SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
Malabe

Fundamentals of Data Mining
IT 3051
Mini Project
Segmenting Consumers on Bath Soap

PREPARED BY

IT16073838 - Dilini H.O. Pathirana
IT16071094 - G.C.L. Chandrasiri
IT13112660 - M.D.S.Jayasekara
IT16038288 - D.P.Madanayake
IT16017702 - H.L.K.Tharushika
IT16017498 - A.A.N.H.Amarasinghe
IT16030954 - Nuzrath M.I.Z.S
IT16161252 - D.H.Vishara Dilmi

Contents

Introduction

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and nondurable). In one major research project, CRISA tracks about 30 product categories (e.g., detergents), and within each category, about 60 to 70 brands. To track purchase behavior, CRISA constituted about 50,000 household panels in 105 cities and towns in India, covering about 80% of the Indian urban market. (In addition to this, there are 25, 000 sample households selected in rural areas, but we are working only with urban market data). The households are carefully selected using stratified sampling. The strata are defined on the basis of socioeconomic status and the market (a collection of cities). CRISA has both transaction data (each row is a transaction) and household data (each row is a household), and for the household data, maintains the following information:

- Demographics of the households (updated annually).
- Possession of durable goods (car, washing machine, etc., updated annually; an "affluence index" is computed from this information)
- Purchase data of product categories and brands (updated monthly).

CRISA has two categories of clients: (1) advertising agencies that subscribe to the database services, obtain updated data every month, and use the data to advise their clients on advertising and promotion strategies; (2) and consumer goods manufacturers, which monitor their market share using the CRISA database.

Understanding the business problem & objectives

Business Objectives

The data must be analyzed by segmenting the variables into different groups according to criteria other than demographics. Customers display different levels of brand loyalty based on price, selection criteria, promotions, affluence, social and economic status, etc. If we can segment the customers based on certain important variables as shown in the dataset, we can target them more in particular by proposing personalized branding and promotion strategies. Therefore, the company's business objective is to train customer segments that exhibit similar buying behavior and are affected the same way by any type of sales proposal or promotional campaigns so that the segments can be targeted in particular for branding and promotion activities.

Data Mining Objectives

To divide variables into clusters or segments based on the following:

- Purchase behavior (quantity, frequency, vulnerability to discounts and brand loyalty)
- Purchasing basis (price, sales proposition)
- Variables describing purchase behavior and purchase basis

We can use the demographic variable combination to find the best segmentation for these clusters.
There is a limit on the number of clusters due to the number of promotions.You can run 4, so ideal clustering should not exceed 4 clusters.

Data Preparation

The given dataset has many missing values. Hence,

- Replaced missing values

Column	Old Value	New Value
SEX	**0**	**2**
FEH	**0**	**3**
MT	**0**	**5**
EDU	**0**	**5**
CS	**0**	**1**

- Deriving the brand loyalty index.
  Brand loyalty index is measured by using 3 criterias.

  - **No. of Brands**-As the number of brands increases, the possibility of switching between brands increases, so the fewer the number of brands, the better. Therefore, we assign lower scores to fewer brands, indicating higher brand loyalty.

  - **Brand Runs**-The lower the number of brand runs, the better. As the number of brand runs increases, the probability of multiple brands' brand runs increases, indicating that the switching behavior is higher. Therefore, assign lower scores to rows with fewer brand runs.

  - **Volume of purchases attributed to each brand** - Taking maximum values out of the variables Br. Cd. 57,144; Br.Cd. 55; Br. Cd. 272Cd.286; Br. Cd.24; Br. Cd.481; Br. Cd.352, Br. Cd.5.

Cluster Analysis using R

Clustering Method

Selecting a clustering method is one of the most important things to do before doing cluster analysis. There are mainly 3 methods to clustering.

1. Partitioning Method
2. Hierarchical Method
3. Density-based Method

Here in this analysis, we focus on customer segmentation. then it was best to select partition method. Hierarchical method best to use when we need to analyze nested clusters. Also before the analysis, we have done preprocessing. One of the advantages of doing preprocess is reducing noises. Then we are not going to use Density-bases method as long as we don't need to worry about noise values.

There are two methods we can do Partition Clustering.

- K – mean
- K medoids

To select the most suitable Partition clustering method we run the cluster analysis based on Demographics using both K-mean and K medoids. Results are mentioned below.

