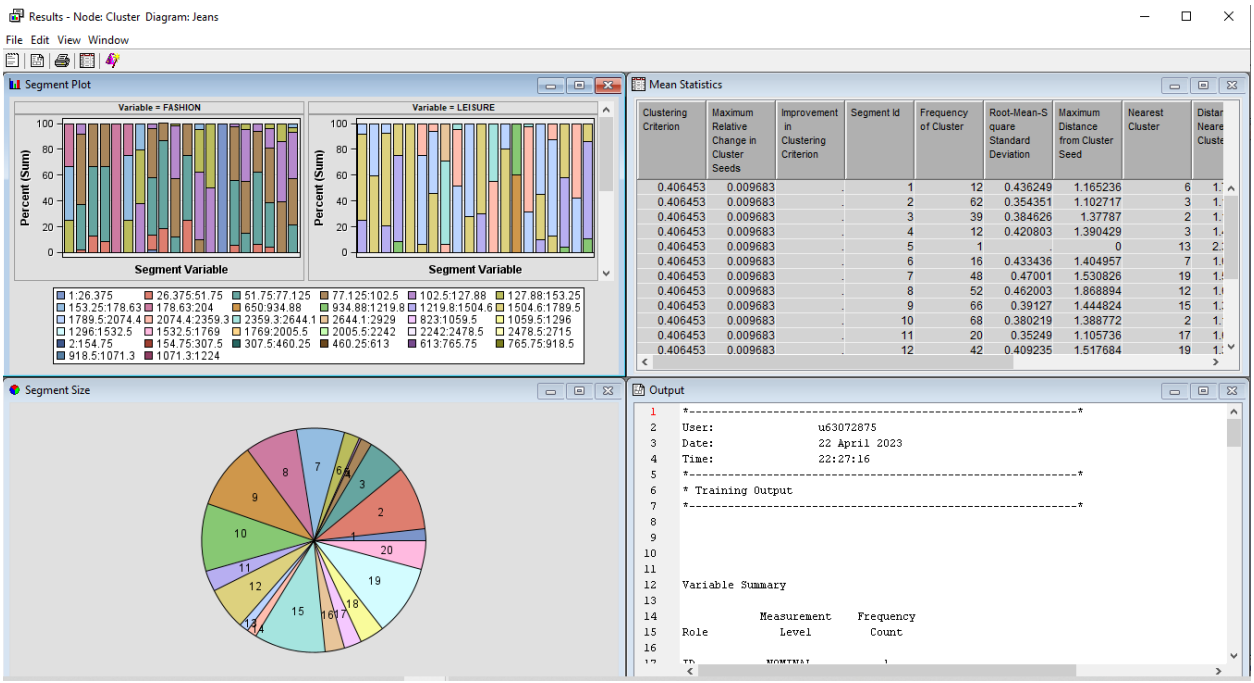


Figure 6. Clustering Results (Auto)

There are too many clusters created and hence a solution to fix this problem would be to self assign the number of clusters to create. The Cluster node's automatic number of cluster specification method seems to generate an excessive number of clusters. The problem with having this many clusters is that it makes data harder to read and gives a greater chance that some points are grouped in the wrong clusters. By having fewer clusters, it provides a lesser chance of error for grouping one point in the wrong cluster.

11. Specify a maximum of six clusters and rerun the Cluster node. How does the number and quality of clusters compare to that previously obtained?

