Assignment 4: Market Segmentation

Kompella Kaushik

Adusumelli Jaideep

Sushaanth Srirangapathy

**Preprocessing of data:**

Attribute transformation:

|  |  |
| --- | --- |
| **Attribute** | **Transformation** |
| EDU\_1 | If(EDU==1,1,0) |
| EDU\_2 | If(EDU==2,1,0) |
| EDU\_3 | If(EDU==3,1,0) |
| EDU\_4 | If(EDU==4,1,0) |
| EDU\_5 | If(EDU==5,1,0) |
| EDU\_6 | If(EDU==6,1,0) |
| EDU\_7 | If(EDU==7,1,0) |
| EDU\_8 | If(EDU==8,1,0) |
| EDU\_9 | If(EDU==9,1,0) |
| MT\_English | If(MT==3,1,0) |
| MY\_Gujarati | If(MT==4,1,0) |
| MT\_Hindi | If(MT==5,1,0) |
| MT\_Kannada | If(MT==6,1,0) |
| MT\_Konkani | If(MT==8,1,0) |
| MT\_Malyalam | If(MT==9,1,0) |
| MT\_Marathi | If(MT==10,1,0) |
| MT\_Punjabi | If(MT==12,1,0) |
| MT\_Rajasthani | If(MT==13,1,0) |
| MT\_Sindhi | If(MT==14,1,0) |
| MT\_Tamil | If(MT==15,1,0) |
| MT\_Telugu | If(MT==16,1,0) |
| MT\_Other | If(MT==19,1,0) |
| MT\_Urdu | If(MT==17,1,0) |
| MT\_Not\_Specified | If(MT==0,1,0) |
| Maxpurchase | max(BrCd24,BrCd272,BrCd286,BrCd352,BrCd481,BrCd5,BrCd55,BrCd55And144) |
| CS\_New | if(CS==0||CS==2,0,1) |
| Maxprop | max([PropCat 10],[PropCat 11],[PropCat 13],[PropCat 12],[PropCat 14],[PropCat 15],[PropCat 5],[PropCat 6],[PropCat 7],[PropCat 8],[PropCat 9]) |
| Maxpropnum | if(MaxProp==[PropCat],5,if(MaxProp==[PropCat],6,if(MaxProp==[PropCat],7,if(MaxProp==[PropCat 8],8,if(MaxProp==[PropCat],9,if(MaxProp==[PropCat10],10,if(MaxProp==[PropCat 11],11,if(MaxProp==[PropCat 12],12,if(MaxProp==[PropCat 13],13,if(MaxProp==[PropCat 14],14,if(MaxProp==[PropCat 15],15,0))))))))))) |

“Maxpurchase” attribute was created to explain the Brand Loyalty.

“EDU\_X”, and “MT\_XXXXX” attribute were created to transform the attributes “mother Tongue” and “Education” to binary.

**1. 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?**

**[Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)].**

The variables used by us are:

Number of Brands, Brand Runs, Total Volume, Number of Transactions, Value, Average Price, Maximum Purchase, Others999.

The attribute “Others999” was used to depict the share to other Brands.

Brand Loyalty:

The attribute “Maxpurchase” was created to depict brand Loyalty. “Maxpurchase” is the maximum value of the Brand columns (not including Others999) for a given row and is an indicator of predilection of a household to a given brand. Here, we have not included the individual brand purchase percentages to cover for the fact that we are considering brand loyalty and not brand loyalty to a given brand. Among the above attributes which were used in this process – “Others999” depicts the lack of loyalty, hence this also should be used for calculating Brand Loyalty. Number of Brands in the purchase indicates to some extent the lack of brand loyalty and we consider the temporal nature of brand loyalty by including Brand Runs, the consecutive buying runs for a given set of brands.

Given below are the performances for different clusters with k values ranging from 2 to 5, generated using K-means Clustering technique.

For k=2:

Performance Matrix:

Avg. within centroid distance: -0.180

Avg. within centroid distance\_cluster\_0: -0.187

Avg. within centroid distance\_cluster\_1: -0.169

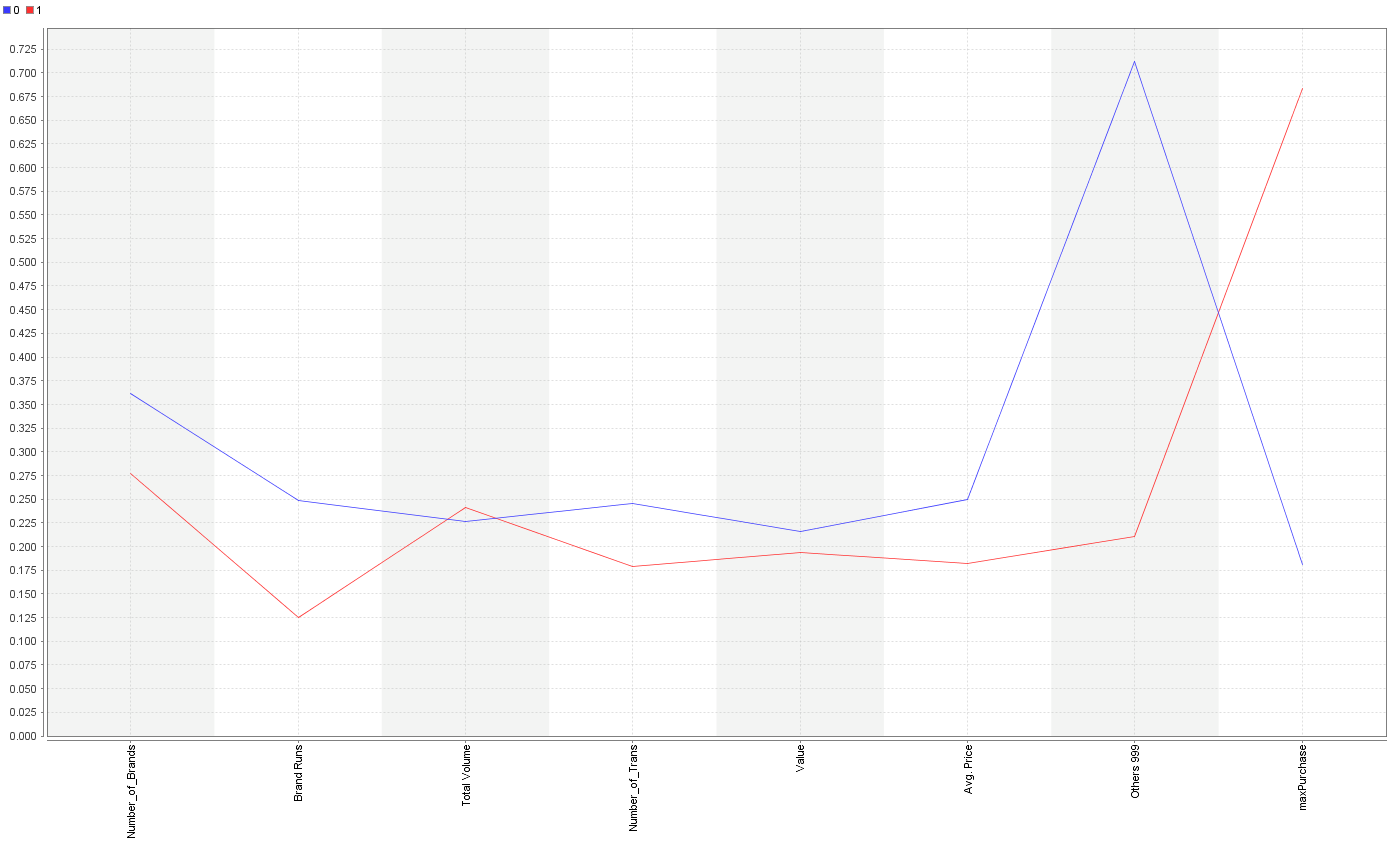
Davies Bouldin: -1.062

Size of Clusters: (373,227)

**Inter Cluster Distance**

|  |  |  |
| --- | --- | --- |
| **First** | **Second** | **Inter Cluster Distance** |
| 1.0 | 2.0 | 0.731 |

The below graph shows the dependence of the attributes on the differences between the clusters. From the graph, it is clear that most of the difference is contributed by others999 and maxpurchase attributes.



