IDS 572
Assignment – 4
Market Segmentation

Bala Rohit Yeruva

Fall 2016

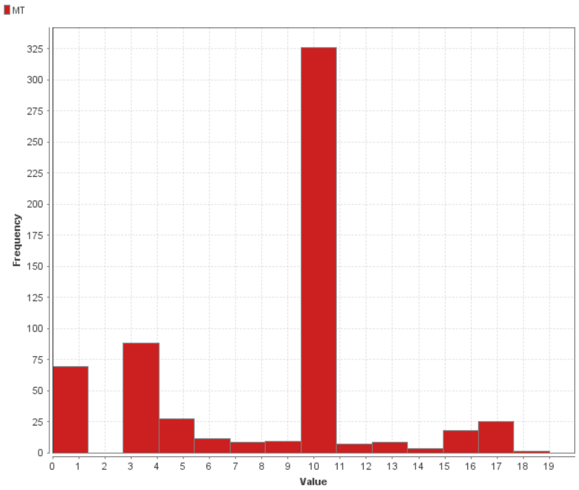
Market Segmentation – Consumers of Bath Soap in India

CRISA, a leading market research agency which specializes in tracking consumer purchase behaviour want to segment Indian market based on the consumption of bath soap.

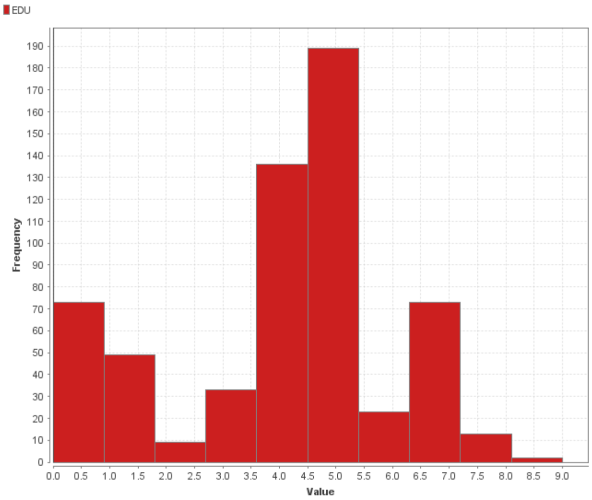
Variable Transformations

After observing the variables in the dataset, we have transformed a few categorical variables into their corresponding indicator variables. The following are the such variables:

i. Mother Tongue (MT): Based on the distribution of the mother tongue among the dataset, we could observe that more than 50% of the population speaks Marathi. Hence, we have created a binary indicator variable, ‘Marathi’, which takes the value ‘1’ when the household speaks Marathi and ‘0’ when they won’t.



ii. Housewife Education (EDU): From the distribution of household of Housewife Education, we could see that majority of them have education either up to 4 years or between 5–9 years in school. Therefore, we have created four binary indicator variables. The first variable indicates whether the housewife illiterate or has no formal schooling. The second indicator variable indicates if they have an education up to 4 years of school. The third indicates if they have 5-9 years of schooling. The last variable refers if they have a graduate or a professional degree.



iii. Cable (CS): According to the code book available in the dataset, this variable is a binary variable indicating if the household has a television cable connection. However, after observing the data, we could observe that a few data points have a third value which doesn't refer or mean anything. Therefore, we have created a new binary variable which has only two values. It takes '1' if they have a cable connection and '0' if they don't.

iv. Sex: Similar to the above variable, after observing the data, we were able to observe three values for the gender variable. Therefore, we have created another binary indicator, 'Gender', which takes the value '1' if the homemaker is female and '0' if the homemaker is a male.

New Variables:

We have created the following variables which are used in cluster analysis based on purchase behaviour and the basis of purchase and pro:

1. Brand Loyalty: A person's brand loyalty was calculated based on the maximum brand percentage that they have purchased among the brands with code list: 5, 24, 55, 57 or 144, 272, 286, 352 and 481.
2. Max Selling Proposition: Similar to Brand Loyalty, even this variable contains the maximum proposition type percentage that a household has purchased.

1. Cluster Analysis

1A) Our first analysis is to segment the market based on the household purchase behaviour. Following are the variables used to generate the clusters:

- Average Price
- Brand Runs
- Number of Brands
- Number of Transactions
- Total Volume
- Value
- Brand Loyalty
- Others 999

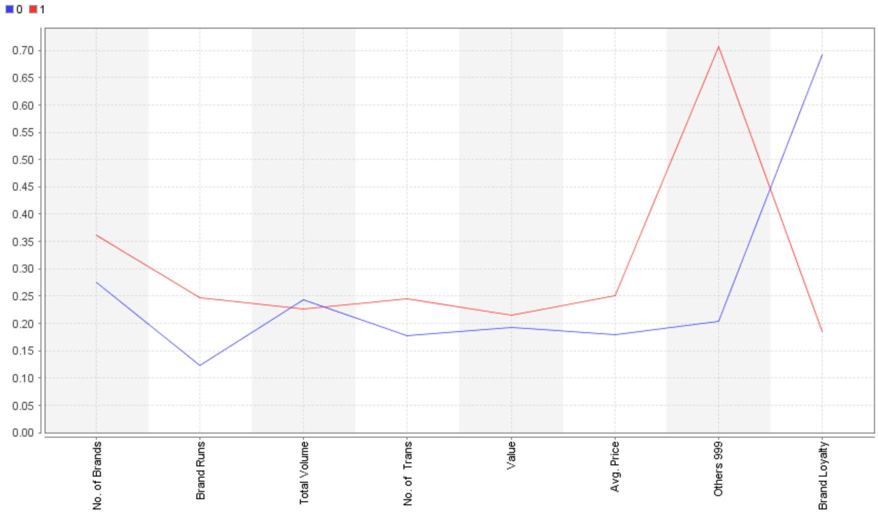
We haven't used the percentages of total purchases comprised by various brands because we have already taken the maximum of them as the Brand Loyalty variable. We have included Others 999 variable which indicates the percentage of total purchases comprising of all the other brands. This variable indicates the lack of loyalty. Regarding the selection of K value, we have tried to analyse the clusters formed when we range the

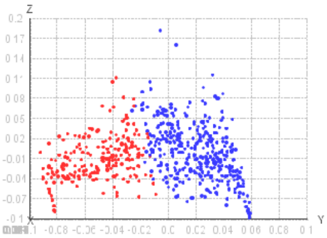
K value from 2 to 5 since we could support as many different promotional approaches and then decided on the best value for K.