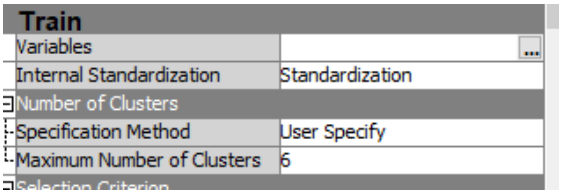


Figure 7. Max Number of Clusters Setting

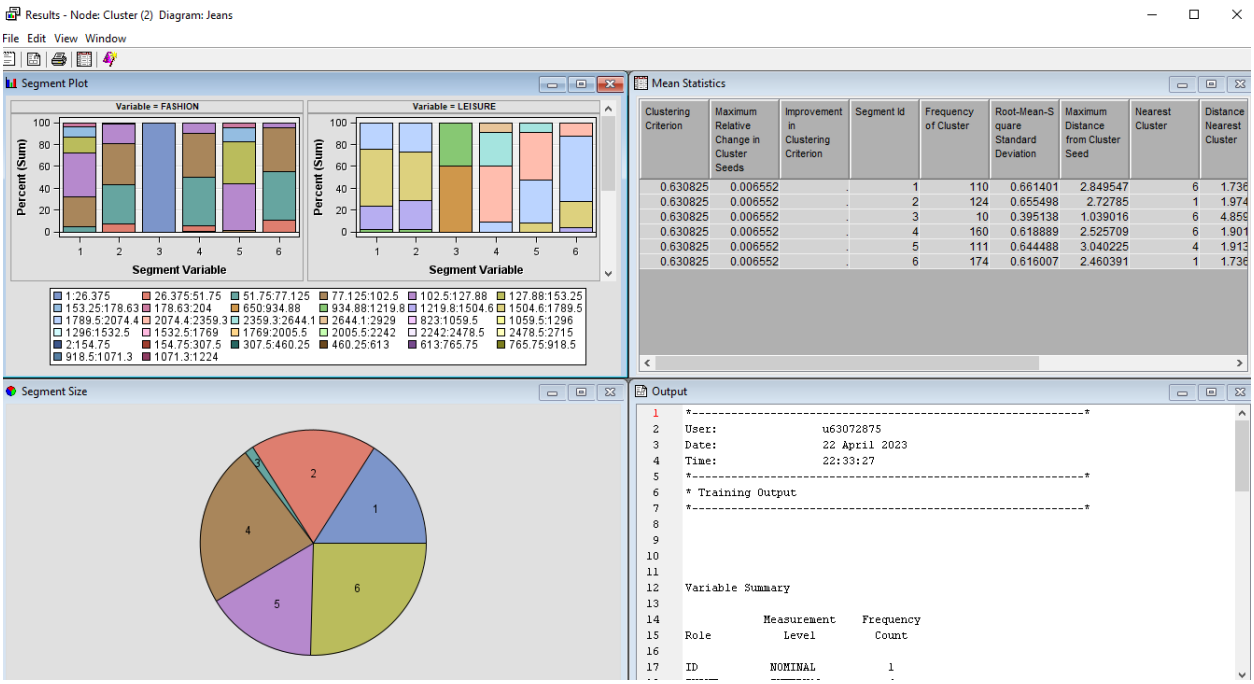


Figure 8. Clustering Results (User Specify)

This update really brings out a good result as the clusters seem much more reasonable and all are sized well. There is one anomaly in the 3rd cluster which seems to not have a lot of population but otherwise the quality and quantity are both acceptable.