For k=3:

Performance Matrix:

Avg. within centroid distance: -0.140

Avg. within centroid distance\_cluster\_0: -0.142

Avg. within centroid distance\_cluster\_1: -0.160

Avg. within centroid distance\_cluster\_2: -0.119

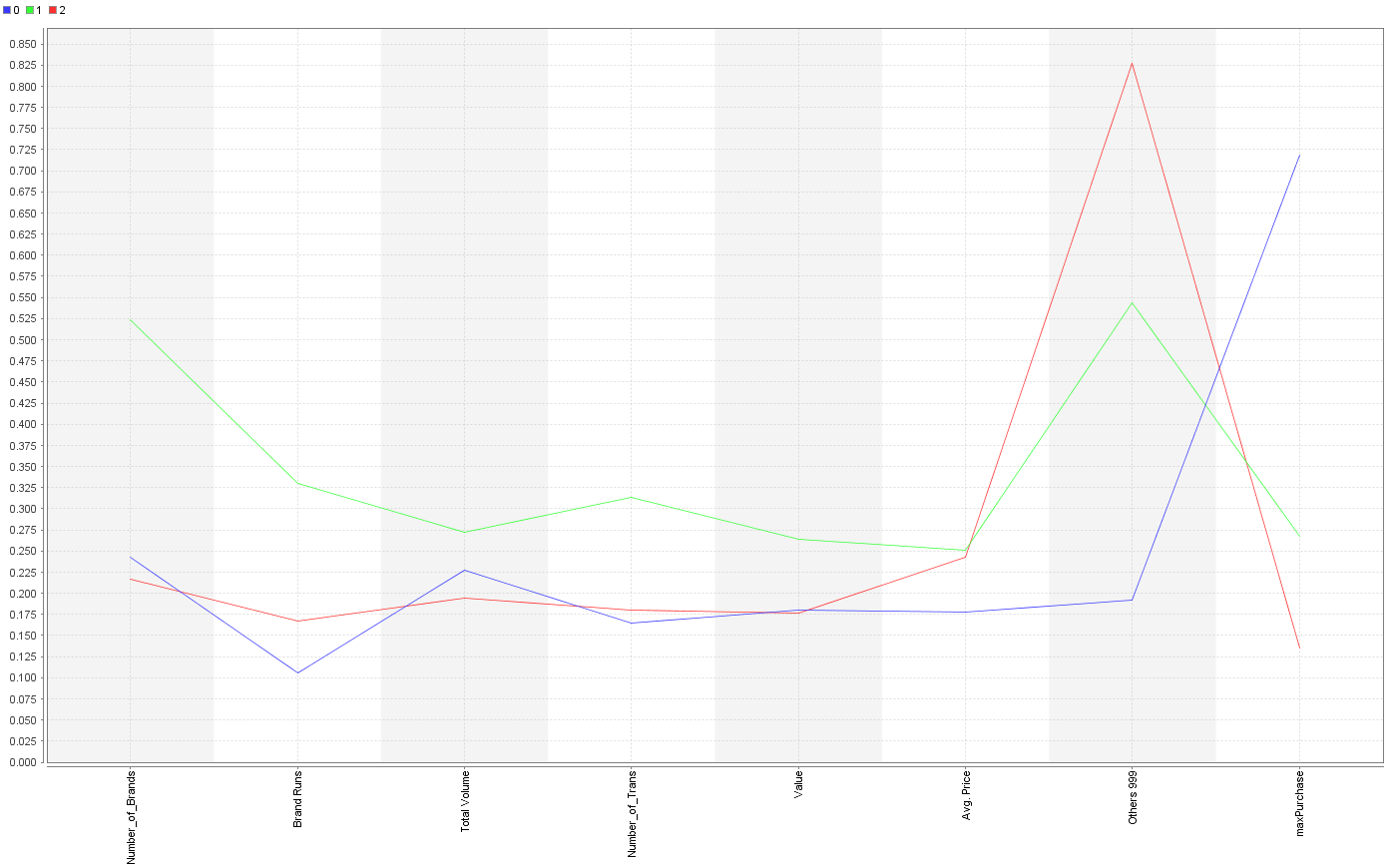
Davies Bouldin: -1.255

Size of Clusters: (197, 204, 199)

**Inter Cluster Distance**

|  |  |  |
| --- | --- | --- |
| **First** | **Second** | **Inter Cluster Distance** |
| 2.0 | 3.0 | 0.500 |
| 1.0 | 2.0 | 0.701 |
| 1.0 | 3.0 | 0.868 |

The below graph shows the dependence of the attributes on the differences between the clusters. From the graph, it is clear that most of the difference is contributed by others999 and maxpurchase attributes.

****

For k=4:

Performance Matrix:

Avg. within centroid distance: -0.127

Avg. within centroid distance\_cluster\_0: -0.200

Avg. within centroid distance\_cluster\_1: -0.116

Avg. within centroid distance\_cluster\_2: -0.125

Avg. within centroid distance\_cluster\_3: -0.121

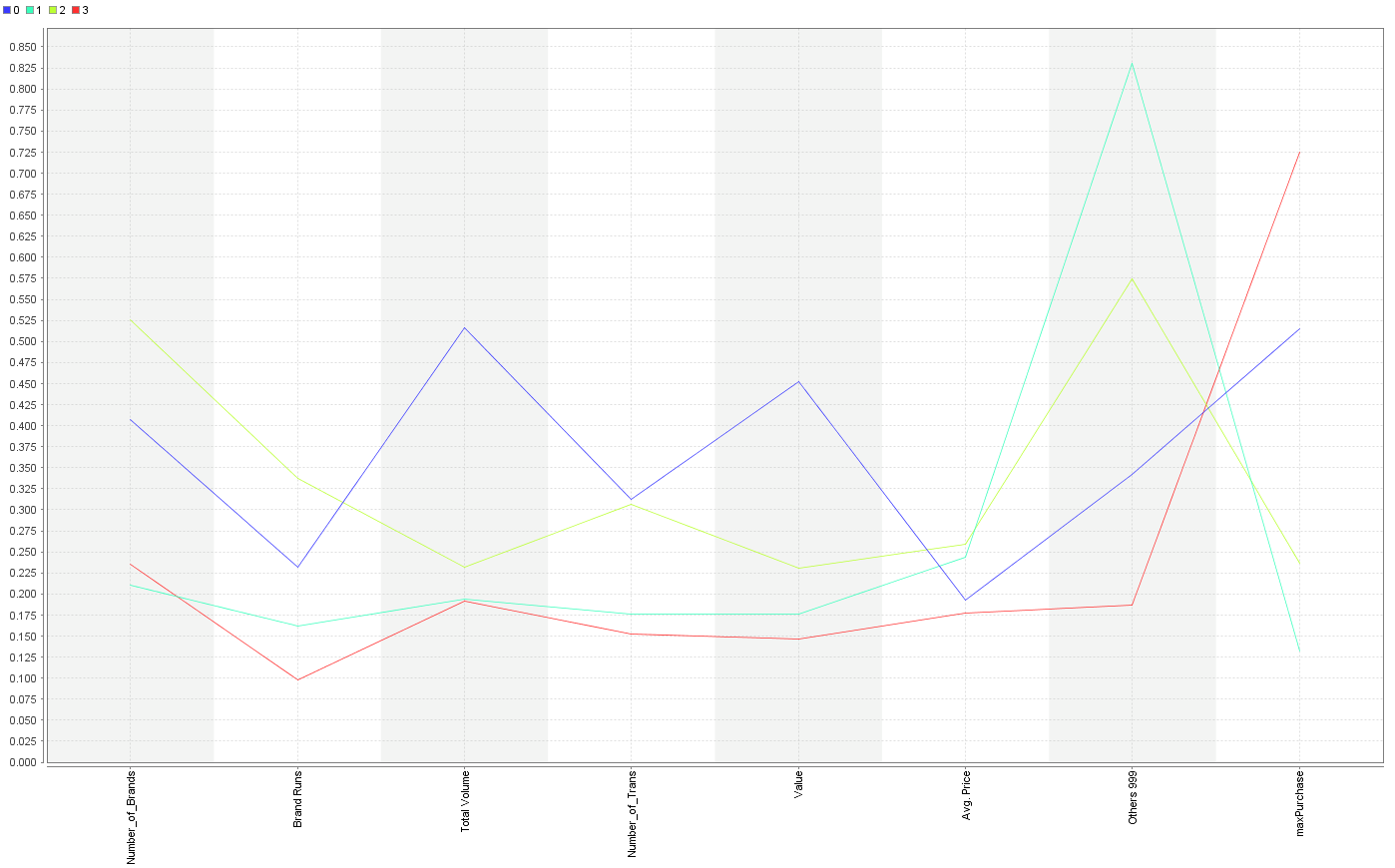
Davies Bouldin: -1.345

Size of Clusters: (51, 191, 181, 177)

**Inter Cluster Distance**

|  |  |  |
| --- | --- | --- |
| **First** | **Second** | **Inter Cluster Distance** |
| 1.0 | 2.0 | 0.794 |
| 1.0 | 3.0 | 0.539 |
| 1.0 | 4.0 | 0.583 |
| 2.0 | 3.0 | 0.478 |
| 2.0 | 4.0 | 0.880 |
| 3.0 | 4.0 | 0.753 |

The below graph shows the dependence of the attributes on the differences between the clusters. From the graph, it is clear that most of the difference is contributed by others999 and maxpurchase attributes.



