

SEGMENTING CONSUMERS OF BATH SOAP

Introduction:

The IMRB household data is explored to gain information on demographic attributes associated with purchase behavior and brand loyalty of consumers. The data set has 47 variables pertaining to the household demographics, purchase summary, purchase within promotion, brand-wise purchases, price-wise purchases and selling proposition-wise purchases.

Demographic variables like socioeconomic status, food habits, spoken language, sex, education level, children etc are categorical variables. Affluence index, age of homemaker and household size are continuous variables.

The Cluster analysis is performed to enable clients of IMRB design cost effective promotions targeted at appropriate segment of consumers.

We use the K-means cluster algorithm using all the continuous variables to segment the data into clusters.

Data auditing shows that we do not have missing values. But many flag values in sex variable are zero in addition to 1 for males and 2 for females. We chose to remove these data records since the data description did not have the clause 'sex not specified'. The data set was reduced to 532 records from initial 600 records. All the variables are z-score normalized using the auto data prep node in SPSS.

Exploratory Data Analysis:

Distribution of Socio Economic Class

Value ▲	Proportion	%	Count
1.000		25.0	150
2.000		25.0	150
3.000		25.0	150
4.000		25.0	150




All 4 socio economic classes are equally represented in the data.

Distribution of Food Eating Habits

Value ▲	Proportion	%	Count
0.000		11.5	69
1.000		27.5	165
2.000		5.67	34
3.000		55.33	332


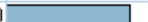
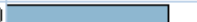

Little more than 50% of consumers are non vegetarian, less than 30% are vegetarians

Distribution of Sex

Value ▲	Proportion	%	Count
0.000		11.33	68
1.000		3.5	21
2.000		85.17	511











About 85% of homemakers are female and only 3 % male. In order to not throw away 11% of data, and although 0 level is not stated in the data description, we are assuming that 0 value corresponds to consumers who did not specify sex.

Distribution of Age

Value ▲	Proportion	%	Count
1.000		2.5	15
2.000		21.5	129
3.000		28.17	169
4.000		47.83	287

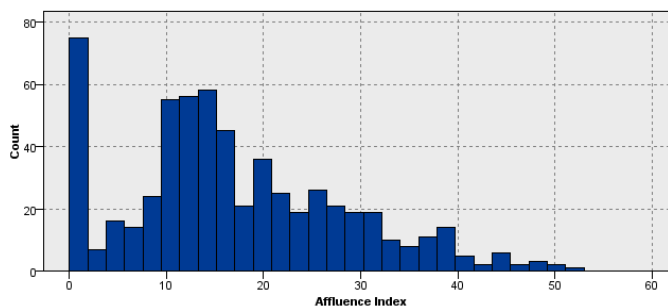
About 50% of consumers are in the over 45 yrs age group. Young people less than 24% account for only 2.5% of shoppers. Similar proportion of shoppers in the 25-44 yrs age group.

Distribution of Education

Value ▲	Proportion	%	Count
0.000		12.17	73
1.000		8.17	49
2.000		1.5	9
3.000		5.5	33
4.000		22.67	136
5.000		31.5	189
6.000		3.83	23
7.000		12.17	73
8.000		2.17	13
9.000		0.33	2

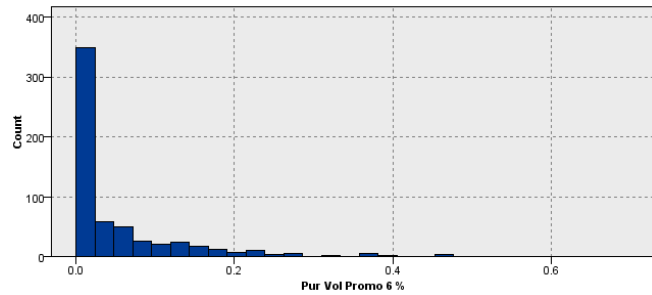
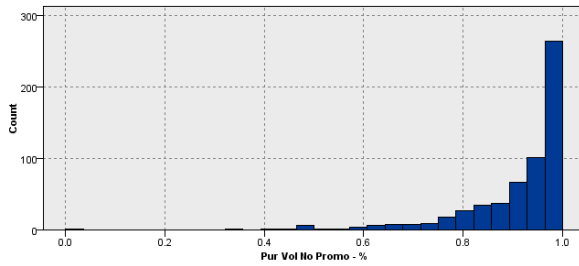
About 50 % of shoppers have between 5- 12 yrs of schooling. A very small percentage of 0.33% are highly educated shoppers, whereas illiterate people account for 8 % of shoppers. About 20% of shoppers have some form of college education.

Histogram of Affluence Index

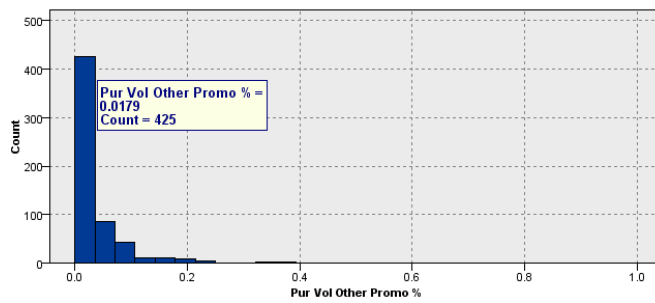


Basis for Purchase:

Histogram of Purchase Vol with No Promotion Histogram of Purchase Vol Promotion 6

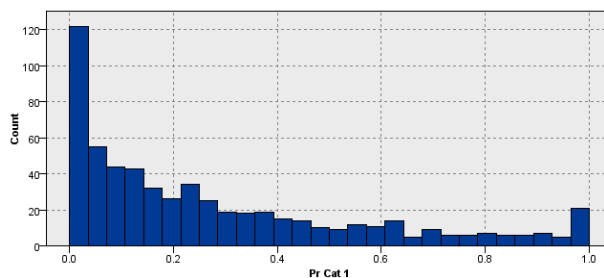


Histogram of Purchase Vol Other promotions

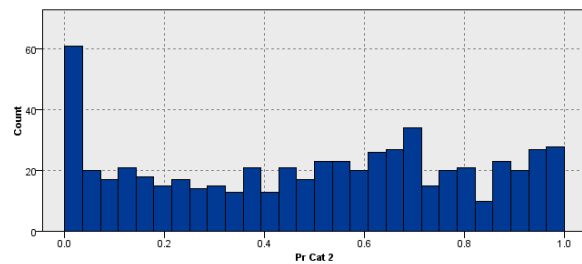


The 3 histograms of purchase volumes show that promotion under code 6 is most popular.

