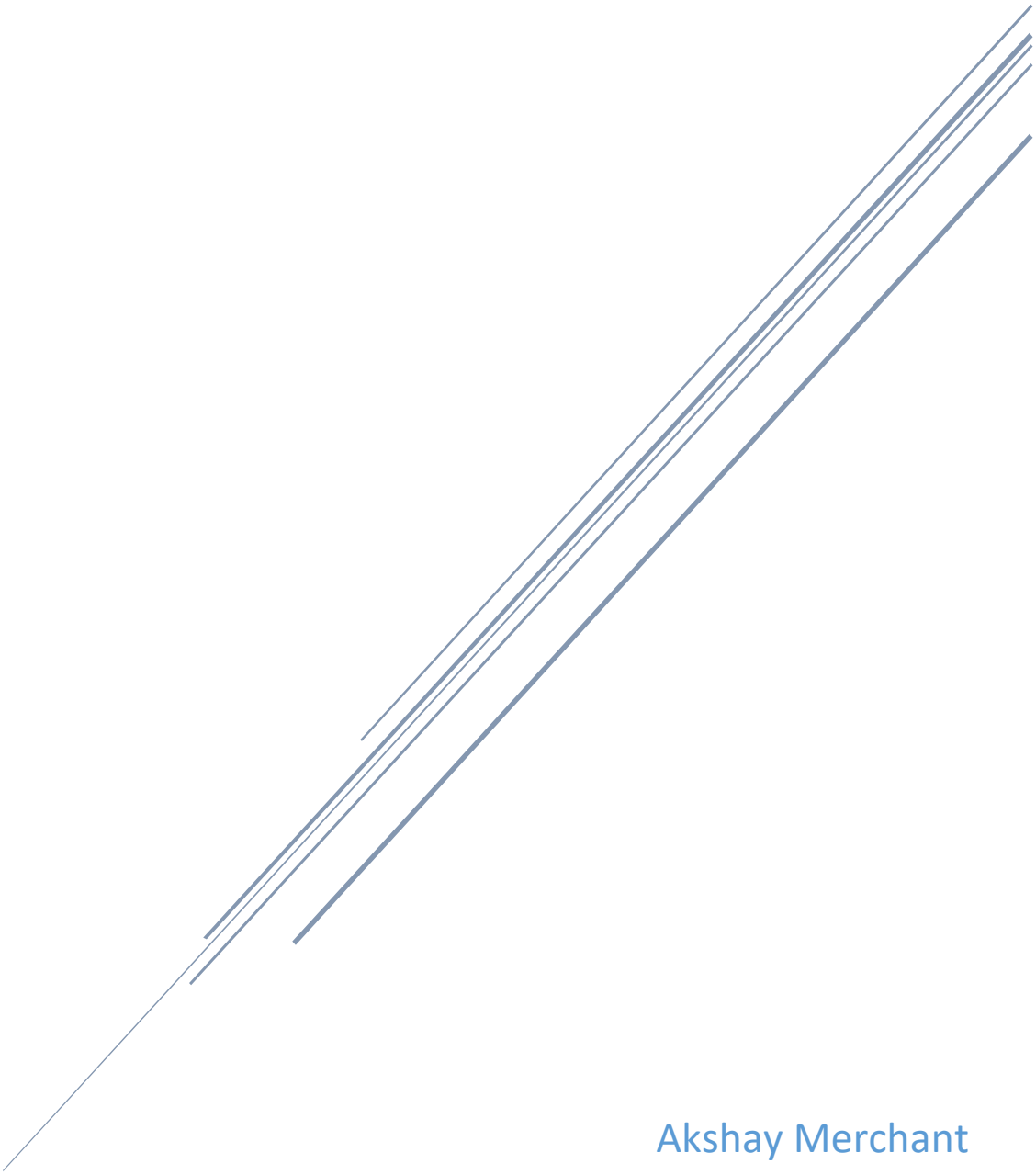
# MARKET SEGMENTATION

## Assignment 4



Akshay Merchant
Ashuthosh Gowda
Laavanya Ganesh

**Assignment 4**

**Answer 1.a.**

We initially used the variables from the dataset that describe purchase behaviour. They were

| #brands | brand runs | total volume | #transactions | value | avg. price | share to other brands |
|---|---|---|---|---|---|---|

We also developed a variable called brand loyalty which was also used to describe purchase behaviour.

We used the brand wise purchase variables and the #brands variable to make brand loyalty. There were 9 brand wise purchase variables:

| BrCd57,144 | BrCd55 | BrCd272 | BrCd286 | BrCd24 | BrCd481 | BrCd352 | BrCd5 | BrCd999 |
|---|---|---|---|---|---|---|---|---|

From these BrCd999(other) could include 1 brand or 100 brands or anything in between. Since we did not know how many brands were included in BrCd(999) we used the following procedure to develop brand loyalty.

1. We counted the number of brands used by each household from the 1st 8 brand wise purchase variables.

| BrCd57,144 | BrCd55 | BrCd272 | BrCd286 | BrCd24 | BrCd481 | BrCd352 | BrCd5 |
|---|---|---|---|---|---|---|---|
| 38% | 13% | 0% | 0% | 0% | 0% | 0% | 0% |

From the data above we can see that this particular household used BrCd(57,144) 37% of the time and BrCd(55) 13% of the time. Thus the number of brands used by this household from the 8 different brands mentioned above is 2.