Attribute	K	Cluster Size(Original)	Cluster Size(Using K mean)	Cluster Size (Using K medoids)	Dissimilarity(K-mean)	Dissimilarity(K-medoids)
SEC	4	150	207	117	57	33
		150	116	196	34	46
		150	35	70	115	80
		150	242	217	92	67
FEH	3	165	234	272	69	107
		34	64	256	30	222
		401	302	72	99	329
SEX	2	579	456	336	123	243
		21	144	264	123	243
EDU	9	49	56	97	7	48
		9	7	73	2	64
		33	52	64	19	31
		136	110	110	26	26
		262	114	80	148	182
		23	97	69	74	46
		73	70	73	3	0
		13	30	11	17	2
		2	64	23	62	21
CHILD	5	59	19	181	40	122
		145	176	97	31	48
		61	78	68	17	7
		267	219	99	48	168
		68	108	155	40	87
AGE	4	15	35	117	20	102
		129	116	196	13	67
		169	207	70	38	99
		287	242	217	45	70
CS	2	542	456	336	86	206
		58	114	264	86	206

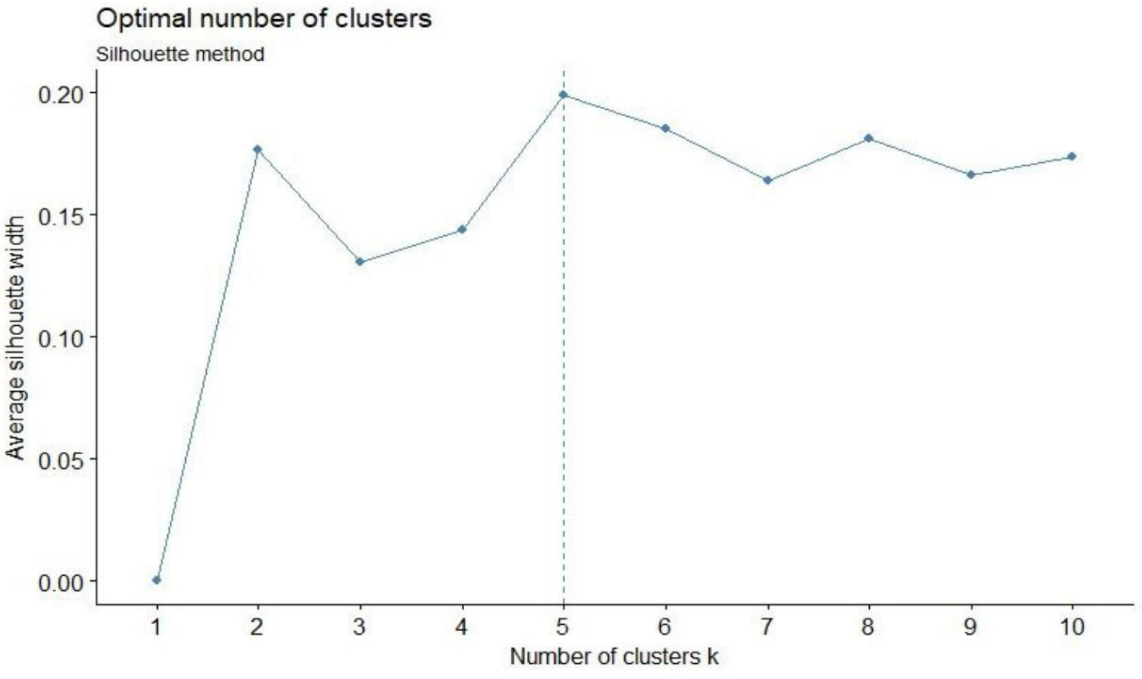
According to the above results, 72.42% said K – mean has fewer dissimilarities than K medoids compared with original cluster size.
In keeping this result, from this onwards we use K-mean algorithm as cluster algorithm.

K- Mean Algorithm

I. To begin this method, no.of clusters should be defined
II. Assign one subject (randomly) to each of the K clusters
III. Determine the cluster centroid
IV. Calculate the distances from (N-K) subjects to cluster centroids
V. Assign (N-K) subject to it's closet cluster
VI. Repeat II to IV until convergence

Clustering based on demographics

All the demographic variables were used in clustering. In all there are 10 variables.
Below diagrams show the cluster analysis.
Optimal number of clusters : 5.



