Jitendra Patel

Renu Singh

Assignment 3

Market Segmentation

**Q1)** **Use k-means clustering to identify clusters of households based on**

**a. The variables that describe purchase behavior (including brand loyalty). How do you evaluate brand loyalty?**

**b. The variables that describe basis-for-purchase.**

**c. The variables that describe both purchase behavior and basis of purchase.**

**d. Try k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering too. You do not need to try all techniques on all combinations in (a)-(c) above; you may pick one set of variables in (a) thru (c) that you feel may be best suited for addressing the segmentation need. How do different parameter values for the different techniques affect the clusters obtained? Are the clusters obtained from the different procedures similar? What might be some reasons for differences in clusters obtained using different procedures? Which would you pick as your 'best' and why?**

**a)** Variables that describes Purchase behavior (including Brand loyalty)

1. Brands
2. No. Of Transactions
3. Brand Runs
4. Total Volume
5. Value
6. Avg. Price
7. Others 999

**Brand loyalty: -**

1. The customers who stick to comparatively less number of brands with maximum brand repetitions would be considered as loyal towards particular brand.
2. Loyalty will also reflect when a customer considers buying a same products for number of time even if the price of the product is on the higher side and there are no discounts offered on the product/Brand.
3. If the Other 999 attribute contains the major chunk of customer purchase then those customers are not the brand loyal i.e. those customers don’t focus on a particular brand while purchasing.

**Cluster Model Based on Purchase Behavior (K=3):-**

Cluster Model Distance between Clusters

|  |  |
| --- | --- |
| Cluster 0 | 198 Items |
| Cluster 1 | 191 Items |
| Cluster 2 | 211 Items |

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.494 |
| Cluster 0 | Cluster 2 | 0.625 |
| Cluster 1 | Cluster 2 | 0.553 |

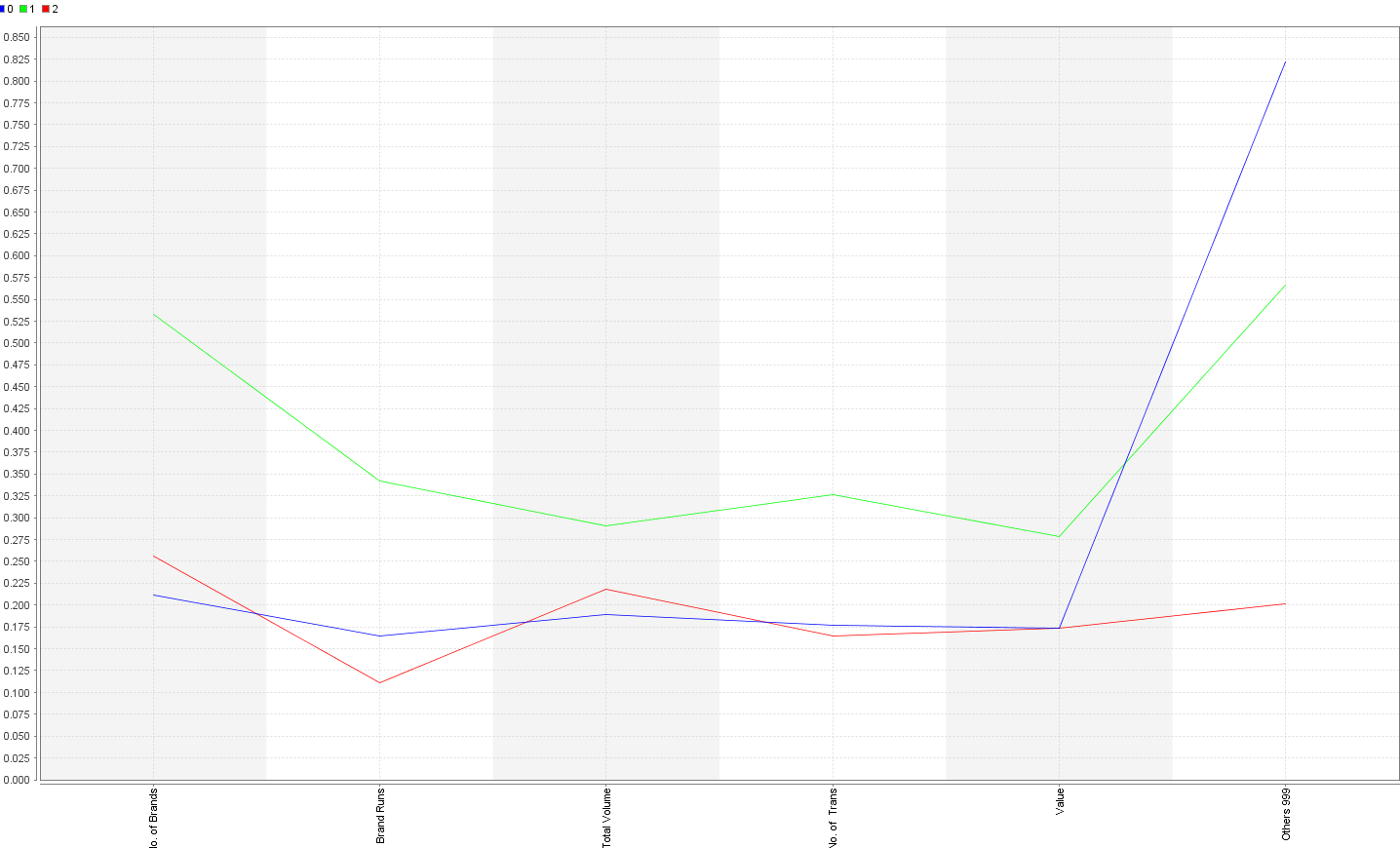
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.101 |
| Avg. within centroid distance cluster 0 | 0.081 |
| Avg. within centroid distance cluster 1 | 0.136 |
| Avg. within centroid distance cluster 2 | 0.089 |
| Davies Bauldin | 1.166 |

Centroid Table:-

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster 0 | Cluster 1 | Cluster 2 |
| No. of brands | 0.211 | 0.533 | 0.257 |
| Brand runs | 0.164 | 0.343 | 0.110 |
| Total volume | 0.189 | 0.291 | 0.219 |
| No of Transactions | 0.177 | 0.327 | 0.164 |
| Values | 0.174 | 0.279 | 0.174 |
| Others 999 | 0.821 | 0.566 | 0.201 |

**Centroid Graph:-**

****

**Cluster Model Based on Purchase Behavior (K=4):**

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.566 |
| Cluster 0 | Cluster 2 | 0.489 |
| Cluster 0 | Cluster 3 | 0.535 |
| Cluster 1 | Cluster 2 | 0.626 |
| Cluster 1 | Cluster 3 | 0.637 |
| Cluster 2 | Cluster 3 | 0.728 |

Distance between Clusters

Cluster Model:-

|  |  |
| --- | --- |
| Cluster 0 | 172 Items |
| Cluster 1 | 198 Items |
| Cluster 2 | 194 Items |
| Cluster 3 | 36 Items |

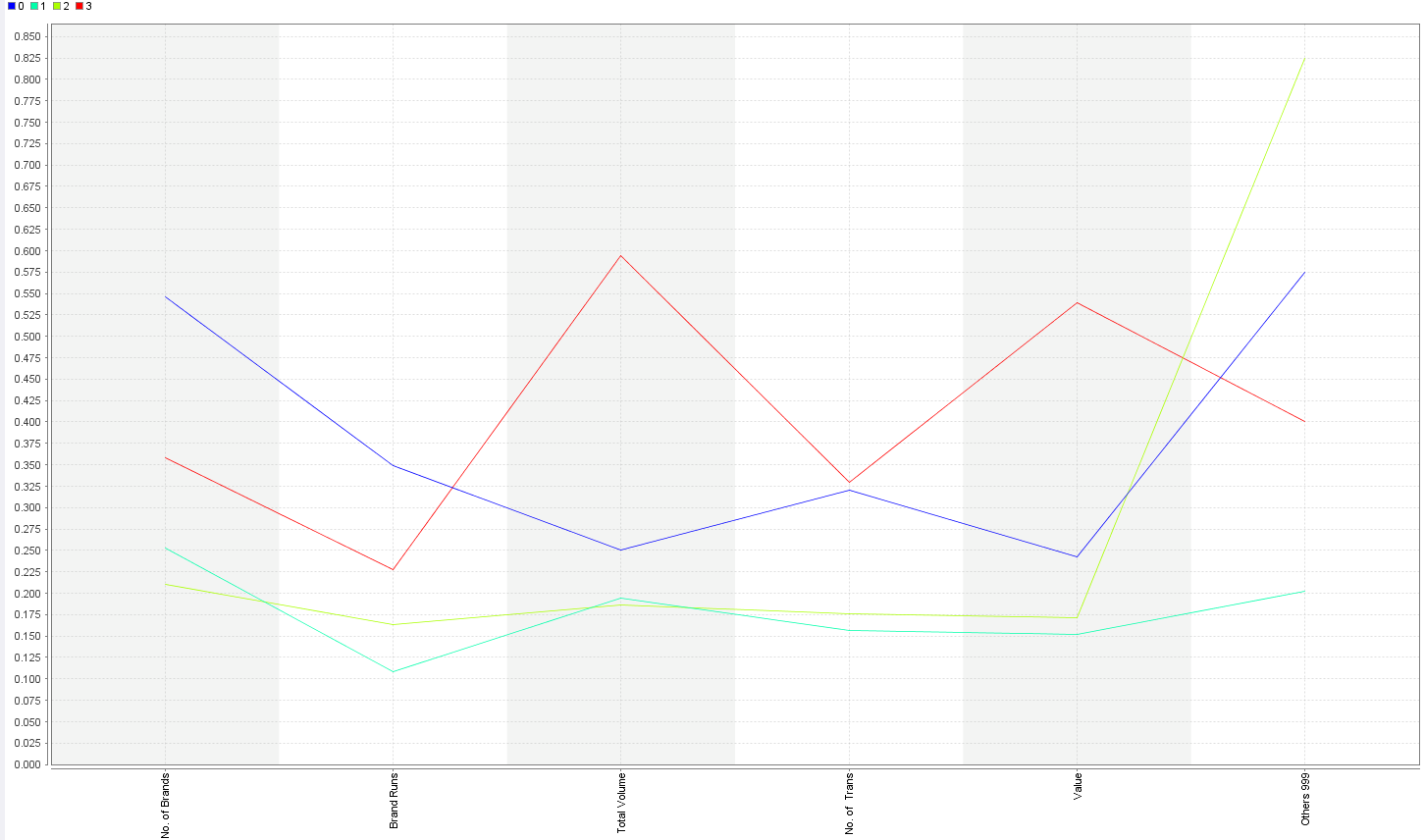
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.089 |
| Avg. within centroid distance cluster 0 | 0.101 |
| Avg. within centroid distance cluster 1 | 0.071 |
| Avg. within centroid distance cluster 2 | 0.078 |
| Avg. within centroid distance cluster 3 | 0.186 |
| Davies Bauldin | 1.185 |

