IDS 572: Assignment 4

Market Segmentation - Segmenting Consumers of Bath Soap

Archan Patel - 661105271 - apate381@uic.edu

Jasbir Singh - 651837003 - jasingh3@uic.edu

Jeetyog Rangnekar - 656052696 - jrangn2@uic.edu

Assignment Question 1

(a)

- To identify the clusters on households, initially we haven't considered the demographic variables.
- To consider the brand loyalty, out of all Purchase Data variables, We found that variables which can be considered for this purpose would be as follow.

Avg. Price	No. of Trans	No. of Brands
Others 999	Trans/Brand Runs	Vol/Tran

- Since, average no. of consecutive transactions before brand switch, average volume of the transaction, total no. of brands purchased & price are all major factor in deciding the brand loyalty, all above mentioned variables were selected.
- Variables which were giving the similar information as above *(example no. of brand runs will provide similar information as Trans/Brand runs & No. of Trans)* have not been considered here.
- Addition to this, we generated a new variable 'MaxLoyaltyBrand' using 'Br. Cd.' variables list - which will reflect the maximum percent of purchase out of all. This should be a significant factor in deciding brand loyalty - higher the percentage, higher the loyalty.
- And to consider the percentage volume purchase of remaining brands, 'Other 999' has been included.
- After selecting the variables, we tried K-means clustering with different values of K (2,3,4,5..). Below is the output of performance for K = 3 & 4.

- **K = 3** ***(Best model for k=3 has been explained)***
  - Parameters
    - Max runs: 15
    - Measure types: MixedMeasures
    - Mixed Measures: MixedEuclideanDistance
    - Max Optimization Steps: 150

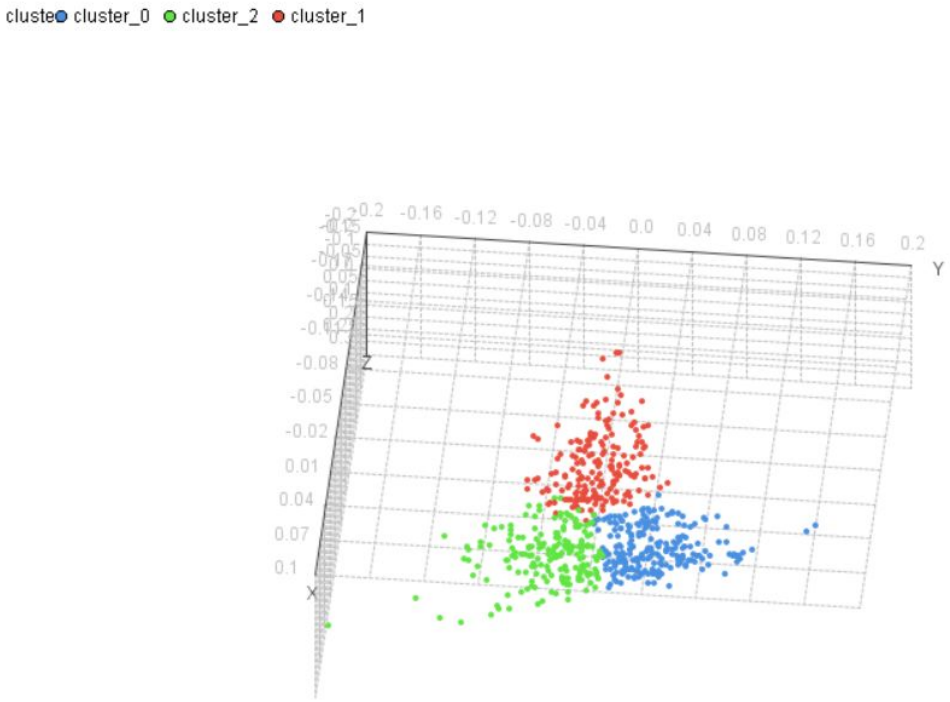
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -4.349
Avg. within centroid distance_cluster_0: -4.144
Avg. within centroid distance_cluster_1: -5.786
Avg. within centroid distance_cluster_2: -3.168
Davies Bouldin: -1.440
```

Cluster Model

```
Cluster 0: 222 items
Cluster 1: 188 items
Cluster 2: 190 items
Total number of items: 600
```

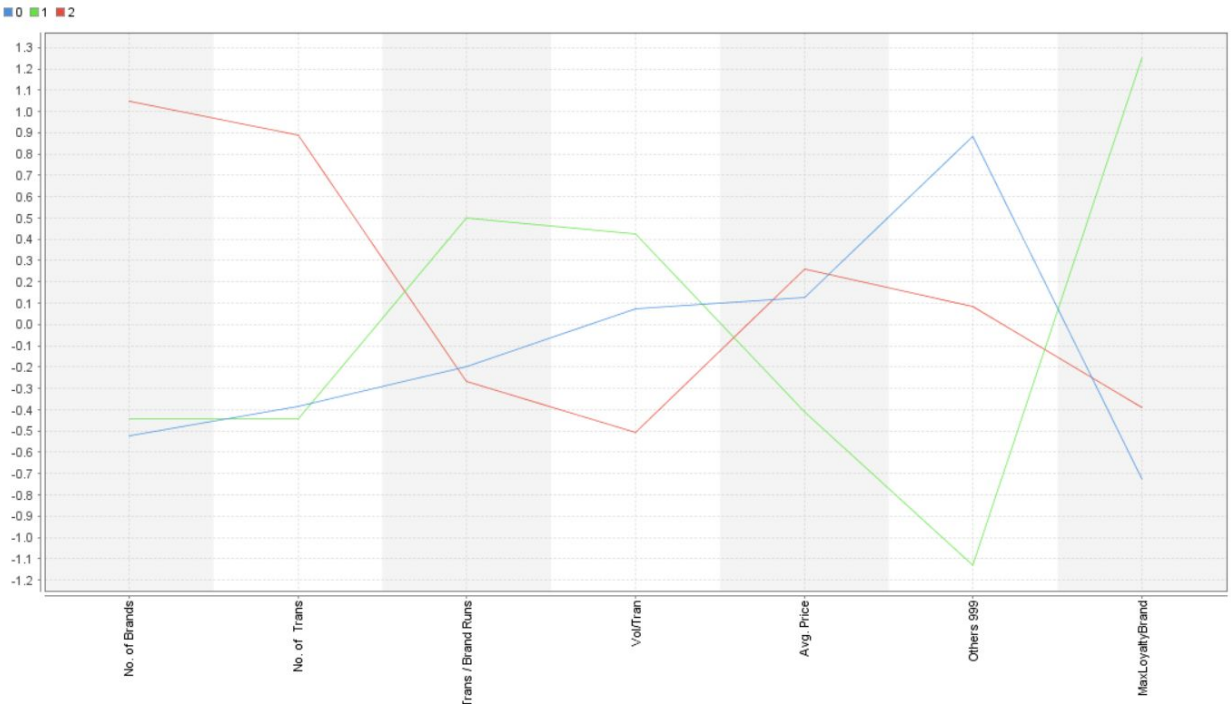
- Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	2.979
1.0	3.0	2.281
2.0	3.0	3.175

- Cluster Plot



- Here, we see a very clear distinguish of clusters for almost all variables. Cluster 0 & 1 show some interaction at variables like No. of Trans & No. of Brands.
- But overall, this gives a clear picture of clusters and it goes with the logical assumption as well.
  - Household in Cluster 2 have high number of brands purchase, so ideally their Loyalty should be low which can clearly seen from the plot.
  - Similarly, households who have high Trans/Brand Runs ratio, should have higher loyalty - which is evident from Cluster 1.

- **K = 4** ***(Best model for k=4 has been explained)***
  - Parameters
    - Max runs: 20
    - Measure types: MixedMeasures
    - Mixed Measures: MixedEuclideanDistance
    - Max Optimization Steps: 150

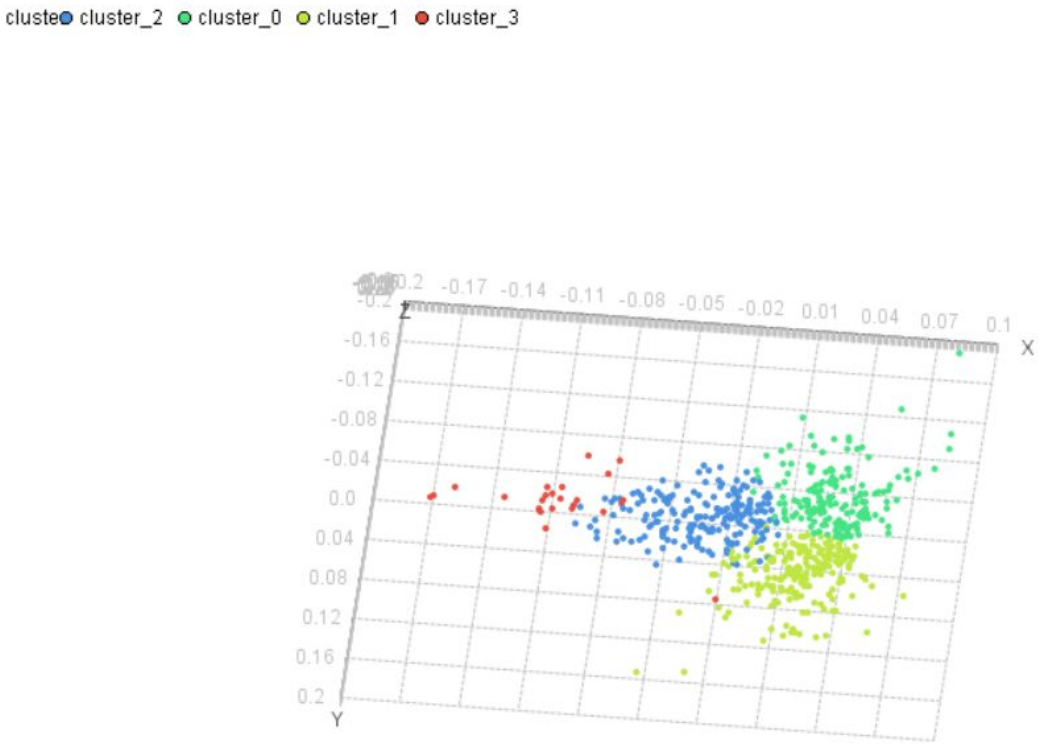
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -3.748
Avg. within centroid distance_cluster_0: -3.084
Avg. within centroid distance_cluster_1: -4.044
Avg. within centroid distance_cluster_2: -3.581
Avg. within centroid distance_cluster_3: -7.252
Davies Bouldin: -1.284
```