2. We then subtracted this number (2) from #brands. This gave us the number of BrCd999(other) brands used by this household.

No of brands used from 8 BrCd Variables : 2

#brands: 3

No of BrCd999(other) brands used: 3-2=1

3. We then divided the value in BrCd999(other) by No of BrCd999(other) brands used. This was done to get a rough estimate of the proportion of each brand that was used.

BrCd999(other): 49%

No of BrCd999(other) brands used: 1

Proportion of each BrCd999(other) brand used: 49/1=49%

4. We then took a standard deviation of all 9 BrCd Variables. The value given by standard deviation, if low, means that the household is less loyal. If the deviation value is high then it means that the household is more loyal.

| BrCd57,144 | BrCd55 | BrCd272 | BrCd286 | BrCd24 | BrCd481 | BrCd352 | BrCd5 | BrCd999 |
|---|---|---|---|---|---|---|---|---|
| 38% | 13% | 0% | 0% | 0% | 0% | 0% | 0% | 49% |

Standard Deviation: 19%

5. We then used min-max normalization to scale the brand loyalty values between 0 and 1. The closer the brand loyalty value is to 1, the less loyal a household is likely to be.

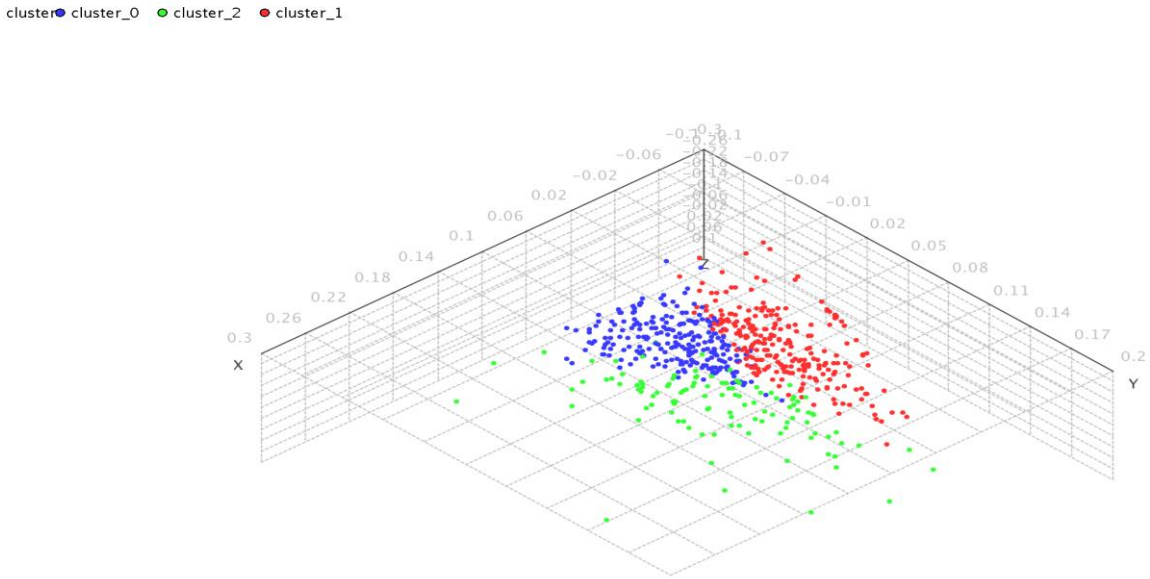
This is how we developed brand loyalty.

We normalized all the variables using Z-transform.

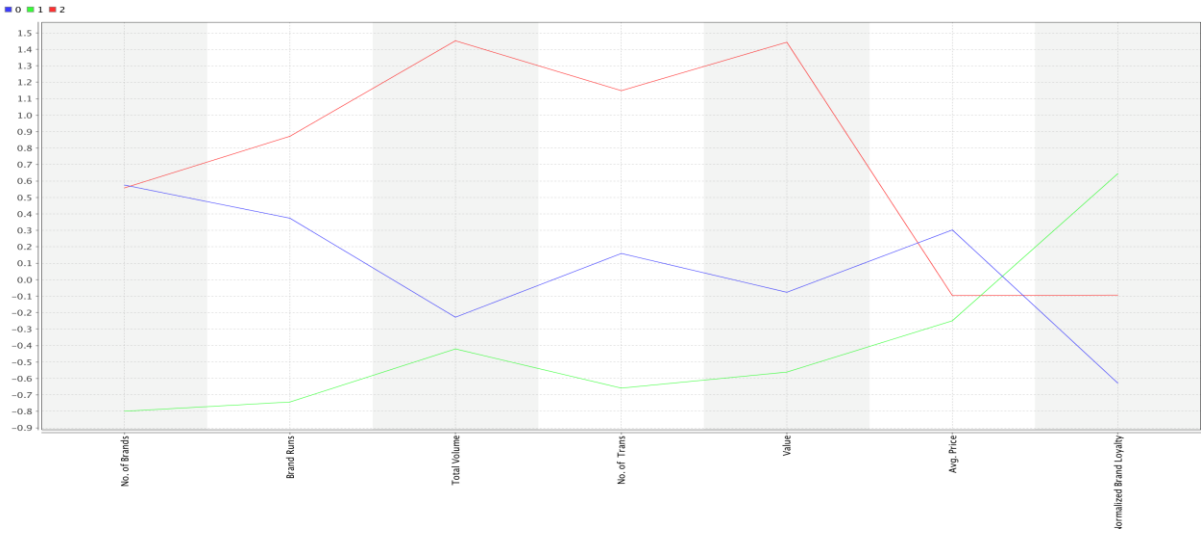
We then performed k-means clustering on the variables that described purchase behaviour. We used clustering to help discover distinct groups within purchase behaviour and partition that data into groups and then assign labels to those groups.