For k=5:

Performance Matrix:

Avg. within centroid distance: -0.110

Avg. within centroid distance\_cluster\_0: -0.080

Avg. within centroid distance\_cluster\_1: -0.116

Avg. within centroid distance\_cluster\_2: -0.093

Avg. within centroid distance\_cluster\_3: -0.113

Avg. within centroid distance\_cluster\_4: -0.217

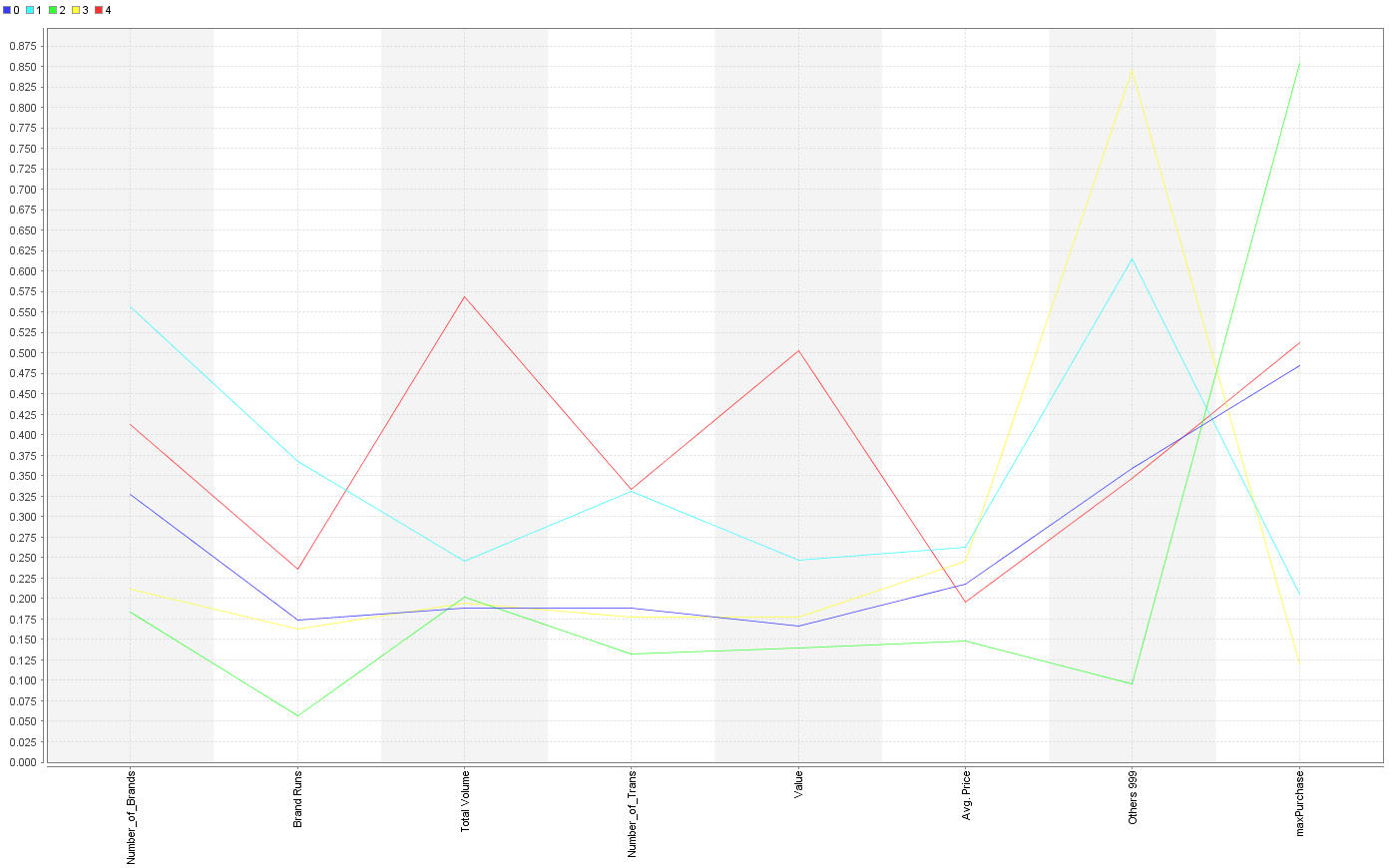
Davies Bouldin: -1.251

Size of Clusters: (135, 145, 100, 180, 40)

**Inter Cluster Distance**

|  |  |  |
| --- | --- | --- |
| **First** | **Second** | **Inter Cluster Distance** |
| 1.0 | 2.0 | 0.515 |
| 1.0 | 3.0 | 0.499 |
| 1.0 | 4.0 | 0.619 |
| 1.0 | 5.0 | 0.539 |
| 2.0 | 3.0 | 0.996 |
| 2.0 | 4.0 | 0.502 |
| 2.0 | 5.0 | 0.615 |
| 3.0 | 4.0 | 1.061 |
| 3.0 | 5.0 | 0.757 |
| 4.0 | 5.0 | 0.849 |

The below graph shows the dependence of the attributes on the differences between the clusters. From the graph, it is clear that most of the difference is contributed by Value, Total Volume, others999 and maxpurchase attributes.



**Choosing the best value of K:**

Choosing the best value of K for the given purpose has both objective and subjective criteria. Objective criteria include consistent group sizes, large inter cluster distance and small values of intra cluster distance. Davies Bouldin index is another criterion used to measure the overall clustering quality. Table above shows these metrics for K=2, 3, 4 and 5. Subjective criterion includes the clusters that lend themselves to good explanation.

Based on good group sizes and decent values of intra-cluster distance and davies bouldin index, we select **K=3 as our best model.**

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

**[Variables: pur\_vol\_no\_promo, pur\_vol\_promo\_6, pur\_vol\_other, all price categories, selling propositions]**

**[Note – would you use all selling\_propositions? Explore the data.]**

The variables used by us are:

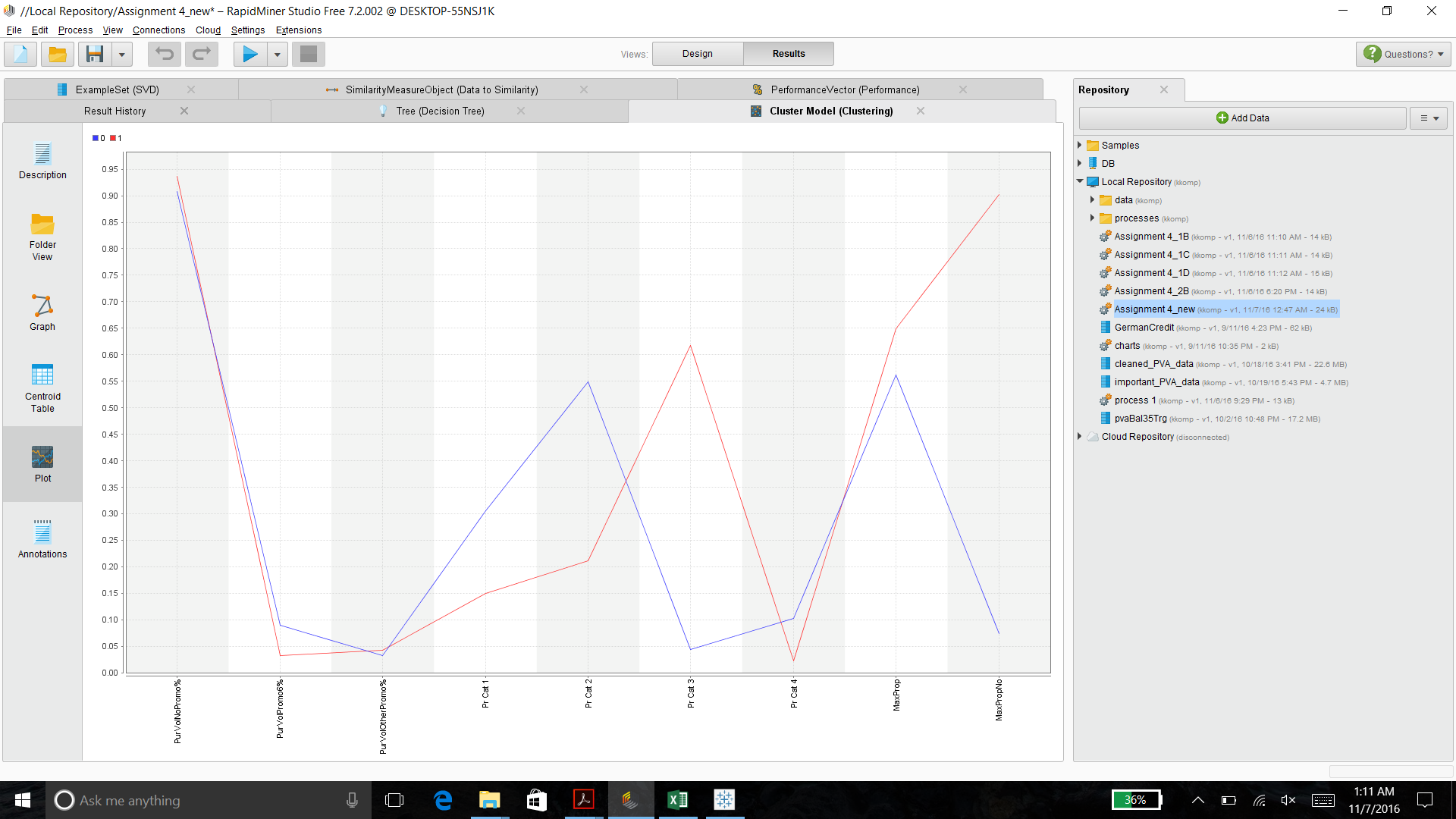
Maxprop, Maxpropno, Pr Cat\_1, Pr Cat\_2, Pr Cat\_3, Pr Cat\_4, Purvolnopromo%, purvolotherpromo%, purvolpromo6%.

All Proposition Categories (1-15) compiled as maxProp which has all maximum proposition value and maxPropNo which has the corresponding Proposition number.