Centroid Table:-l

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 |
| No. of brands | 0.547 | 0.253 | 0.211 | 0.358 |
| Brand runs | 0.349 | 0.108 | 0.163 | 0.228 |
| Total volume | 0.250 | 0.194 | 0.187 | 0.595 |
| No of Transactions | 0.320 | 0.156 | 0.176 | 0.330 |
| Values | 0.243 | 0.152 | 0.172 | 0.539 |
| Others 999 | 0.574 | 0.203 | 0.824 | 0.401 |

**Centroid Graph:-**

****

**Cluster Model Based on Purchase Behavior (K=5):**

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.610 |
| Cluster 0 | Cluster 2 | 0.545 |
| Cluster 0 | Cluster 3 | 0.594 |
| Cluster 0 | Cluster 4 | 0.751 |
| Cluster 1 | Cluster 2 | 0.400 |
| Cluster 1 | Cluster 3 | 0.819 |
| Cluster 1 | Cluster 4 | 0.570 |
| Cluster 2 | Cluster 3 | 0.473 |
| Cluster 2 | Cluster 4 | 0.484 |
| Cluster 3 | Cluster 4 | 0.646 |

Cluster Model Distance between Clusters

|  |  |
| --- | --- |
| Cluster 0 | 35 Items |
| Cluster 1 | 60 Items |
| Cluster 2 | 147 Items |
| Cluster 3 | 170 Items |
| Cluster 4 | 188 Items |

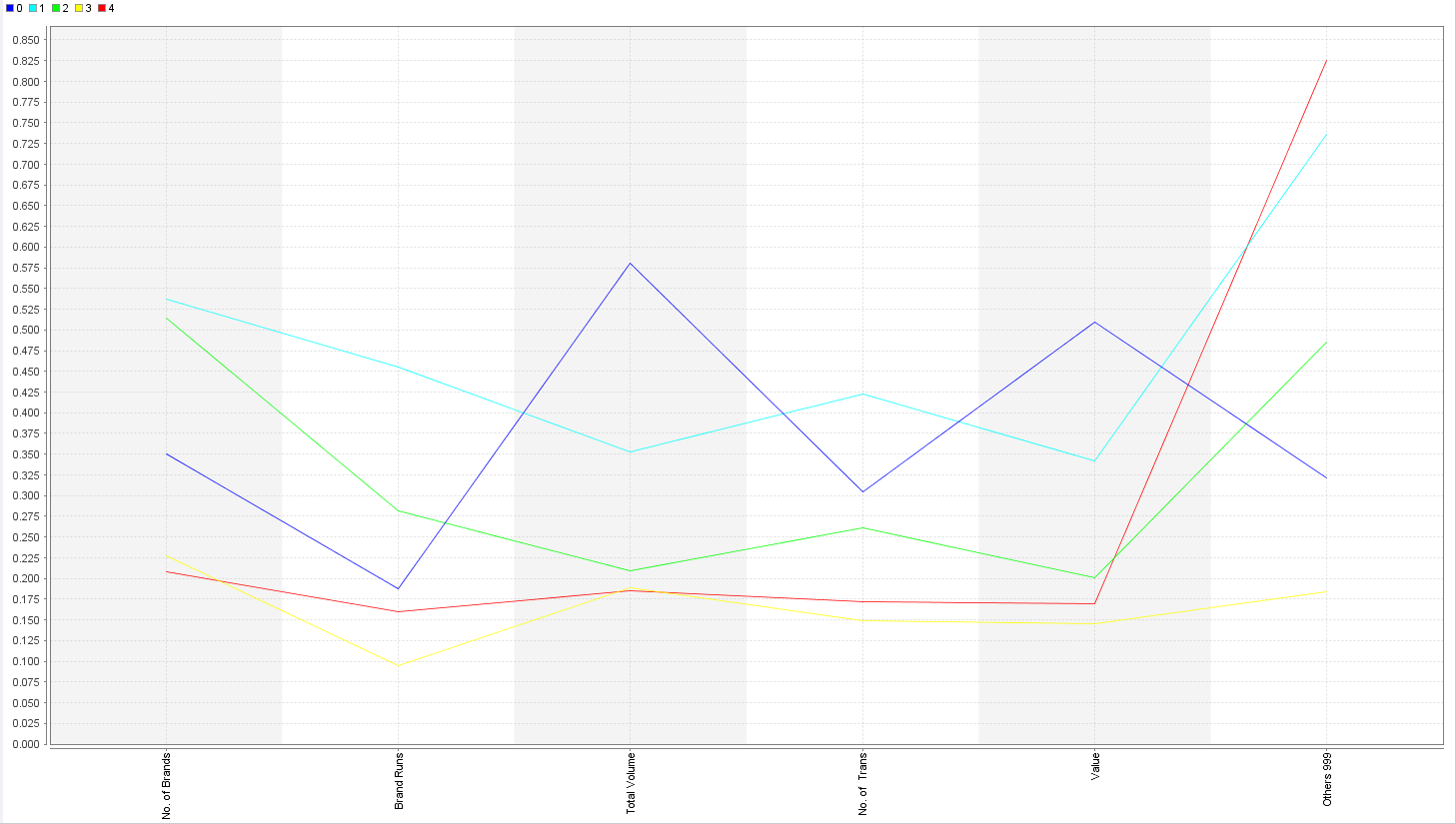
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.080 |
| Avg. within centroid distance cluster 0 | 0.148 |
| Avg. within centroid distance cluster 1 | 0.138 |
| Avg. within centroid distance cluster 2 | 0.064 |
| Avg. within centroid distance cluster 3 | 0.064 |
| Avg. within centroid distance cluster 4 | 0.075 |
| Davies Bauldin | 1.212 |

Centroid Table:-

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| No. of brands | 0.350 | 0.537 | 0.514 | 0.227 | 0.207 |
| Brand runs | 0.188 | 0.456 | 0.281 | 0.094 | 0.159 |
| Total volume | 0.581 | 0.352 | 0.209 | 0.189 | 0.185 |
| No of Transactions | 0.305 | 0.422 | 0.261 | 0.149 | 0.172 |
| Values | 0.509 | 0.342 | 0.201 | 0.145 | 0.169 |
| Others 999 | 0.321 | 0.736 | 0.485 | 0.184 | 0.826 |

**Centroid Graph:-**

****

**b) Selling Propositions:-**

Selling Propositions like Pur Vol Other Promo, PropCat15, PropCat13, PropCat12, PropCat11, PropCat10, and PropCat9 have less significance in data as they contain less than 3% Information within the category.

**Cluster Model Based on the Basis of Purchase (K=3):**

Cluster Model

|  |  |
| --- | --- |
| Cluster 0 | 143 Items |
| Cluster 1 | 378 Items |
| Cluster 2 | 79 Items |

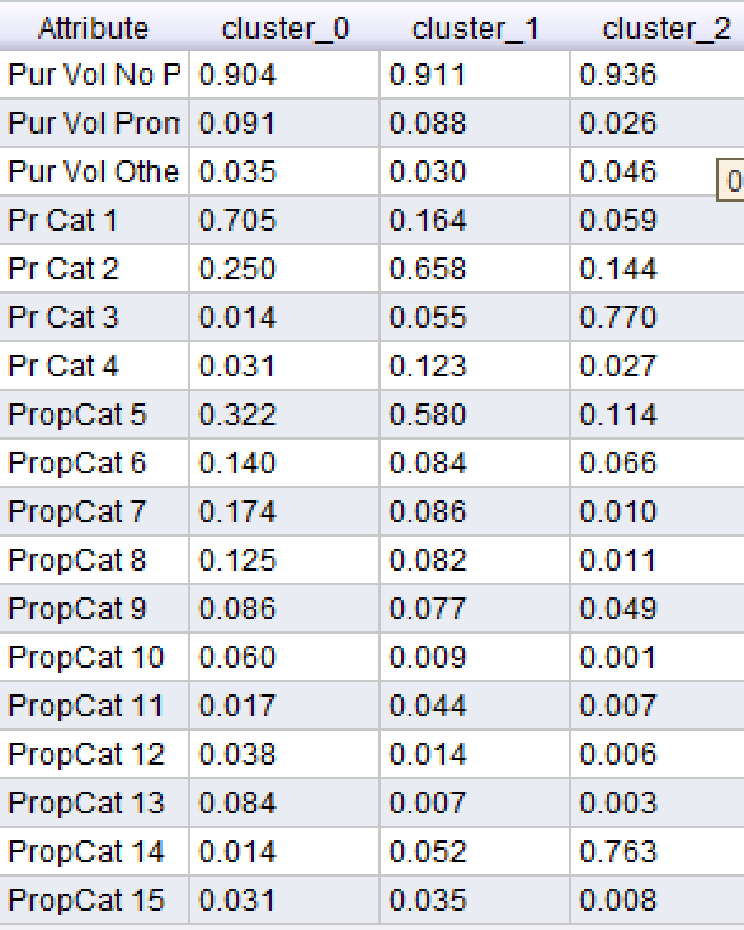
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.390 |
| Avg. within centroid distance cluster 0 | 0.452 |
| Avg. within centroid distance cluster 1 | 0.411 |
| Avg. within centroid distance cluster 2 | 0.176 |
| Davies Bauldin | 1.357 |