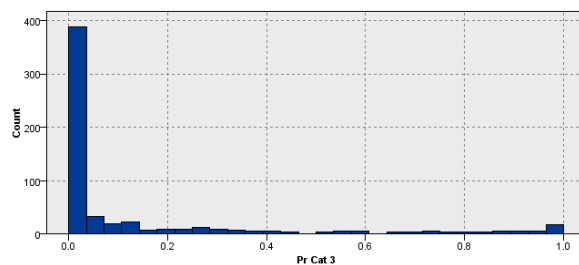
Histogram Price Category 1



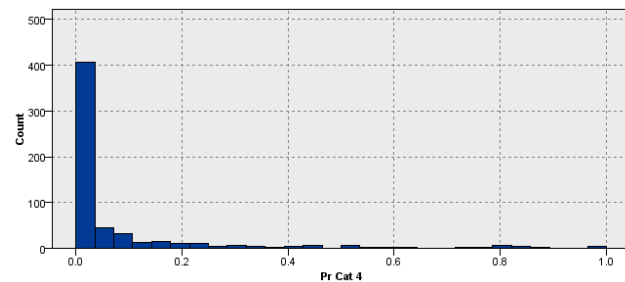
Histogram Price Category 2



Histogram Price Category 3

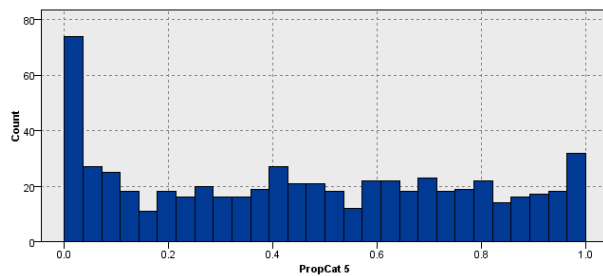


Histogram Price Category 4

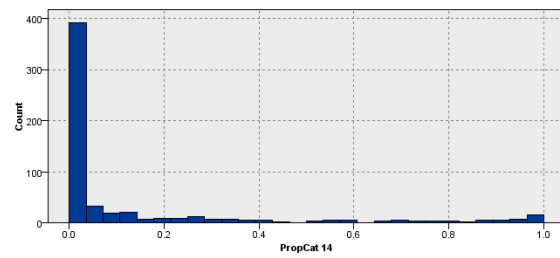


The price category histograms show that most purchases were under price categories 2 and 4.

Histogram Selling Prop 5



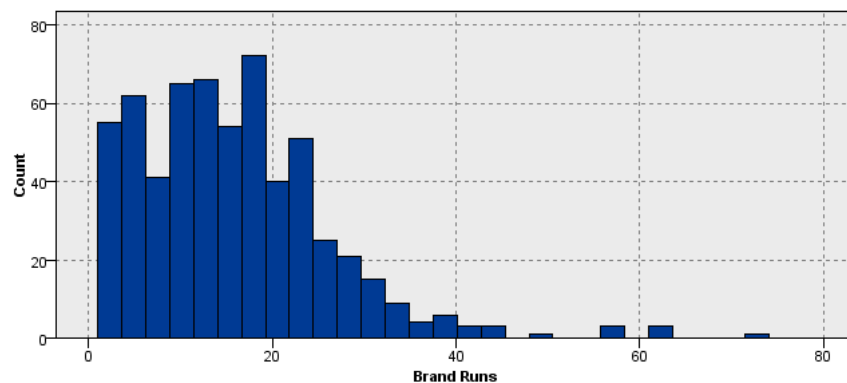
Histogram Selling Prop 14



With respect to different selling propositions most consumers seem to be buying under selling proposition 5 and 14 categories. Proposition 5 is most popular with fairly even distribution.

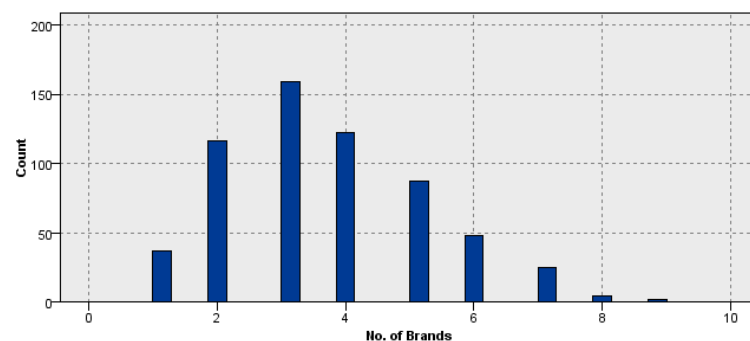
Purchase Behavior:

Histogram of Brand Runs



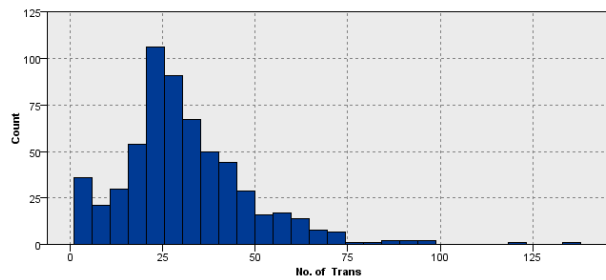
Brand loyalty measured by number of instances of consecutive purchases shows a range of values from 2 to 50 times with highest number of runs about 18.

Histogram of Brands

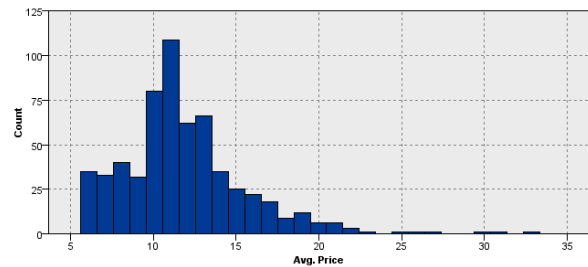


Histogram shows that most products have 3 brands.

Histogram of Transactions

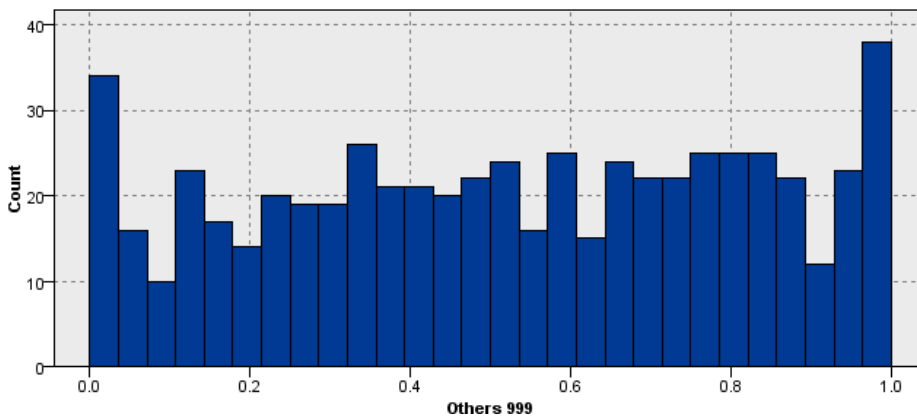


Histogram of Average Price

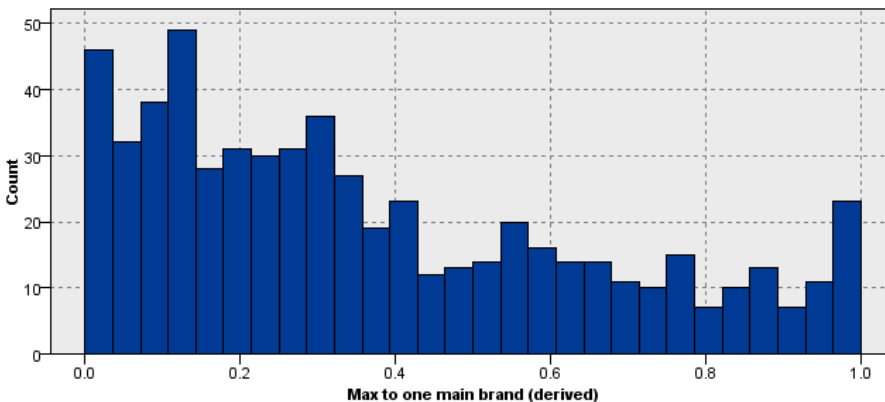


Number of transactions and average price show almost normal distributions. Most consumers pay ₹ 11 per transaction. The mean number of transactions per household is 23.