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

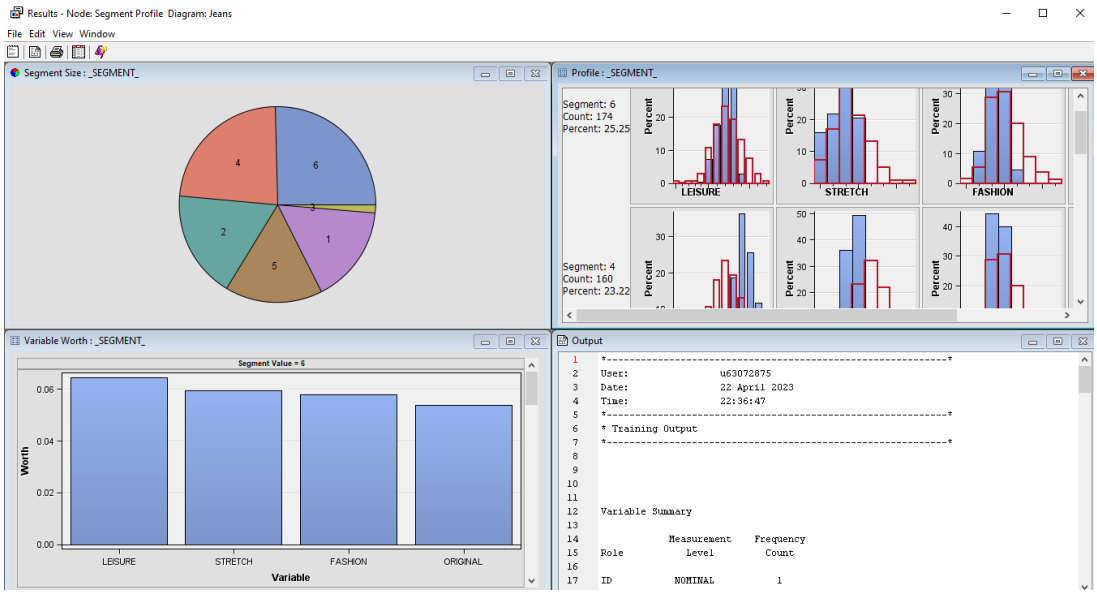


Figure 9. Segment Profile Results

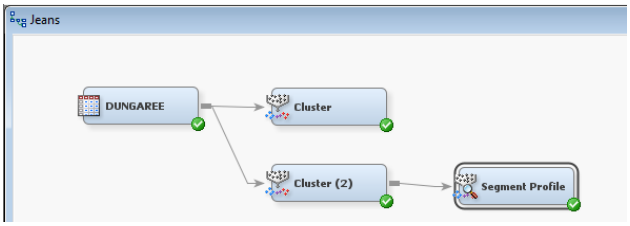


Figure 10. Diagram

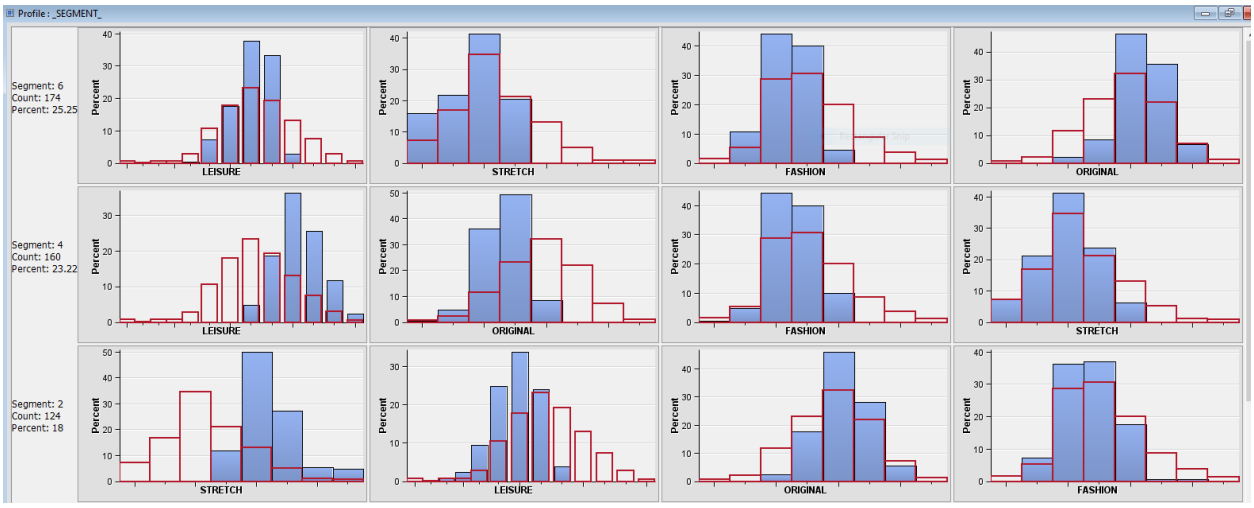


Figure 11. Segments 6,4,2

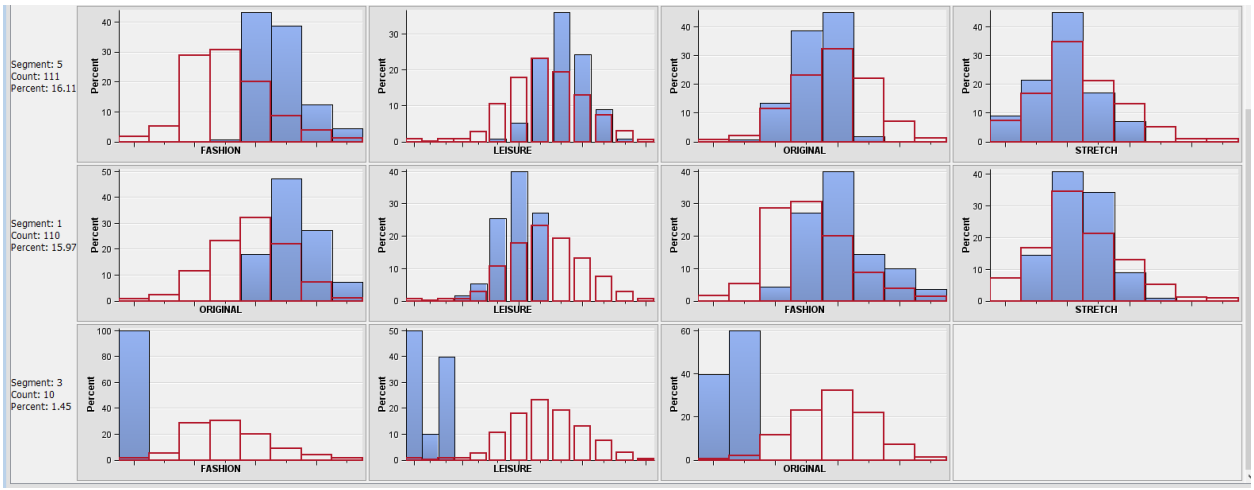


Figure 12. Segments 5,1,3

Observations:

- Segment 1: It can be observed from this segment that it is more prone to selling original jeans than any other kind as can be supported from its skewness.
- Segment 2: It can be observed from this segment that it is more prone to selling stretch jeans than any other kind as can be supported from its skewness. It contains stores selling a higher-than-average number of stretch jeans.

- Segment 3: It is observed from this segment that this store sells the lowest volume of jeans overall.
- Segment 4: It can be observed from this segment that the store is more prone to selling leisure compared to other types.
- Segment 5: It can be observed from this segment that the store is more prone to selling fashion jeans than any other type.
- Segment 6: This store is selling a greater number of jeans in the original category with a very low number of fashion and stretch jeans compared to average.

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.