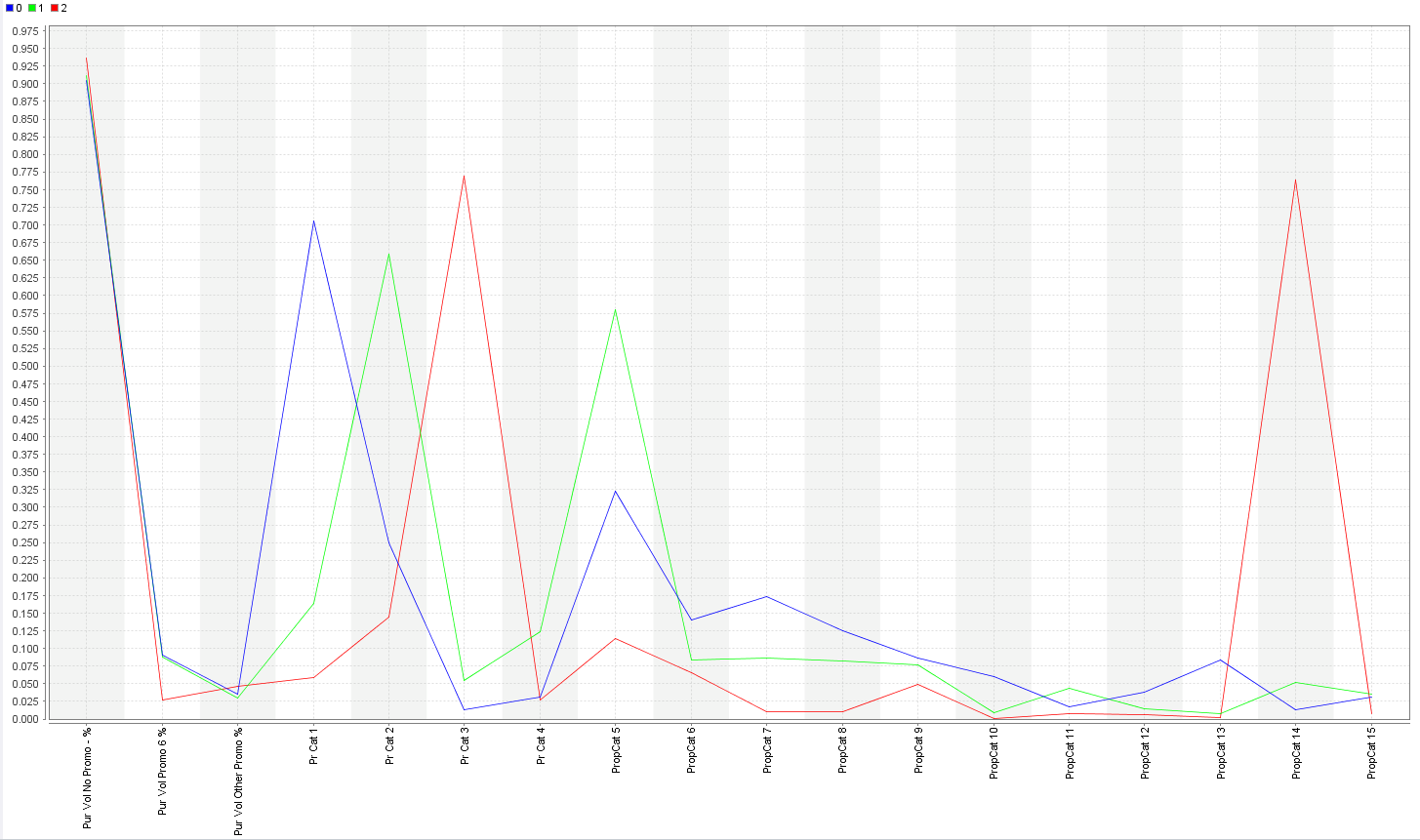
Distance between clusters

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.749 |
| Cluster 0 | Cluster 2 | 1.292 |
| Cluster 1 | Cluster 2 | 1.240 |

**Centroid Table:-**

****

**Centroid Graph:-**

****

**Cluster Model Based on the Basis of Purchase (K=4):**

Cluster Model

|  |  |
| --- | --- |
| Cluster 0 | 235 Items |
| Cluster 1 | 170 Items |
| Cluster 2 | 75 Items |
| Cluster 3 | 120 Items |

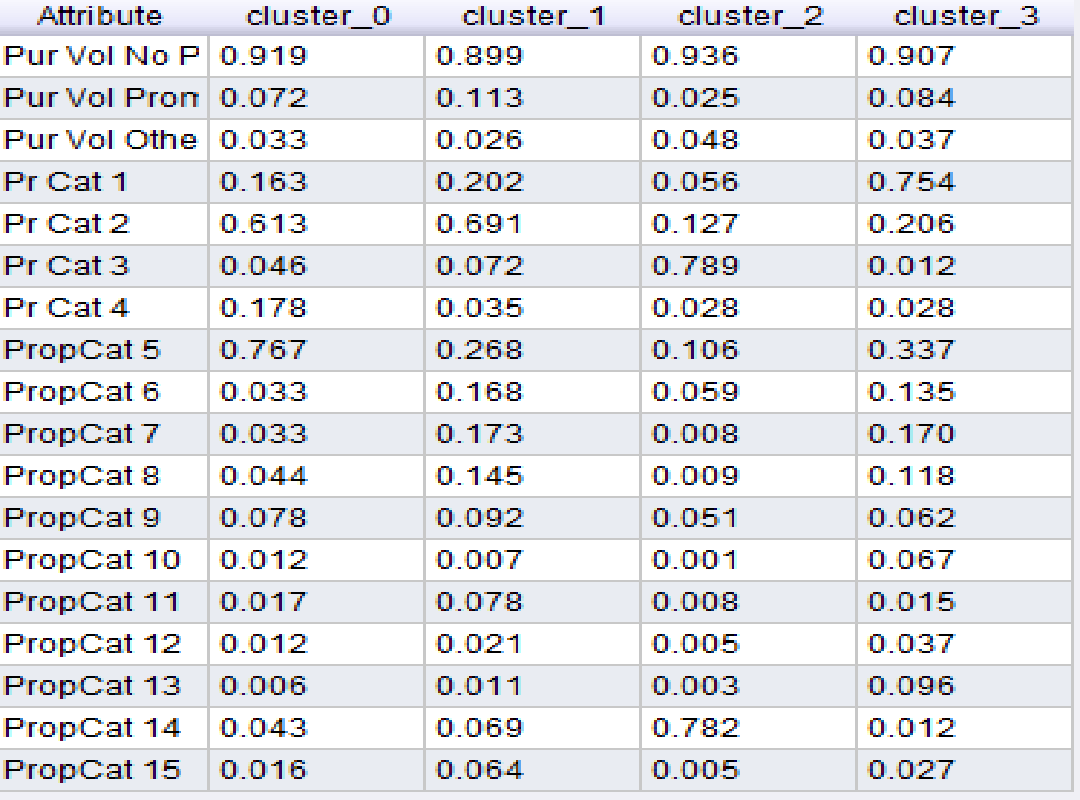
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.338 |
| Avg. within centroid distance cluster 0 | 0.101 |
| Avg. within centroid distance cluster 1 | 0.071 |
| Avg. within centroid distance cluster 2 | 0.078 |
| Avg. within centroid distance cluster 3 | 0.186 |
| Davies Bauldin | 1.563 |

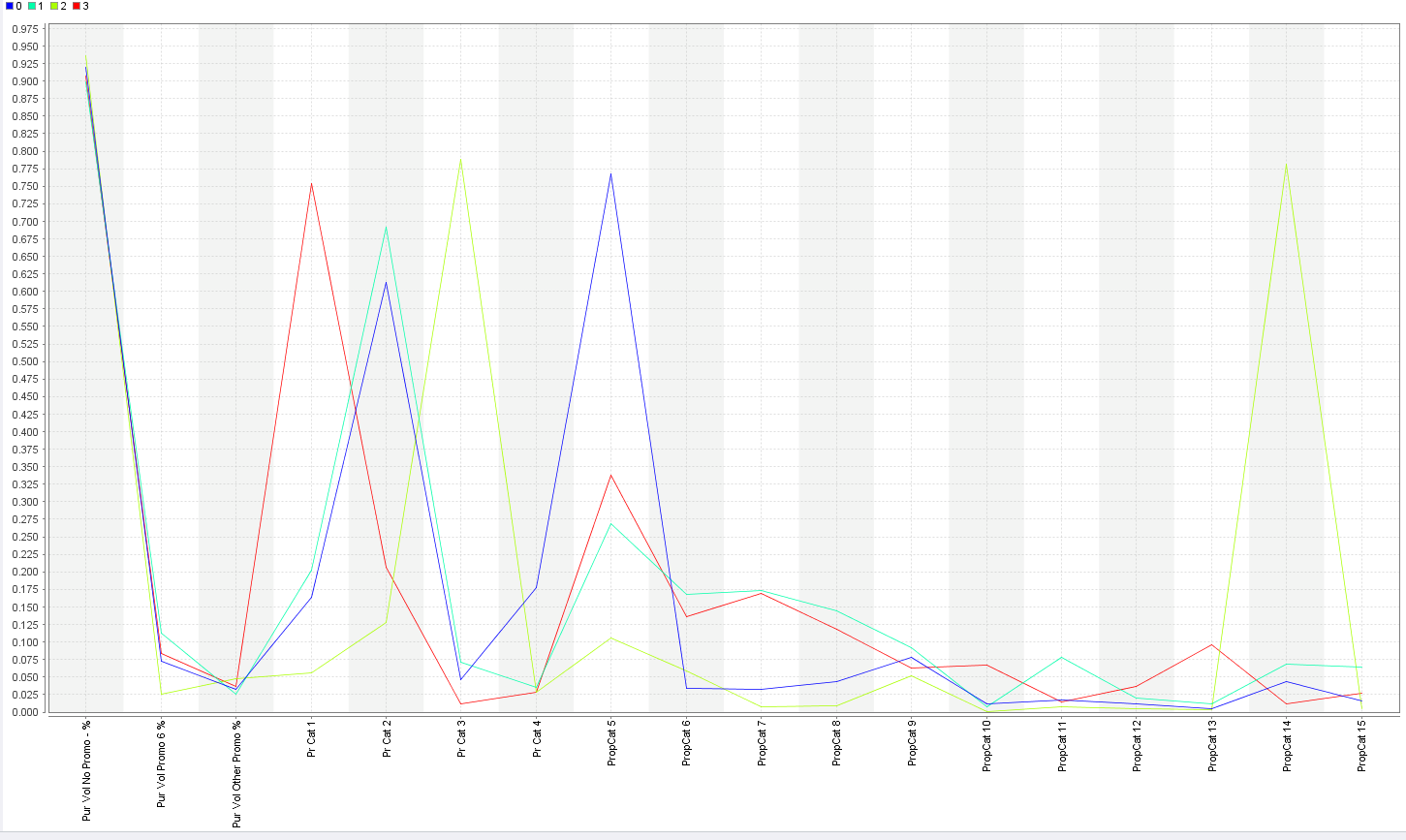
Distance between clusters

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.566 |
| Cluster 0 | Cluster 2 | 0.489 |
| Cluster 0 | Cluster 3 | 0.535 |
| Cluster 1 | Cluster 2 | 0.626 |
| Cluster 1 | Cluster 3 | 0.637 |
| Cluster 2 | Cluster 3 | 0.728 |