(Proposition categories 10-13 have distributions tending to zeros, but we kept them as there are at least some points having valid data and the total data points are only 600)

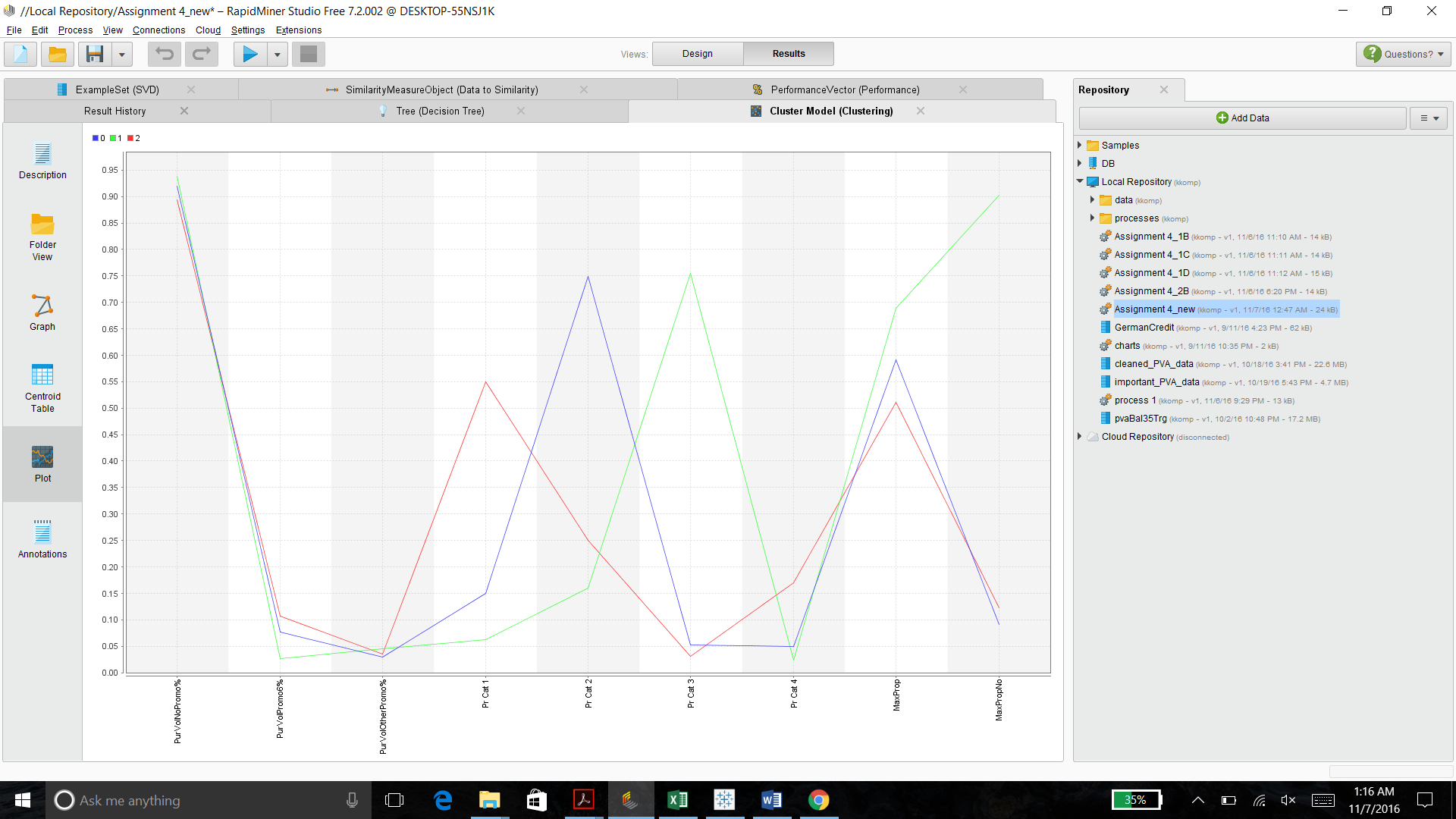
For k=2:

We can notice from the graph below that most of the differences between these two clusters lie with Pr Cat 2, Pr Cat3 and maxPropNo.



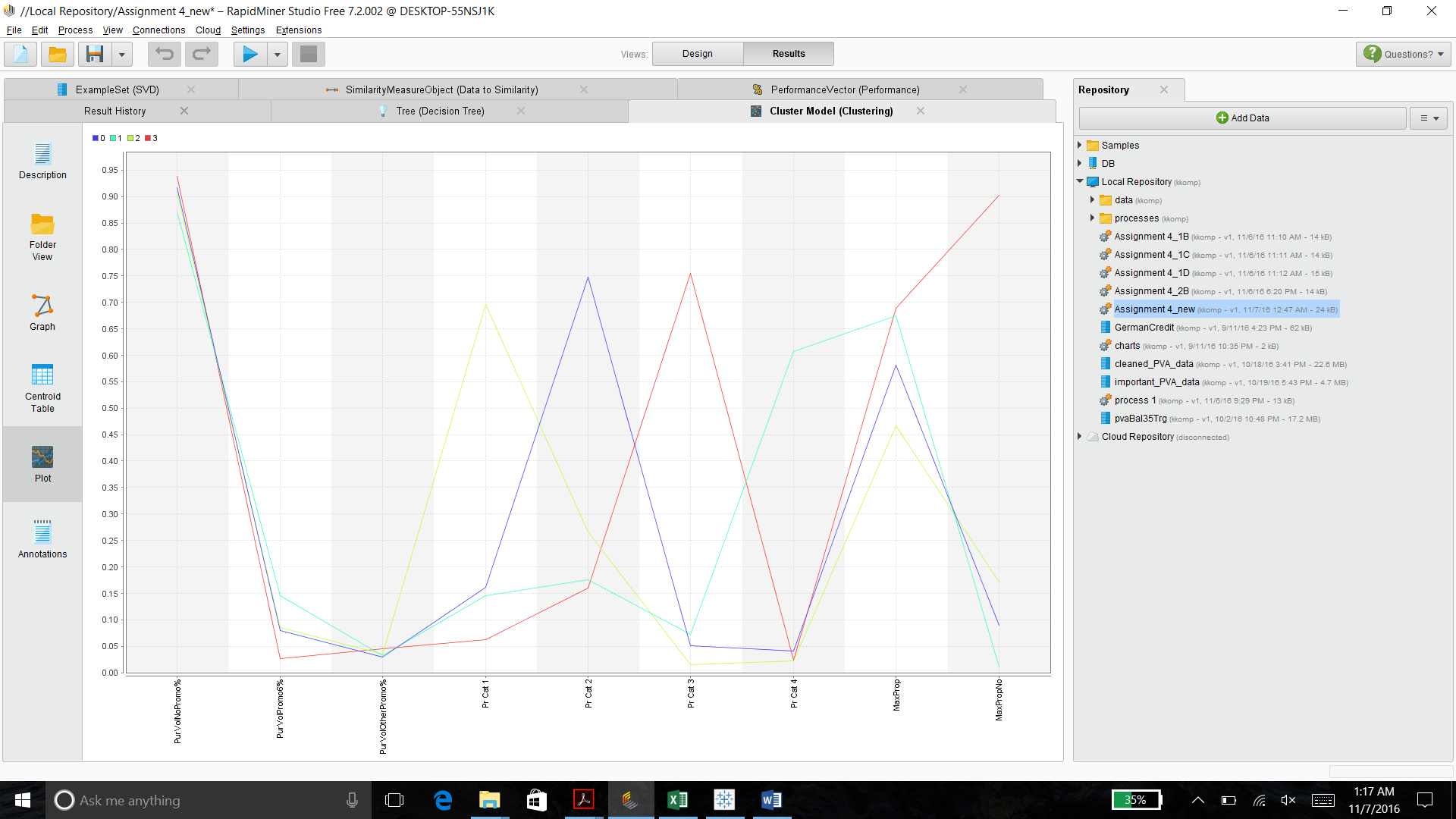
For k=3:

We can notice from the graph below that most of the differences between these three clusters lie with Pr Cat 1, Pr Cat 2, Pr Cat3, maxProp and maxPropNo.