Cluster Summary

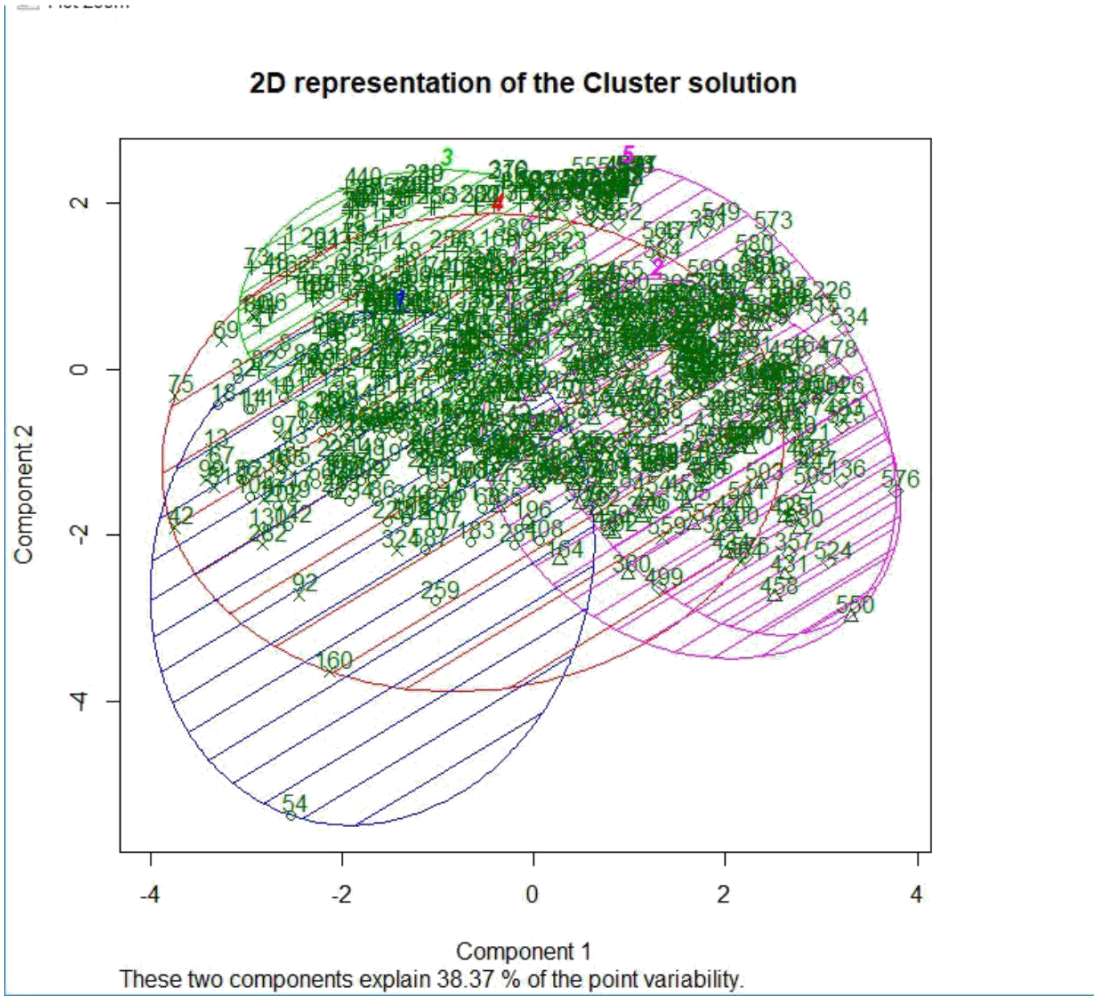
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	18	0.2522	1.1305		5	0.8749
2	223	0.1931	1.3520		4	0.6589
3	40	0.2346	1.2078		2	0.9151
4	159	0.2077	1.3998		2	0.6589
5	160	0.2167	1.3179		1	0.8749

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
SEC	0.37299	0.25609	0.531761	1.135661
FEH	0.44439	0.12182	0.925351	12.396070
MT	0.21261	0.19278	0.183300	0.224440
SEX	0.18393	0.18060	0.042388	0.044264
AGE	0.28850	0.28124	0.055995	0.059316
EDU	0.19940	0.17151	0.265090	0.360711
HS	0.12471	0.12364	0.023671	0.024245
CHILD	0.30428	0.28628	0.120695	0.137262
CS	0.29575	0.19255	0.578946	1.374990
Affluence_Index	0.21528	0.20090	0.135022	0.156099
OVER-ALL	0.27939	0.20818	0.448488	0.813197