K-Means Clustering:

a) K = 2

When K = 2, we have got two clusters with sizes 220 and 380. The below figures tries to depict the two clusters and we could see that both the clusters are reasonably separated with no overlapping. We could also see that the difference between the clusters is prominent based on Brand Loyalty and Other 999 variables. We could see that the first cluster (cluster_0) with high Brand Loyalty and less Other 999 which indicates that the households in this cluster are more brand loyal and would like to stick to one particular brand. However, we must keep in consideration that they do less number of transactions. The second cluster (cluster_1) has a high Other 999 value and low Brand Loyalty which indicates that the households in this cluster would prefer to change the brand of their soap more often and do not stick to one preferred brand. We could also see that the total number of brands purchased by the second cluster is also high than that of the first cluster which again indicates that these households are not brand loyal.

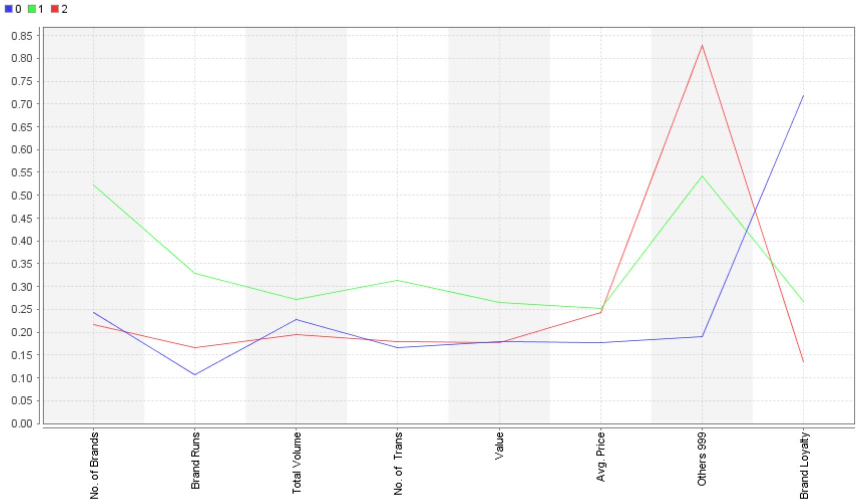


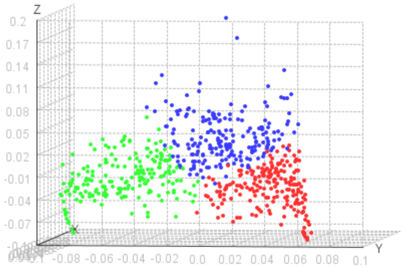


b) K = 3

With K = 3, we have got approximately equal sized clusters: 197, 204 and 199. Further, from the below figures we could see that even now, all the three clusters are reasonably separated with not much overlap. We could also observe that even now, the difference in the three clusters is predominant based on Brand Loyalty and Others 999 variables.

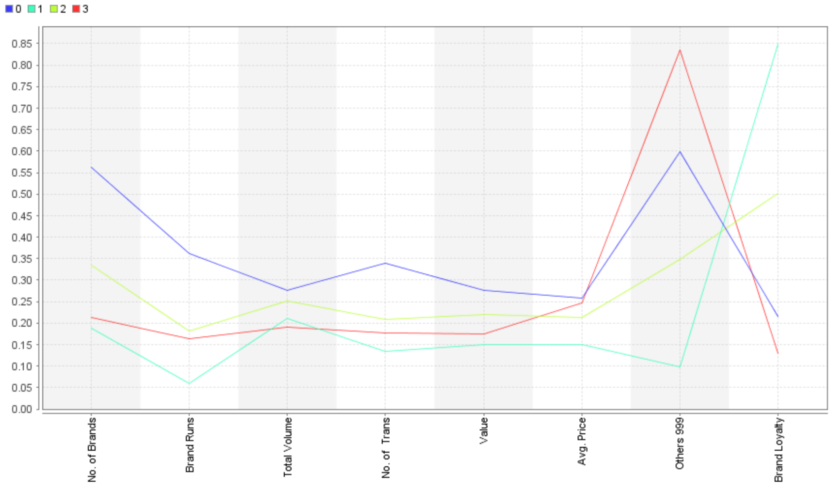
The first cluster (cluster_0) has the highest Brand Loyalty and least Others 999 values. This cluster represents the households that are brand loyal. The second cluster (cluster_1) has average Brand Loyalty as well as Others 999 values but has the highest No. of Brands and Brand Run values. This cluster represents those household which choose a brand and use it for a long time (high brand runs) and then switch to another brand. The third cluster has the least Brand Loyalty and high Others 999 value. This cluster represents those households who prefer to change the brand often and thus aren't loyal to any brand.

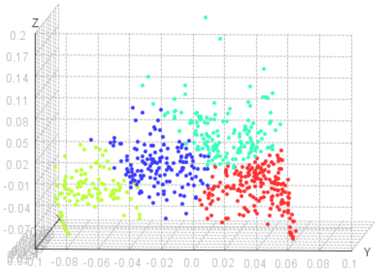




b) K = 4

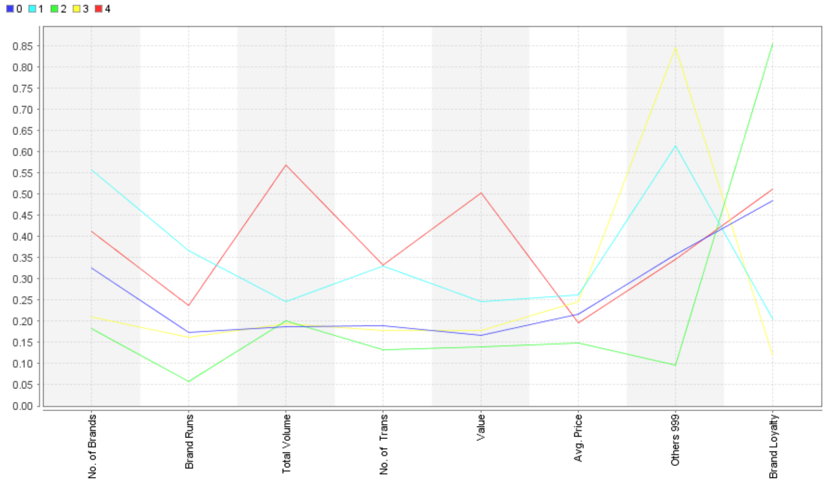
The four cluster sizes are 156, 105, 151 and 188. From the below figures we could see that even now the clusters are well separated. The first cluster (cluster_0) has relatively less Brand Loyalty and high Others 999 values along with highest number of brands and brand runs. This cluster might represent those households who use a brand for a long time and then change the brand. The second cluster (cluster_1) has the highest Brand Loyalty and least Others 999 values along with least number of Brands, Brands Runs and No. of transactions. This cluster might represent the households which are Brand Loyal but do less number of transactions. The third cluster (cluster_2) has average values for almost all variables. The final cluster (cluster_3) has least Brand Loyalty and highest Others 999 and average values for all the other variables. This cluster represents those households which frequently change their soap brand. In this model, we were able to distinguish all the clusters except for the third cluster (cluster_2).