Cluster Model

```
Cluster 0: 172 items
Cluster 1: 203 items
Cluster 2: 200 items
Cluster 3: 25 items
Total number of items: 600
```

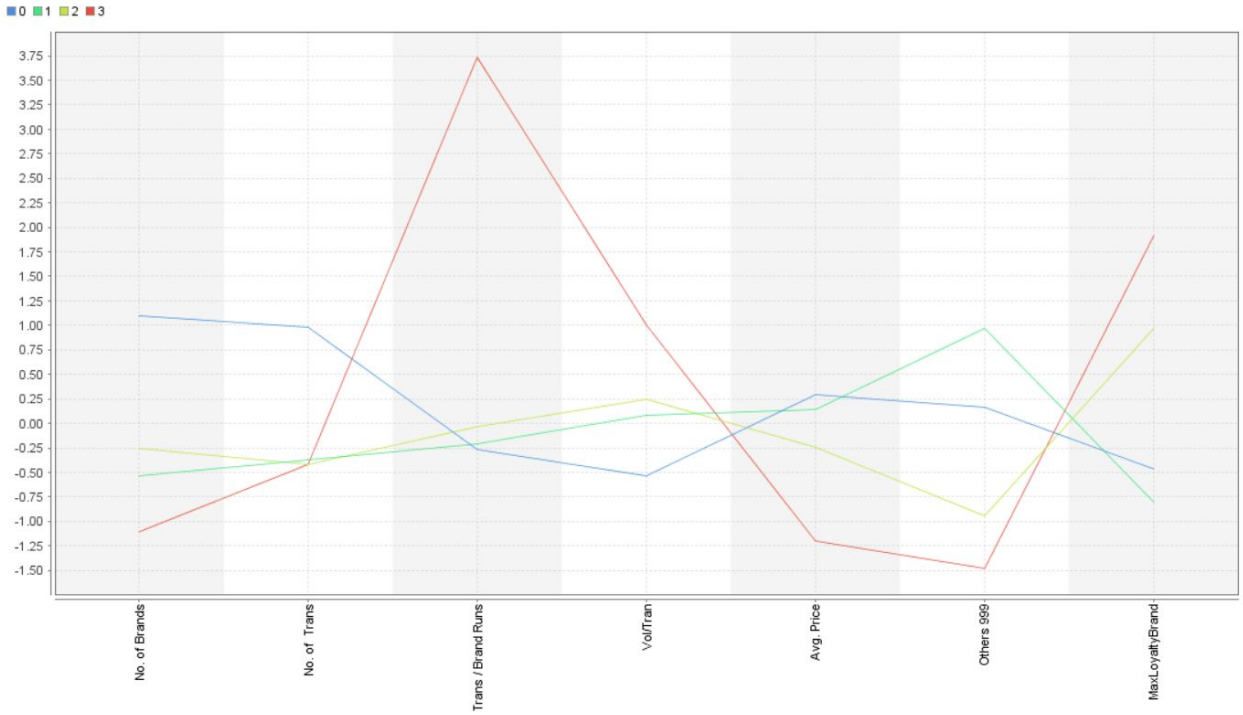
- Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	2.382
1.0	3.0	2.836
1.0	4.0	5.982
2.0	3.0	2.664
2.0	4.0	5.648
3.0	4.0	4.194

- Cluster Plot



- Here, we can see the distinction between Trans/ Brand Runs, but compared to the previous cluster plot (k=3), there are lot of interactions which indicates similar things.
- Moreover, Cluster 1 & Cluster 2 are almost same for majority of the parameters, which disregards the purpose of having different clusters.

- From the comparison of 2 clusters above, we can say that model achieved with K = 3 is a better model. Because,
  - The clusters are quite dense
  - The clusters are distinct with very less intersections
  - Good amount of distribution in each cluster *(k=4 model has only 25 items in Cluster 3)*

(b)

- For clustering model on basis-for-purchase, we have used following variables.

Pr Cat 1	Pr Cat 2	Pr Cat 3	Pr Cat 4
Pur Vol No promo %	Pur Vol Other Promo %	Pur Vol Promo 6 %	---

- To identify the basis of purchase, the price category should be an important factor - as well as what percentage of volume was purchased under promotion. Hence, above variables are selected.

- Since, the product's proposition category also has an important role in basis for purchase, we have generated 2 new variables in addition.
    - MaxProp - which takes the maximum value out of percentage of volume for proposition category. This variables takes the maximum value out of only those proposition categories - for which the individual brand percentage purchase is available. (Example: Brand 55 - Pr 14, Brand 272 - Pr 8, …)(Prop Cat 5,6,7,8,11,13,14)
    - OtherProp - This sums up the percentage volume of proposition category for remaining ones *(the ones which were not included in MaxProp).* (Prop Cat 9,10,12,15)
- After selecting the variables, we tried K-means clustering with different values of K (2,3,4,5..). Below is the output of performance for K = 3 & 4.

- **K = 3** ***(Best model for k=3 has been explained)***
    - Parameters
        - Max runs: 20
        - Measure types: MixedMeasures
        - Mixed Measures: MixedEuclideanDistance
        - Max Optimization Steps: 100

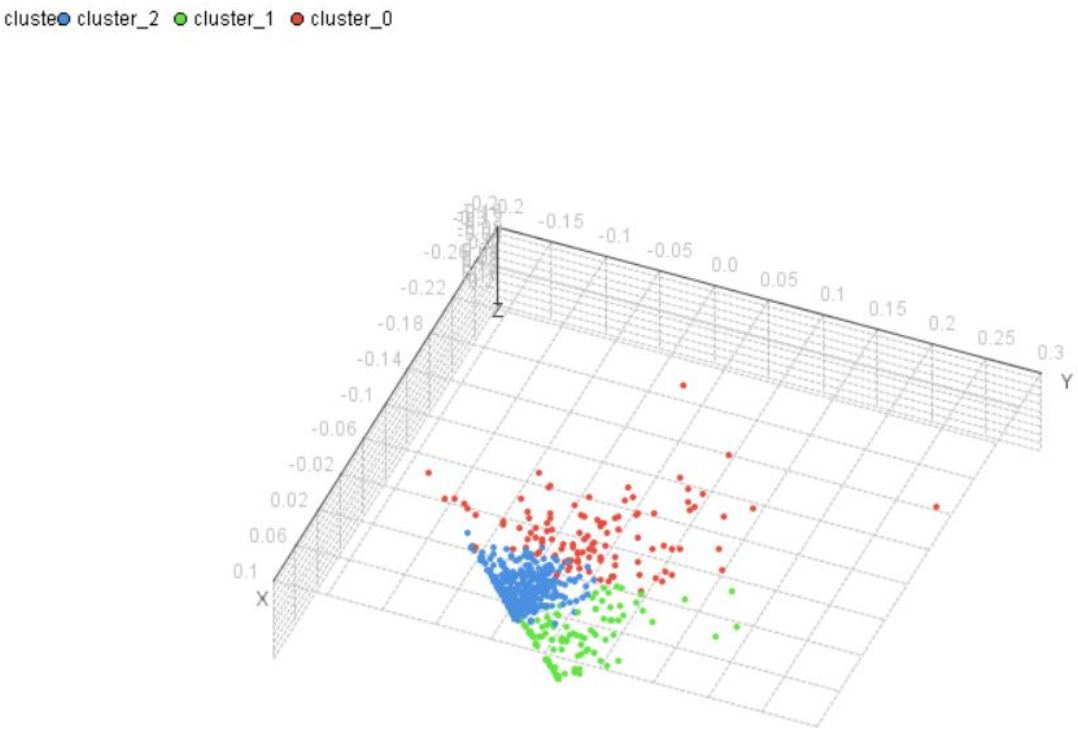
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -6.395
Avg. within centroid distance_cluster_0: -13.081
Avg. within centroid distance_cluster_1: -7.710
Avg. within centroid distance_cluster_2: -4.007
Davies Bouldin: -1.635
```

Cluster Model