Pseudo F Statistic =	120.96

Approximate Expected Over-All R-Squared =	0.38944

Cubic Clustering Criterion =	10.891

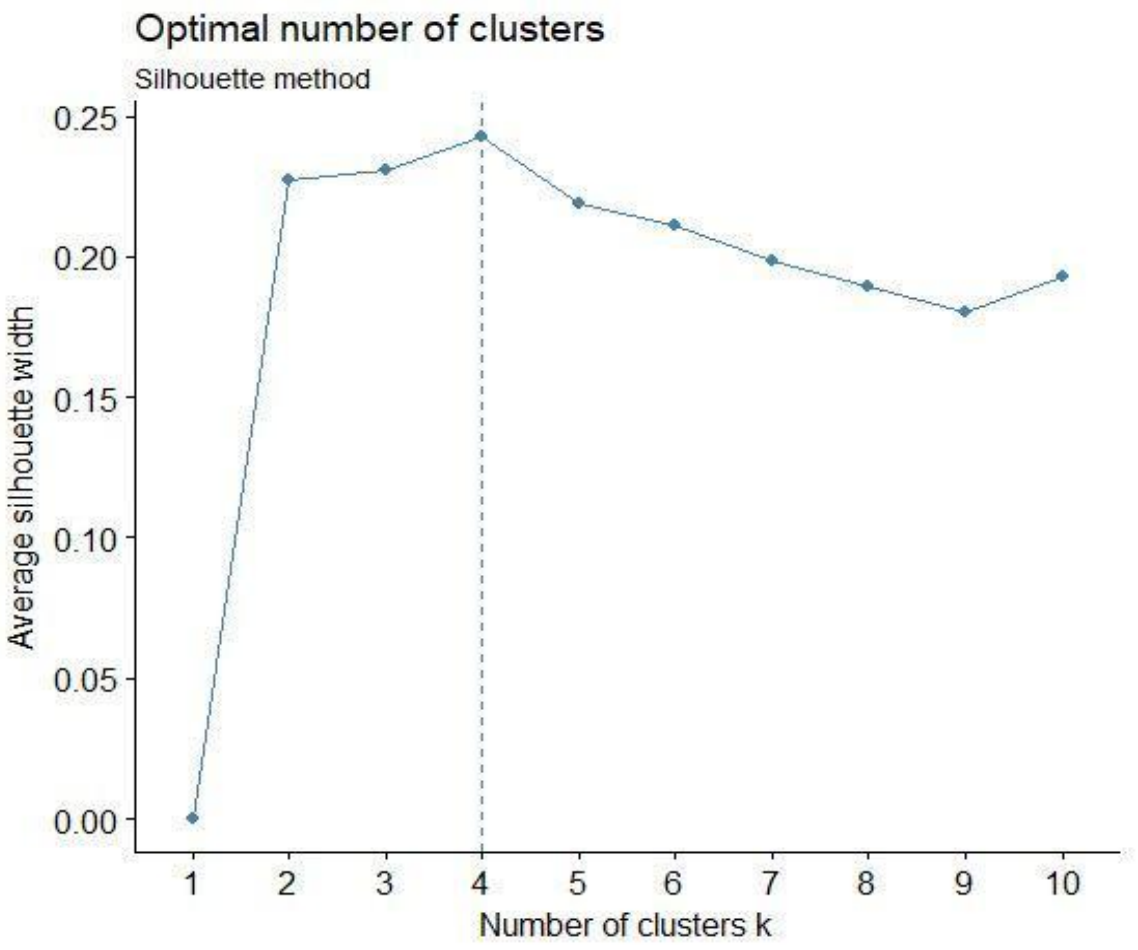


Clustering based on purchase behavior

Variables used for this process are:

- No.of Brands
- Brand Runs
- Total Volume
- No. of Transactions
- Value
- Avg Price
- Others 999(this gives the share of transactions towards other brands which indicates that a customer is not brand loyal.)
- Maximum brand loyalty (Maximum brand loyalty is obtained by taking maximum values out of the variables Br. Cd. 57,144; Br.Cd. 55; Br. Cd. 272Cd.286; Br. Cd.24; Br. Cd.481; Br. Cd.352, Br. Cd.5.)

Optimal number of clusters : 4



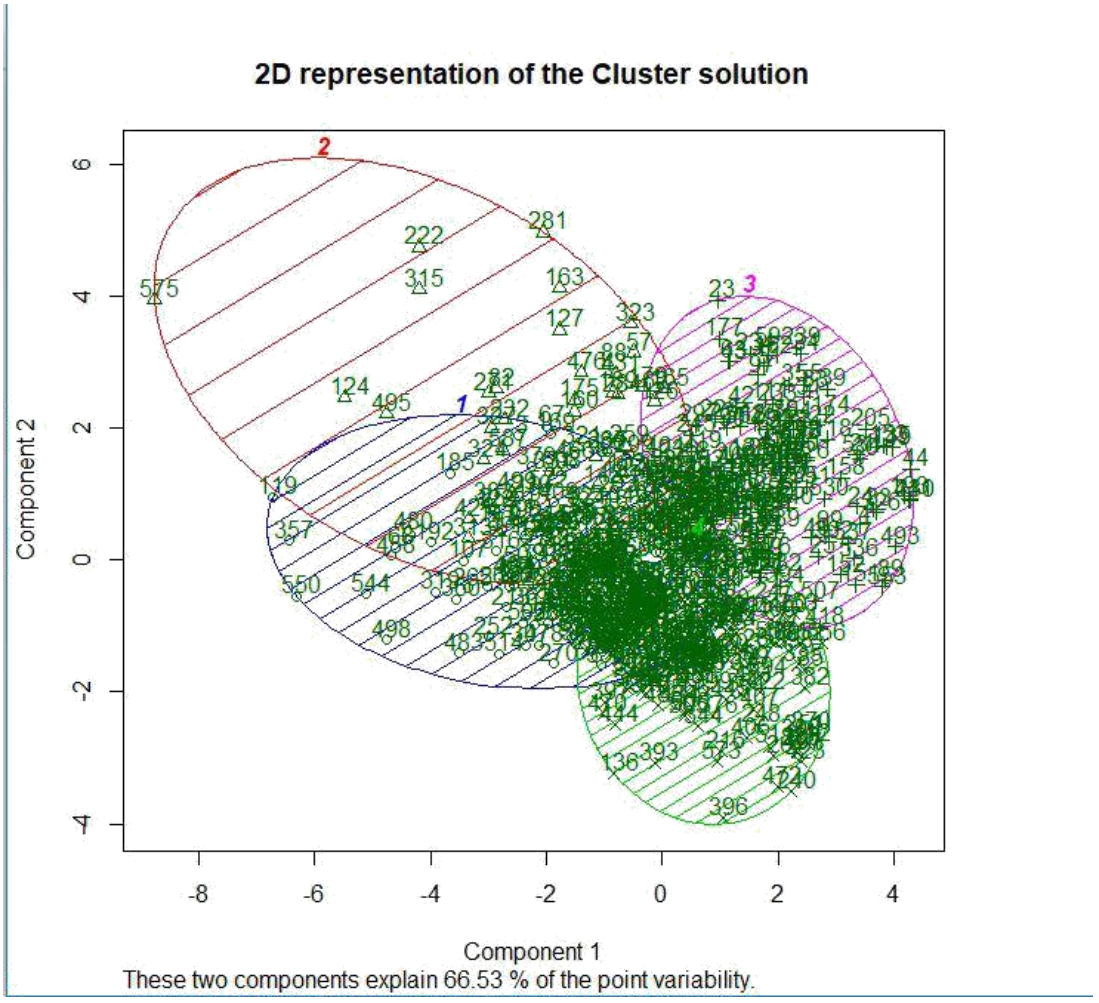
Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	140	0.1489	1.3065		4	0.5932
2	201	0.0992	1.0852		4	0.5057
3	70	0.1085	0.9399		1	0.9117
4	189	0.1150	1.2821		2	0.5057