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

Figure 13. Variable Selection

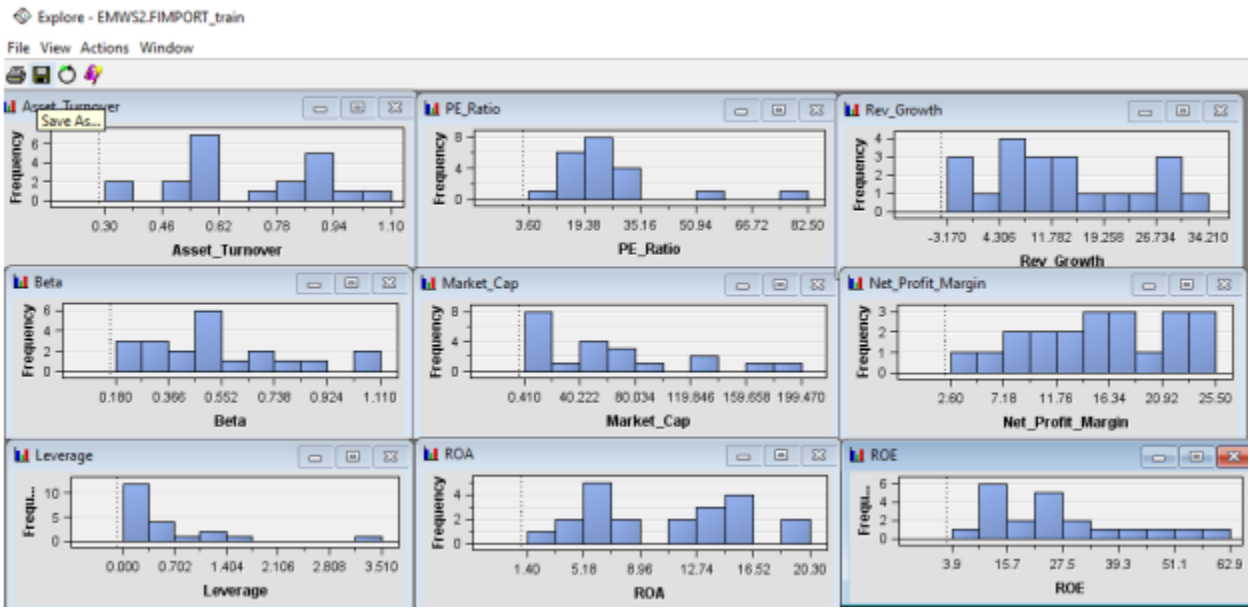


Figure 14. Explore Selected Variables

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

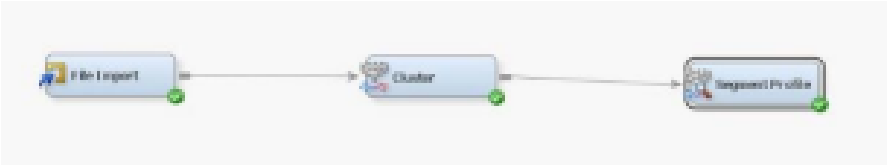


Figure 15. Diagram

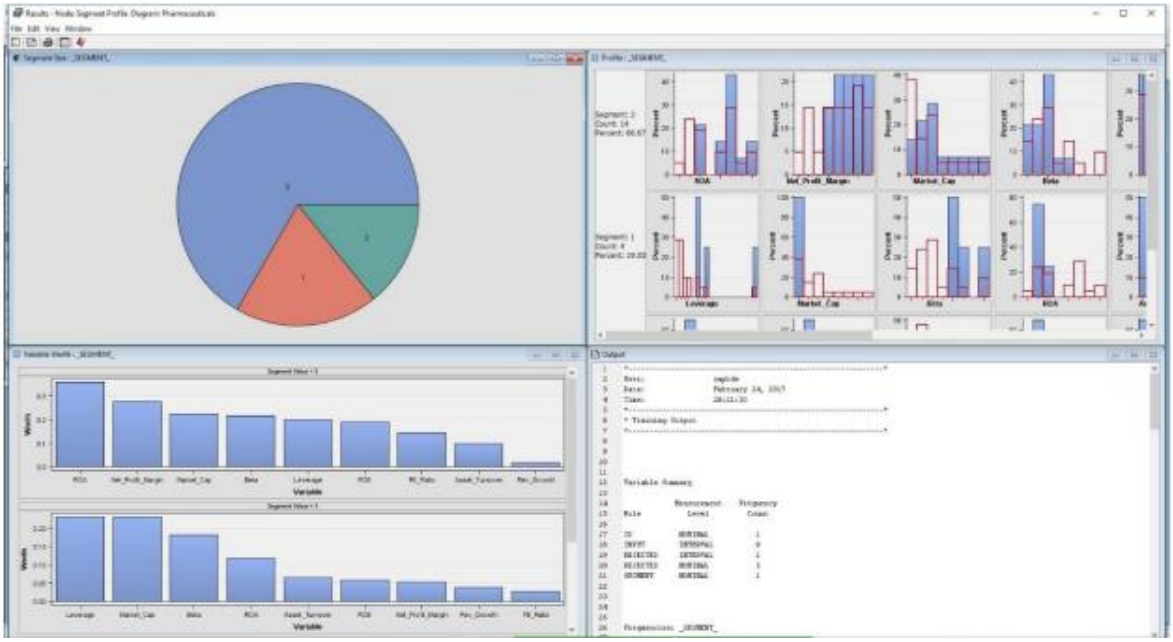


Figure 16. Clustering Result (I)