The data shown below is for the clustering model we got using the purchase behaviour variables.

Shown below is the 3-d scatter plot for k=3.




Shown below is the variable plot for k=3



| k | Average intra-cluster distances |
|---|---|
| 2 | 0.922 |
| 3 | 0.718 |
| 4 | 0.542 |
| 5 | 0.483 |

We have chosen k=3 as the number of clusters due to the following reasons:

The variable plot for k=3 clearly defined each cluster while variable plots for k=4,5 had values in attributes overlapping to some extent. Thus, even though the average intra cluster distance was lower for k=4,5 we preferred k=3 as each cluster in k=3 clearly defined the values for the purchase behaviour attributes.

**Answer 1.b.**

We then used k-means clustering on the variables that described basis-for-purchase.

The variables used to describe basis-for-purchase were

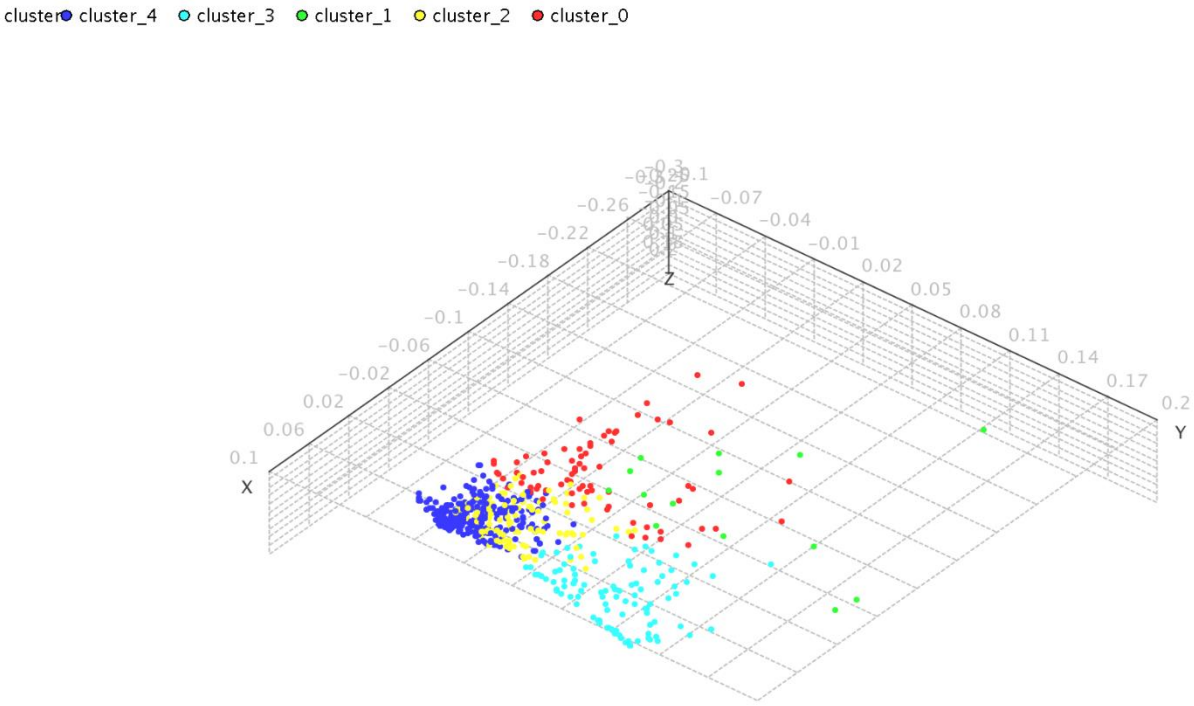
| Pur_vol_no_promo | Pur_vol_promo_6 | Pur_vol_other | Price_Categories | Selling_Proposition |
|---|---|---|---|---|

For selling proposition we decided to create a new variable which included the maximum value from all the selling propositions for that household. We used