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
No__of_Brands	0.19746	0.13443	0.538825	1.168374
Brand_Runs	0.14242	0.10578	0.451069	0.821724
Total_Volume	0.15313	0.15079	0.035144	0.036424
No__of__Trans	0.12721	0.10712	0.294483	0.417400
Value	0.13904	0.13286	0.091414	0.100612
Trans___Brand_Runs	0.11839	0.10835	0.166738	0.200103
Vol_Tran	0.10235	0.09883	0.072126	0.077733
Avg__Price	0.13507	0.11902	0.227428	0.294378
Br__Cd__57__144	0.23643	0.17452	0.457890	0.844643
Br__Cd__55	0.25972	0.10721	0.830446	4.897840
Br__Cd__272	0.09438	0.09323	0.029025	0.029893
Br__Cd__286	0.11288	0.11162	0.027028	0.027779
Br__Cd__24	0.07978	0.07935	0.015679	0.015928
Br__Cd__481	0.09952	0.09871	0.021084	0.021538
Br__Cd__352	0.12241	0.11880	0.062850	0.067065
Br__Cd__5	0.06993	0.06945	0.018712	0.019069
Others_999	0.29732	0.15265	0.737709	2.812563
OVER-ALL	0.15914	0.11837	0.449523	0.816605

Pseudo F Statistic =	162.23

Approximate Expected Over-All R-Squared =	0.30601

Cubic Clustering Criterion =	28.668

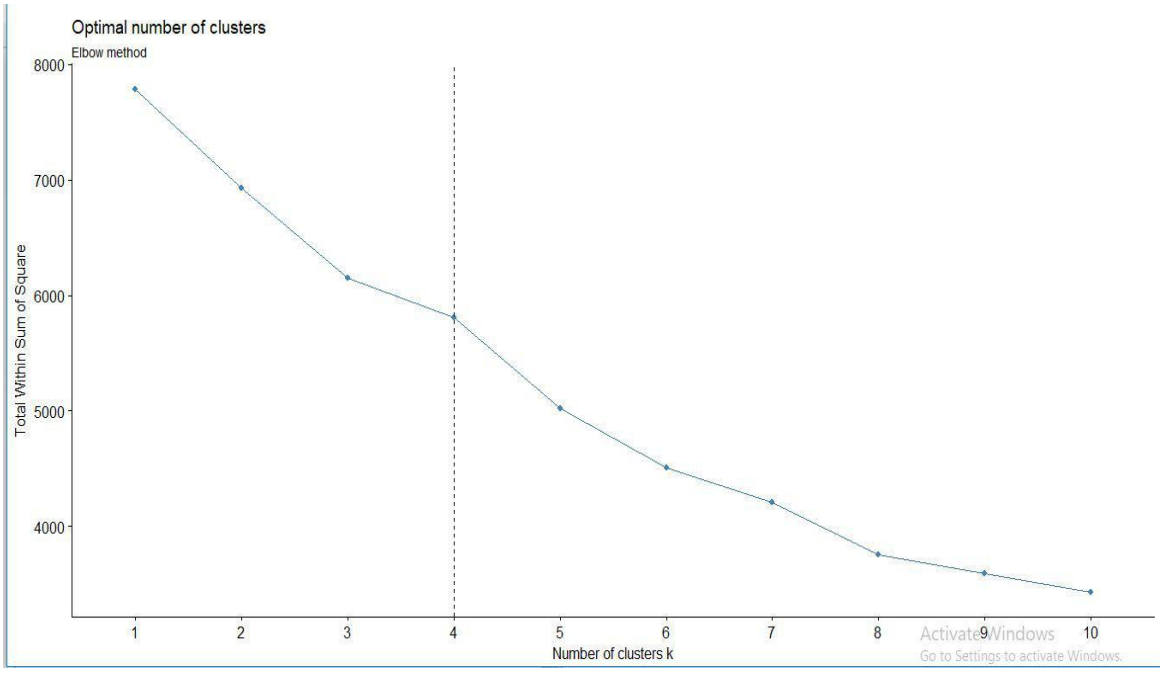

2D representation of the Cluster solution
Component 2
Component 1
These two components explain 66.53 % of the point variability.
1
2
3