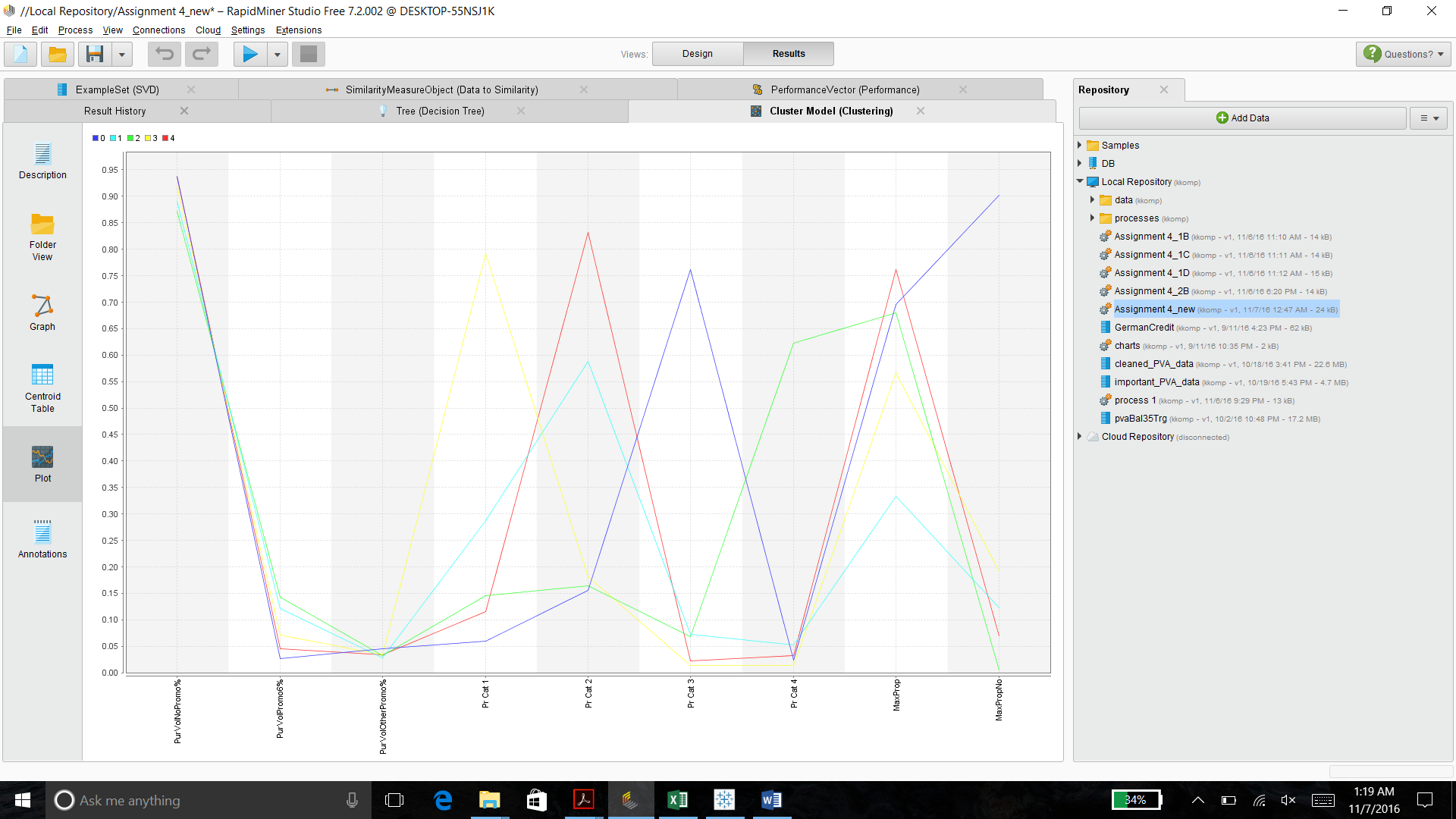
For k=4:

We can notice from the graph below that most of the differences between these four clusters lie with Pr Cat 1, Pr Cat 2, Pr Cat3 and maxProp.



For k=5:

We can notice from the graph below that most of the differences between these five clusters lie with Pr Cat 1 and maxProp.



As explained in question 1 (a), we chose the best model with good group sizes and decent metrics. Along with good explainability **K=5 turns out to be a good model** for the given attributes, with –

Davies Bouldin Index: -1.08

Avg within Centroid distance: -0.168

Avg Inter cluster distance: 0.9959

Size of clusters: (80, 191, 55, 101, 173)

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

**Note: How should k be chosen? Think about how the clusters would be used. It is likely that the marketing efforts would support 2-5 different promotional approaches.**

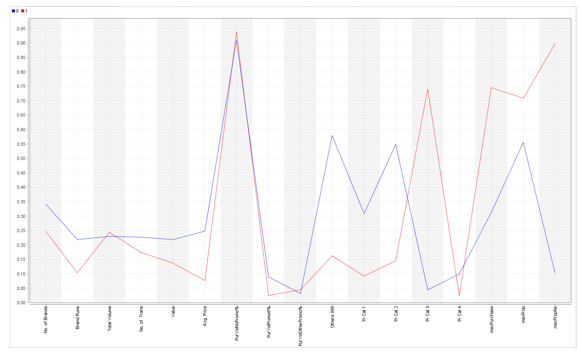
**Note: How should the percentages of total purchases comprised by various brands be treated? Isn’t a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is?**

The variables used by us are:

No.of.Trans, Value, BrandRuns, No.of.Brands, Max Brand Purchase, Others999, All Price Categories (1-4), Pur\_vol\_no\_promo, Pur\_vol\_promo\_6, Pur\_vol\_other, All Proposition Categories (1-15) [in the form of maxProp and maxPropNo]

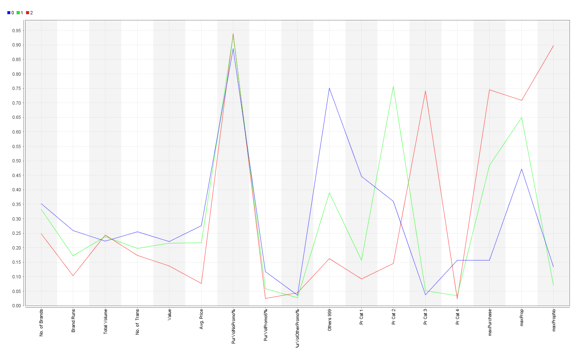
For k=2:

We can notice from the graph below that most of the differences between these two clusters lie with No. of Brands, Others999, PrCat1,Pr Cat2, Pr Cat3 and maxProp.



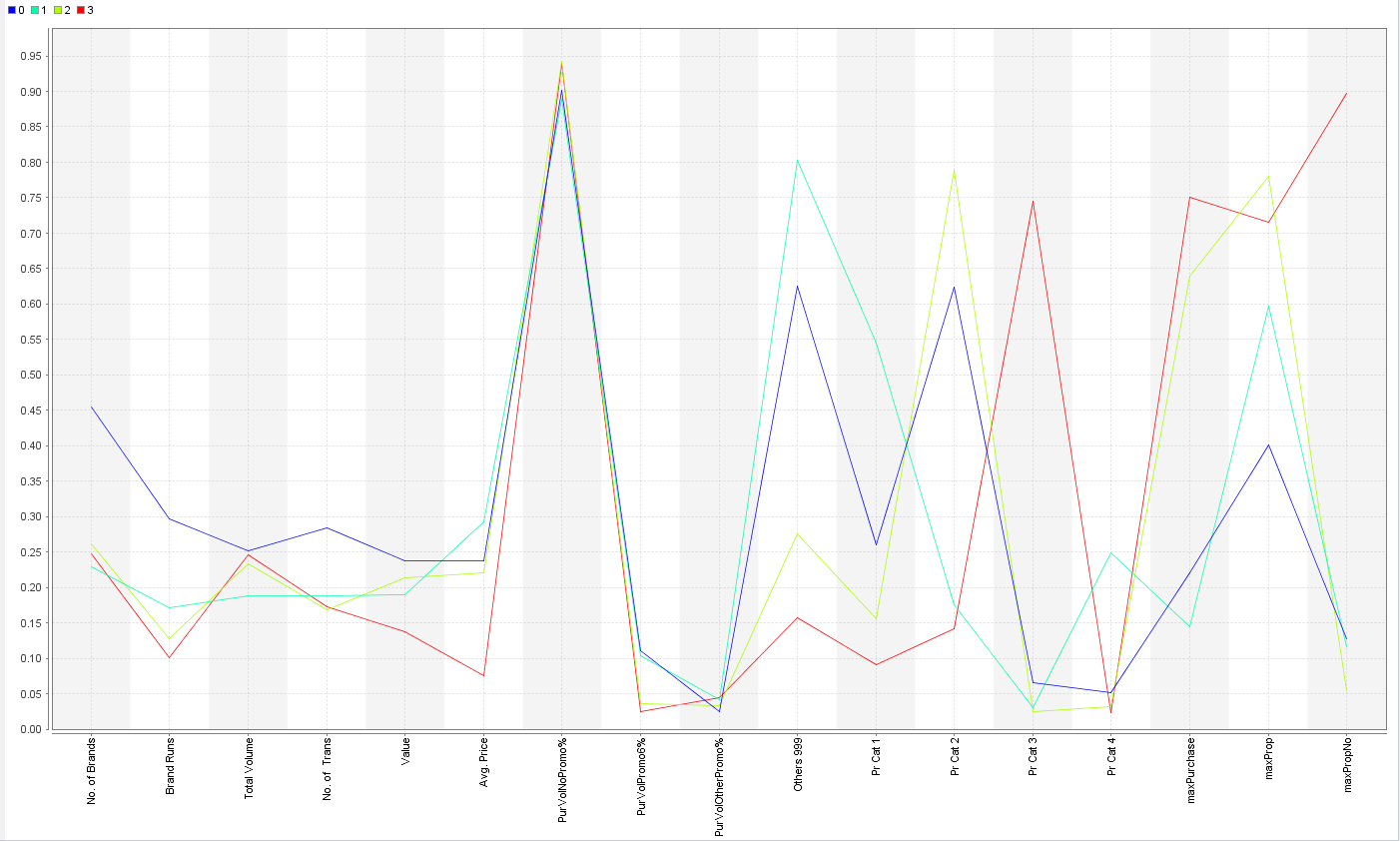
For k=3:

We can notice from the graph below that most of the differences between these three clusters lie with No. of Brands, Others999, PrCat1,Pr Cat2, maxBrandPurchase and maxProp.



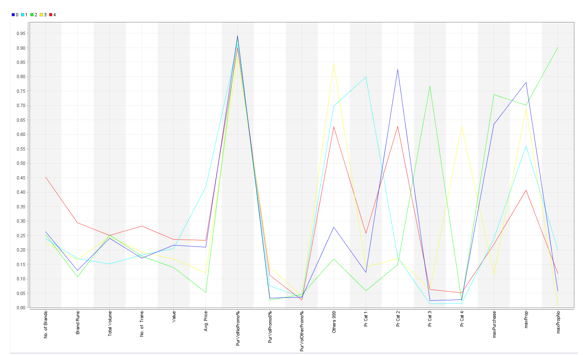
For k=4:

We can notice from the graph below that most of the differences between these four clusters lie with Brand Runs, Others999, PrCat1, maxBrandPurchase and maxProp.