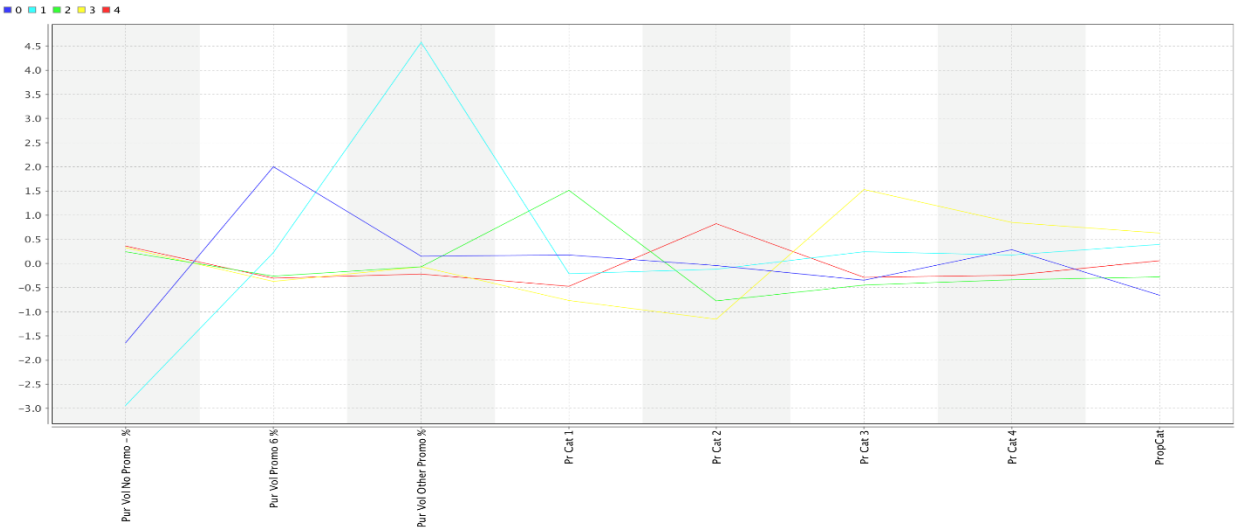
We normalized all the variables using Z-transform.

The data shown below is for the clustering model we got using the basis-for-purchase variables.

Shown below is the 3-d scatter plot for k=5.




Shown below is the variable plot for basis of purchase variables




.

| k | Average intra-cluster distances |
|---|---|
| 2 | 0.823 |
| 3 | 0.759 |
| 4 | 0.645 |
| 5 | 0.503 |

For all values of k, we got similar variable plots, i.e. they all overlapped for certain values of some of the basis-for-purchase variables. Thus, we chose k=5, as it had the lowest average intra cluster distance. This meant that clusters for k=5 were more compact than the clusters for the other k values.

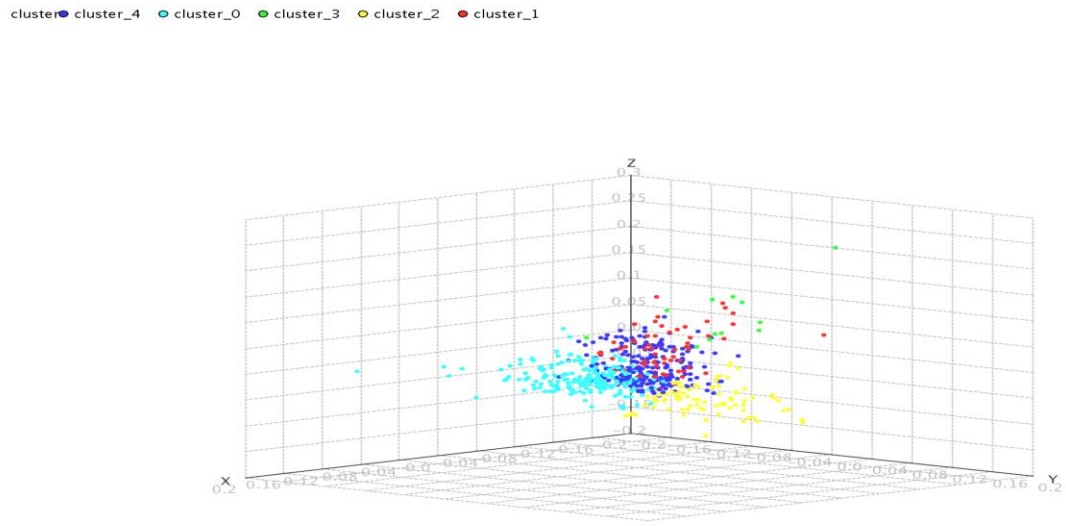
**Answer 1.c.**

We then combined the variables for purchase behaviour and basis for purchase and performed k-means clustering.