```
Cluster 0: 113 items
Cluster 1: 110 items
Cluster 2: 377 items
Total number of items: 600
```

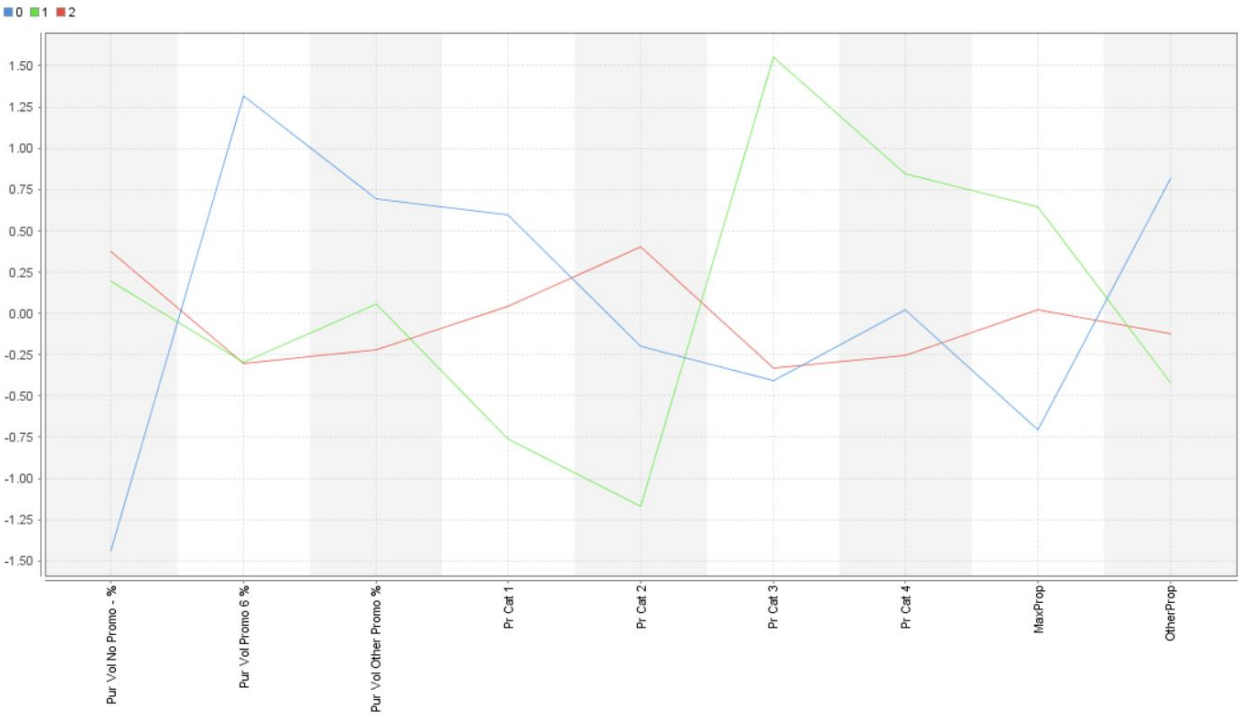
    - Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	4.040
1.0	3.0	2.986
2.0	3.0	2.905

- Cluster Plot



- Again here, we see a very clear distinguish of clusters for almost all variables. At some variables, we can see some interaction between cluster., But overall, this gives a good distinguished picture of clusters.

- **K = 4** ***(Best model for k=4 has been explained)***
  - Parameters
    - Max runs: 20
    - Measure types: MixedMeasures
    - Mixed Measures: MixedEuclideanDistance
    - Max Optimization Steps: 150

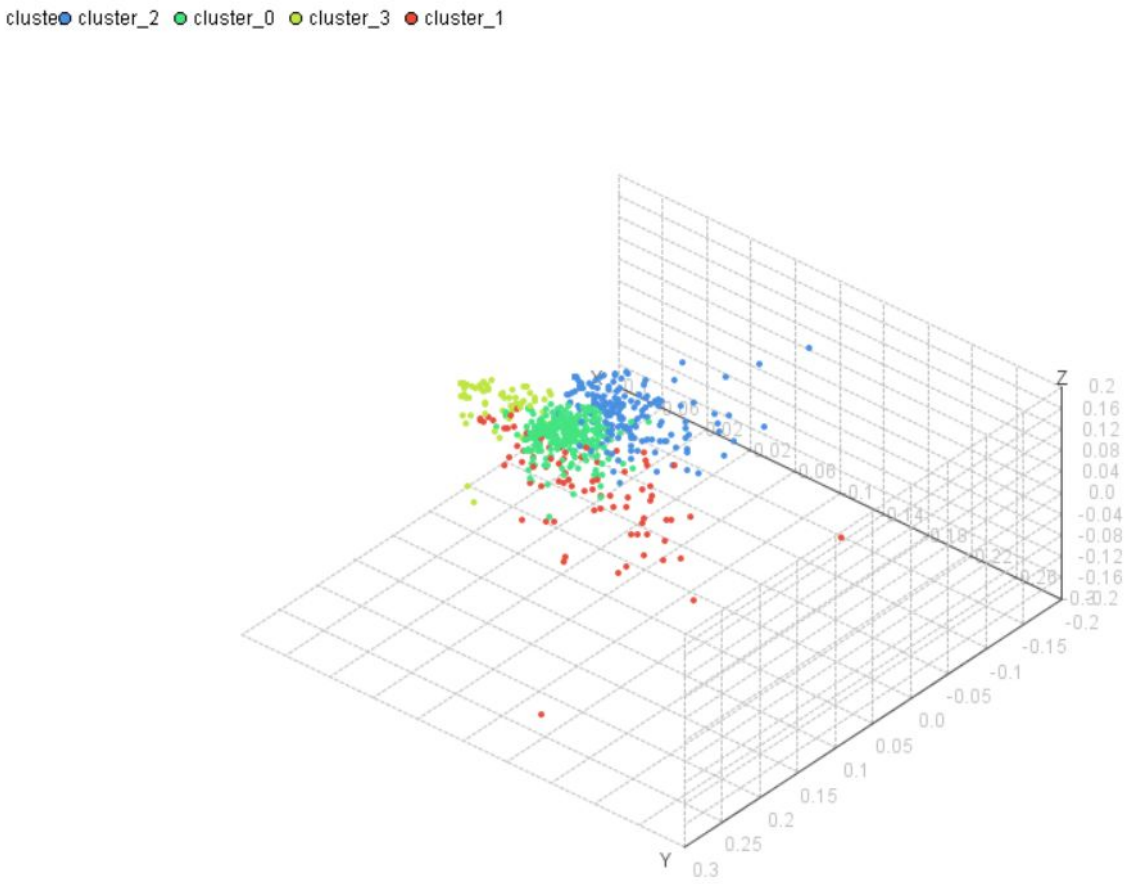
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -5.380
Avg. within centroid distance_cluster_0: -2.792
Avg. within centroid distance_cluster_1: -14.780
Avg. within centroid distance_cluster_2: -5.895
Avg. within centroid distance_cluster_3: -2.917
Davies Bouldin: -1.443
```

Cluster Model

```
Cluster 0: 274 items
Cluster 1: 83 items
Cluster 2: 177 items
Cluster 3: 66 items
Total number of items: 600
```