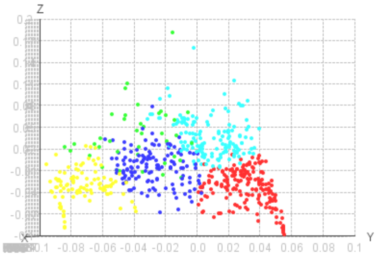




b) K = 5

The five clusters sizes are 135, 145, 100, 180 and 40. We could see that the fifth clusters have relatively lesser number of data points. Further, from the below figures we could see that the fifth cluster (green) is very sparse. We could also see that a couple of the clusters have similar values for most of the variables and it is hard to interpret all the clusters based on the variables.



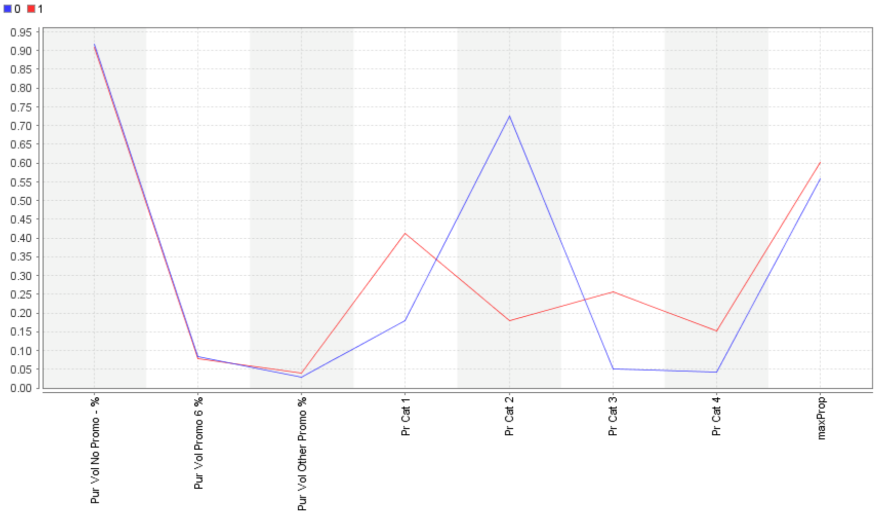


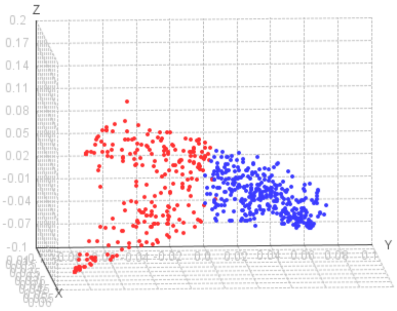
Best ‘K’ value: From our analysis of the clusters, we could observe that when K = 3 we were able to generate reasonably dense clusters with appropriately equal cluster sizes and most importantly, we were able to name all the clusters based on their characteristics. We have then integrated a decision tree model to our clusters and observed an accuracy of 98.33% which indicates that the segmentation of the market into these three clusters is quite justified.

1B) Our next analysis is to segment the market based on the basis of purchase. Following are the variables used to generate the clusters:

- % of Pur Vol No Promo
- % of Pur Vol Other Promo
- % of Pur Vol Promo 6
- Maximum Proposition (maxProp)
- Price Category 1 (Any Premium Soaps)
- Price Category 2 (Any Popular Soaps)
- Price Category 3 (Any Economic Soaps)
- Price Category 4 (Any SubPopular Soaps)

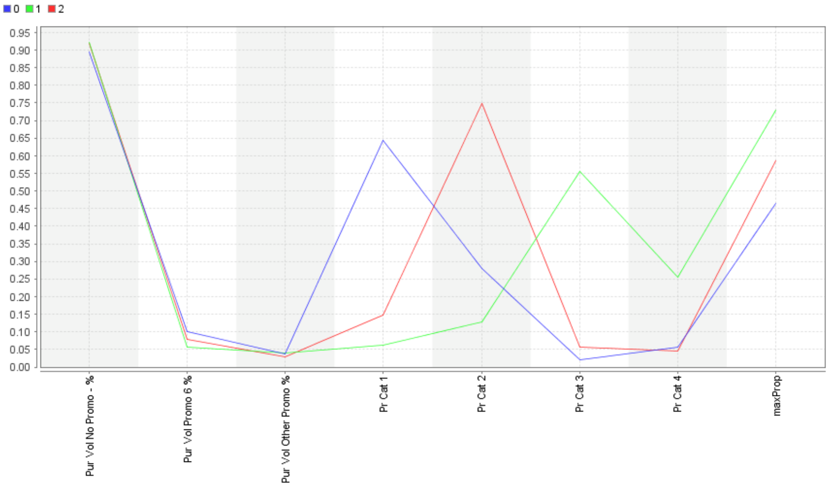
a) K = 2

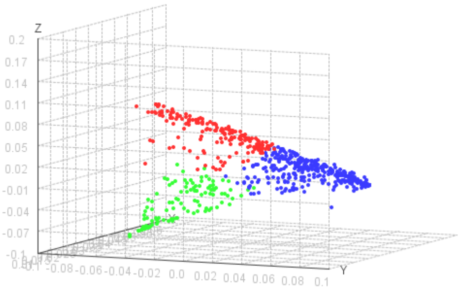




From the above figures, we could see that the two clusters (sizes: 344, 256) formed are well separated. However, the characteristics of the cluster is not distinguishable as we could see that they have almost similar values relative to all the variables except for Price Category 2.

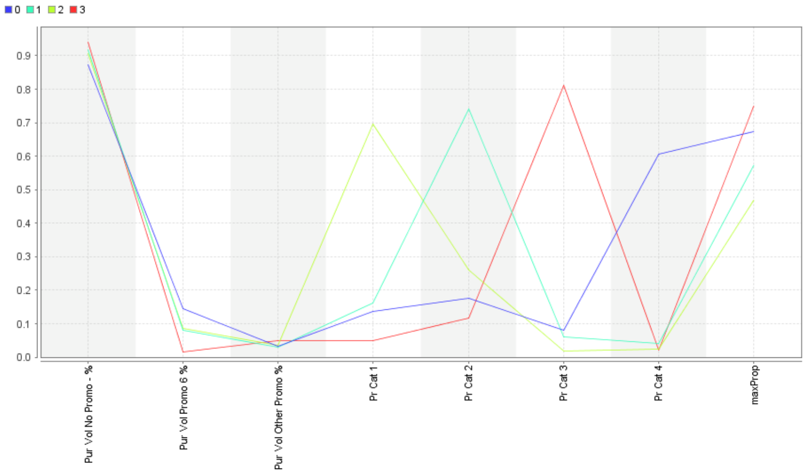
b) K = 3

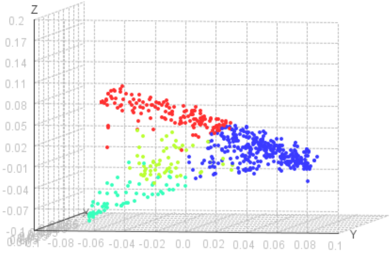




The three clusters (sizes: 178, 113, 309) are again reasonably separated. We could also see that the three clusters have similar values for the promotion variables but have unique values for all the price categories. The first cluster (cluster_0) has the highest Pr Cat 1, the second cluster (cluster_1) has the highest value for Pr Cat 3 and Pr Cat 4 while the third cluster (cluster_2) has the highest value for Pr Cat 2. Thus we can attribute the first cluster to those who prefer premium soaps, the second to those who prefer sub-popular and economic soaps and the third cluster to those who prefer popular soaps.