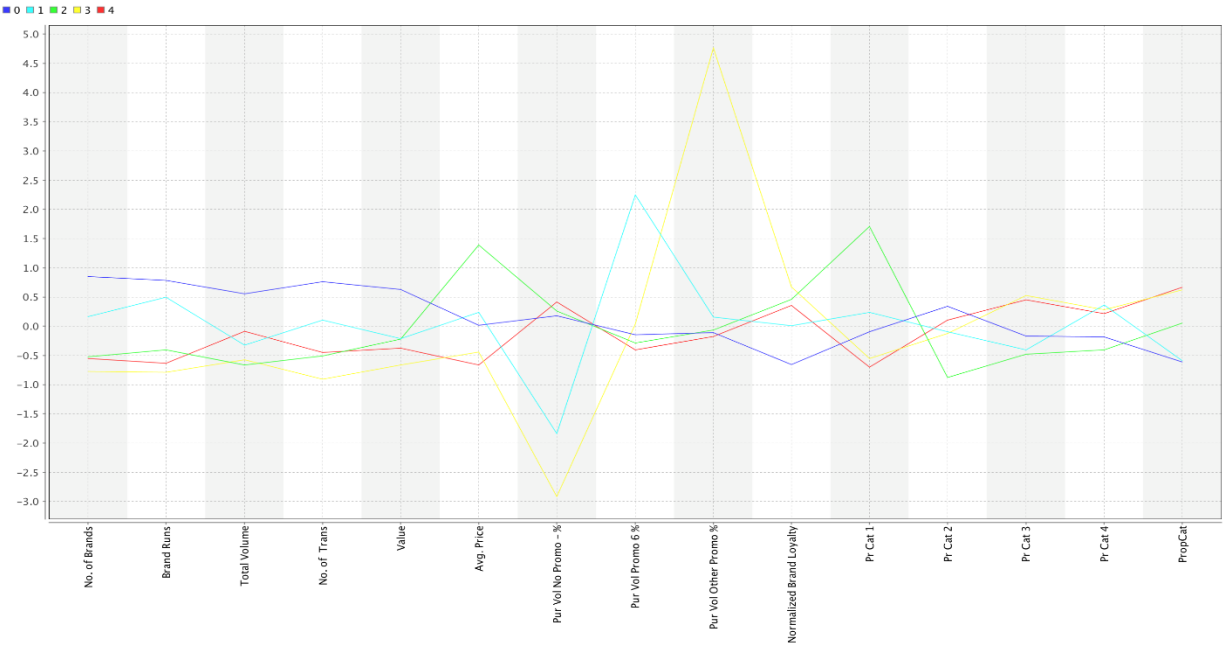
We normalized all the variables using Z-transform.

The data shown below is for the clustering model we got using the combined variables.

Shown below is a 3-d scatter for k=5




Shown below is the variable plot



As you can see it is not easy to differentiate values of the variables for each cluster since they overlap.

| k | Average intra-cluster distances |
|---|---|
| 2 | 0.838 |
| 3 | 0.763 |
| 4 | 0.676 |
| 5 | 0.638 |

For all values of k, we got similar variable plots, i.e. they all overlapped for certain values of some of the variables. Thus, we chose k=5, as it had the lowest average intra cluster distance. This meant that clusters for k=5 were more compact than the clusters for the other k values.

For k-means clustering used on the different data segments mentioned above, we used a value of k between 2 and 5. This was done so that we could then segregate the data into different clusters depending on the k values.

**Answer 1.d.**

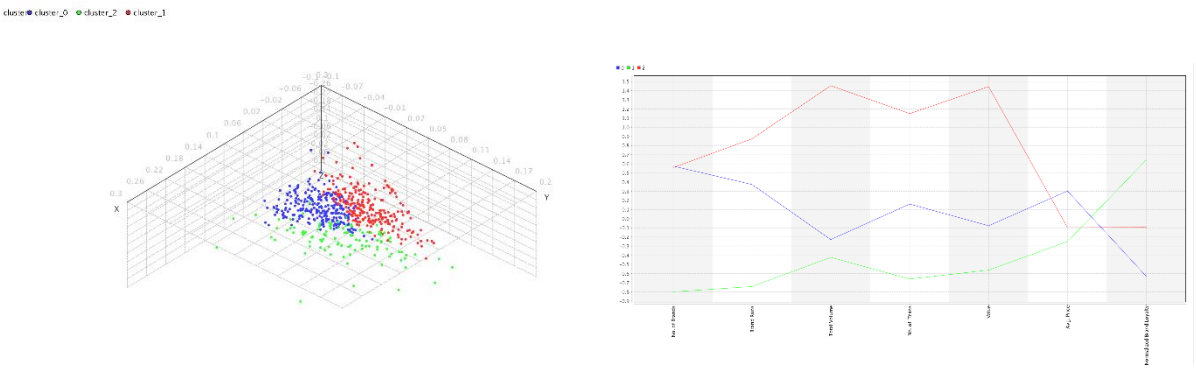
We got out best k-means model in 1(a), i.e. using k-means clustering to describe purchase behaviour. We thought that this was the best model as it distinctly defined each attribute in a cluster better than the other two did, namely 1(b) i.e. variables describing the basis for purchase and 1(c) i.e. a combination of the two segments above.

We then used different clustering techniques on the variables used to describe purchase behaviour.

We compared the models using Accuracy from the Decision Tree.

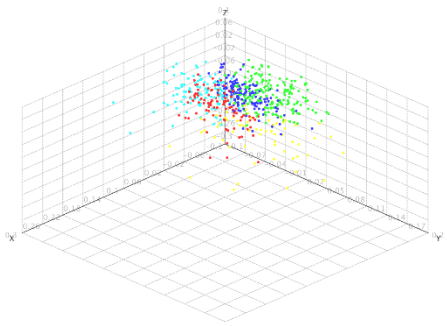
| Technique | Description | Accuracy |
|---|---|---|
| K-Means | K=3, Max runs=100 | 94.66% |
| K-Medoids | K=5, Max runs=100 | 89.82% |
| Kernel k-means | K=3, kernel type=dot | 94.54% |
| Agglomerative | Complete linkage, flattening=7 | 94.32% |
| DBSCAN | Epsilon= 1, Min Points=10, dimension=3 | 87.67% |