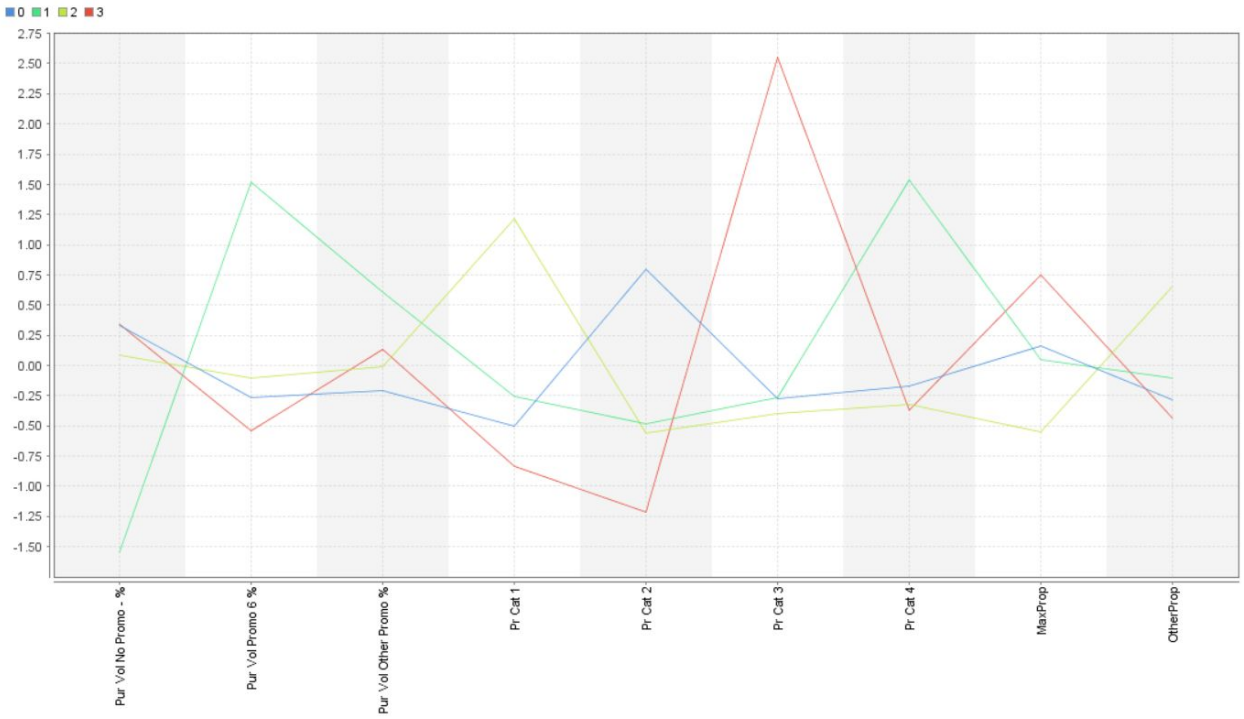
  - Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	3.465
1.0	3.0	2.517
1.0	4.0	3.572
2.0	3.0	3.492
2.0	4.0	4.585
3.0	4.0	4.060

- Cluster Plot



    - Here, we can see the distinction for variables like Pr. Cat 3, Pur Vol Promo 6 % etc. But compared to the previous cluster plot (k=3), there are lot of interactions which indicates similar things.
    - There very few clear and distinguish clusters.

- From the comparison of 2 clusters above, we can say that model achieved with K = 3 is a better model. Because,
    - Each cluster clearly shows a specific price category & proposition category
    - Clusters are more compact & with little interaction.

(c)

- For clustering model for both - purchase behaviour & basis-for-purchase, we have combined both set of variables used in above two models. So, all variables used in above two models are used for this.
- Then again, we tried K-means clustering with different values of K (2,3,4,5..). Below is the output of performance for K = 3 & 4.

- **K = 3** ***(Best model for k=3 has been explained)***
    - Parameters
        - Max runs: 15
        - Measure types: MixedMeasures
        - Mixed Measures: MixedEuclideanDistance
        - Max Optimization Steps: 150

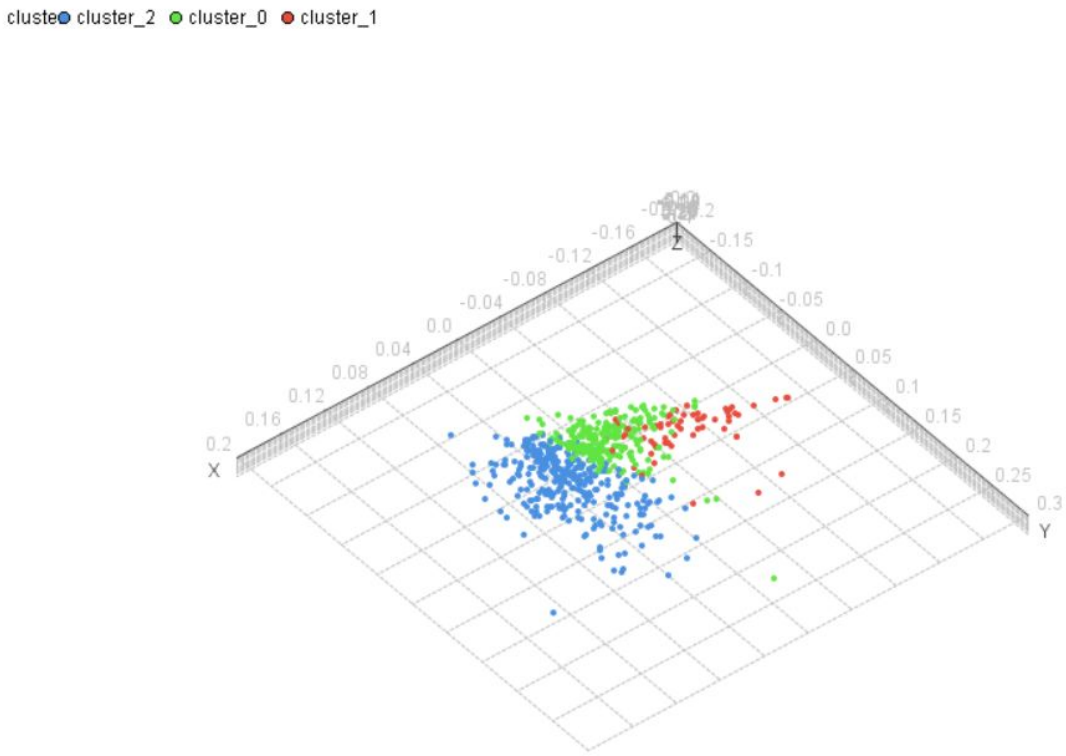
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -11.245
Avg. within centroid distance_cluster_0: -9.412
Avg. within centroid distance_cluster_1: -10.095
Avg. within centroid distance_cluster_2: -12.736
Davies Bouldin: -1.867
```

Cluster Model