For k=5:

We can notice from the graph below that most of the differences between these five clusters lie with Avg. Price, Others999, PrCat1, maxBrandPurchase.



Similar to question 1(a), we have selected k=4 as our best model with –

Davies Bouldin index: -1.526

Avg within Centroid Distance: -0.415

Avg Inter-Cluster Distance: 1.128

Size of clusters: (241, 139, 139, 81)

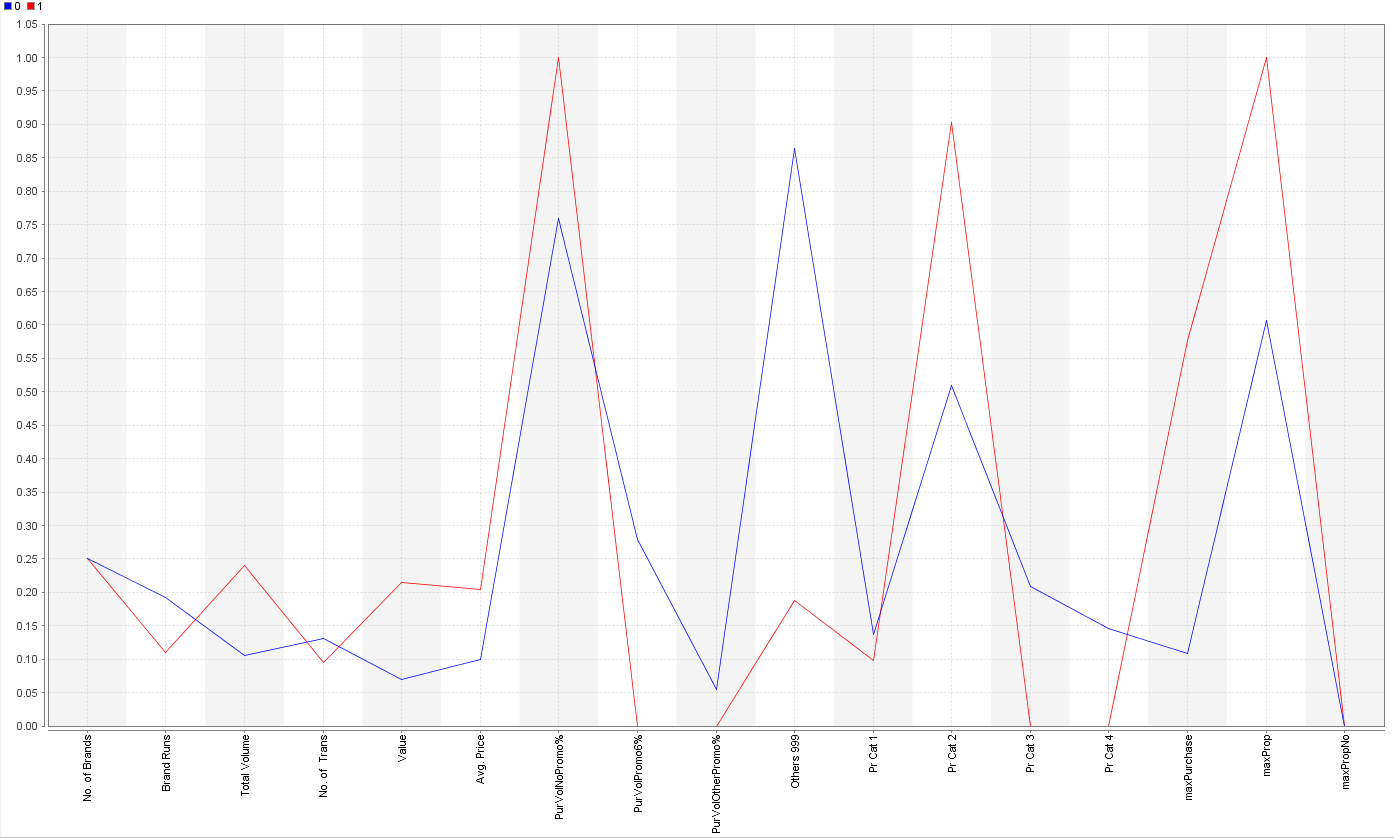
**d. Try k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering. You do not need to try all techniques on all combinations in (a)-(c) above; you may pick one set of variables in (a) thru (c) that you feel may be best suited for addressing the segmentation need. How do different parameter 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?**

For K-Medeoids:

We have tried the various methods as mentioned in the question, the first being **k-mediods**. We have taken purchase behavior and basis of purchase variables for all these methods.

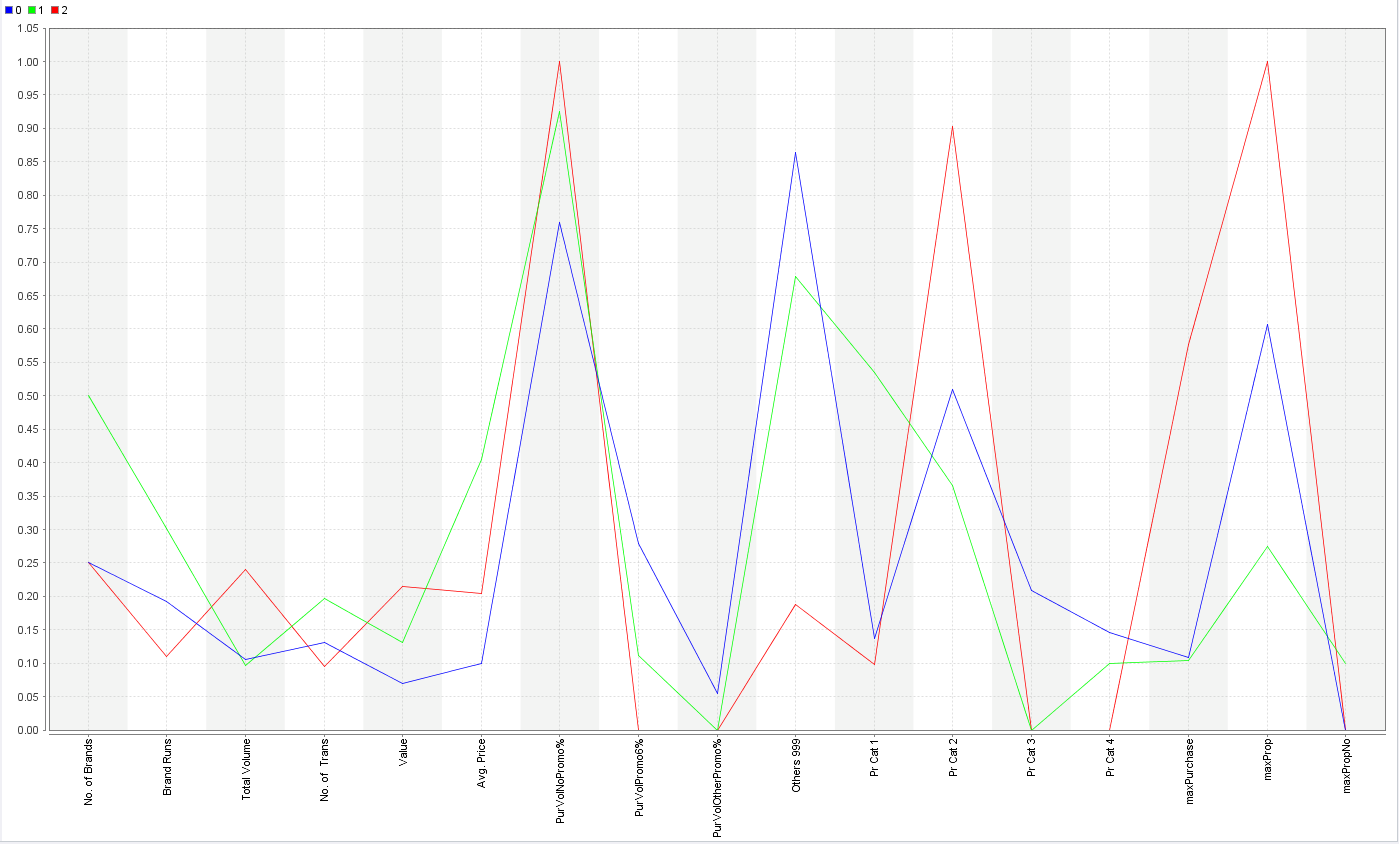
For k=2:

We can notice from the graph below that most of the differences between these two clusters lie with Total Volume, Value, PurchaseVolNoPromo %, Others999, PrCat2 and maxProp.