Histogram of Share to other Brands

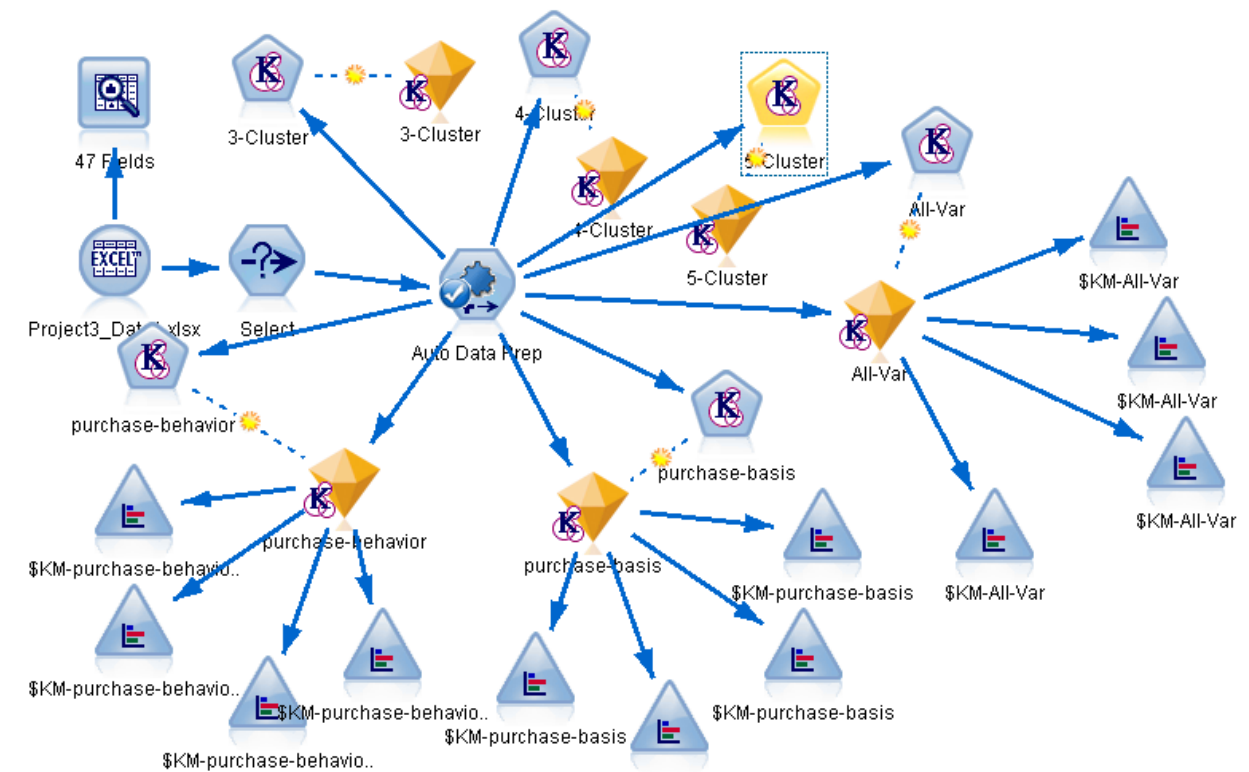


Histogram of Max to One Main Brand



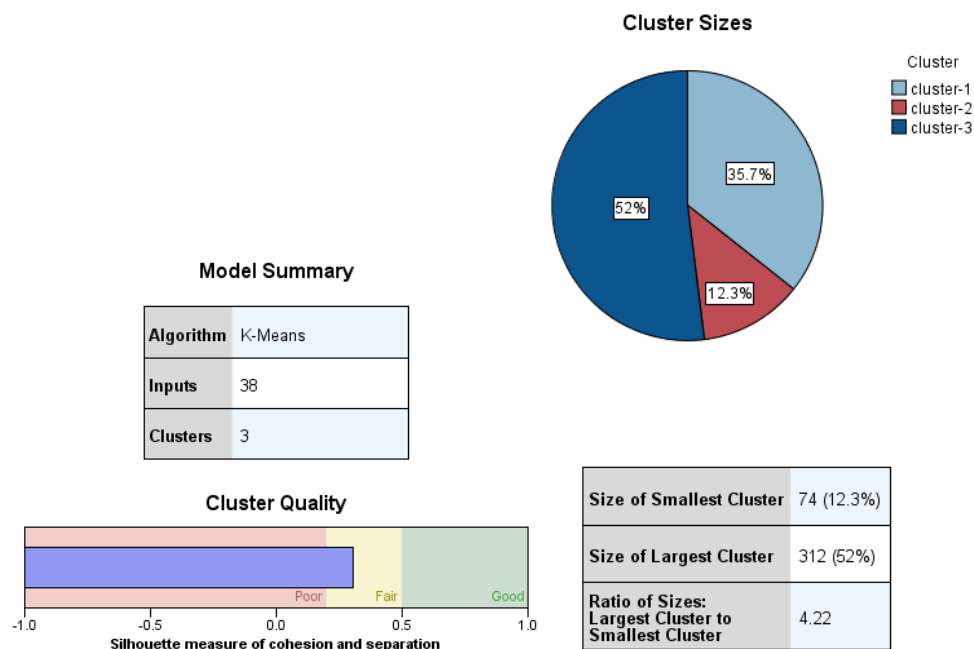
Histograms of Share to other brands and Max to one main brand show sizable distributions. Making them important variables in identifying consumer purchase behavior.

Cluster Model – Determining K

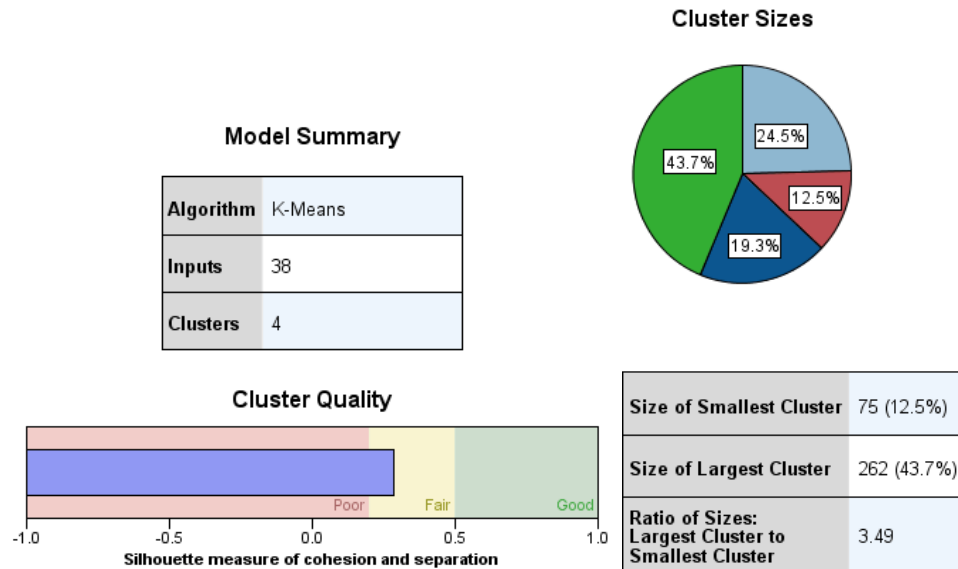


In order to decide which number of clusters to divide the data to get maximum information, we initially built models with 3, 4 and 5 clusters. All the models are same quality in the fair range.