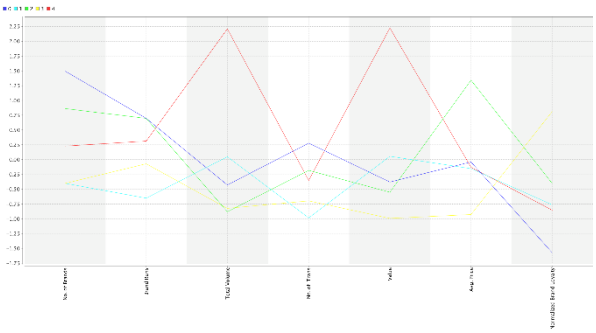
K-Means Clustering:



K-Medoid Clustering:

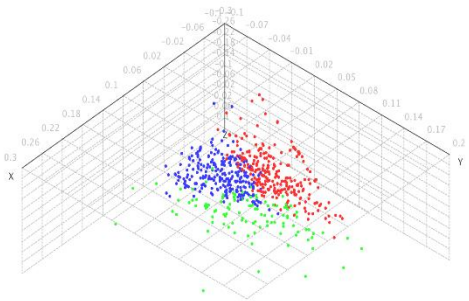






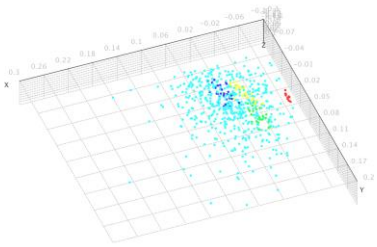
Kernel K-Means Clustering:





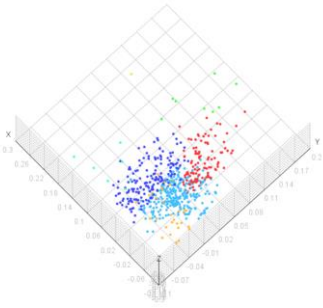
Density Based Spatial Clustering of Applications with Noise(DBSCAN)





Agglomerative Clustering:





Parameter Changes

Density Based Spatial Clustering of Applications with Noise(DBSCAN):

| Epsilon | Min Points | Average Within Cluster Distance |
|---|---|---|
| 0.5 | 5 | 1885.33 |
| 1 | 5 | 1007.325 |
| 1 | 10 | 874.604 |

In DBSCAN as we increase Epsilon and Min Points, the Average Within Cluster Distance decreases.

K-Medoids Clustering:

| K | Intra Cluster Distance | Inter Cluster Distance Range |
|---|---|---|
| 2 | 7.326 | 3.802 |
| 3 | 6.313 | 2.967- 4.497 |
| 4 | 5.430 | 2.315- 4.865 |
| 5 | 4.903 | 2.013-4.865 |

As we increase k, the intra cluster distance decreases and the inter cluster distance range increases.

Agglomerative Clustering:

| Number of Clusters | Single linkage | Complete linkage | Average linkage |
|---|---|---|---|
| 3 | 0.98 | 0.417 | 0.895 |
| 4 | 0.983 | 0.627 | 0.897 |
| 5 | 0.987 | 0.741 | 0.935 |
| 7 | 0.99 | 0.993 | 0.97 |

As the number of clusters increases, the performance vector for single, complete and average linkage increases.

Kernel K-Means Clustering:

| K | Average Within Cluster Distance | |
|---|---|---|
| | Dot Kernel | Radial Kernel |
| 3 | 338.844 | 685.349 |
| 5 | 562.464 | 559.835 |
| 8 | 174.266 | 403.470 |

As K increases the average distance within cluster for Dot Kernel and Radial Kernel decreases.

As you can see, the clusters we got from different techniques were different. We think this is because each clustering method has a different basis for computation of the clusters. Thus each clustering method will give a different outcome.

K-means uses means to find the centres for the clusters.

K-medoids uses medoids as the centres for the clusters. Also K-means is not as robust to outliers as K-medoids.

DBSCAN groups together points that are closely packed together (points with many nearby neighbours).

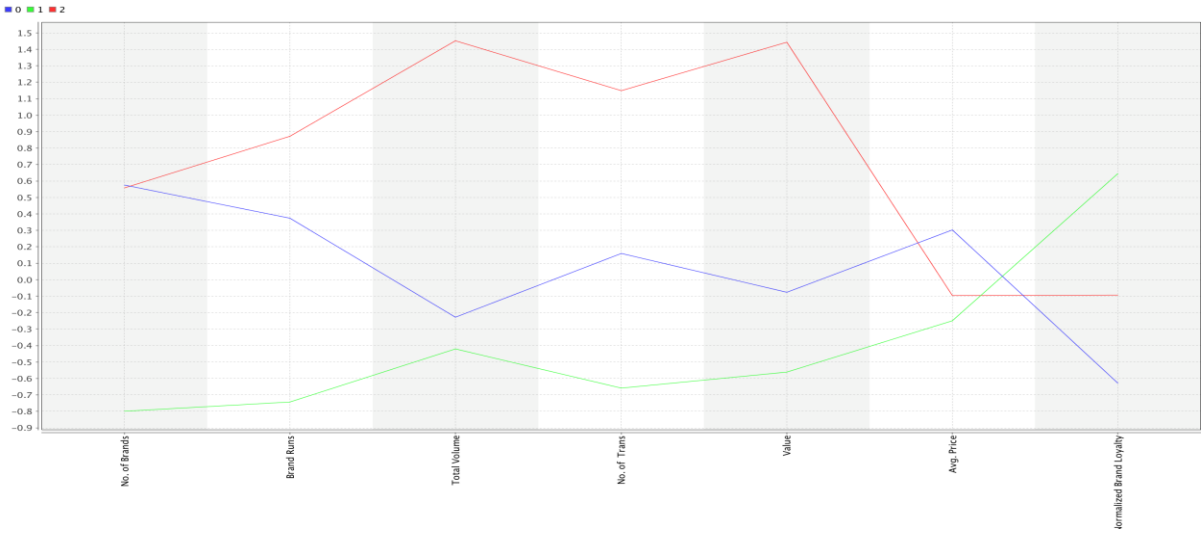
Kernel K-Means works in the same way as K-Means, but uses a kernel method to calculate distance instead of Euclidian distance.