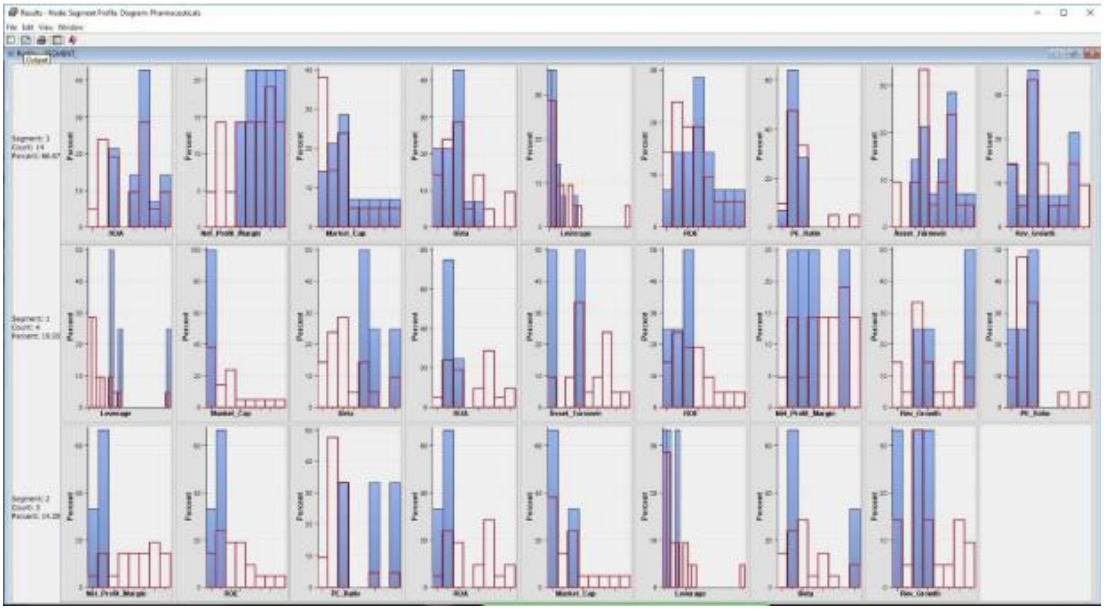


Figure 17. Clustering Result (II)

Total of 3 clusters are generated. The analysis of the clusters considering quantitative variables is as follows:

Segment 3:

- ROA:
    - Return on Assets contributes is maximum with a worth of 0.3612
    - It is slightly higher than the entire sample set
    - The mean is 13.285.
- Net_Profit_Margin:
    - Net Profit Margin has a worth of 0.2778
    - It is higher than the entire sample. The distribution of the data points is also denser compared to sample set.
    - The mean is 18.65.
- Market_Cap:
    - Market Capitalization has a worth of 0.2254
    - It is pretty similar compared to sample set. The frequency of the higher Market
    - Capitalization values are somewhat higher.
    - The mean is 80.355.
- Beta:
    - Beta has a worth of 0.2169
    - It is lower when compared to sample set.
    - The mean is 0.4136.
- Leverage:
    - Leverage has a worth of 0.2001
    - The frequency of the lower Leverage values are somewhat higher.
    - The mean is 0.3136.
- ROE:
    - Return on Equity has a worth of 0.1897
    - It is higher compared to sample set. The frequency of the higher ROA values is somewhat higher compared to sample.
    - The mean is 31.4.
- PE_Ratio:
    - Price/earnings Ratio has a worth of 0.1451
    - The frequency of the lower values is somewhat high
    - The mean is 20.69.
- Asset_Turnover:
    - Asset Turnover has a worth of 0.0989
    - It is somewhat higher compared to overall sample set. The frequency of the higher
    - values is slightly higher.
    - The mean is 0.771.
- Rev_Growth:
    - Estimate Revenue Growth has a worth of 0.0187.

    - The spread of the data is analogous to the overall sample set.
    - The mean is 12.49.

Overall, Segment 3 has a 66.67% share of the entire distribution, Maximum worth is held by Return on Assets followed closely by NPM and Market Capitalization. The Projected Revenue Growth for the Segment is low and holds minimum worth.