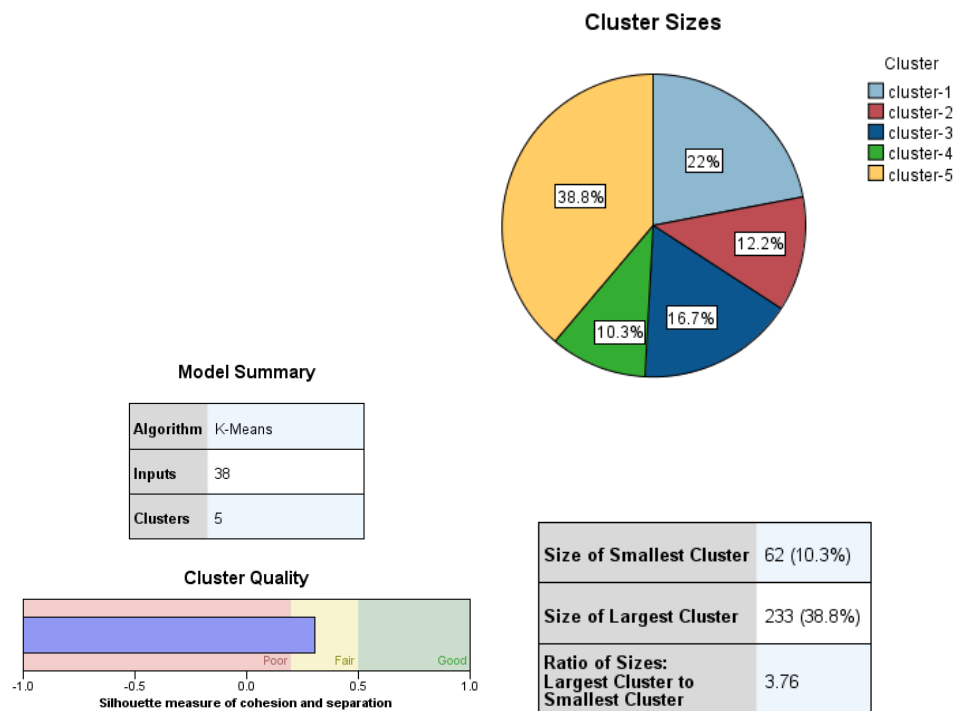
3-Cluster



4-Cluster

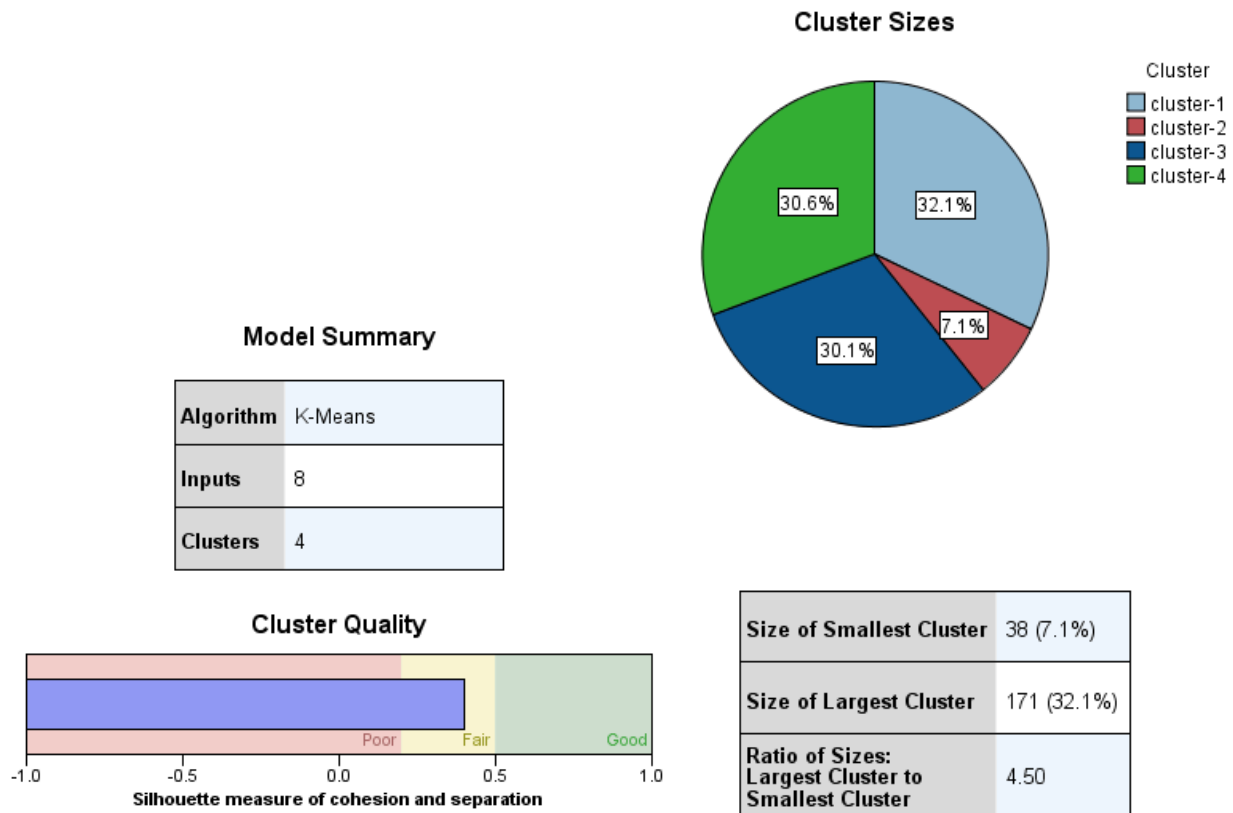


5-Cluster

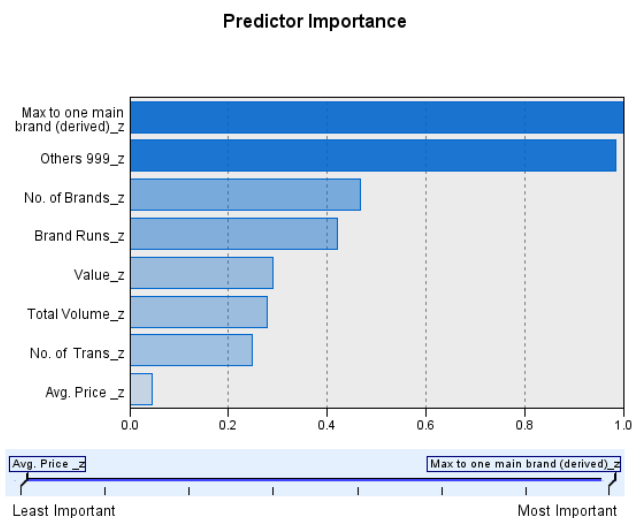


Choosing $k = 4$ as the best cluster size, we proceed to build cluster models based on variables that describe purchase behavior, basis-for-purchase separately and together.

A. Purchase Behavior Model



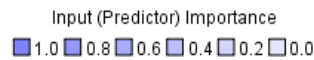
The model shows we have 4 clusters induced by 8 inputs for purchase behavior. The model falls in fair range. Cluster 1 has 171 records, cluster 2 has 38 records, cluster 3 has 160 records and cluster 4 has 163 records.



The most important predictors in the model summary table are Max to one main brand and Share to Other (brands) 999 .