In Agglomerative Clustering all observations start in their own cluster and pairs of clusters are merged as one moves up the hierarchy.

**Answer 2.a**

We think that the best segmentation is the one with purchase behaviour variables. When we performed clustering on this data using k-mean clustering we got distinct values for each attribute that explained their corresponding clusters.

k-means variable plot for purchase behaviour variables




Cluster 0(blue): in cluster 0 we see that households that buy a high number of brands, have a moderate number of brand runs, purchase in small volume with not very high product values and have a modest number of transactions have a low brand loyalty.

Cluster 1(green): in cluster 1 we see that households that buy fewer brands, have lower brand runs, purchase in small volume with smaller product values and have lower number of transactions have a high brand loyalty.

Cluster 2(red): in cluster 2 we see that households that buy many brands, have a high number of brand runs, purchase in high volumes with high cost products and have a high number of transactions have a medium brand loyalty.

These purchase behaviour variables help identify the group of households that should be targeted for specific campaigns.

For example, it does not make sense to send advertising to households (cluster 1) that have a large brand loyalty but buy few brands, but it might help to send them discounts and vouchers for those brands that they buy.

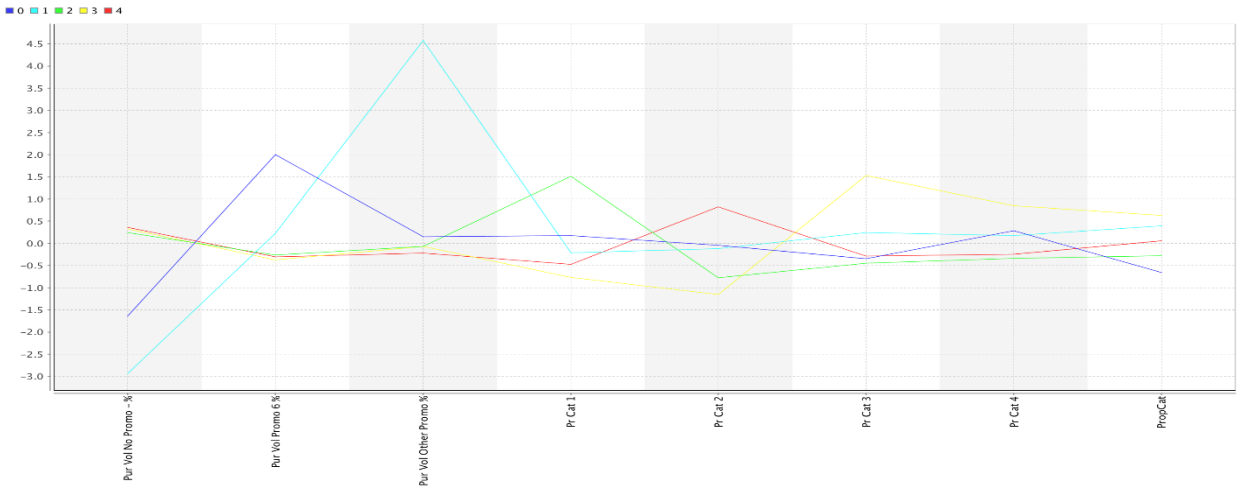
Targeted advertising and promotions could instead be sent to those households (cluster 2) that purchase in large volumes and are somewhat loyal as it would justify the advertising cost for the company.

For the other households (cluster 3), one could use common in-store or online discounts to predict whether their buying power and brand loyalty justifies the cost of sending advertisements and promotions.

We believe that clustering based on purchase behaviour could help companies predict which households to target. This would directly relate to an improvement in their marketing and promotional campaigns. This fact combined with the distinctness and individuality of the clusters made us choose the purchase behaviour variables as the best segmentation.

Using basis-for-purchase variables would help define the households to target for certain promotional offers.

As you can see from the variable plot above, the households in cluster 0 (dark blue) purchase the highest volume during the promo code 6. This information could be used to provide them early access to future deals like promo code 6 deals.




We could also use a combination of the two segments above, but the clusters we obtained were not as clear as their individual clusters, thus we preferred clustering these segments individual to gain more information.