Segment 1:

- Leverage:
    - It has a worth of 0.2322
    - It is higher than the overall sample set
    - The mean is 0.174
- Market_cap:
    - Market Capitalization has a worth of 0.2322
    - It is lower than the overall sample set
    - The mean is 1.248
- Beta:
    - It has a worth of 0.1814
    - It is higher than the overall sample set.
    - The mean is 0.8325
- ROA:
    - Return on Assets has a worth of 0.1179
    - It is lower than the overall sample set.
    - The mean is 5.4.
- Asset_Turnover
    - Asset Turnover has a worth of 0.066
    - It is lower than the overall sample set and concentrated on the lower end.
    - The mean is 0.45
- ROE:
    - Return on Equity has a worth of 0.0581
    - It is lower than the overall sample set.
    - The mean is 17.95
- Net_Profit_Margin:
    - Net Profit Margin has a worth of 0.0536
    - It is lower than the overall sample. The frequency of the higher lower values and higher values is more and central values are lesser than the overall sample set
    - The mean is 13.28
- Rev_Growth:
    - Estimated Revenue Growth has a worth of 0.0385
    - The frequency of the higher values is higher and the frequency of the moderately low values are also greater.

    - The mean is 21.24
- PE_Ratio:
    - Price/earnings ratio has a worth of 0.02951
    - The frequency of the lower values is higher. The Price/earnings ratio for is lower than the overall sample set.
    - The mean is 19.525

Overall, Segment 1 has a 19.05% share of the total distribution which is significantly lower than Segment 3. The maximum worth is shared equally by Leverage and Market_cap (0.2322) followed closely by Beta. The Price/earnings ratio holds the least importance with a worth of only 0.02951

Segment 2:

- Net_Profit_Margin:
    - Net Profit Margin has a worth of 0.1306
    - It is higher than the overall sample set but is not consistent throughout the graph
    - The mean is 5.133
- ROE:
    - Return on Equity has a worth of 0.1306
    - It is higher than the overall sample, however it is concentrated on the lower end.
    - The mean is 10.1
- PE_Ratio:
    - Price/earnings ratio has a worth of 0.1020
    - The frequency of the higher values is more
    - The mean is 55.63
- ROA:
    - Return on Assets has a worth of 0.1020
    - It is lower than the overall sample set.
    - The mean is 4.2
- Market_Cap:
    - Market Capitalization has a worth of 0.0816
    - It is lower than the overall sample set
    - The mean is 26.91
- Leverage:
    - Leverage has a worth of 0.0544
    - It is lower than the overall sample set
    - The Leverage mean is 0.3167
- Beta:
    - Beta has a worth of 0.0449
    - The frequency of the higher lower values and higher values is greater, central values are lesser than the overall sample set

    - The mean is 0.64
- Rev_Growth:
    - Estimated Revenue Growth has a worth of 0.0206
    - It is lower than the overall sample set
    - The mean is 6.99
- Asset Turnover : This does not contribute at all to this segment

Overall, Segment 2 has a 14.29% share of the total distribution. The maximum worth is shared equally by Net_Profit_Margin and ROE (0.1306) followed by PE_Ratio and ROA sharing an importance of 0.1020. The Asset Turnover does not contribute to the segment and hence does not contribute to the overall distribution.

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

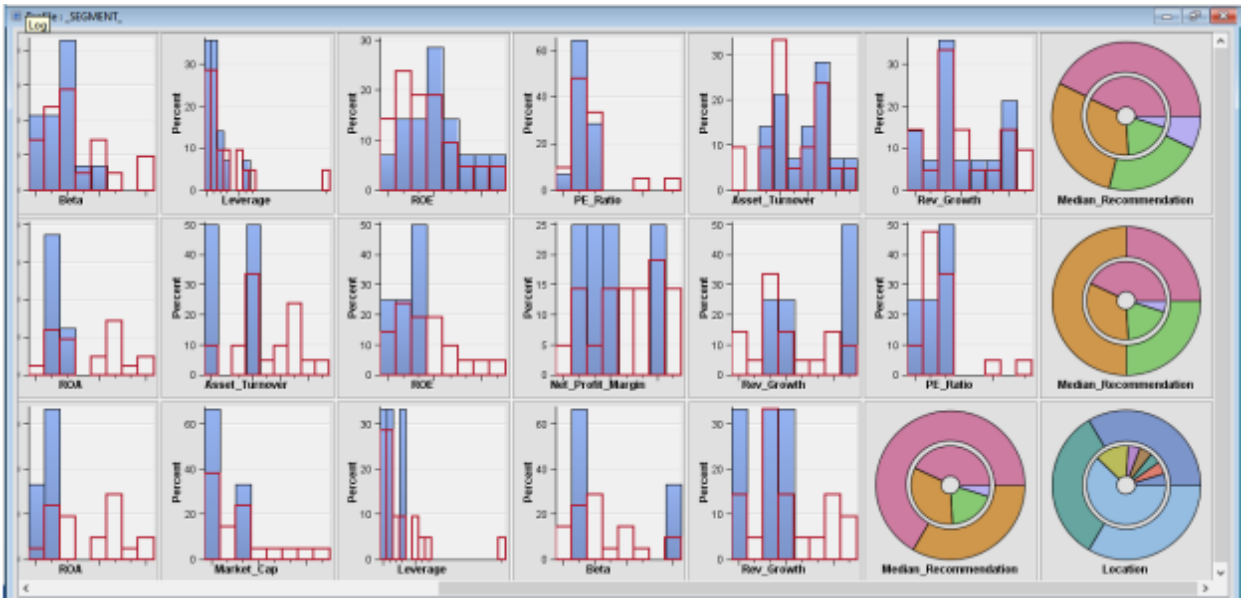


Figure 18. Clustering Results w/ variables 10-12

Cluster 1:

- Median Recommendations: Cluster 1 is comprised of moderate buy, moderate sell and hold recommendations. There are no patterns with respect to this variable.
- Location: This cluster is comprised of US and Ireland based firms. No pattern detected.
- Exchange: This cluster is the only cluster that have firms on the NASDAQ and AMEX exchanges.
- No noticeable patterns detected.