c) K = 4

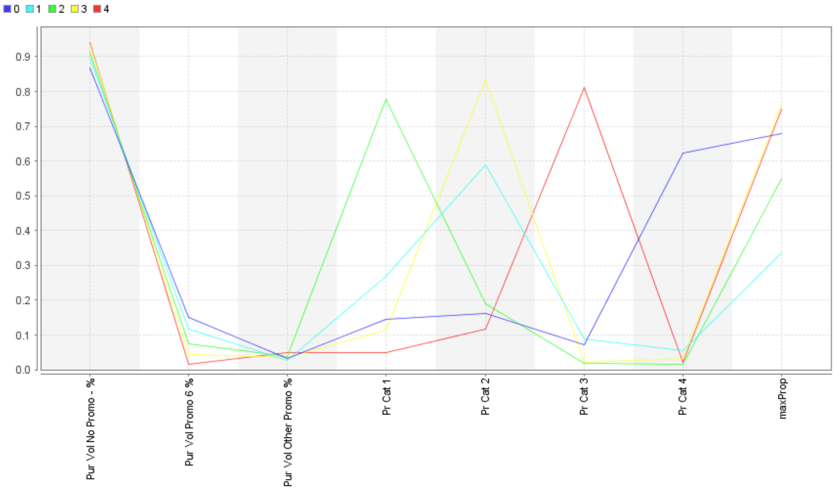


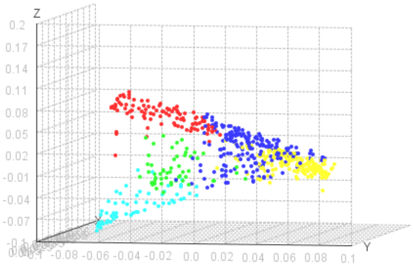


The clusters formed are of sizes: 58, 322, 150 and 70. We could see that there are clusters with relatively less sizes when compared with others. However, we could observe that we could again distinguish all the clusters based on the price categories. The first cluster (cluster_0) has the highest value for Pr Cat 4 which indicates that these household prefer sub-popular soaps. The second cluster (cluster_1) has the highest value for Pr Cat 2 which indicates that these households prefer popular soaps. The third cluster (cluster_2) has the highest value for Pr Cat 1 which indicates that these households prefer premium soaps. The fourth cluster (cluster_3) has the highest value for Pr Cat 3 which indicates that these households prefer economic soaps.

d) K = 5

The clusters (55, 195, 108, 172 and 70) aren't as well separated as the other models. We observe similar segmentation as before (when K=4) however, the second cluster (cluster_1) and the fourth cluster (cluster_3) have relatively high values for Pr Cat 2 which indicates that we have segmented the market for more than what we needed.

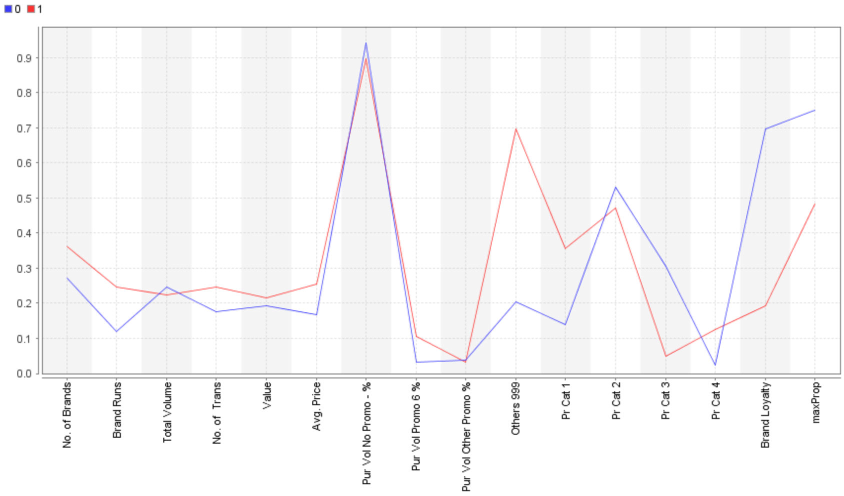


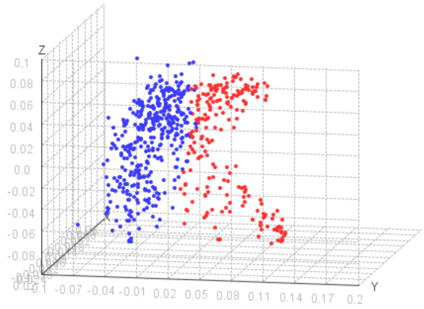


Best ‘K’ value: From our analysis of the clusters, we could observe that when K = 4 we were able to generate reasonably dense clusters and was able to name all the clusters based on their price category characteristics. We have then integrated a decision tree model to this model and observed an accuracy of 99.83% which is strong evidence to state that segmentation of the market into these four clusters is quite appropriate.

1C) We know try to segment the market on both purchase behaviour as well as the basis of purchase. We have used all the variables used above for generating the above clusters.