Cluster Profiles

Clusters



Cluster	cluster-1	cluster-4	cluster-3	cluster-2
Label				
Description				
Size	32.1% (171)	30.6% (163)	30.1% (160)	7.1% (38)
Inputs	Max to one main brand (derived)_z	Max to one main brand (derived)_z	Max to one main brand (derived)_z	Max to one main brand (derived)_z
	Others 999_z 1.04	Others 999_z 0.15	Others 999_z -1.05	Others 999_z -0.58
	No. of Brands_z -0.50	No. of Brands_z 1.05	No. of Brands_z -0.35	No. of Brands_z 0.56
	Brand Runs_z -0.11	Brand Runs_z 0.98	Brand Runs_z -0.59	Brand Runs_z 0.30
	Value_z -0.03	Value_z 0.15	Value_z -0.20	Value_z 2.16
	Total Volume_z -0.06	Total Volume_z -0.02	Total Volume_z 0.02	Total Volume_z 2.14
	No. of Trans_z -0.12	No. of Trans_z 0.71	No. of Trans_z -0.28	No. of Trans_z 0.84
	Avg. Price_z 0.09	Avg. Price_z 0.25	Avg. Price_z -0.39	Avg. Price_z -0.13

Cluster 1: This cluster consists of consumers with Least Max to One Brand value and consequent highest share to Other 999.

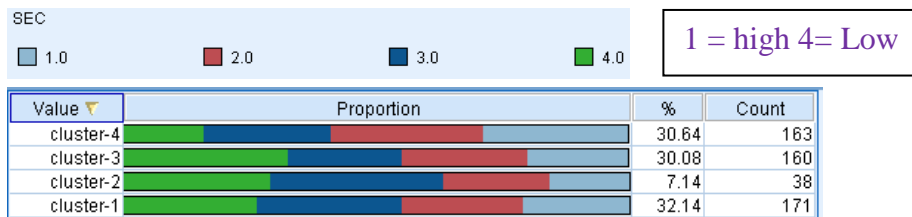
Cluster 2: This cluster consists of consumers with relatively large Max to One Brand and relatively low share to Other 999.

Cluster 3: This cluster consists of consumers with highest Max to One Brand value and consequently lowest share to Other 999.

Cluster 4: This cluster consists of consumers with relatively low Max to One Brand value and relatively high share to Other 999.

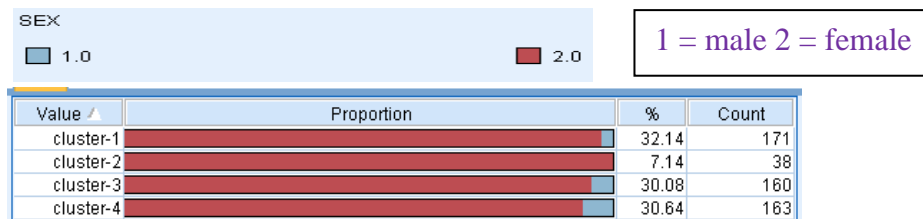
Cluster Demographics

Cluster Distribution by Socio Economic Class



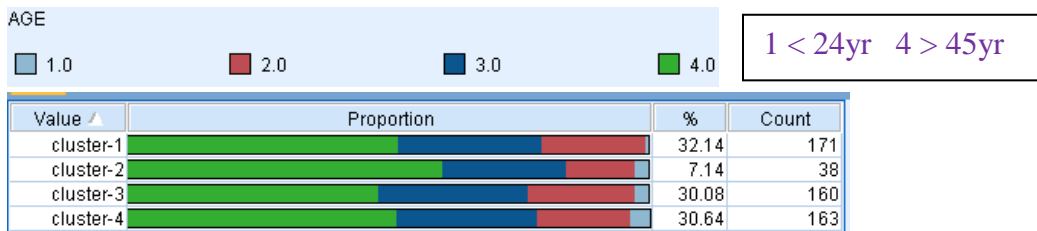
All Clusters consists of similar proportions of people for all socio economic classes.

Cluster Distribution by Sex



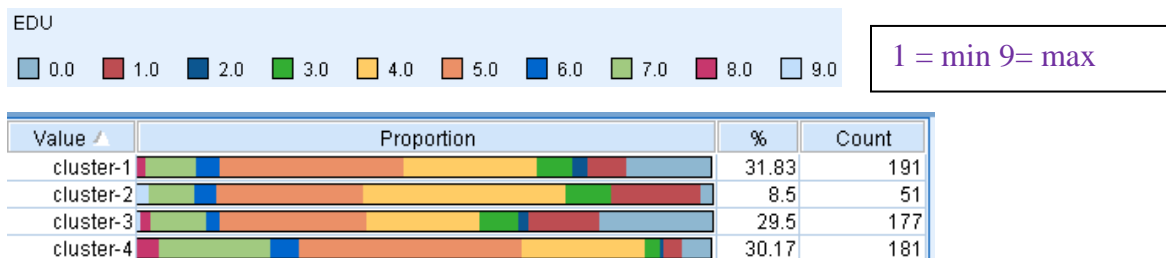
Cluster 2 consists of only females. Since the 85% of shoppers are female all clusters have large proportion of females.

Cluster Distribution by Age



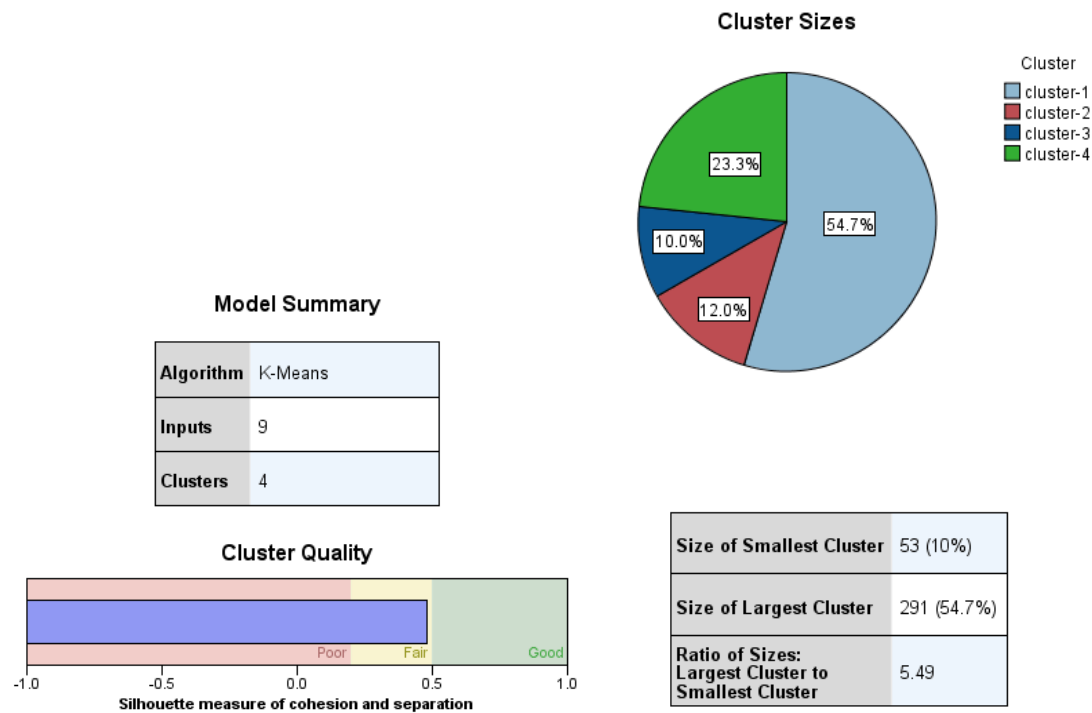
About 50% of all clusters made of older people more than 45yr. Cluster 1 has a very small percentage of young people.