Cluster 2:

- Median Recommendations: Cluster 2 is comprised of moderate buy and hold recommendations. There are no patterns with respect to this variable.
- Location: This cluster is comprised of Canada, Germany and US based firms. No pattern detected.
- Exchange (Exchange): The only pattern observed is the similarity with respect to all the firms trading on the NYSE in this cluster. No noticeable patterns detected.

Cluster 3:

- Median Recommendations: Cluster 3 is comprised of moderate buy, moderate sell, strong buy and hold recommendations. Most of the firms that have hold recommendations are present in here.
- Location: This cluster is comprised of France, Switzerland and US based firms. No pattern detected.
- Exchange: The only pattern observed is the similarity with respect to all the firms trading on the NYSE in this cluster.
- No noticeable patterns detected

EMWS1.Clus_TRAIN

Obs #	Symbol	Name	Market_Cap	Beta	PE_Ratio	ROE	ROA	Asset_Turnover	Leverage	Rev_Growth	Net_Profit_Margin	Median_Recommendation	Location	Exchange	Segm
1	ABT	Abbott Laboratories	68.44	0.32	24.7	26.4	11.8	0.7	0.42	7.54	16.1	Moderate Buy	US	NYSE	
2	AGN	Allergan, Inc.	7.58	0.41	82.5	12.9	5.5	0.9	0.6	9.16	5.5	Moderate Buy	CANADA	NYSE	
3	AHM	Amersham plc	6.3	0.46	20.7	14.9	7.8	0.9	0.27	7.05	11.2	Strong Buy	UK	NYSE	
4	AZN	AstraZeneca PLC	67.63	0.52	21.5	27.4	15.4	0.9	0	15	18	Moderate Sell	UK	NYSE	
5	AVE	Aventis	47.16	0.32	20.1	21.6	7.5	0.6	0.34	26.81	12.9	Moderate Buy	FRANCE	NYSE	
6	BAY	Bayer AG	16.9	1.11	27.9	3.9	1.4	0.6	0	-3.17	2.6	Hold	GERMAN	NYSE	
7	BMY	Bristol-Myers Squibb Comp	51.33	0.5	13.9	34.8	15.1	0.9	0.57	2.7	20.6	Moderate Sell	US	NYSE	
8	CHTT	Chattem, Inc	0.41	0.85	26	24.1	4.3	0.6	3.51	6.38	7.5	Moderate Buy	US	NASDAQ	
9	ELN	Elan Corporation, plc	0.78	1.08	3.6	15.1	5.1	0.3	1.07	34.21	13.3	Moderate Sell	IRELAND	NYSE	
10	LLY	Eli Lilly and Company	73.84	0.18	27.9	31	13.5	0.6	0.53	6.21	23.4	Hold	US	NYSE	
11	GSK	GlaxoSmithKline plc	122.11	0.35	18	62.9	20.3	1	0.34	21.87	21.1	Hold	UK	NYSE	
12	IVX	IVAX Corporation	2.6	0.65	19.9	21.4	6.6	0.6	1.45	13.99	11	Hold	US	AMEX	
13	JNJ	Johnson & Johnson	173.93	0.46	28.4	28.6	16.3	0.9	0.1	9.37	17.9	Moderate Buy	US	NYSE	
14	MRX	Medicis Pharmaceutical Cor	1.2	0.75	28.6	11.2	5.4	0.3	0.93	30.37	21.3	Moderate Buy	US	NYSE	
15	MRK	Merck & Co., Inc	132.56	0.46	18.9	40.6	15	1.1	0.28	17.35	14.1	Hold	US	NYSE	
16	NVS	Novartis AG	96.65	0.19	21.6	17.9	11.2	0.5	0.06	-2.69	22.4	Hold	SWITZER	NYSE	
17	PFE	Pfizer Inc	199.47	0.65	23.6	45.6	19.2	0.8	0.16	25.54	25.2	Moderate Buy	US	NYSE	
18	PHA	Pharmacia Corporation	56.24	0.4	56.5	13.5	5.7	0.8	0.35	16	7.3	Hold	US	NYSE	
19	SGP	Schering-Plough Corporatio	34.1	0.51	18.9	22.6	13.3	0.8	0	8.58	17.6	Hold	US	NYSE	
20	WPI	Watson Pharmaceuticals, In	3.26	0.24	18.4	10.2	6.8	0.5	0.2	29.18	15.1	Moderate Sell	US	NYSE	
21	WYE	Wyeth	48.19	0.63	13.1	54.9	13.4	0.6	1.12	0.36	25.5	Hold	US	NYSE	

Figure 19. 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.