**Centroid Table:-**

****

**Centroid Graph:-**

****

**Cluster Model Based on the Basis of Purchase (K=5):**

Cluster Model

|  |  |
| --- | --- |
| Cluster 0 | 168 Items |
| Cluster 1 | 75 Items |
| Cluster 2 | 184 Items |
| Cluster 3 | 118 Items |
| Cluster 4 | 55 Items |

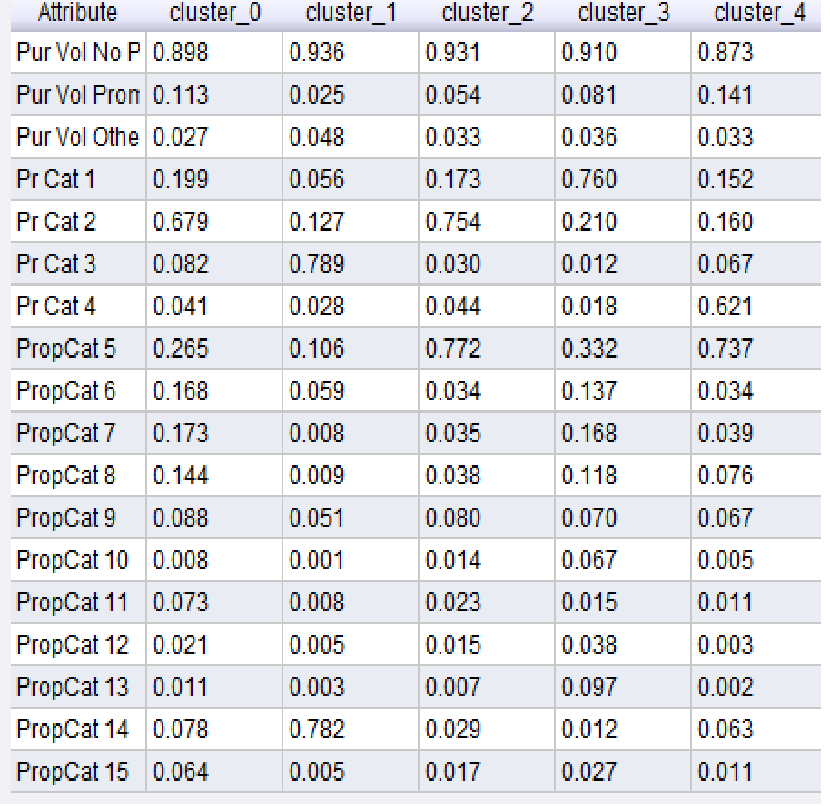
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.290 |
| Avg. within centroid distance cluster 0 | 0.402 |
| Avg. within centroid distance cluster 1 | 0.161 |
| Avg. within centroid distance cluster 2 | 0.167 |
| Avg. within centroid distance cluster 3 | 0.439 |
| Avg. within centroid distance cluster 4 | 0.221 |
| Davies Bauldin | 1.399 |

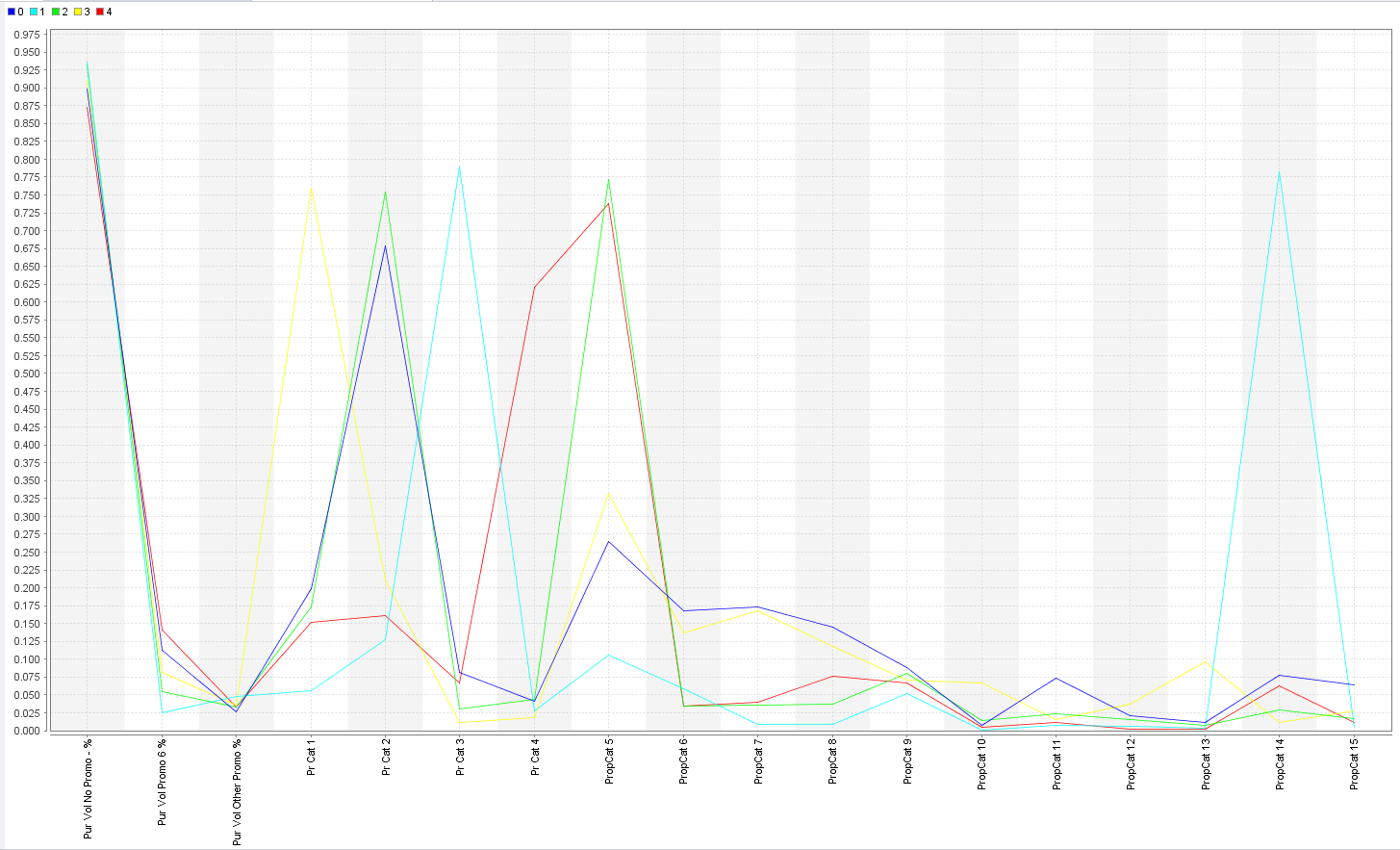
Distance between clusters

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 1.193 |
| Cluster 0 | Cluster 2 | 0.571 |
| Cluster 0 | Cluster 3 | 0.753 |
| Cluster 0 | Cluster 4 | 0.938 |
| Cluster 1 | Cluster 2 | 1.413 |
| Cluster 1 | Cluster 3 | 1.346 |
| Cluster 1 | Cluster 4 | 1.350 |
| Cluster 2 | Cluster 3 | 0.939 |
| Cluster 2 | Cluster 4 | 0.838 |
| Cluster 3 | Cluster 4 | 0.976 |

**Centroid Table:-**

****

**Centroid Graph:-**

****

**Cluster Model Based on Purchase Behavior & Basis of Purchase (K=3):**

Cluster Model

|  |  |
| --- | --- |
| Cluster 0 | 194 Items |
| Cluster 1 | 331 Items |
| Cluster 2 | 75 Items |