Cluster Distribution by Education

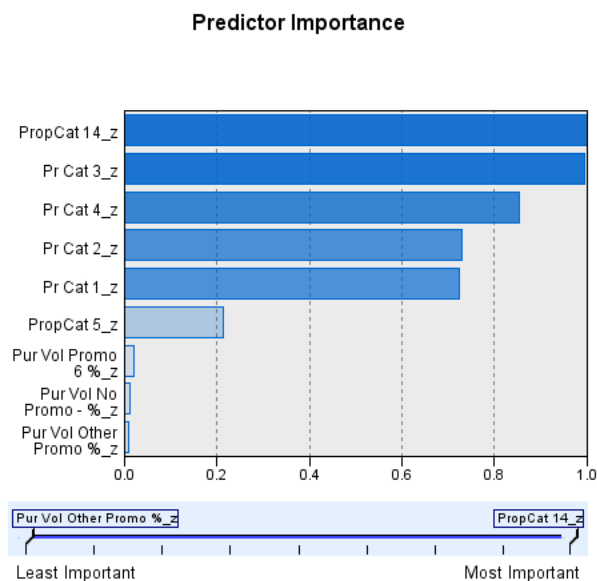


Although illiterate to college educated professionals form a part of all clusters, the higher educated people form a very small percentage of the smallest cluster.

B. Basis for Purchase Model



The model shows we have 4 clusters induced by 9 inputs for purchase behavior. The model falls very close to good range. Cluster 1 has 291 records, cluster 2 has 64, cluster 3 has 53 cluster 4 has 124 records.

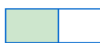





The most important predictors in the model are selling proposition PropCat14 (any carbolic soaps) and price category Pr Cat 3 (any economy/carbolic soaps).

Cluster profiles

Clusters

Input (Predictor) Importance
 1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-1	cluster-4	cluster-2	cluster-3
Label				
Description				
Size	 54.7% (291)	 23.3% (124)	 12.0% (64)	 10.0% (53)
Inputs	PropCat 14_z -0.32	PropCat 14_z -0.45	PropCat 14_z 2.34	PropCat 14_z -0.29
	Pr Cat 3_z -0.32	Pr Cat 3_z -0.46	Pr Cat 3_z 2.34	Pr Cat 3_z -0.28
	Pr Cat 4_z -0.24	Pr Cat 4_z -0.36	Pr Cat 4_z -0.29	Pr Cat 4_z 2.73
	Pr Cat 2_z 0.80	Pr Cat 2_z -0.65	Pr Cat 2_z -1.14	Pr Cat 2_z -1.03
	Pr Cat 1_z -0.42	Pr Cat 1_z 1.40	Pr Cat 1_z -0.77	Pr Cat 1_z -0.45
	PropCat 5_z 0.29	PropCat 5_z -0.42	PropCat 5_z -1.11	PropCat 5_z 0.90
	Pur Vol Promo 6 %_z 0.01	Pur Vol Promo 6 %_z 0.07	Pur Vol Promo 6 %_z -0.34	Pur Vol Promo 6 %_z 0.46
	Pur Vol No Promo - %_z	Pur Vol No Promo - %_z	Pur Vol No Promo - %_z	Pur Vol No Promo - %_z
	Pur Vol Other Promo %_z	Pur Vol Other Promo %_z	Pur Vol Other Promo %_z	Pur Vol Other Promo %_z

Cluster 1: Consumers who chose relatively low value of Prop Cat 14 and Pr Cat 3 but low value of purchase with promotion code 6 as their basis for purchase.

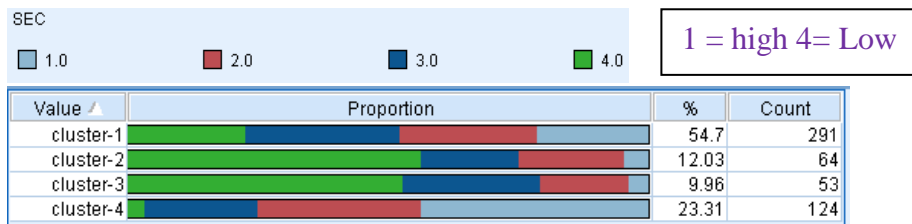
Cluster 2: Consumers who chose high value of Prop Cat 14 and low Pr Cat 3 but lowest value of purchase with promotion code 6 as their basis for purchase.

Cluster 3: Consumers who chose relatively low value of Prop Cat 14 and Pr Cat 3 and highest value of purchase with promotion code 6 as their basis for purchase.

Cluster 4: Consumers who chose lowest value of Prop Cat 14 and Pr Cat 3 but relatively high value of purchase with promotion code 6 as their basis for purchase

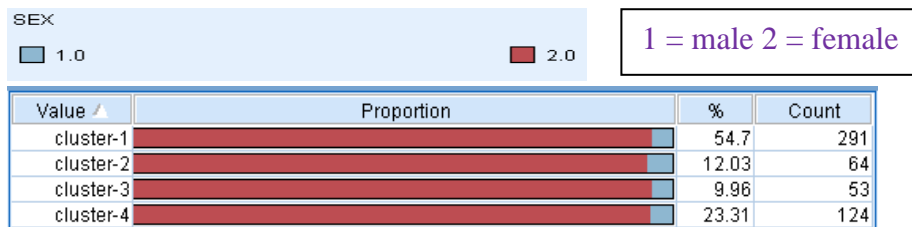
Cluster Demographics

Cluster Distribution by Socio Economic Class



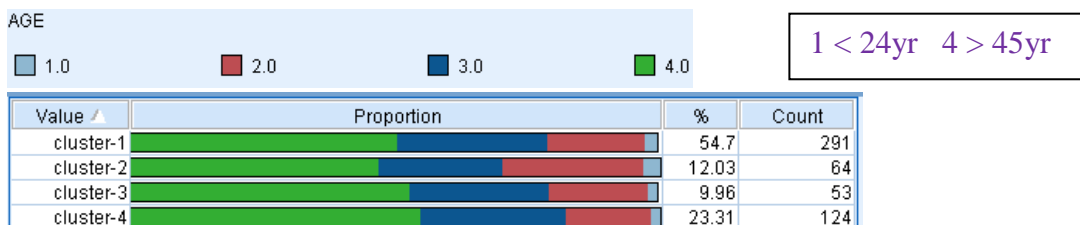
Although people from highest socio economic class form sizable proportion of Cluster 1,2 and 3 their proportion in cluster 4 is very small. and Cluster 4 show opposite trend in the socioeconomic status. More than 50 % of.

Cluster Distribution by Sex



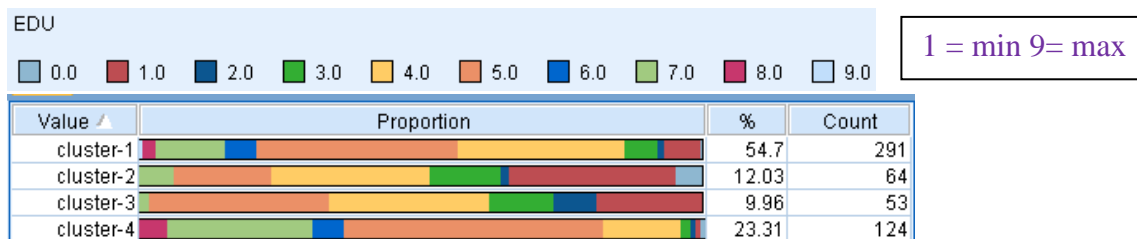
The cluster distributions by sex is very similar for all clusters.