```
Cluster 0: 219 items
Cluster 1: 63 items
Cluster 2: 318 items
Total number of items: 600
```

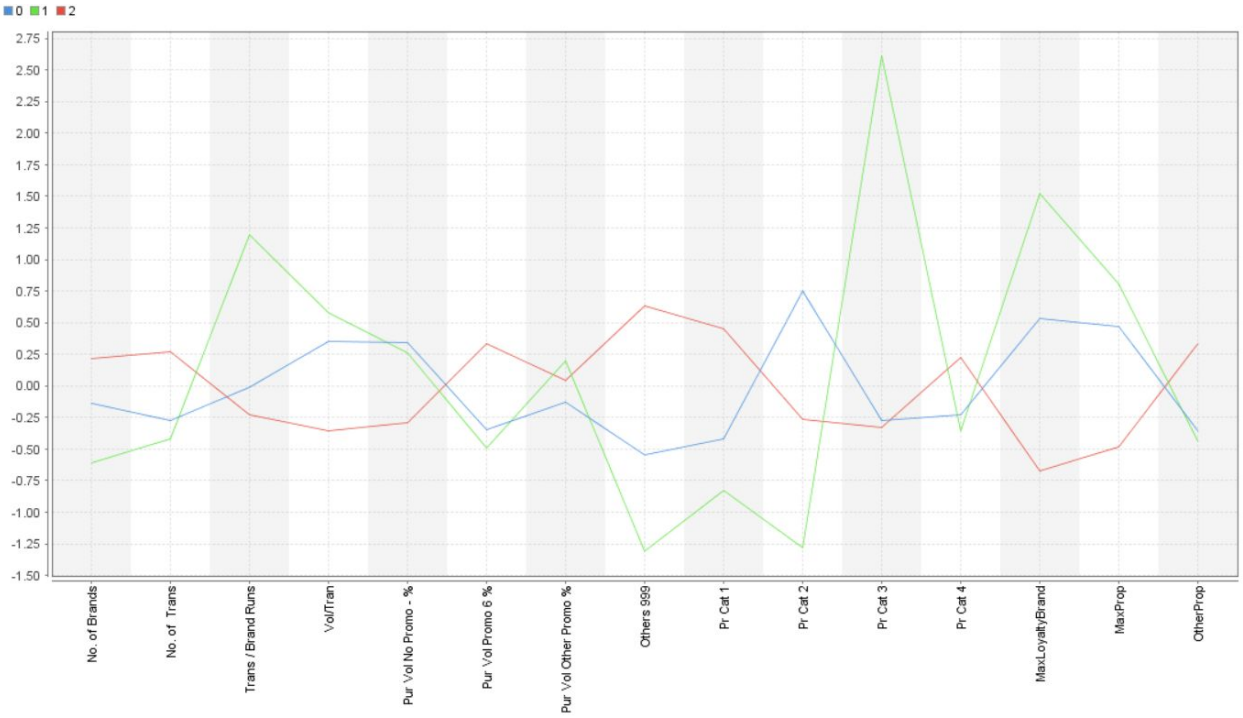
- Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	4.022
1.0	3.0	2.852
2.0	3.0	5.254

- Cluster Plot



- Here, we can see some distinction for Pr. Cat 3, Trans/Brand Runs etc. But, overall there are too many interactions and cluster do not provide a clear & distinguish picture.
- Also, the distance between Cluster 1 & 3 is comparatively less.

- **K = 4** ***(Best model for k=4 has been explained)***
  - Parameters
    - Max runs: 12
    - Measure types: MixedMeasures
    - Mixed Measures: MixedEuclideanDistance
    - Max Optimization Steps: 120

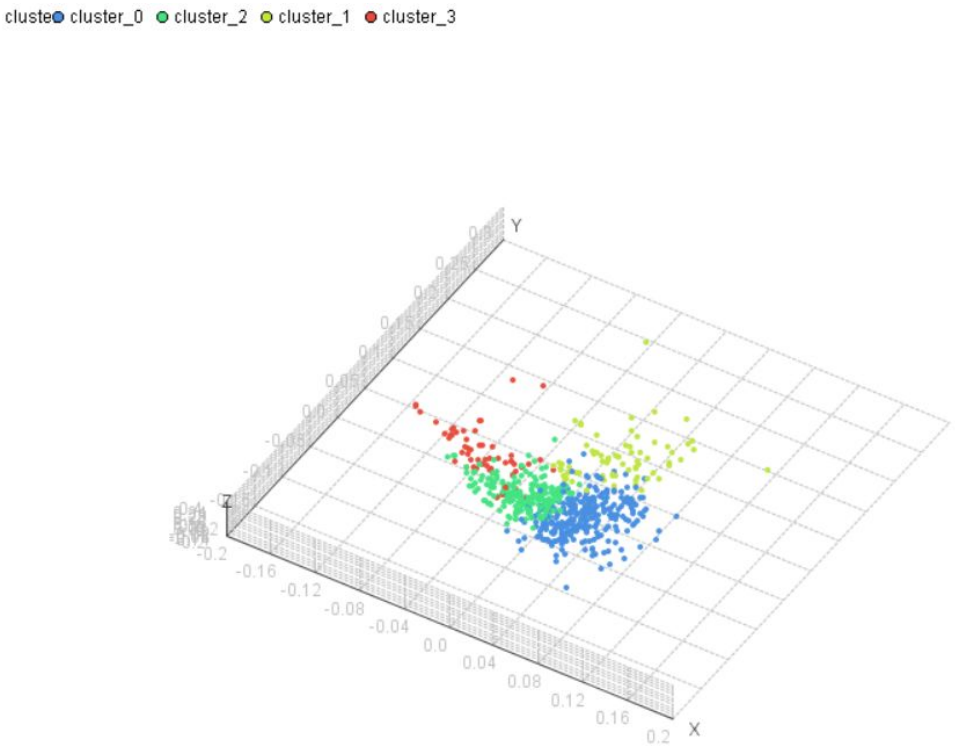
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -10.087
Avg. within centroid distance_cluster_0: -9.606
Avg. within centroid distance_cluster_1: -19.443
Avg. within centroid distance_cluster_2: -7.272
Avg. within centroid distance_cluster_3: -9.856
Davies Bouldin: -1.720
```

Cluster Model

```
Cluster 0: 285 items
Cluster 1: 72 items
Cluster 2: 186 items
Cluster 3: 57 items
Total number of items: 600
```

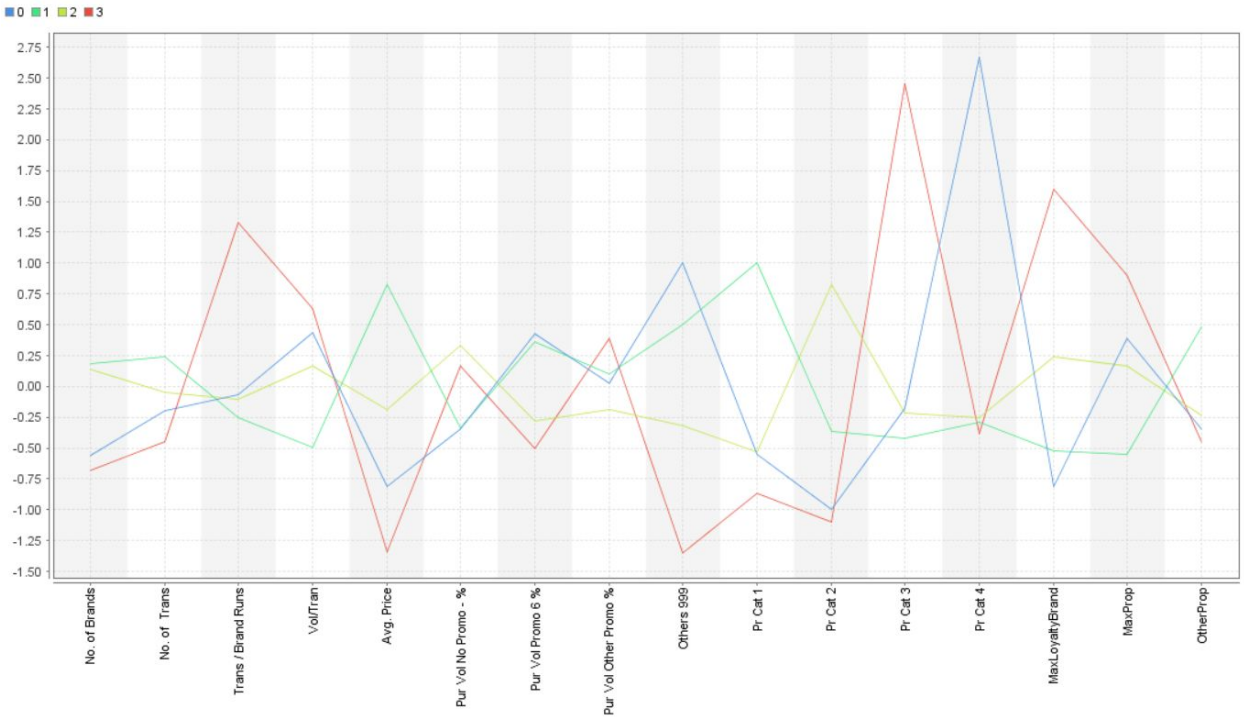
- Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	3.585
1.0	3.0	2.958
1.0	4.0	5.462
2.0	3.0	4.127
2.0	4.0	5.705
3.0	4.0	4.190

- Cluster Plot



- Here, despite having some interaction, we can see distinction for many variables compared to the previous cluster plot (K=3). Plus, no clusters provide the same information as well.
- So, despite not items not equally distributed among the clusters, comparatively this model gives better performance.

- From the comparison of 2 clusters above, we can say that model achieved with K = 4 is a better model. Because,
  - Each cluster can be categorized based on loyalty and basis of purchase comparatively better than other models.
  - Clusters are more distinct with very less interaction.

Assignment Question 2

(a) & (b)

- We have considered the model used in question 1(c) above to perform different clustering models. Because, for the overall market segmentation, basis for purchase & brand loyalty both should have an important role to perform.

- **K-medoids**
  - After trying different values of K for K-medoids model, we found the best model with K=3. Below is the summary for the same.