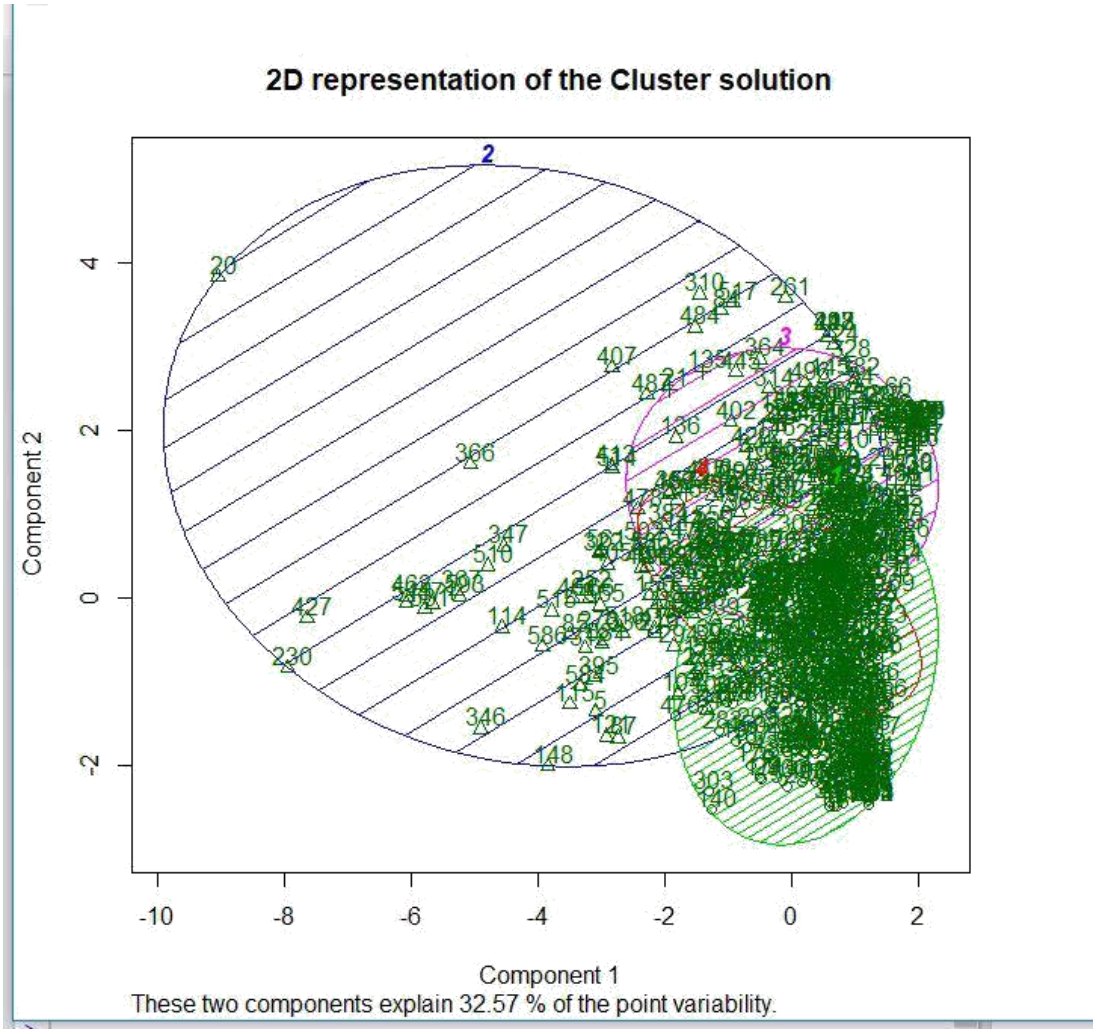
Clustering based on basis of purchase

Variables used for this process are:

- All price categories
- Selling propositions
- Purchase volume with no promotion, promotion 6 and other promotion

We have plotted graphs for all selling propositions and observed that PropCat 9, PropCat 10, PropCat 11,
PropCat 13, PropCat 14 have very less data points against them. We did not observe much distribution.
patterns for this variables. So we have considered only PropCat 5 – 8, PropCat 12, PropCat 15.





Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	204	0.1615	1.2014		3	0.5845
2	82	0.1674	1.3017		1	0.7634
3	198	0.1916	1.2177		1	0.5845
4	116	0.0833	1.1441		1	0.6555

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
PropCat_12	0.07904	0.07815	0.027323	0.028090
PropCat_15	0.10428	0.10201	0.047863	0.050269
Pur_Vol_No_Promo____	0.11949	0.11941	0.006430	0.006471
Pur_Vol_Promo_6__	0.13943	0.13836	0.020240	0.020658
Pur_Vol_Other_Promo__	0.07198	0.07177	0.011018	0.011140
Pr_Cat_1	0.28087	0.18390	0.573431	1.344286
Pr_Cat_2	0.31158	0.17944	0.670013	2.030419
Pr_Cat_3	0.26803	0.23832	0.213354	0.271220
Pr_Cat_4	0.19168	0.16827	0.233206	0.304131
PropCat_5	0.31634	0.23188	0.465391	0.870525
PropCat_6	0.17129	0.15773	0.156297	0.185251
PropCat_7	0.19575	0.18047	0.154312	0.182469
PropCat_8	0.15826	0.15090	0.095368	0.105422
OVER-ALL	0.20255	0.16177	0.365374	0.575731