Cluster Distribution by Age



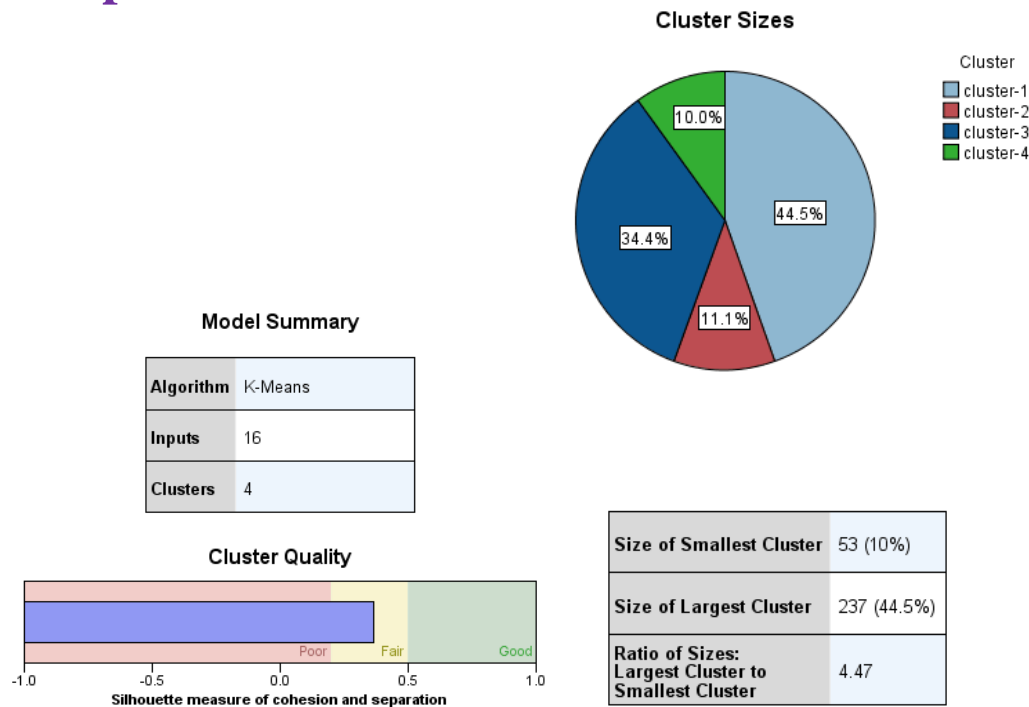
More tha 50 % of all clusters are people > 45yr. Young people form less tha 5% of all clusters.

Cluster Distribution by Education



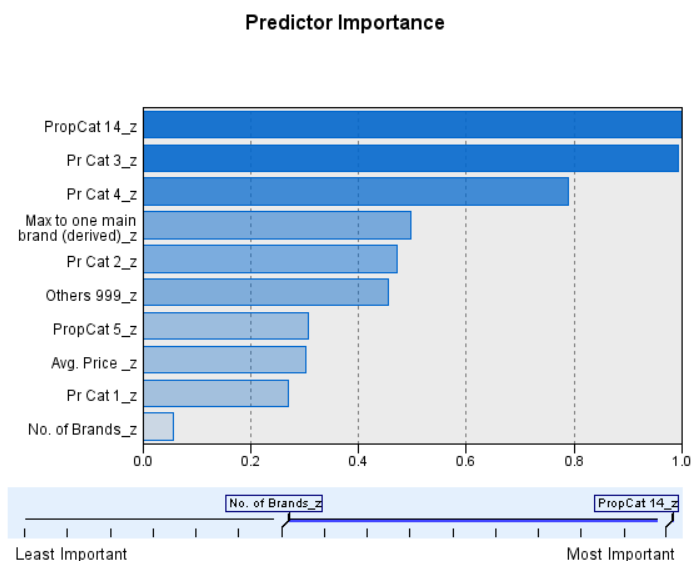
All education levels from illiterate to professionals are reprinted in all 4 clusters

C. Complete Model With All Relevant Variables



In order to do further downstream analysis of consumer characteristics to guide advertising and promotional campaigns, we build a cluster model with all variables relevant for purchase behavior and basis of purchase excluding Brand Runs. We retain Brand Runs as target variable in a classification model to gauge brand loyalty.





The model shows we have 4 clusters induced by all the 16 inputs. The model falls in fair range. Cluster 1 has 237 records, cluster 2 has only 59 records, cluster 3 has 183 records, and cluster 4 has 53 records.



The model shows that Prop Cat14 (any carbolic soaps) and Pr Cat3 (any economy/carbolic soaps) are the most important predictors. The basis for purchase variables dominate the complete model.

Clusters

Input (Predictor) Importance
 1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-1	cluster-3	cluster-2	cluster-4
Label				
Description				
Size	 44.5% (237)	 34.4% (183)	 11.1% (59)	 10.0% (53)
Inputs	PropCat 14_z -0.38	PropCat 14_z -0.31	PropCat 14_z 2.45	PropCat 14_z -0.23
	Pr Cat 3_z -0.38	Pr Cat 3_z -0.31	Pr Cat 3_z 2.45	Pr Cat 3_z -0.20
	Pr Cat 4_z -0.24	Pr Cat 4_z -0.30	Pr Cat 4_z -0.30	Pr Cat 4_z 2.68
	Max to one main brand (derived)_z	Max to one main brand (derived)_z	Max to one main brand (derived)_z	Max to one main brand (derived)_z
	Pr Cat 2_z -0.07	Pr Cat 2_z 0.91	Pr Cat 2_z -1.22	Pr Cat 2_z -1.03
	Others 999_z 0.57	Others 999_z -0.62	Others 999_z -1.13	Others 999_z 1.06
	PropCat 5_z -0.35	PropCat 5_z 0.63	PropCat 5_z -1.16	PropCat 5_z 0.87
	Avg. Price_z 0.54	Avg. Price_z -0.12	Avg. Price_z -1.30	Avg. Price_z -0.80
	Pr Cat 1_z 0.60	Pr Cat 1_z -0.51	Pr Cat 1_z -0.78	Pr Cat 1_z -0.49
	No. of Brands_z 0.39	No. of Brands_z 0.04	No. of Brands_z -0.36	No. of Brands_z -0.50
	Pur Vol Promo 6 %_z 0.23	Pur Vol Promo 6 %_z -0.24	Pur Vol Promo 6 %_z -0.37	Pur Vol Promo 6 %_z 0.48
	No. of Trans_z 0.45	No. of Trans_z -0.08	No. of Trans_z -0.10	No. of Trans_z -0.10
	Pur Vol No Promo - %_z	Pur Vol No Promo - %_z	Pur Vol No Promo - %_z	Pur Vol No Promo - %_z
	Value_z 0.22	Value_z 0.23	Value_z -0.30	Value_z -0.15
	Total Volume_z -0.06	Total Volume_z 0.23	Total Volume_z 0.47	Total Volume_z 0.29
	Pur Vol Other Promo %_z	Pur Vol Other Promo %_z	Pur Vol Other Promo %_z	Pur Vol Other Promo %_z

Code description

Any Carbolic

Any Economy/Carbolic

Any Sub-Popular

Any Popular Soap

Any Beauty

Any Premium Soap

Cluster Profiles

Premium Brand Shopper- Cluster 1: This customer purchases mainly premium brands and is and pays the highest average cost. This cluster with 44.5% (237 records) of total consumer in survey consists of consumers who chose the lowest value of Prop Cat 14 and Pr Cat 3 but relatively high value of purchase with promotion code 6 as their basis for purchase. With respect to purchase behavior they have relatively low value (negative) for Max to One Main Brand and relatively high Share to Other 999 variable.