- Parameters
  - Max runs: 15
  - Measure types: MixedMeasures
  - Mixed Measures: MixedEuclideanDistance
  - Max Optimization Steps: 180

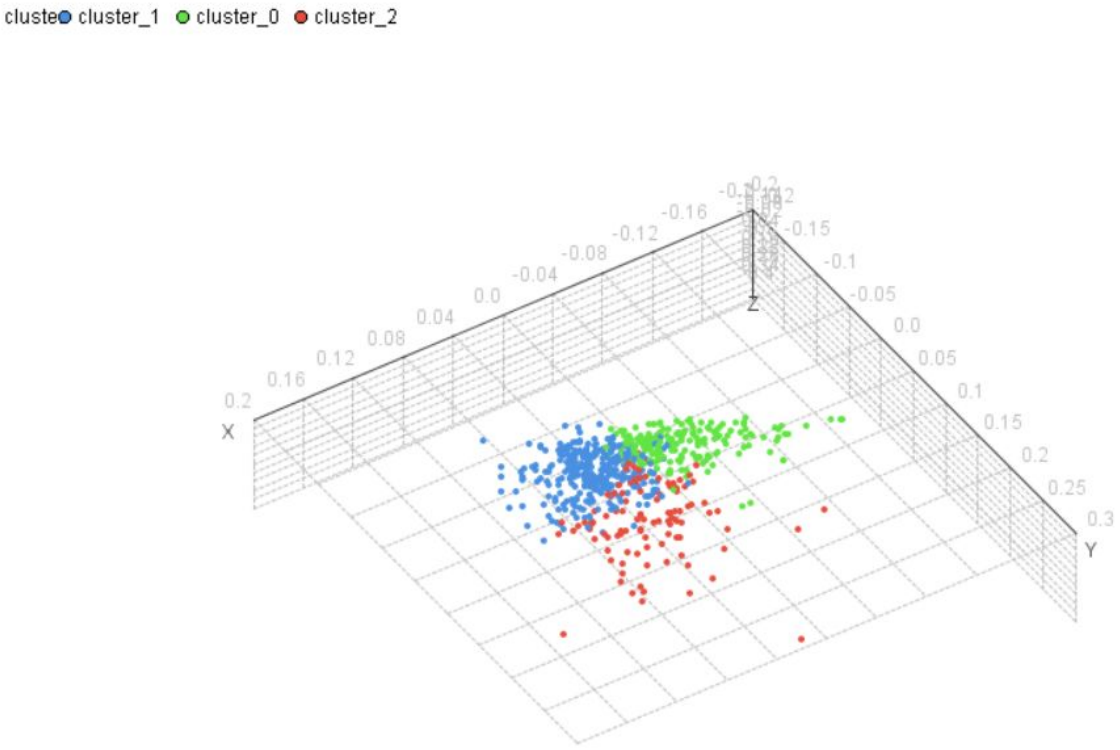
Performance Vector

```
PerformanceVector:
Avg. within centroid distance: -14.692
Avg. within centroid distance_cluster_0: -16.955
Avg. within centroid distance_cluster_1: -11.318
Avg. within centroid distance_cluster_2: -20.050
Davies Bouldin: -1.966
```

Cluster Model

```
Cluster 0: 181 items
Cluster 1: 304 items
Cluster 2: 115 items
Total number of items: 600
```

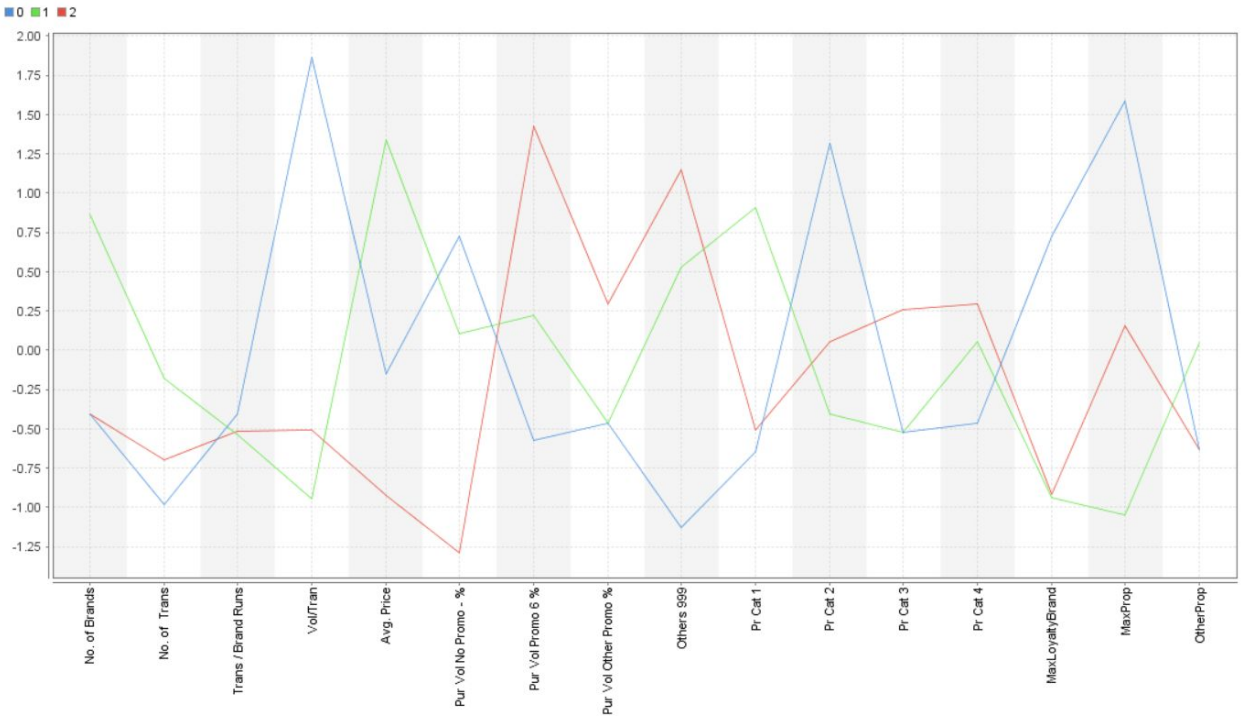
- Scatter Plot - 3D



- Centroid Distance

First	Second	Distance
1.0	2.0	5.452
1.0	3.0	5.204
2.0	3.0	3.351

- Cluster Plot



- The performance of K-medoids cluster is almost similar to k-means clustering for different parameters. The cluster distance and the interactions between variables behave in similar fashion as k-means clustering here.

- **Agglomerative**
    - After trying different values of, we found the best model with no. of clusters - 3. Below is the summary for the same.
    - Parameters
        - Mode: AverageLink
        - Measure types: NumericalMeasures
        - Numerical Measures: Camberra Distance

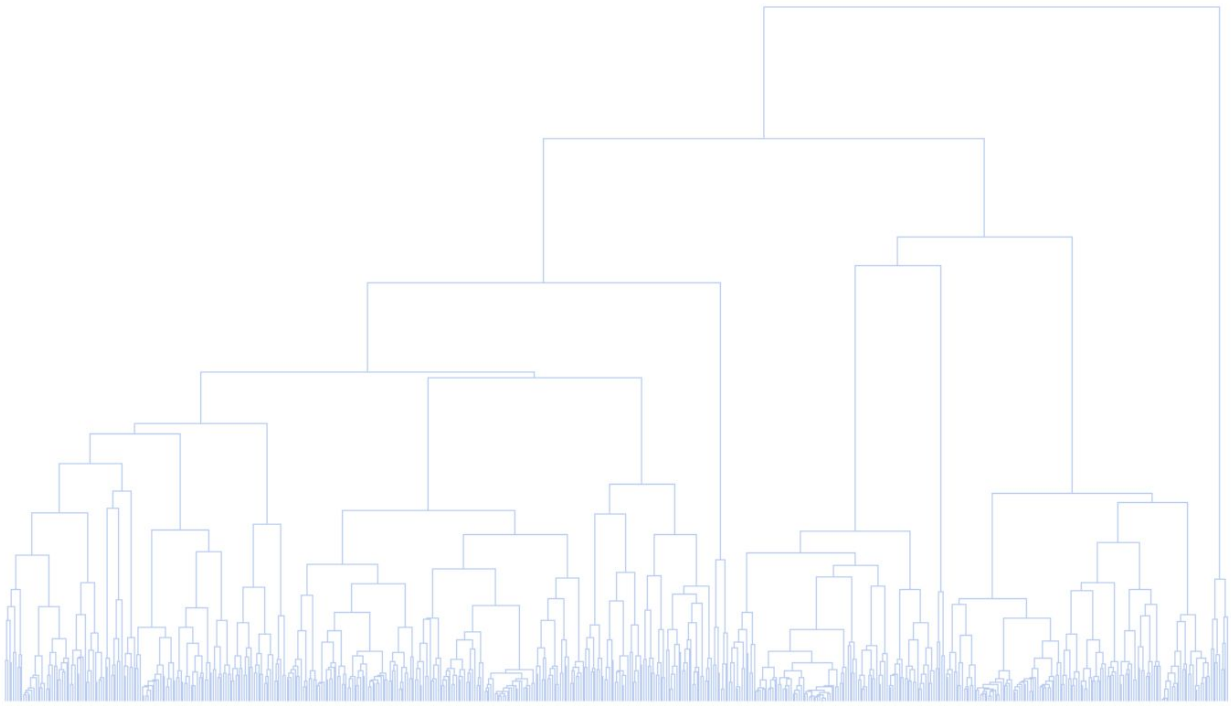
Performance Vector

```
PerformanceVector:
Number of clusters: 3.000
Cluster Number Index: 0.995
```