Pseudo F Statistic =	114.38

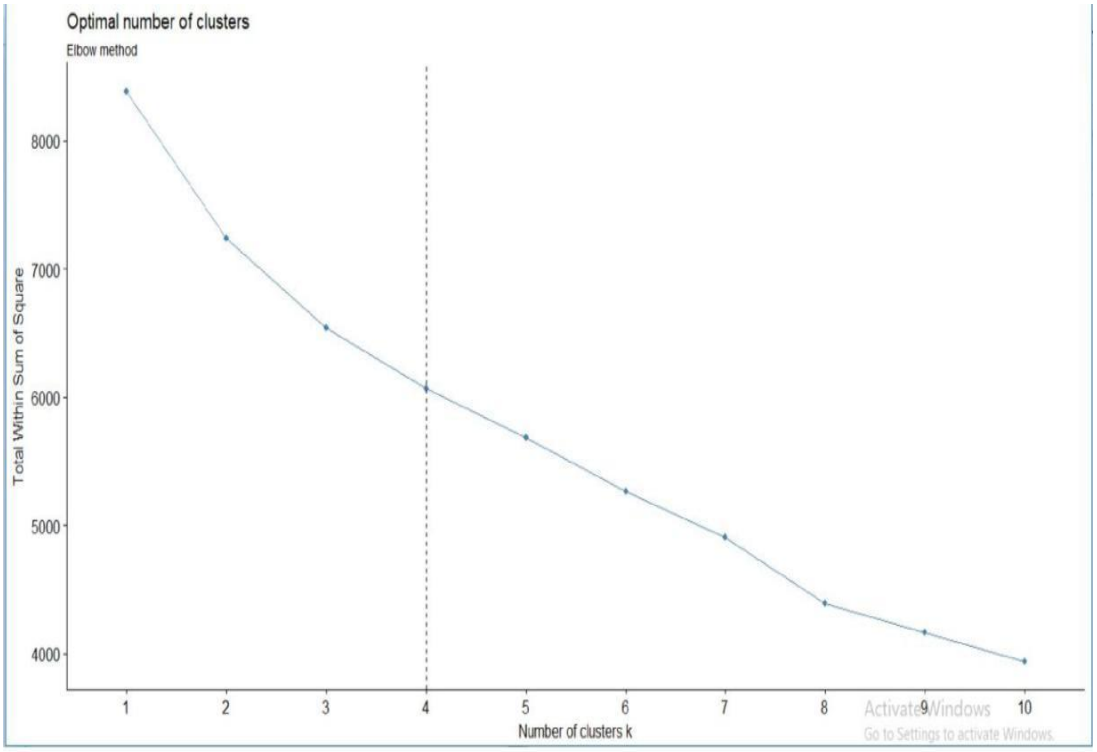
Approximate Expected Over-All R-Squared =	0.31923

Cubic Clustering Criterion =	8.257

Clustering based on both purchase behaviors and basis of purchase.

Here we have combined both customer purchase behavior and basis of purchase together.

Optimal number of clusters : 4.