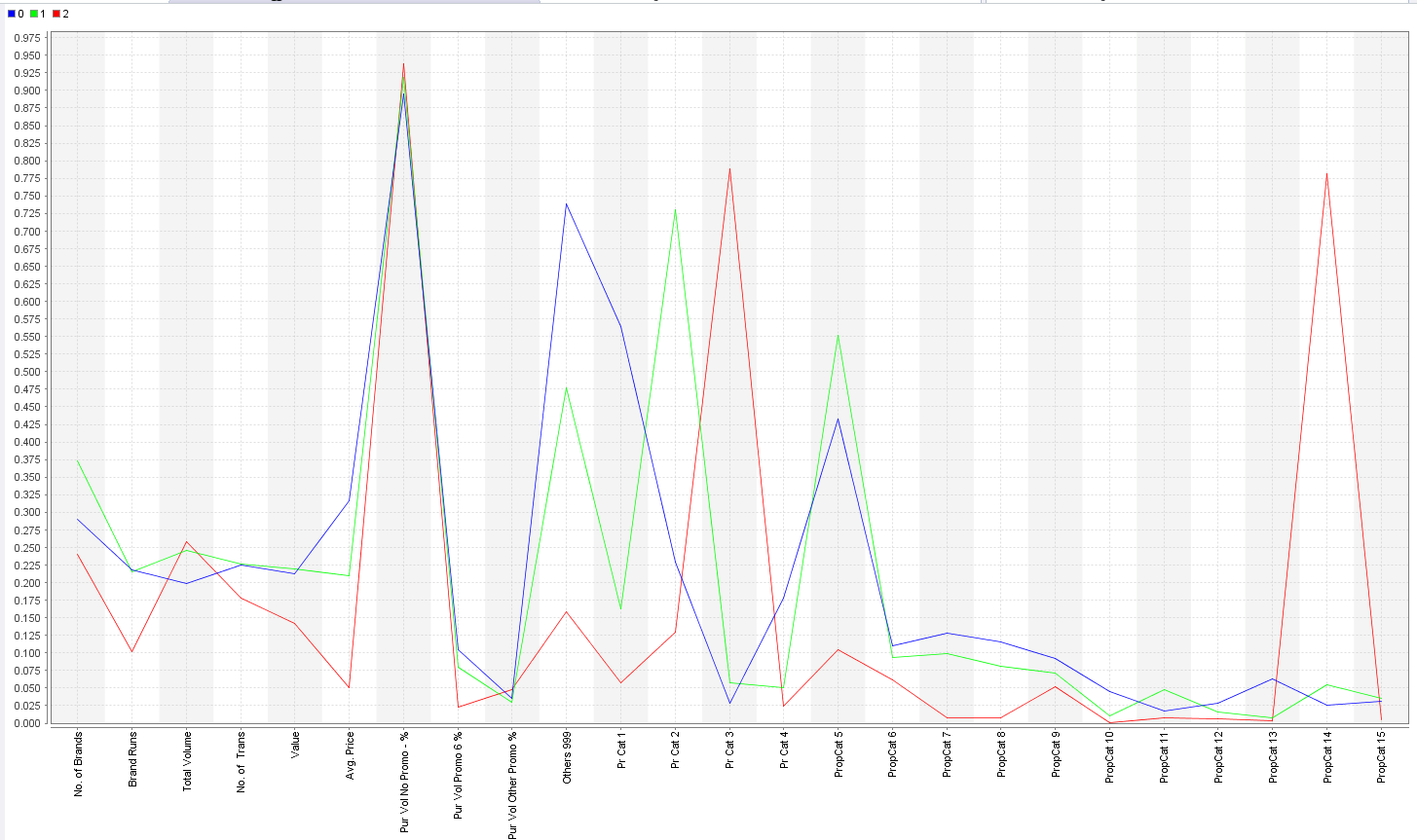
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.574 |
| Avg. within centroid distance cluster 0 | 0.761 |
| Avg. within centroid distance cluster 1 | 0.534 |
| Avg. within centroid distance cluster 2 | 0.267 |
| Davies Bauldin | 1.695 |

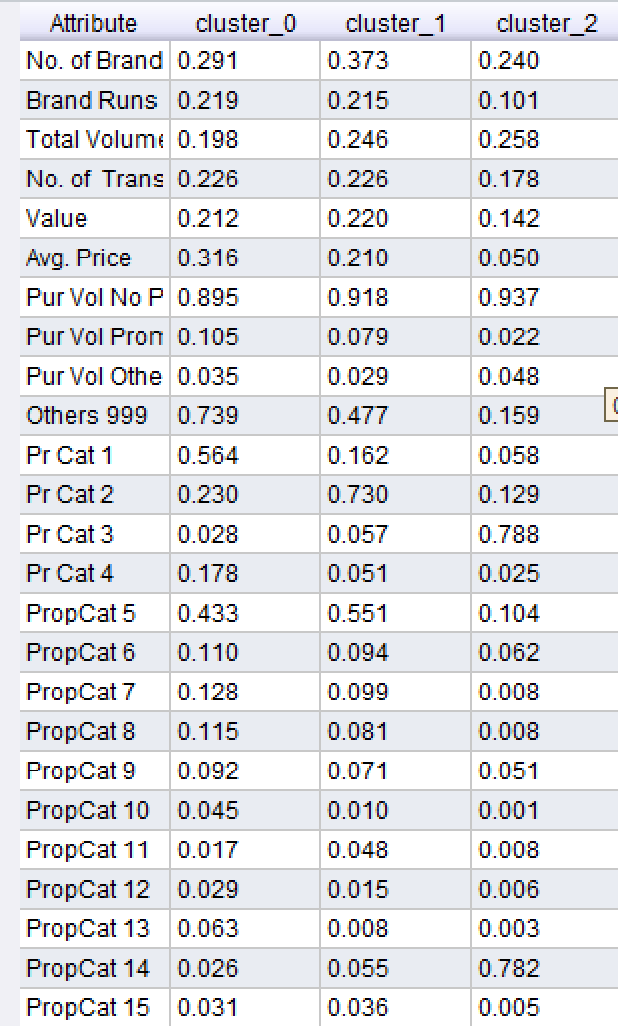
Distance between clusters

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.736 |
| Cluster 0 | Cluster 2 | 1.423 |
| Cluster 1 | Cluster 2 | 1.350 |

**Centorid Graph:-**

****

**Centroid Table:-**

****

**Cluster Model Based on Purchase Behavior & Basis of Purchase (K=4):**

Cluster Model

|  |  |
| --- | --- |
| Cluster 0 | 136 Items |
| Cluster 1 | 62 Items |
| Cluster 2 | 75 Items |
| Cluster 3 | 327 Items |

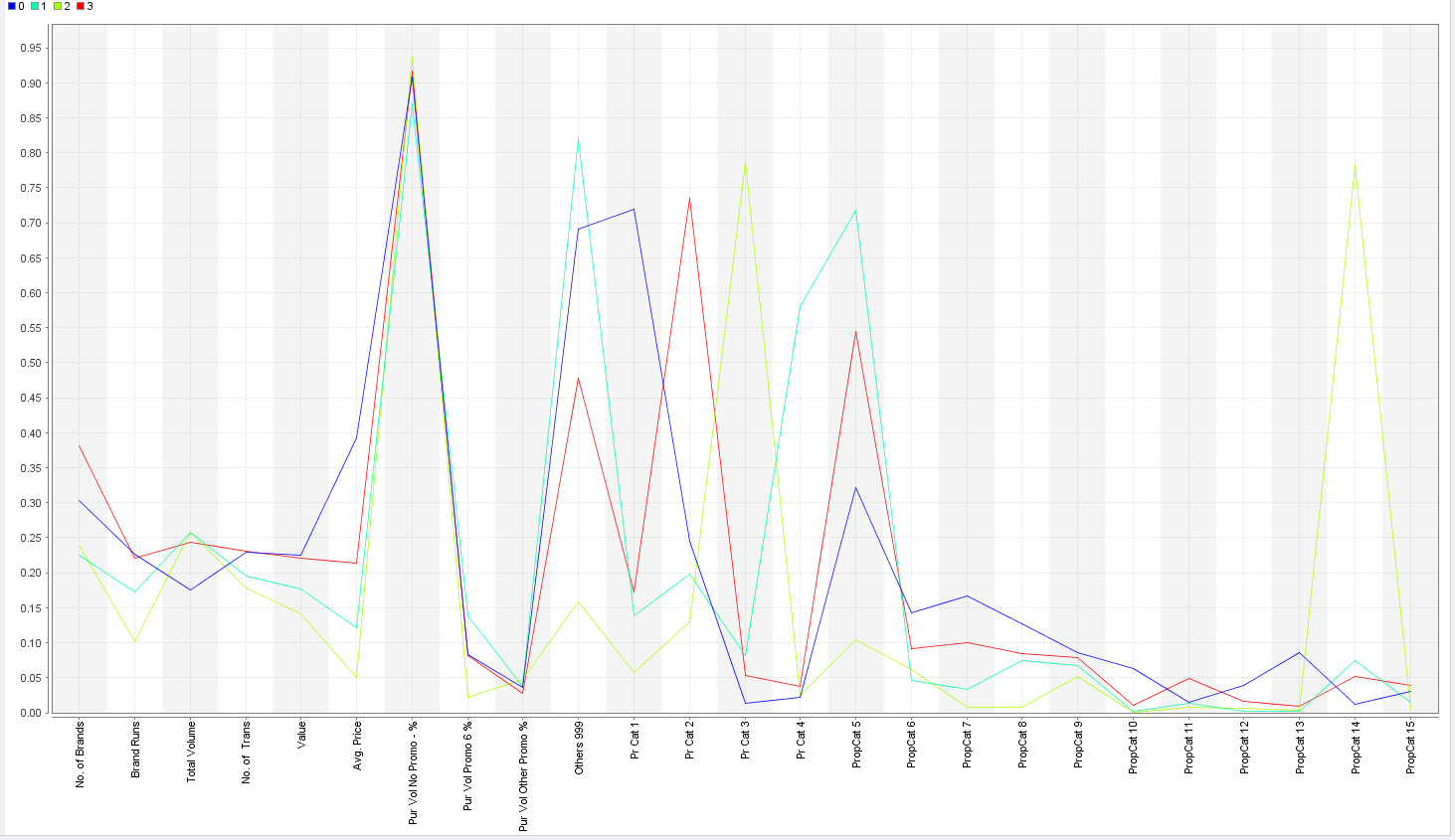
Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.509 |
| Avg. within centroid distance cluster 0 | 0.667 |
| Avg. within centroid distance cluster 1 | 0.357 |
| Avg. within centroid distance cluster 2 | 0.267 |
| Avg. within centroid distance cluster 3 | 0.528 |
| Davies Bauldin | 1.453 |