**Answer 2.b**

To calculate demographic distribution to the clusters, we have bucketed the Variables as:

| Demographics | Transformation |
|---|---|
| SEC | No Transformation |
| FEH | 0 - Non Veg |
| MT | 1-Gujarati |
| | 2-Marathi |
| | 0- Everything Else |
| SEX | 0- Male |
| | 1-Female |
| AGE | No change |

| | |
|---|---|
| EDU | 0:4 - 0 |
| | 5 -1 |
| | 6:9 - 2 |
| HS | No change |
| Child | No change |
| CS | No change |
| Affluent Index | 0:10 – 1 |
| | 10:20 – 2 |
| | 20:30 – 3 |
| | 30:40 – 4 |
| | 40:50 – 5 |

We have plotted the distribution of each demographic attribute among the clusters

Based on this information, we can make conclusions about each clusters dominant demographics:

| Demograph | Cluster_0 | Cluster_1 | Cluster_2 |
|---|---|---|---|
| SEC | Majority is SEC > 3 | Majority if SEC = 4 | Majority is SEC <3 |
| FEH | All FEH are there | Majority is Non- Veg | Majority are Veg |
| Mother Tongue | All languages are present | Majority are other languages | Majority is Marathis |
| EDU | Evenly distributed | Majority are EDU <4 | Majority have EDU > 6 |
| HS | HS > 6 are dominant | HS less than 3 | HS between 3 and 6 |
| Age | Evenly distributed | majority AGE = 2 or 3 | Evenly distributed |
| Affluent Index | Evenly distributed | AI= 1,2 | AI = 3,4,5 |

Distribution of Social Economic Condition:

| Cluster | SEC1 | SEC2 | SEC3 | SEC4 |
|---|---|---|---|---|
| cluster_0 | 13% | 14% | 19% | 20% |
| cluster_1 | 43% | 39% | 45% | 58% |
| cluster_2 | 44% | 47% | 36% | 22% |
| Grand Total | 150 | 150 | 150 | 150 |

Distribution of Food Eating Habits:

| Cluster | Veg | Egg/Veg | Non Veg |
|---|---|---|---|
| cluster_0 | 10% | 12% | 19% |
| cluster_1 | 43% | 44% | 48% |
| cluster_2 | 47% | 44% | 33% |
| Grand Total | 165 | 34 | 401 |

Distribution of Mother Tongue:

| Cluster | Others | Marathi | Gujarati |
|---|---|---|---|
| cluster_0 | 15% | 19% | 8% |
| cluster_1 | 55% | 41% | 46% |
| cluster_2 | 29% | 40% | 46% |

| | | | |
|---|---|---|---|
| Grand Total | 191 | 326 | 83 |

Education:

| Cluster | Sum of EDU0 | Sum of EDU1 | Sum of EDU2 |
|---|---|---|---|
| cluster_0 | 19% | 15% | 12% |
| cluster_1 | 56% | 39% | 32% |
| cluster_2 | 25% | 46% | 56% |
| Grand Total | 300 | 189 | 111 |

Numeric Variables

| Cluster | Average of HS | Average of CHILD |
|---|---|---|
| cluster_0 | 6.09 | 2.90 |
| cluster_1 | 3.49 | 3.45 |
| cluster_2 | 4.21 | 3.10 |

Distribution of Age:

| Cluster | Sum of AGE 1 | Sum of AGE 2 | Sum of AGE 4 | Sum of AGE 3 |
|---|---|---|---|---|
| cluster_0 | 20% | 9% | 22% | 14% |
| cluster_1 | 40% | 57% | 41% | 47% |
| cluster_2 | 40% | 34% | 37% | 39% |
| Grand Total | 15 | 129 | 287 | 169 |

Distribution of Affluence Index:

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|
| cluster_0 | 6% | 20% | 19% | 16% | 22% | 67% | 99 |
| cluster_1 | 74% | 45% | 30% | 32% | 33% | 0% | 278 |
| cluster_2 | 21% | 35% | 51% | 52% | 44% | 33% | 223 |
| Total | 136 | 260 | 110 | 73 | 18 | 3 | |

**Answer 3.**
We used the Clusters assigned for each Data points as labels and built a decision tree model using the demographic attributes.
The decision tree gave an accuracy of 93% for the differentiation of clusters.

Performance Matrix:

| accuracy: 93.00% | true cluster_1 | true cluster_2 | true cluster_0 | class precision |
|---|---|---|---|---|
| pred. cluster_1 | 298 | 19 | 9 | 91.41% |
| pred. cluster_2 | 5 | 180 | 5 | 94.74% |
| pred. cluster_0 | 3 | 1 | 80 | 95.24% |
| class recall | 97.39% | 90.00% | 85.11% | |

The decision tree helps in identifying the demographic attributes which will decisively distinguish the data points to be bucketed into clusters. Further, this model could be run on any new data points to determine which cluster it would most likely fall into. Now, we can use the Marketing strategy assigned for the cluster.

A decision tree can be used to distinguish between clusters, however it must be done carefully:
1. We first run our clustering algorithm to get the cluster labels.

2. We then partition the data points into train and validation.

3. Build a decision tree model on the train data, use the same model on validation.

4. Now, we can compare the different Clustering techniques using the accuracy.