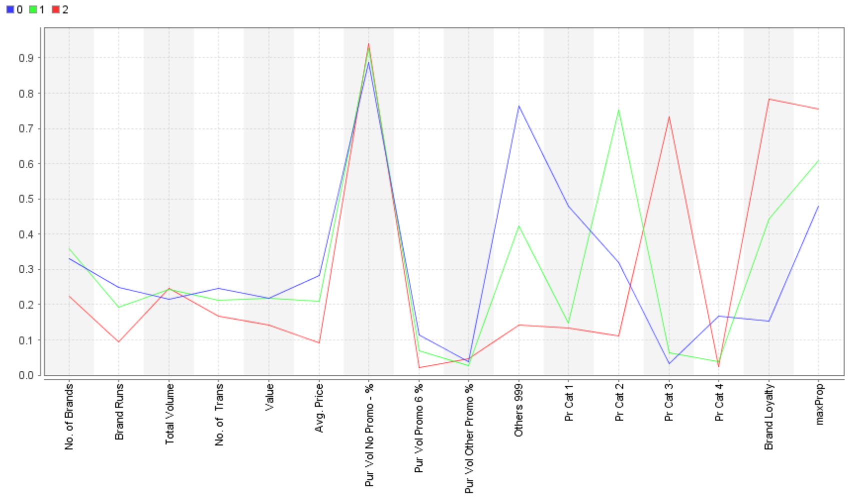
a) K = 2

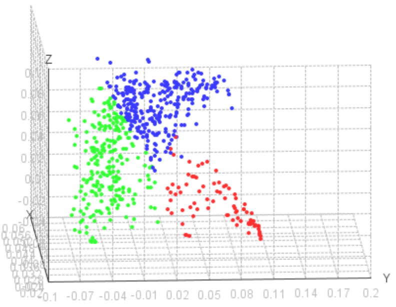




The clusters (sizes: 213 and 387) are well separated. However, based on their attribute values, we could see that except for Brand Loyalty, Others 999 and MaxProp values, they have similar values for all the other variables.

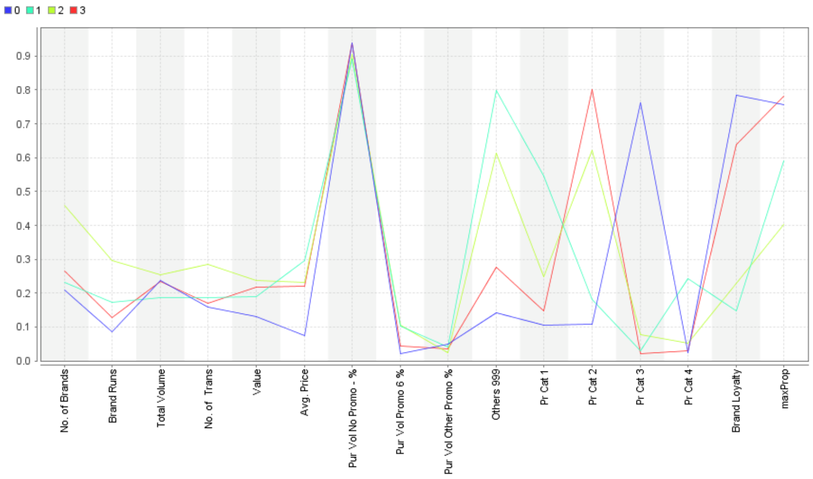
b) K = 3

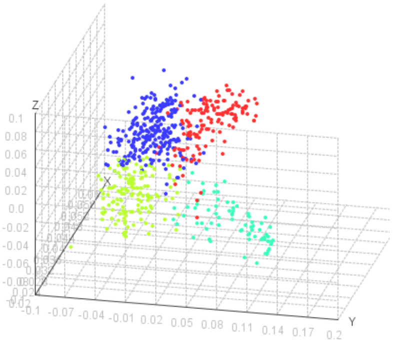




The three clusters (241, 280 and 79) appear to be well formed with no overlap. The first cluster (cluster_0) has highest Others 999 as well as Pr Cat 1 and least Brand Value. This indicates that these household prefer using premium soaps and frequently change their brand based on the brand premium value. The second cluster (cluster_1) has the highest Pr Cat 2 and average values for Others 999 and Brand Loyalty which indicates that they prefer popular brands and might change their brand based on the change in the popularity of the brand. The third cluster (cluster_2) has the highest Brand value as well as Pr Cat 3 which indicates that they prefer soaps that are economical and stick to them.

c) K = 4

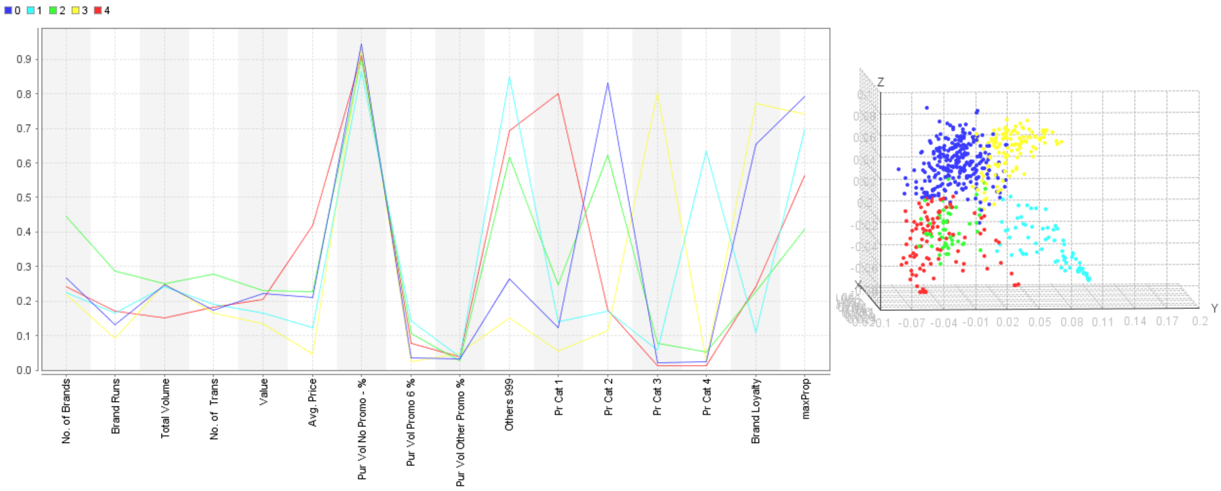




The clusters again seemed to be formed well with appropriate separation. The first cluster (cluster_0) has the highest Brand Loyalty as well as Pr Cat 3 which indicates that they prefer economical soaps and stick with these brands. The second cluster

(cluster_1) has least Brand Loyalty and high Pr Cat 1 and Pr Cat 4. This indicates that these households prefer premium and sub-popular brands and change the brand as frequently as possible based on the brand value. The third cluster (cluster_2) has the relatively least Brand Loyalty and has the highest value for Pr Cat 2. This indicates that these households prefer to use only popular brands and change the brand based on the change in their popularity. The fourth cluster (cluster_3) has relatively high brand loyalty value along with Pr Cat 2 which indicates that they prefer popular brands and stick to these brands.