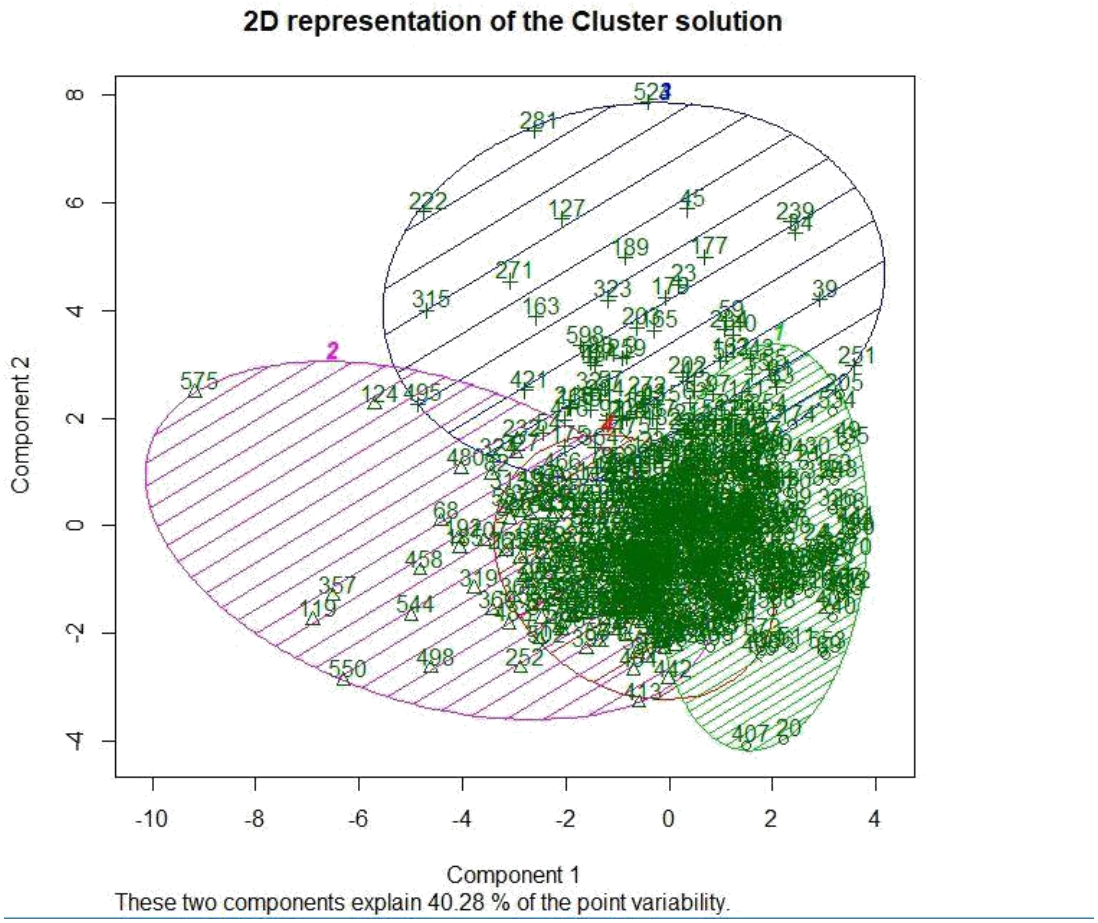
Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	102	0.1138	1.4514		2	0.7977
2	263	0.1381	1.4876		3	0.7534
3	108	0.1593	1.7091		2	0.7534
4	127	0.1800	1.7337		2	0.8979

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
PropCat_12	0.07904	0.07839	0.021290	0.021753
PropCat_15	0.10428	0.10314	0.026645	0.027375
Pur_Vol_No_Promo____	0.11949	0.11696	0.046810	0.049109
Pur_Vol_Promo_6__	0.13943	0.13398	0.081218	0.088398
Pur_Vol_Other_Promo__	0.07198	0.07205	0.003201	0.003212
Pr_Cat_1	0.28087	0.17123	0.630210	1.704241
Pr_Cat_2	0.31158	0.25833	0.316050	0.462094
Pr_Cat_3	0.26803	0.21246	0.374833	0.599572
Pr_Cat_4	0.19168	0.17338	0.185914	0.228372
PropCat_5	0.31634	0.20036	0.600846	1.505299
PropCat_6	0.17129	0.16562	0.069812	0.075051
PropCat_7	0.19575	0.18702	0.091783	0.101058
PropCat_8	0.15826	0.15573	0.036544	0.037930
No__of_Brands	0.19746	0.19264	0.053068	0.056042
Brand_Runs	0.14242	0.13184	0.147362	0.172831
Total_Volume	0.15313	0.15170	0.023424	0.023985
No__of__Trans	0.12721	0.12498	0.039555	0.041185
Value	0.13904	0.13758	0.025746	0.026426
Trans___Brand_Runs	0.11839	0.11121	0.122033	0.138995
Vol_Tran	0.10235	0.09873	0.074167	0.080109
Avg__Price	0.13507	0.10578	0.389773	0.638734
Br__Cd__57__144	0.23643	0.14521	0.624662	1.664262
Br__Cd__55	0.25972	0.20403	0.385975	0.628597
Br__Cd__272	0.09438	0.09349	0.023763	0.024342
Br__Cd__286	0.11288	0.11092	0.039155	0.040751
Br__Cd__24	0.07978	0.07561	0.106094	0.118686
Br__Cd__481	0.09952	0.09781	0.038881	0.040454
Br__Cd__352	0.12241	0.11558	0.113070	0.127485
Br__Cd__5	0.06993	0.06737	0.076595	0.082949
Others_999	0.29732	0.21302	0.489261	0.957948
OVER-ALL	0.17925	0.14825	0.319380	0.469249

Pseudo F Statistic =	93.22

Approximate Expected Over-All R-Squared =	0.18584

Cubic Clustering Criterion =	40.235



Hence, we can see that a combination of both basis of purchase and purchase behavior gives the minimum overall R-Squared value out of all the segmentations. Hence segmenting with the combination of both basis of purchase and purchase behaviour is the best segmentation criteria.

Conclusion

By comparing all the segmentations done so far, segmentation of customers by considering both customer purchase behaviour and basis of purchase, gives the best segmentation.
Based on the segment profile of this segmentation basis, we can say that the segments have the following membership.

Cluster Number	**Characteristics**
1	Less than average brand loyalty, Second highest average price, Highest volume purchase.
2	Highest brand loyalty, Less than second highest average price, Average volume purchase.
3	Average brand loyalty, Highest average price,Least volume purchase.
4	Least brand loyalty, Least average price, Less than average volume purchase.