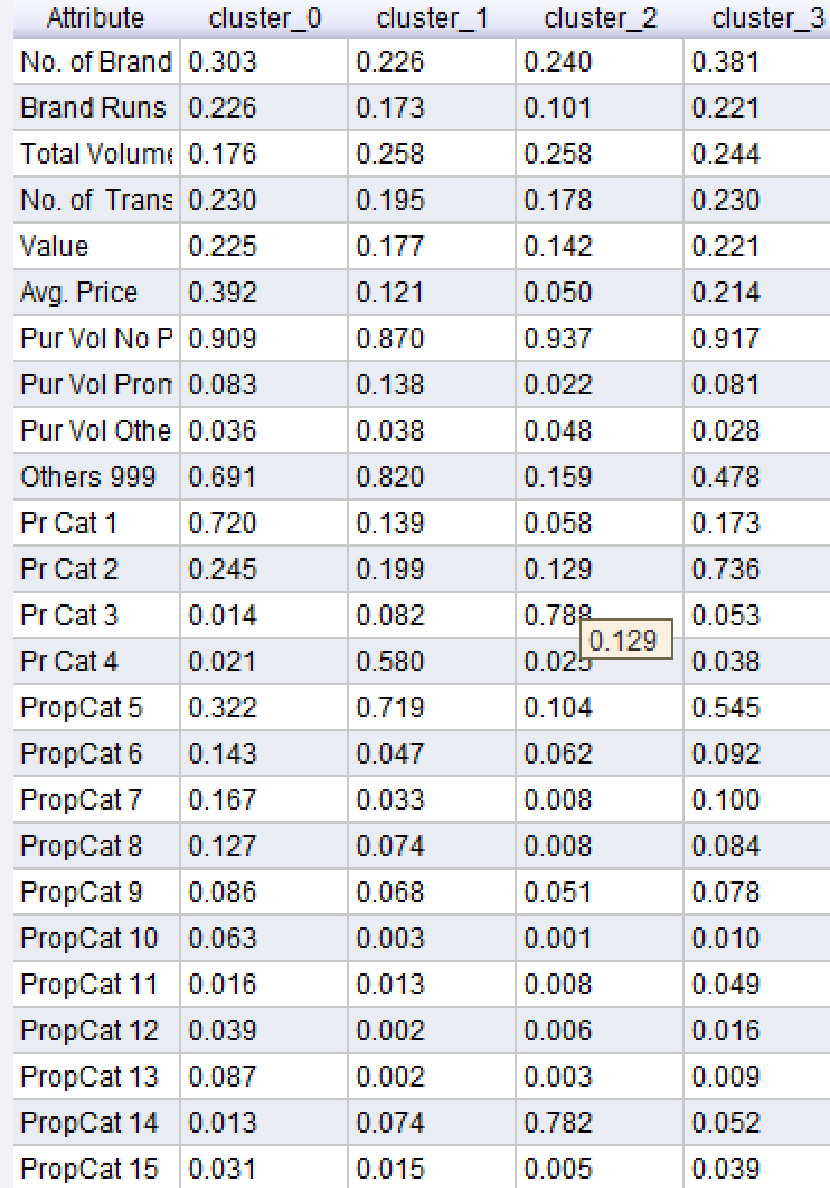
Distance between clusters

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 0.987 |
| Cluster 0 | Cluster 2 | 1.480 |
| Cluster 0 | Cluster 3 | 0.837 |
| Cluster 1 | Cluster 2 | 1.473 |
| Cluster 1 | Cluster 3 | 0.885 |
| Cluster 2 | Cluster 3 | 1.358 |

**Centroid Graph:-**

****

**Centroid Table:-**

****

**Cluster Model Based on Purchase Behavior & Basis of Purchase (K=5):**

Cluster Model

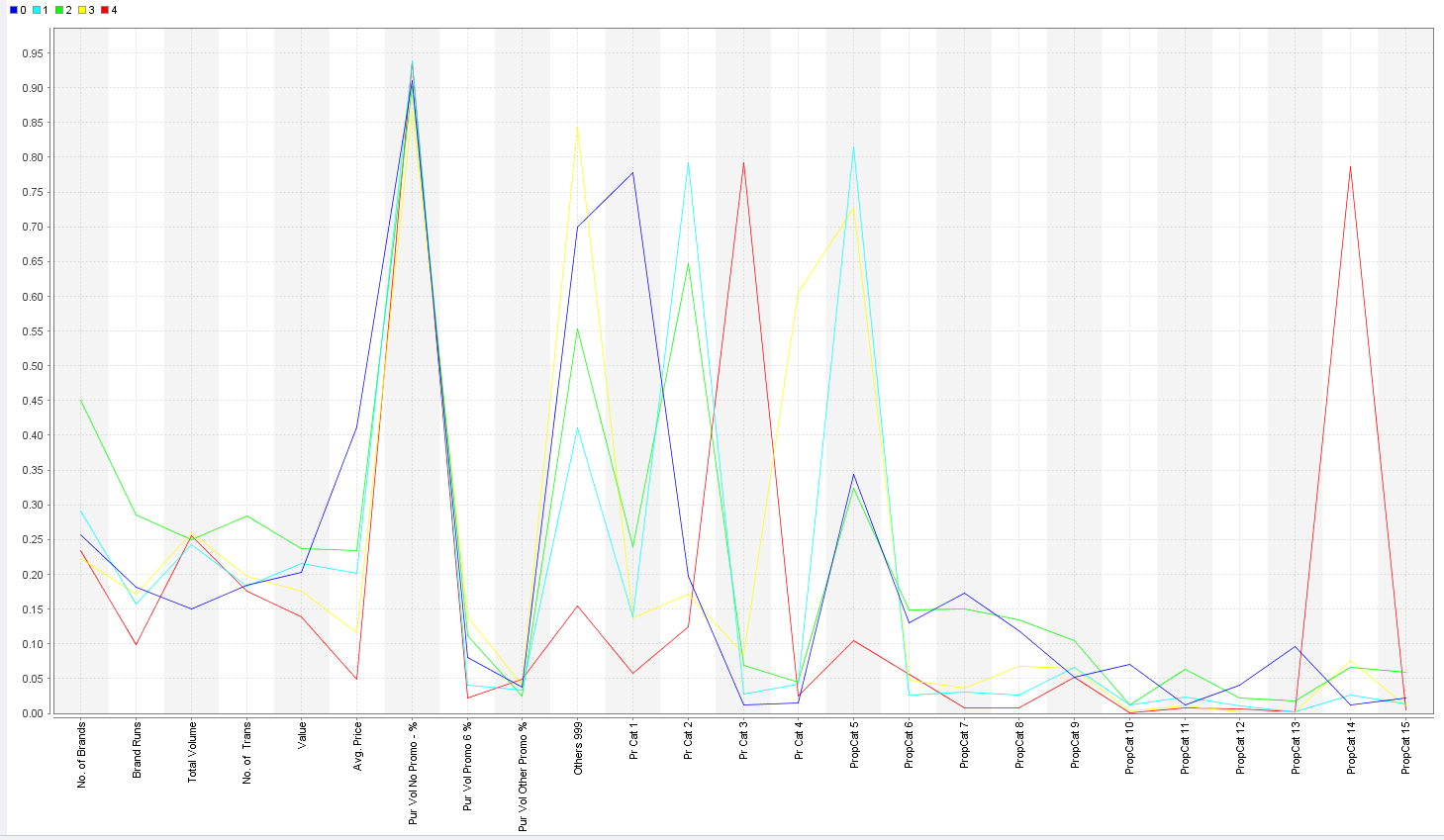
|  |  |
| --- | --- |
| Cluster 0 | 107 Items |
| Cluster 1 | 146 Items |
| Cluster 2 | 217 Items |
| Cluster 3 | 56 Items |
| Cluster 4 | 74 Items |

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 1.085 |
| Cluster 0 | Cluster 2 | 0.801 |
| Cluster 0 | Cluster 3 | 1.042 |
| Cluster 0 | Cluster 4 | 1.518 |
| Cluster 1 | Cluster 2 | 0.635 |
| Cluster 1 | Cluster 3 | 0.968 |
| Cluster 1 | Cluster 4 | 1.491 |
| Cluster 2 | Cluster 3 | 0.962 |
| Cluster 2 | Cluster 4 | 1.325 |
| Cluster 3 | Cluster 4 | 1.499 |

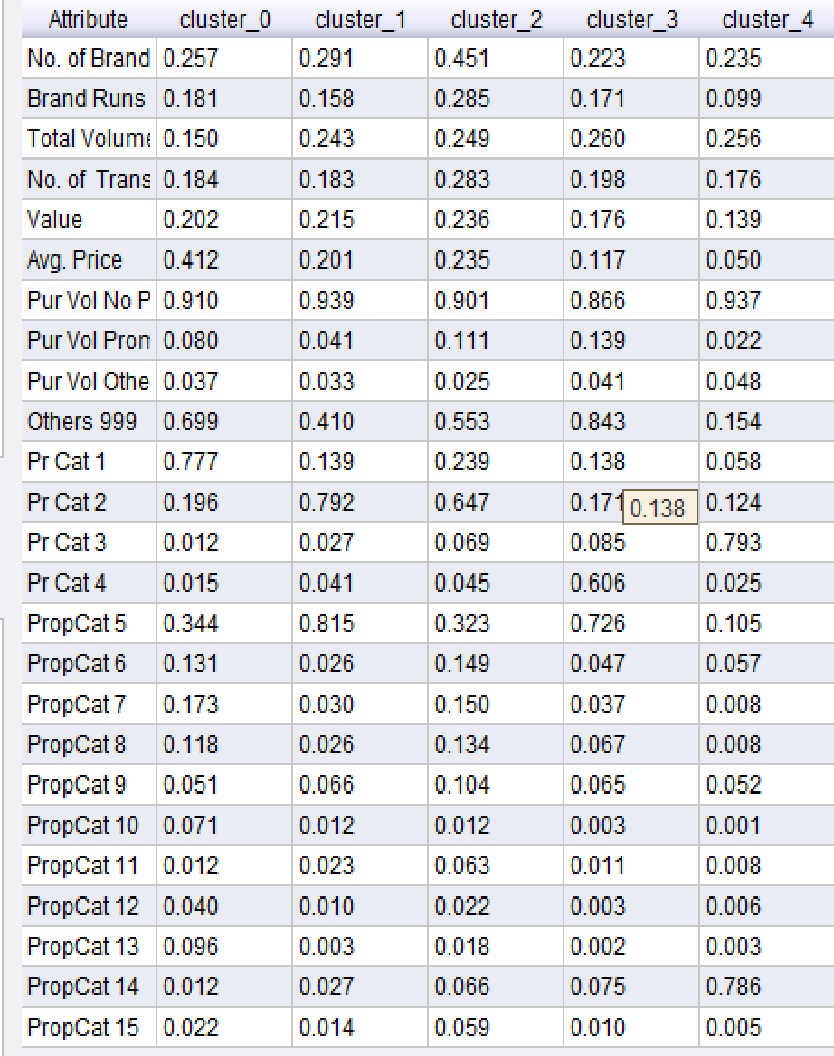
Performance Vector Distance between Clusters