d) K = 5



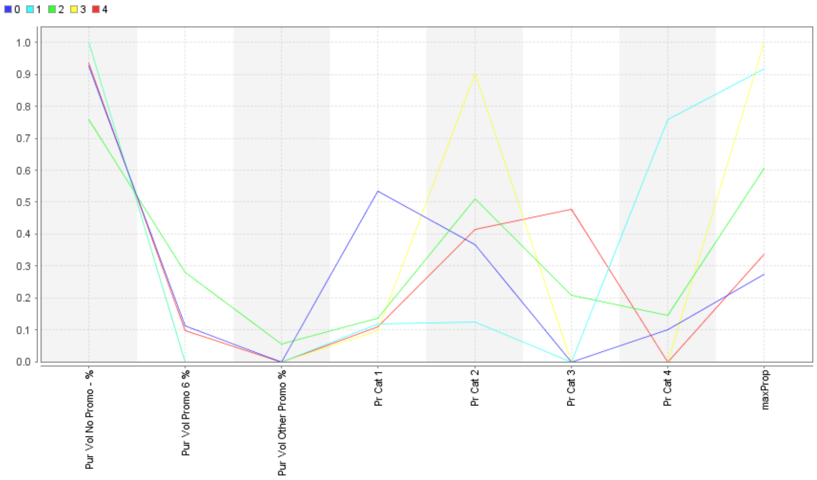
The clusters (123, 53, 256, 71, 97) doesn't seemed to be formed as clear as before. We could observe that all the clusters formed above when K=4 still persist along with characteristics and the new cluster (cluster_1) has the lowest Brand Loyalty and high Pr Cat 4 which indicates that these households prefer brands that are sub-popular and frequently change their brand frequently again may be based on the brand popularity. This cluster can thus be considered similar to cluster (cluster_2) which represents the household who prefer popular soaps and frequently change the brand.

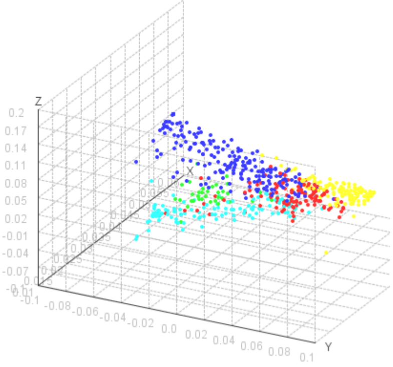
<u>Best 'K' value</u>: From our analysis of the clusters, we could observe that once again, when K = 4 we were able to generate reasonably dense clusters and was able to name all the clusters based on their brand loyalty as well as their basis of purchase characteristics. We then integrated a decision tree model and observed an accuracy of 96.00% which is reasonably good evidence to conclude that our segmentation of the market into these four clusters is quite appropriate.

1D) We know try to segment the market based on the basis of purchase using K-Mediods, Kernel K-Means, Agglomerative clustering and DBSCAN.

K-Medoids Clustering:

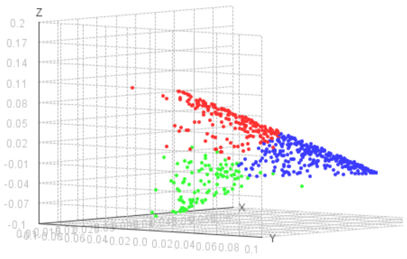
After trying the K values from 2 to 5, we could observe that when K = 5, we were able to generate reasonably meaningful clusters.

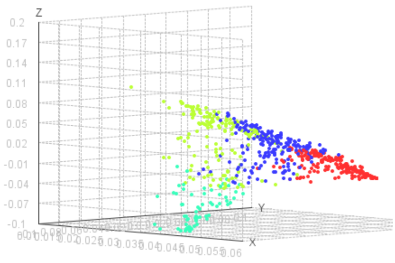


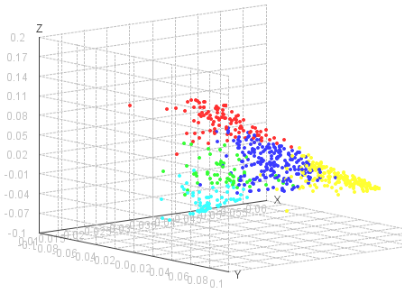


The clusters (sizes: 202, 36, 106, 137, 119) seemed to have quite a few overlaps. The first cluster has high Pr Cat 1 and low maxProp which indicates that they prefer premium soaps and purchase quite different type of soaps. The second cluster has high Pr Cat 4 value along with high maxProp value which indicates that they prefer sub-popular brands and purchase similar type of soaps. The third cluster has high Pr Cat 2, a medium maxProp value and relatively high percentage of Promo 6% which indicates that they prefer popular soaps and tend to purchase more when there is a promotion sale of 6%. The fourth cluster has high Pr Cat 2 value along with highest maxProp value indicating that they prefer popular brands and purchase the same type of soaps frequently. The fifth cluster has high Pr Cat 3 value and less maxProp value indicating that these households prefer to buy economic brands and would purchase different types of soaps.

Kernel K-Means: After trying the K values from 3 to 5, we could observe from the below figure, that when K=3, we were able to generate clusters with good density and appropriately separated. Changing the Kernel type did have slight effect on the formation of these clusters. However, only radial kernel type gave dense clusters with clear distinction with reasonable cluster sizes.

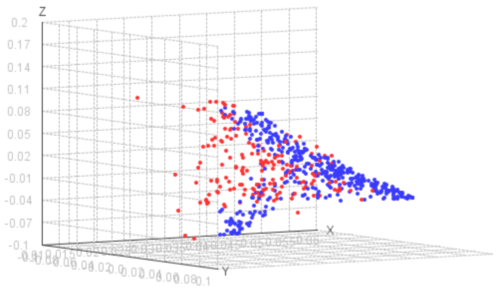






Agglomerative Clustering: We have used the flatten operator in Rapid Miner so as select the number of clusters to be generated to be five. Then, we have tried to use different combinations of cluster distance modes: single linkage, complete linkage and average linkage. However, Single Linkage and Average Linkage resulted in majority clusters with only one data points. Complete Linkage was better with only one cluster resulting in single data point. The cluster sizes when using complete linkage along with distance measure type as Numerical and Euclidean Distance, were 123, 398, 16, 1 and 62. For the dendogram and other SVD visualizations generated by this agglomerative clustering please refer to Appendix A.

DBSCAN: We ran several iterations of DBSCAN and were able to generate 2 clusters with relatively better sizes were when we use the following parameters: Epsilon = 0.1884 and MinPts = 10. Measure type and the numerical measure used were Numerical and Euclidean distance respectively.



Conclusion: We could observe that the clusters obtained from different techniques are different which indicates that the type of clustering technique selected to form the clusters has an effect on the cluster generation. We could also see that we weren't able to generate meaningful clusters when using Agglomerative Clustering and DBSCAN. The reason for these techniques to fail could be due to high density of the data. Among K-Medoids and Kernel K-Means, K-Medoids appears to generate more meaningful clusters because we were able to identify the different characteristics of the clusters.