- Segment 1: Low Scale Industries
- Segment 2: Unperformed Companies
- Segment 3: Most Profitable Companies

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)

Figure 20. Diagram

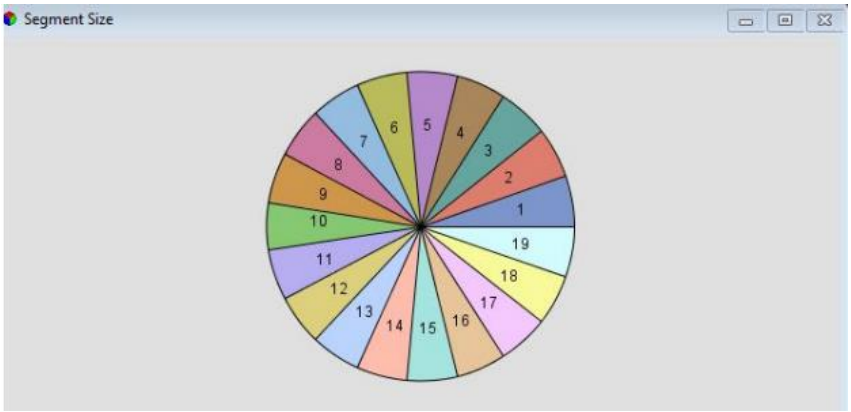


Figure 21. Attempt 1 - Segment Size

The outcome is not so perfect as there are 19 clusters and distance to the nearest cluster also varies a lot.

4 Clusters:

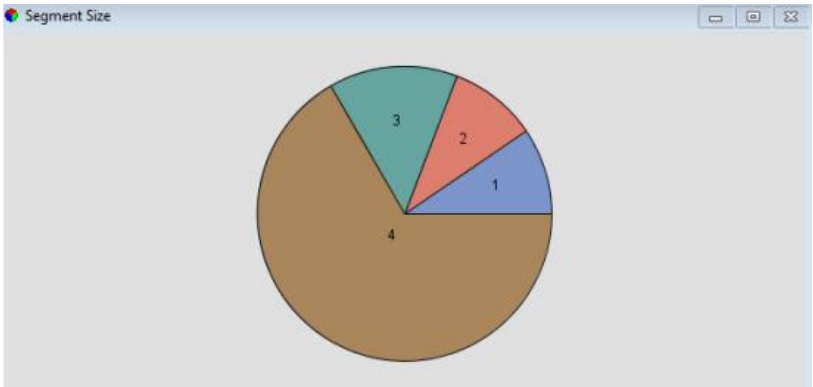


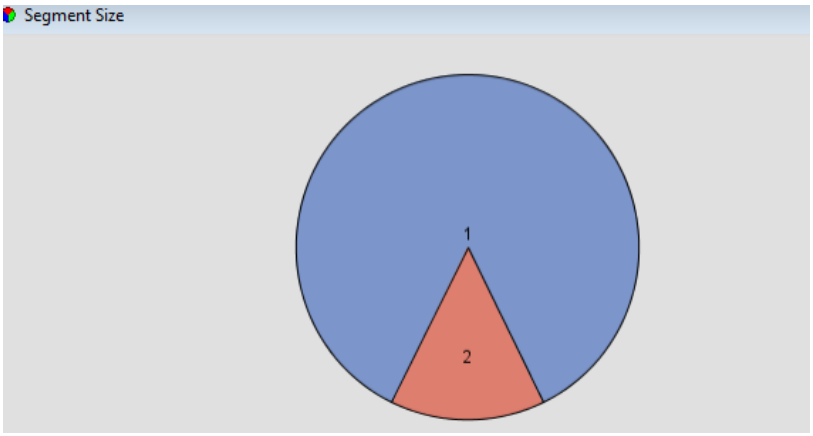
Figure 22. Attempt 2 - 4 Clusters.

Clustering Criterion	Maximum Relative Change in Cluster Seeds	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
0.416386	0	0	1	2	0.422521	1.636415	3	3.131229
Clustering Criterion	0	0	2	2	0.356067	1.37904	4	4.161037
0.416386	0	0	3	3	0.560049	3.158472	1	3.131229
0.416386	0	0	4	14	0.456433	3.283264	3	4.066274

Figure 23. Attempt 2 - 4 Clusters

The number of clusters are changed to 4. The clusters are not even and the distance is also more compared to the original method.

2 Clusters:



Clustering Criterion	Maximum Relative Change in Cluster Seeds	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
0.513318	0	0	1	18	0.54063	4.749337	2	3.636114
0.513318	0	0	2	3	0.531333	2.959212	1	3.636114

Figure 24. Attempt 3 - 2 Clusters

No additional information is provided for the interpretation. So not effective as original method.