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.458 |
| Avg. within centroid distance cluster 0 | 0.643 |
| Avg. within centroid distance cluster 1 | 0.313 |
| Avg. within centroid distance cluster 2 | 0.563 |
| Avg. within centroid distance cluster 3 | 0.344 |
| Avg. within centroid distance cluster 4 | 0.258 |
| Davies Bauldin | 1.598 |

**Centroid Graph:-**

****

**Centroid Table:-**

****

**Comparison Table Based on Purchase Behavior:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Avg. Within Centroid Distance** | **Davis Bouldin Index** | **Distance between Clusters** |
| **K=3** | 0.101 | 1.166 | 0.494,0.625,0.553 |
| **K=4** | 0.089 | 1.185 | 0.566,0.489,0.626,0.637,0.728,0.535 |
| **K=5** | 0.080 | 1.212 | 0.610,0.594,0.545,0.751,0.400,0.819,0.570  0.473,0.484,0.646 |

Observation:-

K=5 Model is best as it has high cluster density & Higher distance between the Clusters.

**Comparison Table Based on Basis of Purchase:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Avg. Within Centroid Distance** | **Davis Bouldin Index** | **Distance between Clusters** |
| **K=3** | 0.39 | 1.357 | 0.749, 1.292, 1.240 |
| **K=4** | 0.338 | 1.563 | 0.566,0.489,0.535,0.626,0.637,0.728 |
| **K=5** | 0.29 | 1.399 | 1.193,0.571,0.753,0.938,1.413,1.346,1.350,  0.939,0.838,0.976 |

Observation:-

K=5 Model is best as it has high cluster density & Higher distance between the Clusters.

**Comparison Table Based on Purchase Behavior & Basis of Purchase:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Avg. Within Centroid Distance** | **Davis Bouldin Index** | **Distance between Clusters** |
| **K=3** | 0.574 | 1.695 | 0.736,1.423,1.350 |
| **K=4** | 0.509 | 1.453 | 0.987,1.480,0.837,1.473,0.885,1.358 |
| **K=5** | 0.458 | 1.598 | 1.085,0.801,1.042,1.518,0.635,0.96,1.491,  0.962,1.325,1.499 |

Observation:-

K=5 Model is best as it has high cluster density & Higher distance between the Clusters.

**d) Considering purchase behavior case for Comparison with other cluster models:-**

**i) k-mediods:**

Cluster Model

|  |  |
| --- | --- |
| Cluster 0 | 128 items |
| Cluster 1 | 193 items |
| Cluster 2 | 176 items |
| Cluster 3 | 94 items |
| Cluster 4 | 9 Items |

Performance Vector

|  |  |
| --- | --- |
| Avg. within centroid distance | 0.477 |
| Avg. within centroid distance cluster 0 | 0.177 |
| Avg. within centroid distance cluster 1 | 0.683 |
| Avg. within centroid distance cluster 2 | 0.472 |
| Avg. within centroid distance cluster 3 | 0.492 |
| Avg. within centroid distance cluster 4 | 0.280 |
| Davies Bauldin | 1.491 |

Distance between Clusters

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | 1.199 |
| Cluster 0 | Cluster 2 | 0.705 |
| Cluster 0 | Cluster 3 | 1.263 |
| Cluster 0 | Cluster 4 | 1.421 |
| Cluster 1 | Cluster 2 | 0.837 |
| Cluster 1 | Cluster 3 | 0.923 |
| Cluster 1 | Cluster 4 | 1.374 |
| Cluster 2 | Cluster 3 | 0.837 |
| Cluster 2 | Cluster 4 | 1.297 |
| Cluster 3 | Cluster 4 | 1.175 |

**ii) DBSCAN:**

For Epsilon=1.0 & min Points=5

|  |  |
| --- | --- |
| Cluster 0 | 0 items |
| Cluster 1 | 600 items |

For Epsilon=0.1 & min Points=5

|  |  |
| --- | --- |
| Cluster 0 | 581 items |
| Cluster 1 | 6 items |
| Cluster 2 | 5 items |
| Cluster 3 | 8 items |

For Epsilon=0.2 & min Points=3

|  |  |
| --- | --- |
| Cluster 0 | 92 Items |
| Cluster 1 | 497 Items |
| Cluster 2 | 5 Items |
| Cluster 4 | 3 Items |
| Cluster 5 | 3 Items |

From The modeling technique used above by changing parameters there will be changes in number of clusters, density and distance between the clusters.

From above results we can observe that there is no proper balance for the number of items in a particular cluster.

In this case by comparing K-Mediods & DBSCAN with K-Means:-

Based on Proper balance of clusters, Density within clusters & distance between the clusters.

We can conclude that K-Mean model performance is better than other 2 models.

**2.** (a) Select what you think is the 'best' segmentation - explain why you think this is the ‘best’. You can also decide on multiple segmentations, based on different criteria -- for example, based on purchase behavior, or basis for purchase,....( think about how different clusters may be useful). (b) Comment on the characteristics (demographic, brand loyalty and basis-for-purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)

1. Best segmentation:-

Criterion for Selection.

* Purchase Behavior
* Basis of Purchase
* Both

Taking in to account the best cases from all the criterions:-

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=5 For Purchase Behavior | K=5 For Basis of Purchase | K=5 For Both |
| Avg. Within Centroid Distance | 0.080 | 0.29 | 0.458 |
| Davis Bouldin Index | 1.212 | 1.399 | 1.598 |
| Distance between Clusters | 0.610,0.594,0.545,0.751,0.400,0.819,0.570  0.473,0.484,0.646 | 1.193,0.571,0.753,0.938,1.413,1.346,1.350,  0.939,0.838,0.976 | 1.085,0.801,1.042,  1.518,0.635,0.96,1.491,  0.962,1.325,1.499 |

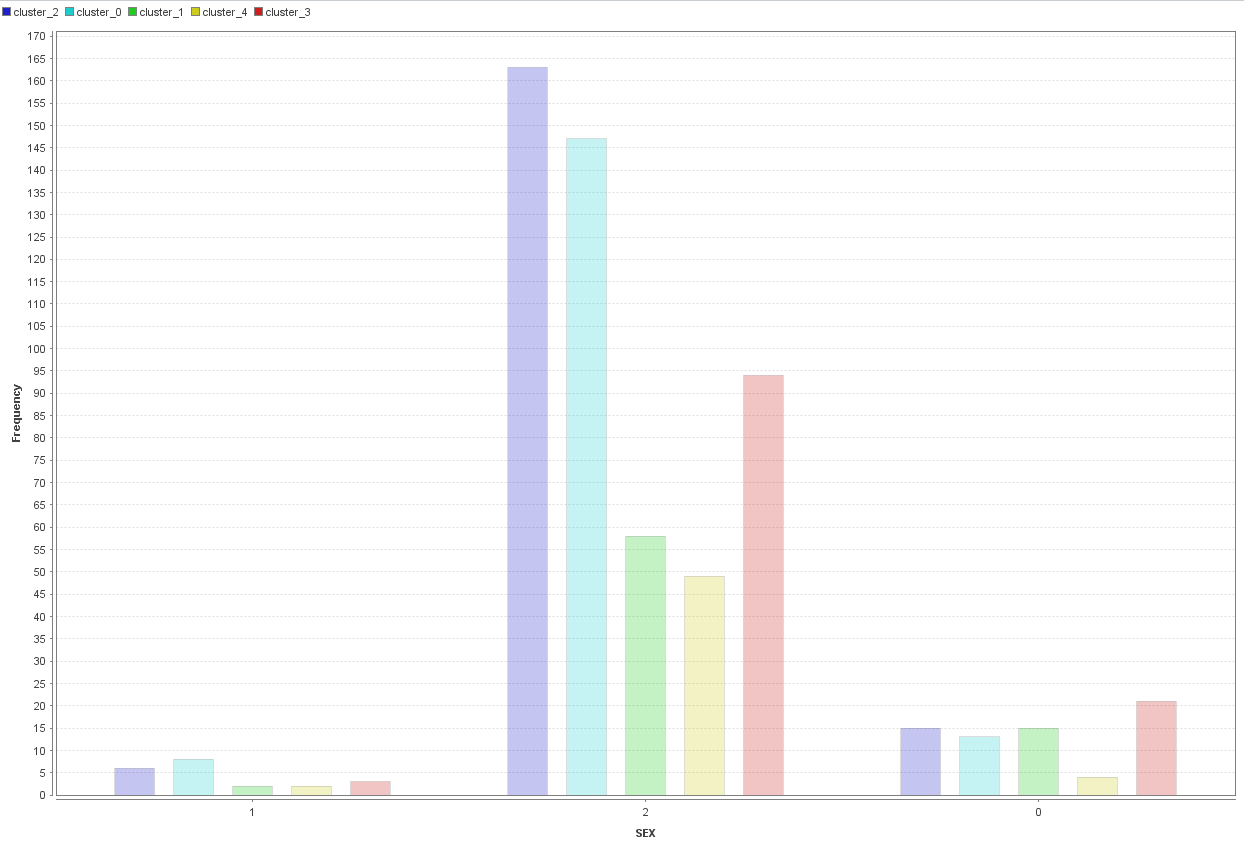
Among 3 best performances, segmentation based on purchase behavior would be the best one as both Avg. Within Centroid Distance and Davies Bouldin Index are the least in Purchase Behavior case and also, distance between pair of clusters is decently high.