2. Best Segmentation

Based on the analysis of all the above clustering models, we believe that our best model is when using K-Means technique with K = 4 and including all the variables related to both purchase behavior and basis of purpose. The following table shows the model performance characteristics of the best model in each analysis. From the table we could clearly see that the above chosen best model has better values for all the cluster evaluation parameters.

Cluster Analysis	*K*	*Davies Bouldin*	*Avg. within centroid distance*	*Cluster sizes*
Purchase Behavior	3	-1.255	-0.14	197, 204, 199
Basis of Purchase	4	-0.955	-0.162	58, 322, 150, 70
Both	4	-1.49	-0.372	75, 143, 246, 136

For our chosen best model, below is the inter cluster distance matrix. We could observe that almost all are well separated.

Clusters	*1*	*2*	*3*	*4*
1	0	1.312	1.255	1.050
2	1.312	0	0.696	1.076
3	1.255	0.696	0	0.746
4	1.050	1.076	0.746	0

Cluster Characteristics:

As described above in the analysis section, we could distinguish the clusters as follows:

First Cluster (cluster_0): Consists of 75 households with high Brand Loyalty along with high preference towards economic soaps. Out of these 75 households, 41 speak Marathi and 55 of them are female housemakers. Majority (46) of them are illiterates or do not have any formal schooling and only 2 of them have a college degree. They also have the least Affluence Index (8.36), number of brands (2.68), avg. price (7.697) and total number of transactions (22.787). However, they have the highest transactions per brand (5.087), volume per transaction (531.803). Hence, this cluster represents the households with majority of illiterate housemakers who stick to economic brand and tend to make high volume transactions.

Second Cluster (cluster_1): Consists of 143 households with less brand loyalty and prefer premium and sub-popular brands. Around 63 of them speak Marathi and 117 of them are female housemakers. Majority (46) of them have school education up to 5-9 years only. Their Affluence index is 16.818 with less number of brands (2.86) and least number of total volume purchased (9661.049) and relatively lesser transaction per brand (2.322) and volume per transaction (379.204). However, they have the highest avg. Price (13.813) value. The households from this cluster can be classified as majority of them being female housemakers with 5-9 years of education who prefer to buy less volume as well as less number of transactions and prefer premium soap brands which tend to be very costly.

Third Cluster (cluster_2): Consists of 246 households with less brand loyalty and prefer popular brands. About 152 of them speak Marathi and 222 of them are female housemakers with 89 of them having 5-9 years of school education and 52 of them have college degree. Around 197 of them have cable connection and they have an affluence index equal to 19.663 which is the highest among the clusters. They have the highest average number of brands (4.667), brand runs (22.549), total volume (13036.309), number of transactions (40.171) and least average transaction per brand (1.89) and volume per transaction (348.771). Thus, this cluster represents those households with majority of Marathi speaking female housemakers with relatively better education. They tend to frequently change their soap brand and use it for quite some time and then switch the brand again. However, they prefer to purchase less volume and in less number of transactions.

Fourth Cluster (cluster_3): Consists of 136 households with high brand loyalty and prefer to buy popular brands. Around 70 of them speak Marathi and 117 of them are female housemakers with 45 of them having 5-9 years of school education and 29 of them having a college degree. About 108 of these households have cable connections and their affluence index is equal to 17.228, the second best. They have relatively high total volume (12079.669), high volume per transaction (508.246) and high average price (11.725). However, they have relatively low number of transactions (24.213). Thus, this cluster represents those households with majority of them being female Marathi speaking homemakers with average education and prefer to buy huge volumes in one transaction and thus do not make many transactions. They stick to the popular brands that they like even if the avg. price is relatively high.

3. Decision Tree in Interpreting the Clusters

We combined our best clustering model with a decision tree with the following parameters: criterion = gain ratio; maximal depth = 20, pruning confidence = 0.1 and no prepruning. Based on the accuracy of the model, we could observe that the decision tree was quite efficient in identifying the clusters appropriately. The below table depicts the confusion matrix:

	True Clust_2	*True Clust_0*	*True Clust_1*	*True Clust_3*	*Precision*
Pred. Clust_2	244	0	15	6	92.08%
Pred. Clust_0	1	75	0	0	98.68%
Pred. Clust_1	0	0	128	1	99.22%
Pred. Clust_3	1	0	0	129	99.23%
Recall	99.19%	100%	89.51%	94.85%	

The following are a few of the rules from the tree:

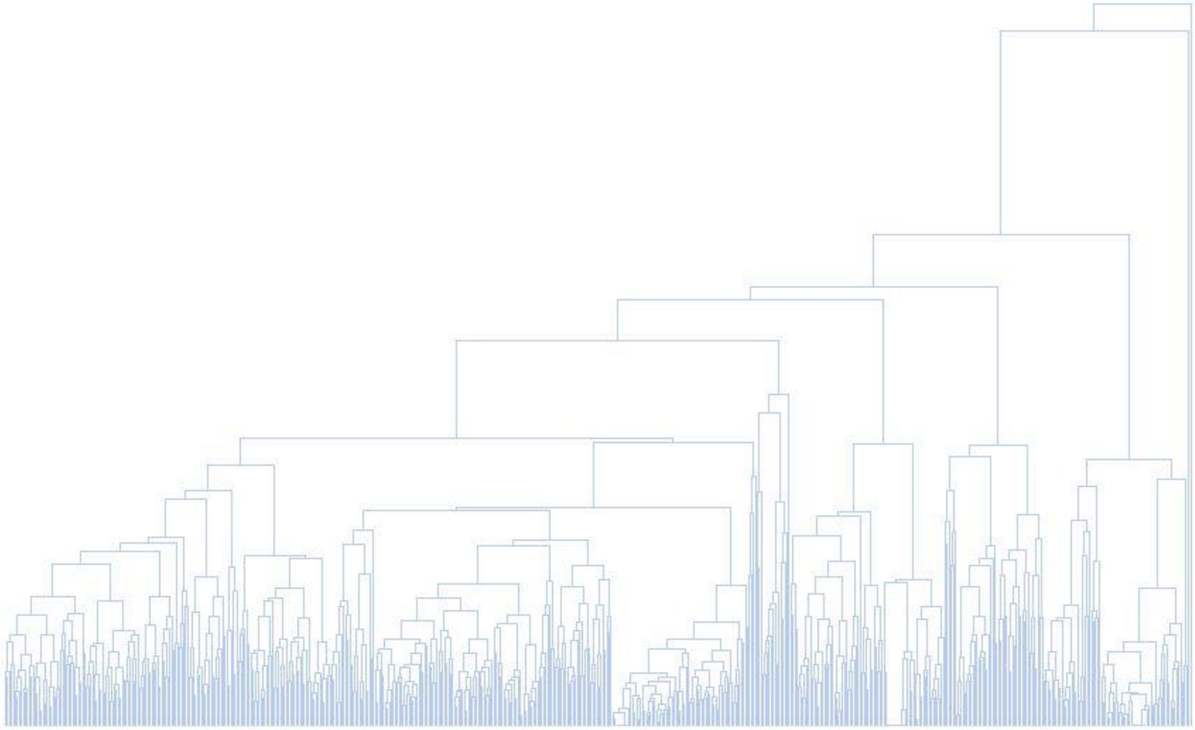
1. If Pr Cat 3 > 0.468 and No. of Brands > 7 then Cluster_2
2. If Pr Cat 3 > 0.468 and No. of Brands <= 7 then Cluster_0