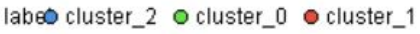
Cluster Model

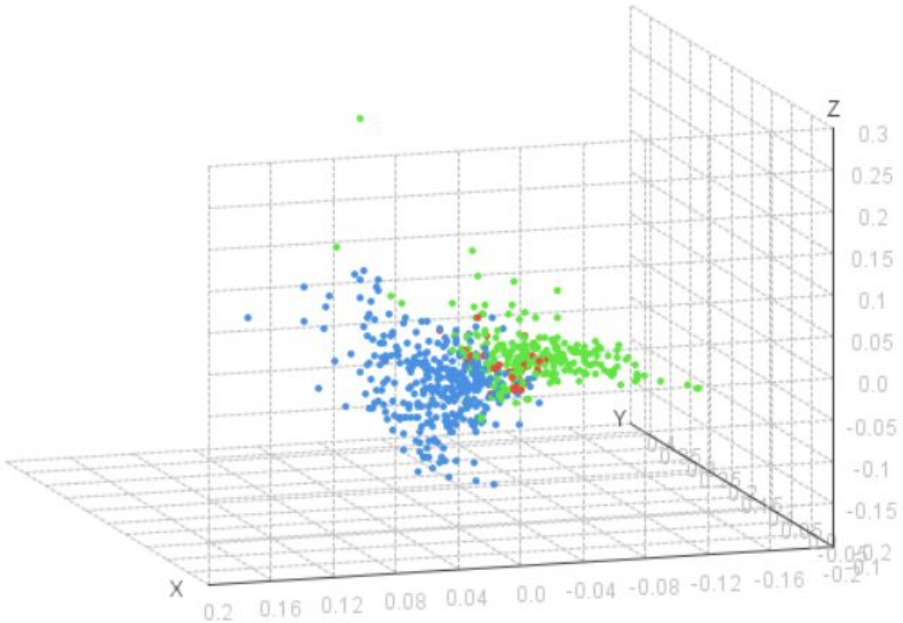
```
Cluster 0: 231 items
Cluster 1: 32 items
Cluster 2: 337 items
Total number of items: 600
```

    - Dendogram



- Scatter Plot - 3D





- For Agglomerative clustering, the single linkage did not give a good model. It was heavily allocating all the items to a single cluster.
- Although, complete linkage was able to distribute items to cluster pretty well, it was unable to distinguish the clusters. The clusters were very close to each other.

- **Kernel k-means**
  - After trying different type of kernels, we found the best model for kernel k-means clustering for dot kernel with k=3. Below is the summary for the same.

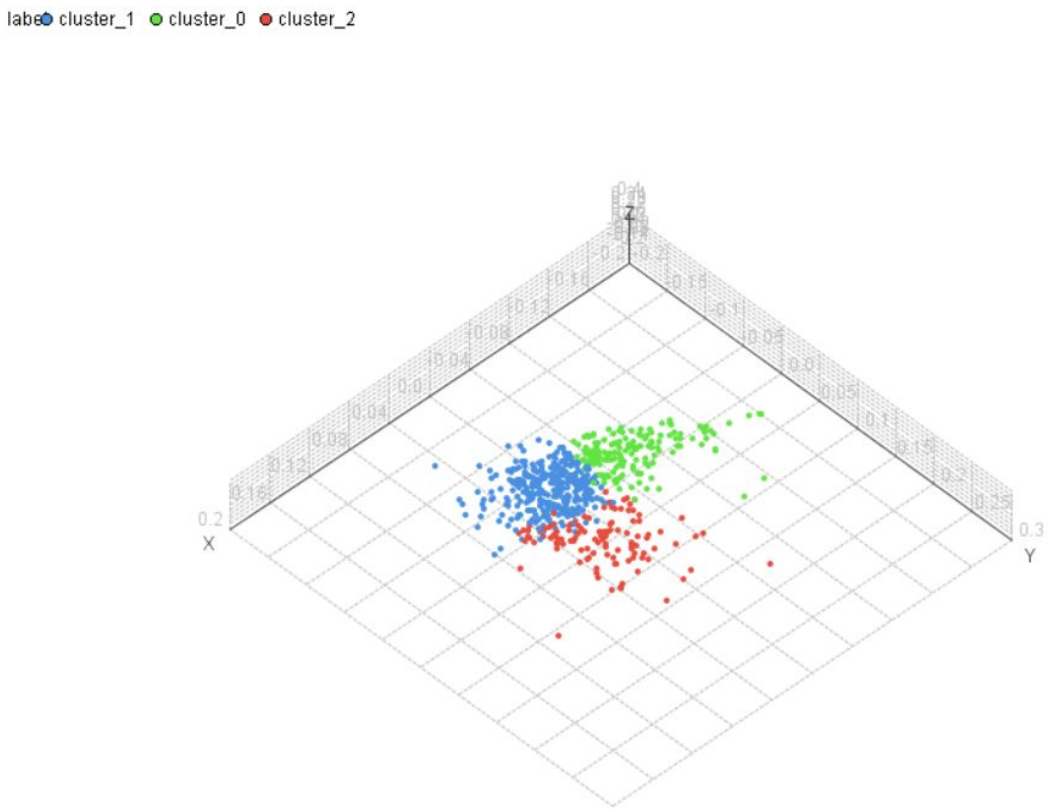
Performance Vector

```
PerformanceVector:
Number of clusters: 3.000
Cluster Number Index: 0.995
```

Cluster Model

```
Cluster 0: 161 items
Cluster 1: 324 items
Cluster 2: 115 items
Total number of items: 600
```

- Scatter Plot - 3D



- While experimenting with different kernel types with different values of k, we found that most of the kernels failed to equally distribute items to clusters. 90-05% of items were allocated to a single cluster.

- Also, those who were able to allocate equally, were unable to distinguish them properly. Dot kernel was the only one who gave better performance.

- **DBSCAN**
  - For this dataset, DBSCAN gives very bad performance for all different parameter values. Different values of epsilon & min points does not change the performance. Below is the summary of the best model among all DBSCAN which gives 2 clusters.
  - Parameters:
    - Epsilon: 2.0
    - Min points: 10
    - Measure types: MixedMeasures
    - Mixed Measures: MixedEuclideanDistance

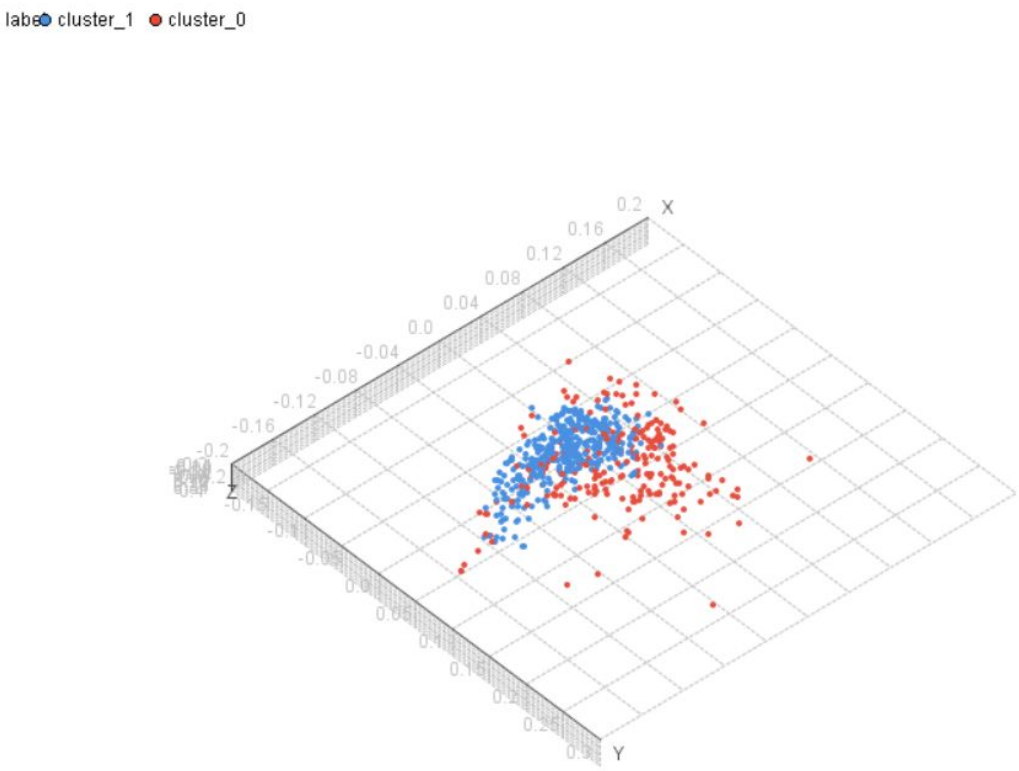
Performance Vector

```
PerformanceVector:
Avg. within cluster distance: -1476.646
Avg. within cluster distance for cluster 0: -1417.622
Avg. within cluster distance for cluster 1: -1511.063
```

Cluster Model

```
Cluster 0: 221 items
Cluster 1: 379 items
Total number of items: 600
```

  - Scatter Plot - 3D



  - For majority of the parameter values, DBSCAN is creating only one cluster. Where it is creating more than 1 cluster, the distribution is not equal.