**Therefore, K=5 on the basis of Purchase behavior is the best segmentation.**

1. **Analysis of Demographics on Selected best cluster:-**

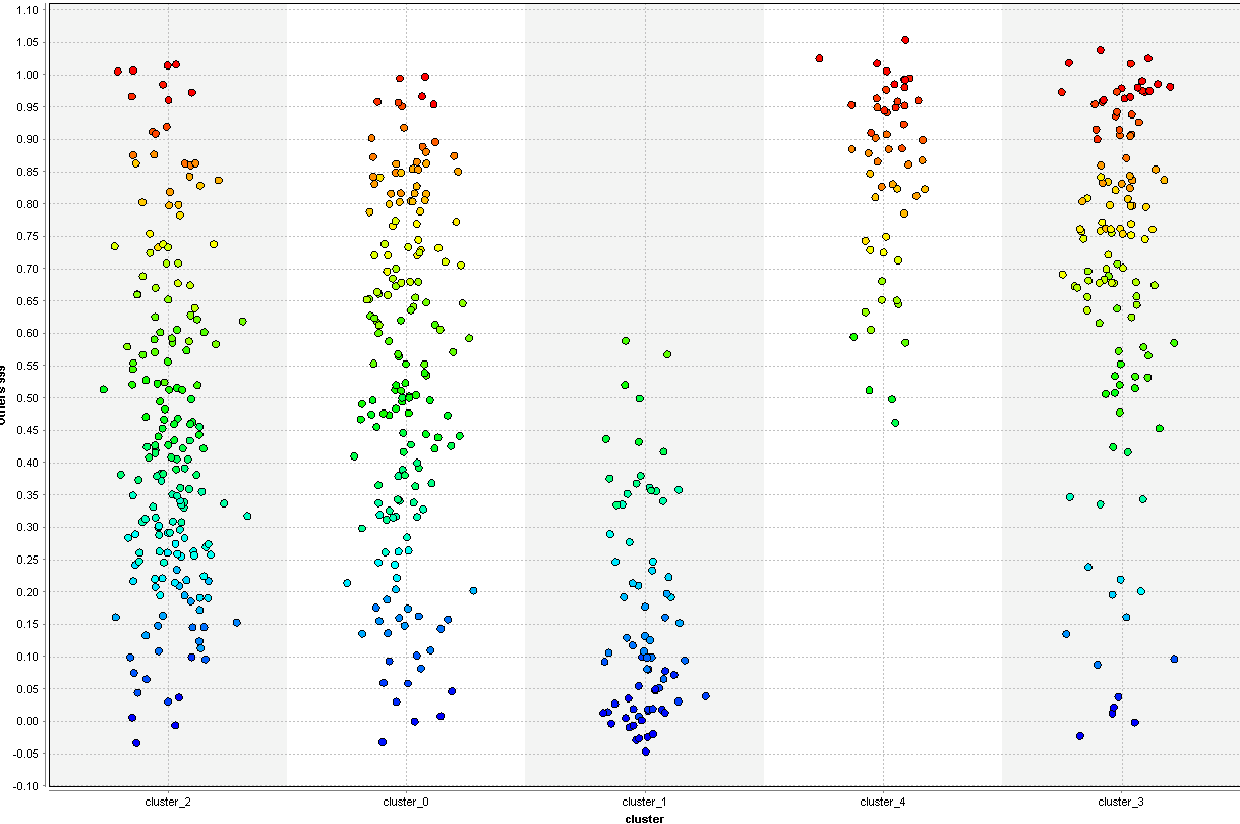
****

Graph above indicates that most customers are of 45+ age .Cluster 0, Cluster 2 & Cluster 3 dominates the Purchase.

****

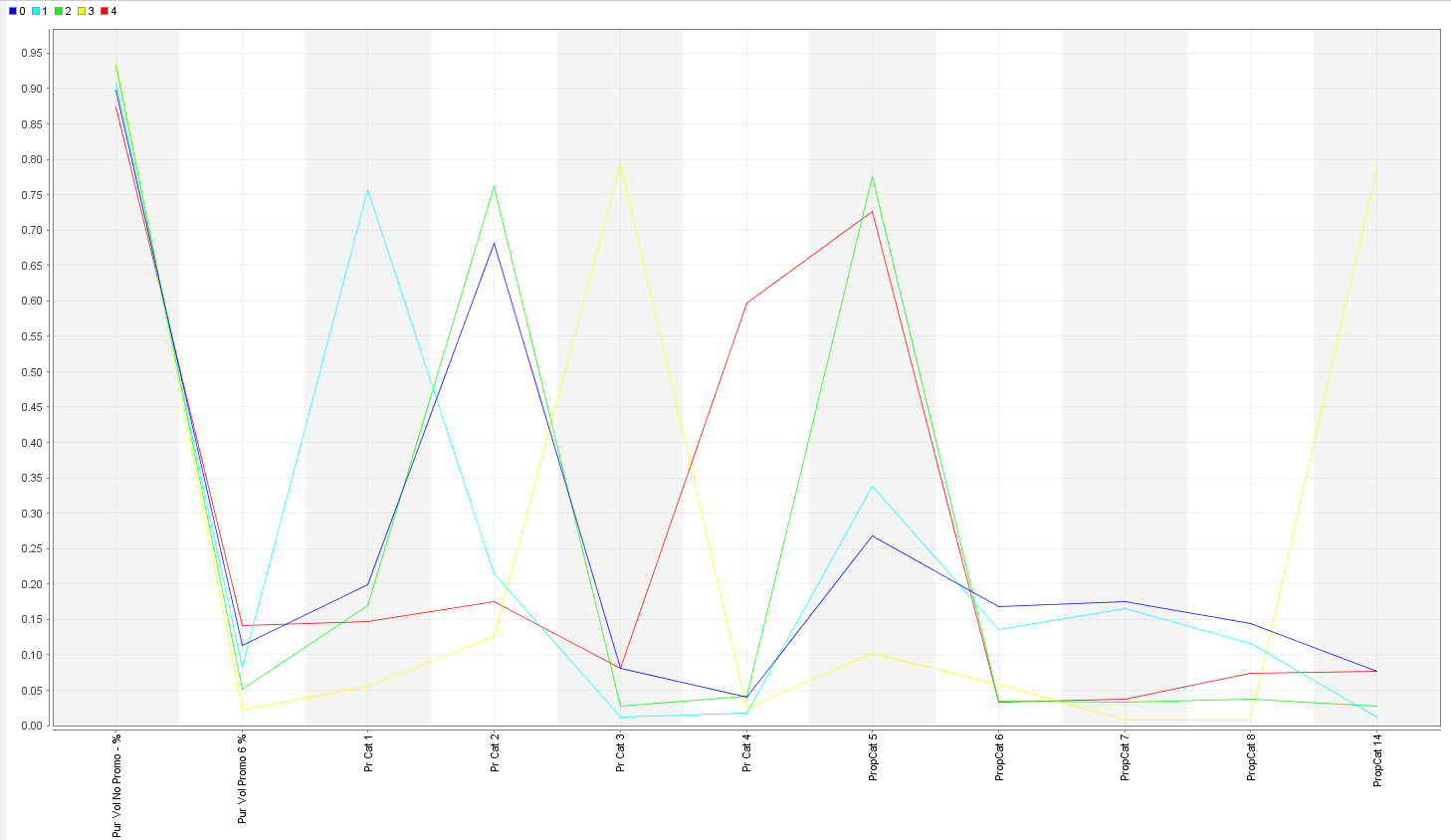
Graph above indicates that most purchasers are females as seen in the graph.

**Analysis of Brand Loyalty on Customers:**

****

We can observe that Purchaser from cluster 2 are most loyal towards the brands they purchase

**Analysis of Basis for Purchase on the Clusters:**

****

We can observe the Purchase basis selection of customers for different selling propositions.

Different color lines indicates the different clusters category.

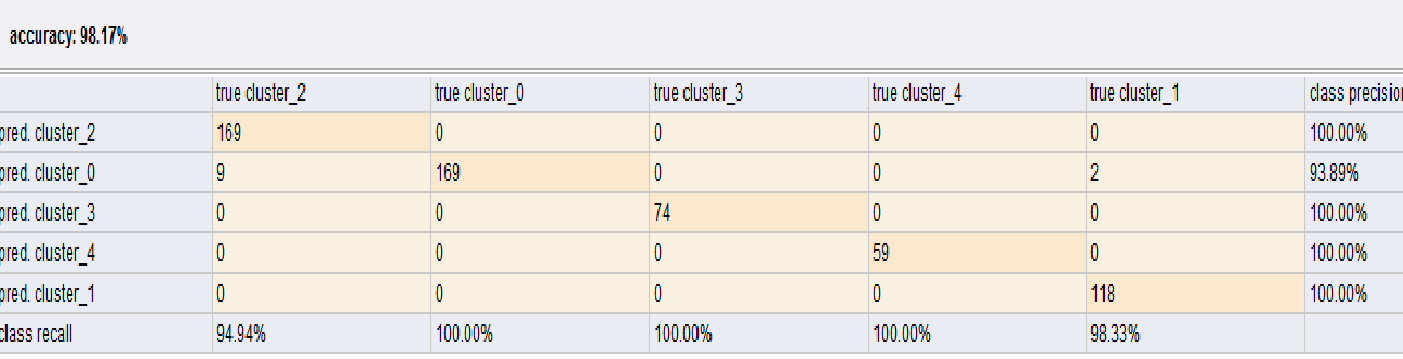
**3.** for the best segmentation, obtain a description of the clusters using a decision tree – how effective is the tree in identifying the different clusters? Does the tree help in explaining/interpreting the different clusters? (Explain why/why not) (You may use a decision tree to help choose the ‘best’ clustering)

By applying a Dicision Tree model for the Performance evaluation

Using Parameters:-

* Gain Ratio
* Depth= 8

Perfomance :-

****

98.17% accuracy and simplified structure of tree makes interpretation and identification of clusters easy. With these cluster categorization we can identify and locate about the particular customer’s behavior and their purchase habits and we can utilize these data categorization for future offers and improve the sales.

**Decision Tree**