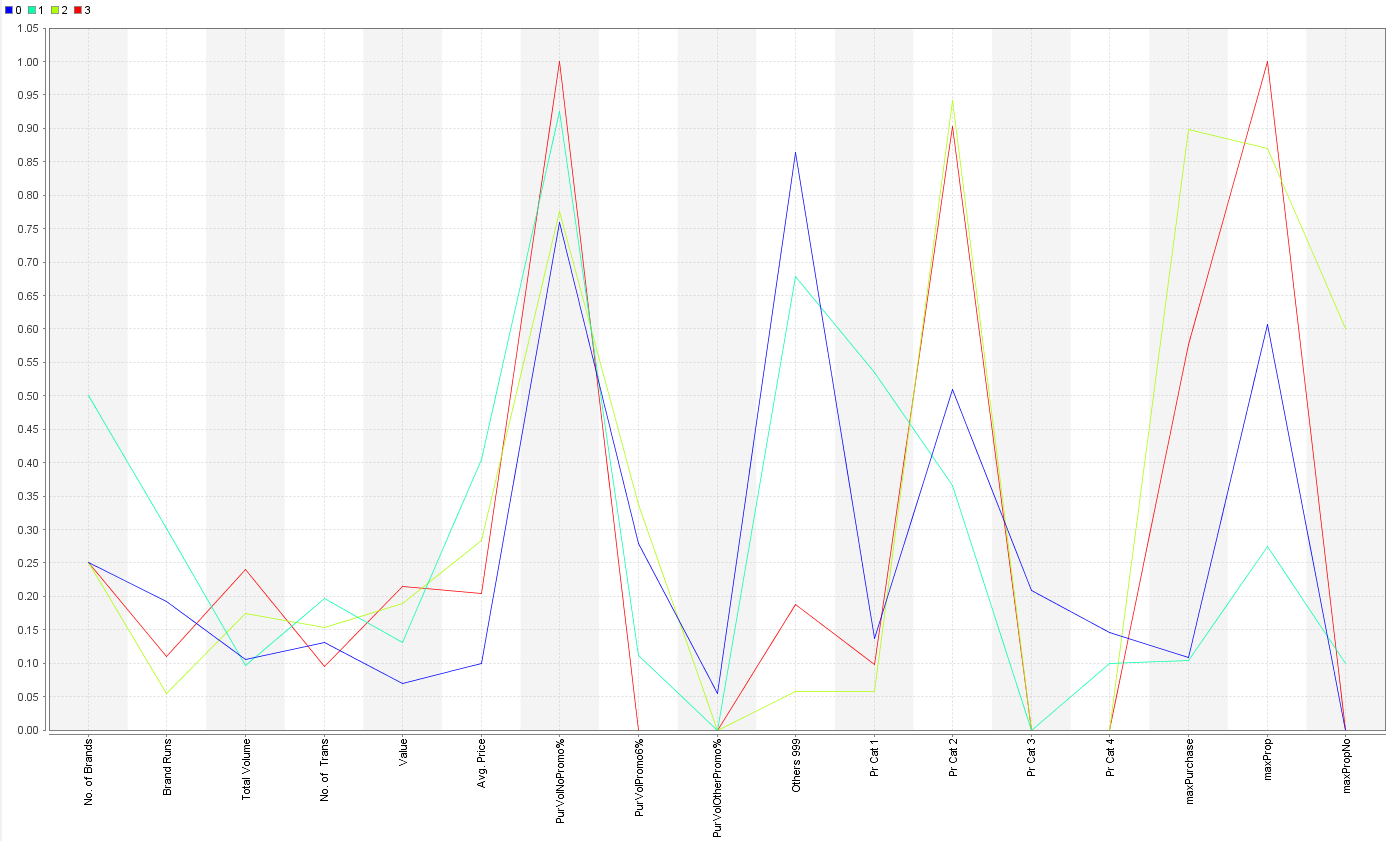
For k=3:

We can notice from the graph below that most of the differences between these three clusters lie with Avg. Price, Others999, PrCat2 and maxProp.



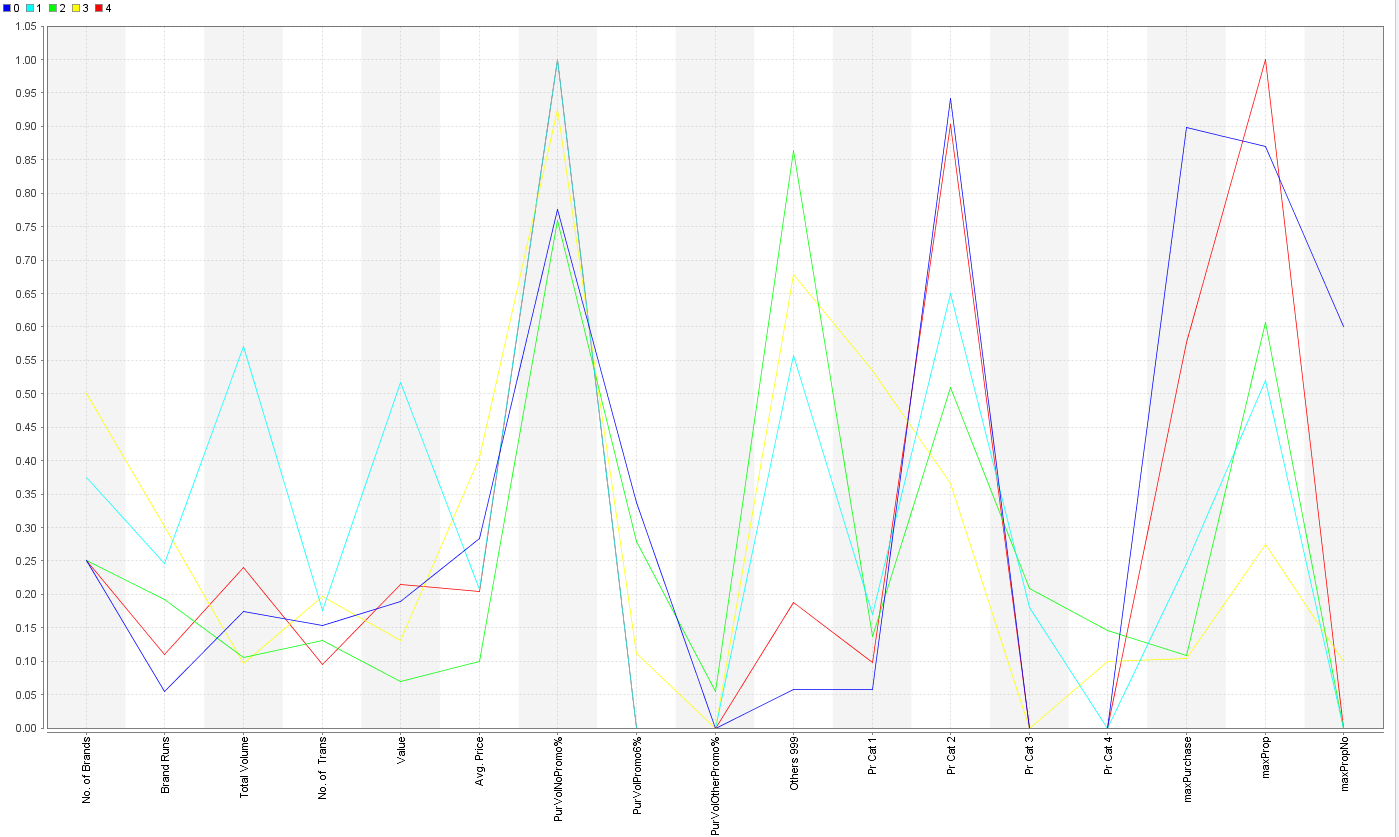
For k=4:

We can notice from the graph below that most of the differences between these four clusters lie with Avg. Price, Others999 and maxProp.



For k=5:

We can notice from the graph below that most of the differences between these five clusters lie with Brand Runs, Value, Others999 and maxProp.



Again, from the explanation in Question 1(a), the best K turns out to be **K=5 for which we get good cluster sizes**, decent cluster metrics and for the clusters.

For K-Means Kernel:

For Kernel k-Means, we tried for different values of K from 2 to 5 and obtained their corresponding clusters.

Kernel K-means just like K-means and K-mediods clustering algorithms are driven by the value of “k”. A change in the “k” indicates a change in the number of clusters. So the cluster density varies by the K values.

But the clusters we obtained were overlapping onto each other. And we couldn’t find completely distinct clusters.

For Agglomerative Clustering:

For the Agglomerative Clustering, we used the flattening operator along with the operator to extract the clusters. Following summarizes the parameters used –

* Mode – Complete Linkage
* Measure Types – Numerical Measures
* Numerical Measure – Euclidean Distance
* Number of clusters (in the flattening cluster operator) – 2 to 5

We did not get good cluster models using this method. This suggests that agglomerative clustering is not a suitable clustering for this data.

For DBSCAN:

We ran several iterations of DBSCAN, with different combinations of values for parameters. But it seemed to have no impact on the output.

We have tried with different values of epsilon, min. points and the distance measures, but we see that most of the data points are going into one cluster and the other clusters have only around 1 or at the maximum 10 data points.