Based on the variables at the top of the decision tree, we were able to identify the variables that are critical in identifying the clusters. A few of the variables that were observed at the top of the tree are Pr Cat 3, No. of Brands, Br_Cd_24, Br_Cd_5, Brand Loyalty, Total Volume and Pr Cat 1. These variables seemed to appear at the top of the tree irrespective of the tree parameters and hence we could deem these variables to be important. Thus, the decision tree model was quite robust and was accurately able to justify the market segmentation by our best model.

Appendix

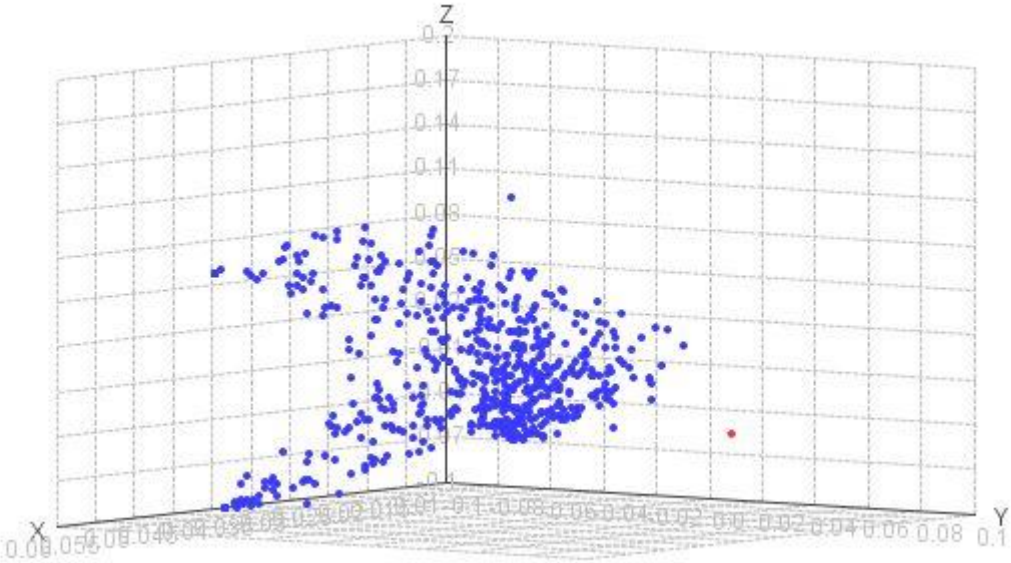
Appendix A

Dendogram:

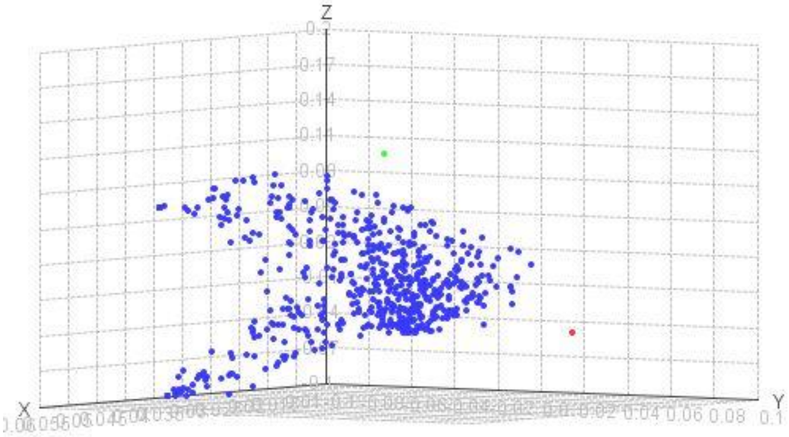


SVD Visualizations:

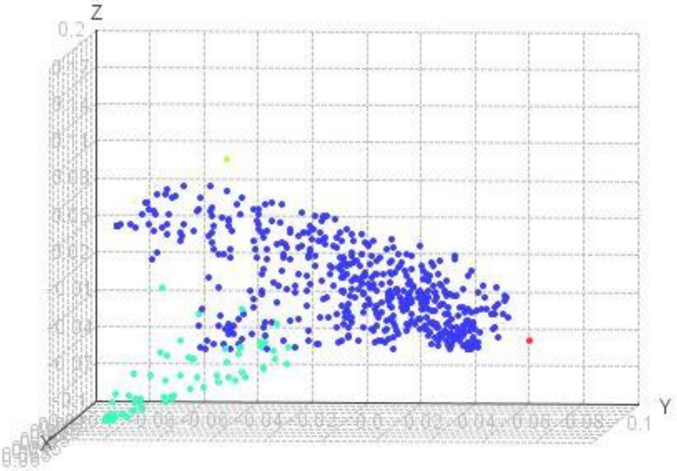
K = 2



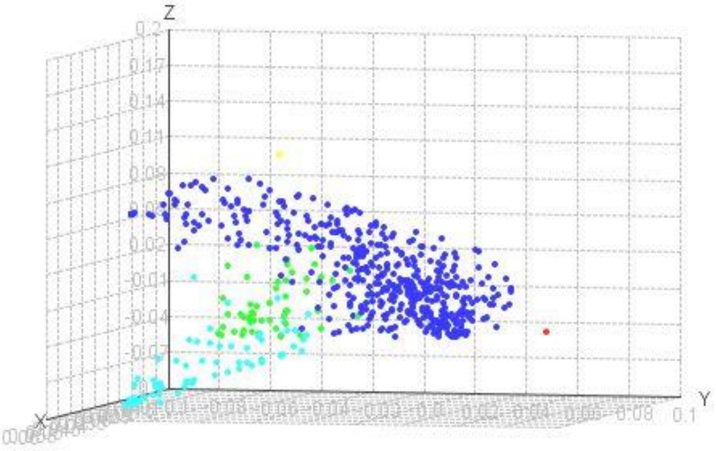
K = 3


Z
0.2
0.17
0.14
0.11
0.08
X
Y
0.02
0.04
0.06
0.08
0.1

K = 4


Z
0.2
Y
0.1
X

K = 5


Z
0.2
0.17
0.14
0.11
0.08
0.02
-0.01
-0.04
Y
0.08
0.1
X