- The above mentioned model is one of the few models which gives better distribution and more than 1 cluster.

- The reason we are getting different clusters for different procedures might be the characteristics of the cluster models.
  - The reason we are not getting good models with DBSCAN is because the current data set has varying density (data points are closer to each other). Whereas DBSCAN forms a separation for a lower density regions between the clusters - which is not true in our case.
  - Agglomerative does not clearly distinguish the clusters. In our dataset, clusters not distinguished hierarchically - that might be the reason for Agglomerative notr performing well.
  - The reason k-means & k-medoids are giving almost same performance is because the similarity in their characteristics.

Assignment Question 2(c) & Assignment Question 3(a)

- In order to identify the best model, we used Decision Trees. We ran all three models for which we were getting better performance (k-Means, k-Medoids & Kernel k-Means) in Decision Tree with Split criteria as Gain_ratio. Below is the summary of our output.
  - **k-Means**

accuracy: 99.17%

	true cluster_2	true cluster_1	true cluster_0	class precision
pred. cluster_2	317	5	0	98.45%
pred. cluster_1	0	213	0	100.00%
pred. cluster_0	0	0	65	100.00%
class recall	100.00%	97.71%	100.00%	

  - **k-Medoids**

accuracy: 83.83%

	true cluster_1	true cluster_0	true cluster_2	class precision
pred. cluster_1	262	12	38	83.97%
pred. cluster_0	42	169	5	78.24%
pred. cluster_2	0	0	72	100.00%
class recall	86.18%	93.37%	62.61%	

  - **Kernel k-Means**

accuracy: 98.33%

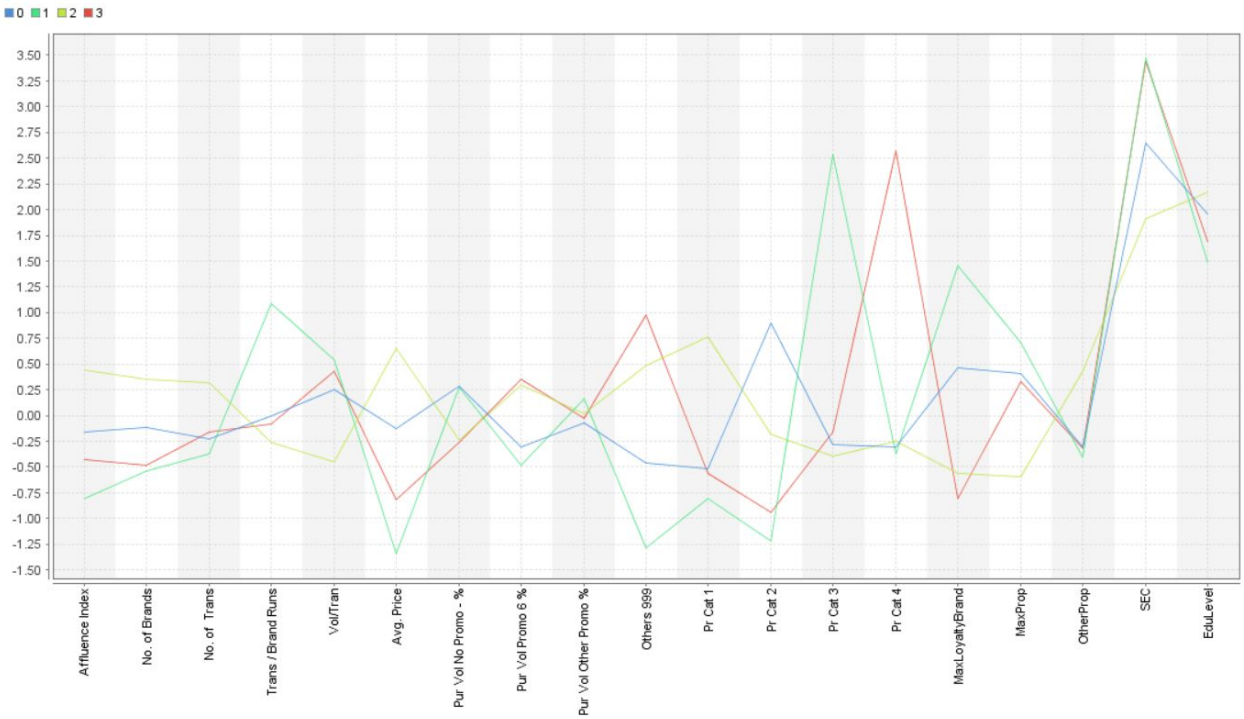
	true cluster_2	true cluster_1	true cluster_0	class precision
pred. cluster_2	331	2	3	98.51%
pred. cluster_1	4	189	1	97.42%
pred. cluster_0	0	0	70	100.00%
class recall	98.81%	98.95%	94.59%	

- Based on the above, we can see that the k-Means gives better accuracy for all 3 clusters. Hence, we can say that k-Means gives us the best clustering model for brand loyalty & basis of purchase.
- Apart from that, as explained in question 1(c), the clusters are very distinct with equal distribution in all. So, we can consider it the best model for this dataset.

Assignment Question 3

(b)

- In order to interpret the characteristics, we joined our model with the demographic variables. All demographic variables were not considered for it.
- SEC was included in the model since Socio Economic Class would affect the purchasing behaviour. Same goes with Affluence Index as well.
- For EDU variable, since many of the codes were providing similar information, we combined them and created a new variable ‘EduLevel’ with three values - 1, 2 & 3
    - 1 - Code ‘0’ or ‘1’ or ‘2’
    - 2 - Code ‘3’ or ‘4’ or ‘5’
    - 3 - Code ‘6’ or ‘7’ or ‘8’ or ‘9’
- Apart from this, various demographic variables were transformed and were experimented in the model.
    - MT was transformed in 3 values - ‘1’ - Marathi, ‘2’ - Gujarati & ‘3’ - Others (Since Marathi & Gujarati had highest numbers comparatively)
    - Other variables which were considered are AGE, FEH, SEX, CHILD.
    - However, we did not find significantly distinct clusters considering these variables. So they are not used in the final model.
- Here is the cluster plot of the model.



- From the plot, we can see that households with lower education level tends to have high brand loyalty. Also, they seems to prefer Economy/Carbolic soaps.
- Households with lower socioeconomic class also tends to have high loyalty. Also, they seem to have low no. of transaction and highest Vol/Trans - which means they prefer to buy in bulk.
- Households with high education level seems to be low on brand loyalty. Their number of brands seems to be high - which means they prefer products of different brands. They also seem to favour Premium soaps.
- Households with Low AI tends to have high brand loyalty and they seem to prefer Economic/Carbolic soaps. They also prefer buying in bulk.
- Households with High AI tends have lower brand loyalty. They prefer premium soaps. They also prefer buying in small chunk since vol/trans is lowest for them and no. of transactions are high.

(c)

- We obtained the best clustering from K-means clustering algorithm which provided us with 4 different segments of customers. The different segments formed were as follows:

  - **Most Loyal Households:** The households belong to the least affluent class of the society having the lowest socioeconomic status. These customers mostly purchase a high proportion of economy/carbolic soaps and they follow a longer streak of buying the same brand of soaps. The customers belonging to this section have a low level of education, mostly illiterate or having no formal

schooling.

- **Moderately Loyal Households:** The customers coming from these households are the ones who have attained schooling and belong to a relatively higher socioeconomic status. These set of households buy the popular soaps having a higher volume of purchases which have no promotion codes. These customers come from a comparatively highly affluent households and are loyal to a particular brand for a considerable amount of time before switching.

- **Less Loyal Households:** The most affluent households are the ones who are comparatively less loyal to the given brands. These customers belong to the highest socioeconomic status and are graduates or professionals. These customers buy premium soaps and switch back and forth between the brands the most. It can be judged that the most affluent customers usually experiment with different brands in the market to choose the best one.

- **Least Loyal Households:** The customers belong to lower socioeconomic households and highly prefer sub-popular soaps which are not included in the given common type of brands. These households being less affluent always try to purchase products which offer some promotion codes.

- As stated in Question 3(a), we compared the different clustering algorithms using a decision tree and found that K-means clustering gave us the best model.
- Decision trees were found to be useful in classifying the observations correctly in different clusters. As we can see from the confusion matrix of K-means algorithm with only 5 incorrectly predicted observations the algorithm has an accuracy of 99.17%.
- With the decision tree with the different clusters acting as terminal nodes we can create a decision rules to find out on what basis the different clusters are formed.