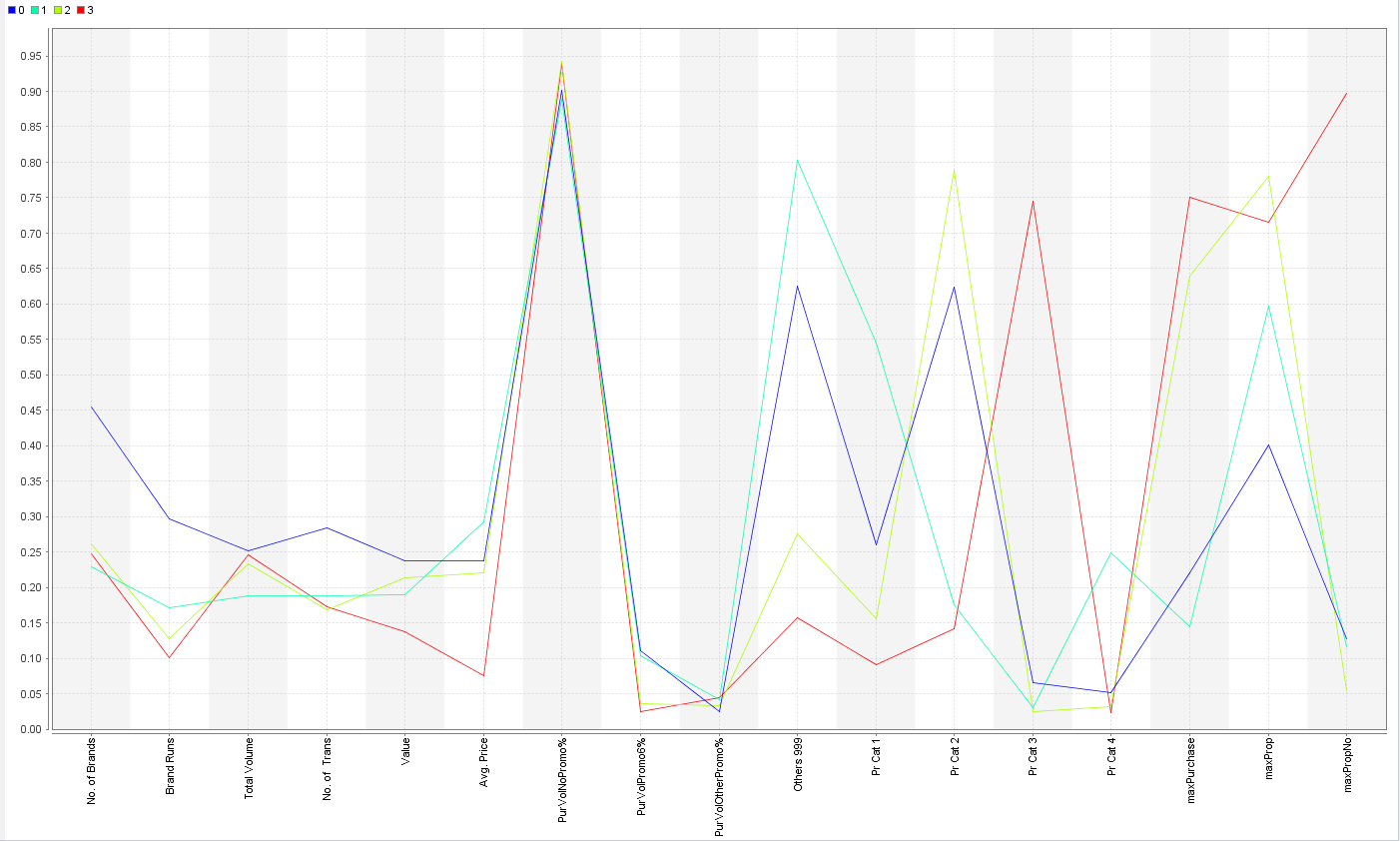
The reason for poor performance of DBSCAN could be because of the high dimensionality of the data. The DBSCAN clustering is not suitable for clustering on this data.

Hence, we could not extract any useful information regarding the model performance from Agglomerative and DBSCAN and k-mean kernel. Between k-medoids and kernel k- means, we would choose k-mediods since it has better formed clusters.

So, comparing all the best models in each case in question 1 with k-medoids clustering, we chose K-means, since it gave a better clustering output when compared to that of the k-medoids.

**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 useful0.**

Based on our analysis using 6 clustering models and multiple iterations including optimizing the parameters in rapid miner, our best model is a k-means clustering model with k = 4. For this model, the intra cluster distance is small meaning that the data points in the cluster are similar and the inter cluster distance between the clusters is good and the size of the clusters is also good. Hence we concluded that this was our best model.



**Households in cluster 0** buy the highest number of brands, have the highest brand runs for a great amount of volume. Most of the purchases for these households seem to be purchases not on promotion. These households have a low brand loyalty towards our selected brands but have a relatively high brand loyalty for the brands in others999.Their major purchases comprise of the popular soaps.

**Households in cluster 1** buy the least number of brands, have a moderate brand runs for the least amount of volume. Most of the purchases for these households seem to be purchases not on promotion. These households have the least brand loyalty for our selected brands and have the highest brand loyalty towards the brands in others999. Their major purchases include premium soaps.

**Households in cluster 2** buy a decent number of brands, have a low number of brand runs but for a great amount of volume. Most of the purchases for these households seem to be purchases not on promotion. These households have a high brand loyalty towards our selected brands and a low brand loyalty for all the other brands. They mostly buy popular soaps.

**Households in cluster 3** buy moderately low number of brands, have the least number of brand runs for almost the same amount of volume in comparison to other clusters. Most of the purchases for these households seem to be purchases not on promotion. These households seem to have the highest brand loyalty for the selected brands and are least loyal towards the others and their maximum purchases for economic/carbolic soaps.

**(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.)**

We “joined” the demographics data to the already existing cluster data to see how the demographics changes in each of the cluster. Since each of the demographic attribute cannot be grouped in the similar manner, we used sum, average, median and mode functions to get values for these attributes.

The following is the table showing the different clusters and their corresponding demographic data –

**Cluster 0 –**

* Households are more likely to be aged above 45 years,
* Have a median affluence index of 18,
* With a median average price of 0.227,
* With children below the age of 14,
* With television or broadcast TV cable,
* Speaking Marati, Gujarati and Assameese,
* Females belonging to a Socio Economic class ‘C’,
* With low level of education (10-12 years),
* And with a house size of around 4 people.

These households have a low brand loyalty towards our selected brands but have a relatively high brand loyalty for the brands in others999.

Most of the purchases for these households seem to be purchases not on promotion with their least purchases on promo code other than 6. Their major purchases comprise of the popular soaps.

**Cluster 1 –**

* Households are more likely to be aged above 45 years,
* Have a median affluence index of 15,
* With a median average price of 0.273,
* With most of the children below the age of 14 and a few house with none or not specified,
* With television or broadcast TV cable,
* Speaking Marati majorly,
* Females belonging to a Socio Economic class ‘A’,
* With low level of education (10-12 years),
* And with a house size of around 3 people.

These households have the least brand loyalty for our selected brands and have the highest brand loyalty towards the brands in others999.

Most of the purchases for these households seem to be purchases not on promotion with their least purchases on promo code other than 6. Their major purchases comprise of the premium soaps.

**Cluster 2 –**

* Households are more likely to be aged above 45 years,
* Have a median affluence index of 15,
* With a median average price of 0.194,
* With most of the children below the age of 14 and a few houses with none or not specified,
* With television or broadcast TV cable,
* Speaking Marati and Gujarati,
* Females belonging to a Socio Economic class ‘A’,
* With low level of education (10-12 years),
* And with a house size of around 4 people.

These households have a high brand loyalty towards our selected brands and a low brand loyalty for all the other brands.

Most of the purchases for these households seem to be purchases not on promotion with their least purchases on promo code 6. Their major purchases comprise of the popular soaps.

**Cluster 3 –**

* Households are more likely to be aged above 45 years,
* Have a median affluence index of 10,
* With a median average price of 0.048,
* With most of the children below the age of 14 and a few house with none or not specified,
* With television or broadcast TV cable,
* Speaking Marati and Assameese,
* Females belonging to a Socio Economic class ‘D/E’,
* And with a house size of around 4 people.

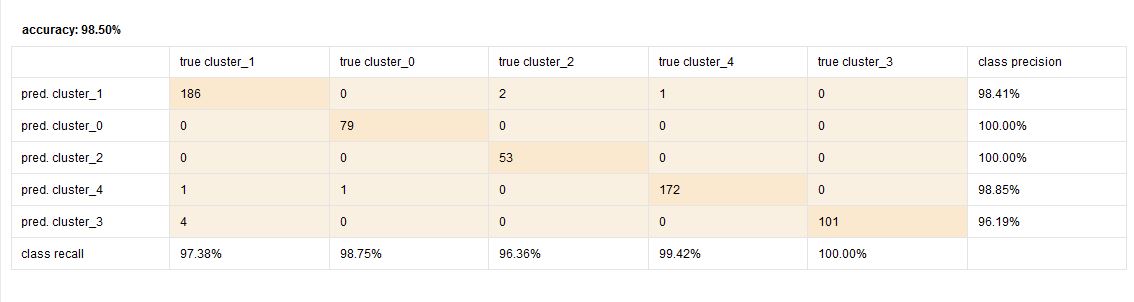
These households seem to have the highest brand loyalty for the selected brands and are least loyal towards the others999.

Most of the purchases for these households seem to be purchases not on promotion with their least purchases on promo code 6. Their major purchases comprise of the economic/carbolic soaps.

**3. For the best segmentation, obtain a description of the clusters. You may base this on attributes describing the clusters (not restricted to attributes used for clustering). You should also build a decision tree to help describe the clusters – 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).**

We ran the decision tree with Gini index criterian and the decision tree was able to identify the clusters distinctly.

Given below is the confusion matrix of the decision tree:



The decision tree appears to be a clearer version as to which attributes determine the forming of which clusters. It is evident from the decision tree that the variables appearing in the top part of the tree are the variables that have a maximum influence in forming the clusters. The variables appearing in the top part of the tree are:

PR Cat\_2

MaxProp

Pr Cat\_1

Maxpropno

We tried different parameters for the decision tree, but the top nodes remained almost consistent.