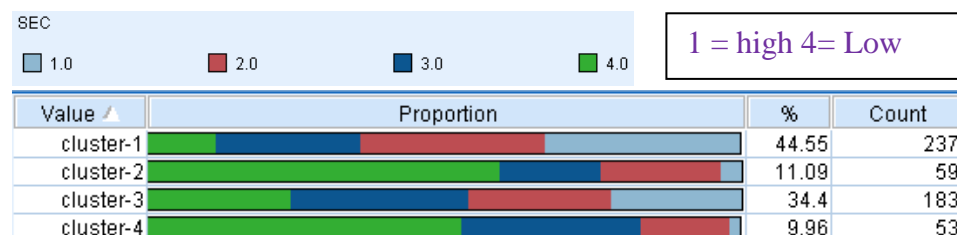
Price Only Shopper- Cluster 2: This customer pays the lowest average price of the 4 clusters and purchases any economy soaps. This cluster with of 11% (59 records) of total consumer in survey consists of the consumer who chose the highest value of Prop Cat 14 and Pr Cat 3 but relatively low value of purchase with promotion code 6 as the basis for purchase. With respect to purchase behavior this consumer has the highest value for Max to One Main Brand and a consequent lowest value for Share to Other 999 variable.

Mainstream/Popular/Beauty Shopper- Cluster 3: This customer is not as price sensitive ad purchases both beauty and popular brand soaps. This cluster with of 34.4% (183 records) of total consumer in survey consists of consumers who chose relatively low value of Prop Cat 14 and Pr Cat 3 and relatively low value of purchase with promotion code 6 as their basis for purchase. With respect to purchase behavior they have relatively high values for Max to One Main Brand and a consequent relatively low value for Share to Other 999 variable.

Value/Beauty Shopper- Cluster 4: This customer is more price sensitive than the mainstream customer and purchases beauty soap brands. This cluster with 10% (53 records) of total consumers in survey consists of consumers who chose low value of Prop Cat 14 and Pr Cat 3 but highest value of purchase with promotion code 6 as their basis for purchase. With respect to purchase behavior they have lowest values for Max to One Main Brand and a consequent high value for Share to Other 999 variable.

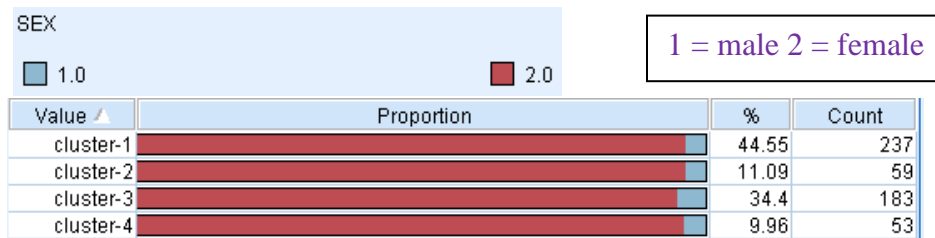
Cluster Demographics

Cluster Distribution by Socio Economic Class



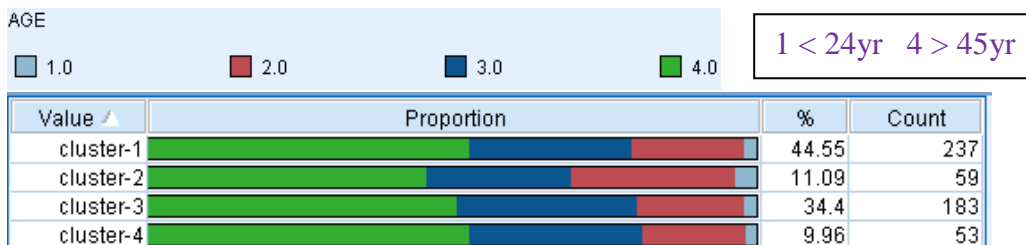
Cluster 1 (Premium) and 3(Mainstream) have higher proportion of people from upper social classes. More than 50% of people in Cluster 2 (Price) and 4 (Value) are from lowest socio economic class.

Cluster Distribution by Sex



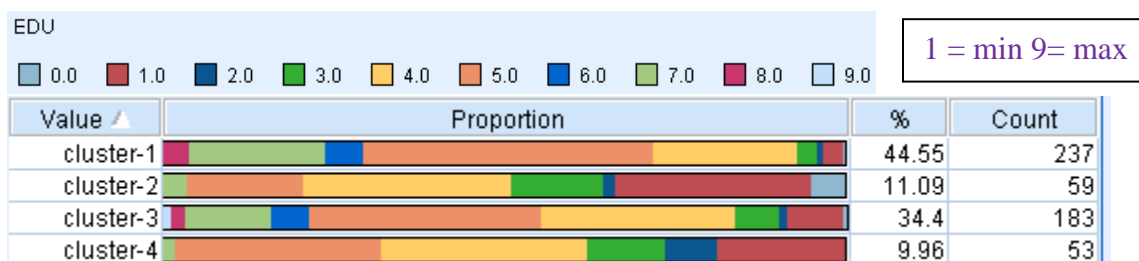
Cluster 1, 2, 3 and 4 are more than 85 % female.

Cluster Distribution by Age



People less than 24 yr make up a very small percentage of all clusters. Almost 50 % of all cluster are ladies older than 45 yr.

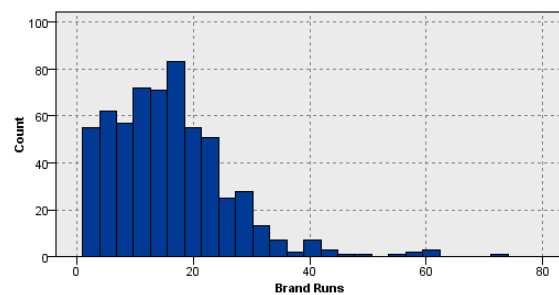
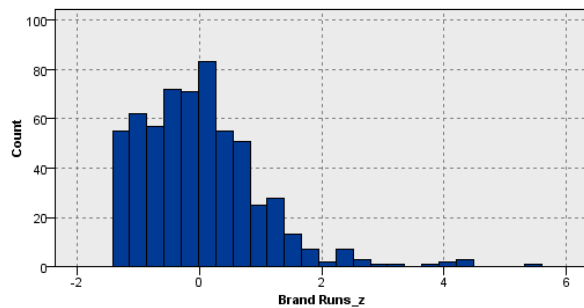
Cluster Distribution by Education



Cluster 4 has does not include college professionals. They are largely members of only cluster 2.

CART Classification - Brand Loyalty Model

In order to do further downstream analysis of the brand loyalty of consumers, we next apply the K-Means clusters as predictors to CART classification algorithm. By doing so we able to target loyal customers. Brand Runs, the number of instances of consequent purchases of brands is a measure of brand loyalty. A new 3 level categorical variable called Cat.Loyalty is derived from Brand Runs variable. The derived variable Cat.Loyalty is the target field in CART algorithm. The details of Cat.Loyalty are show below



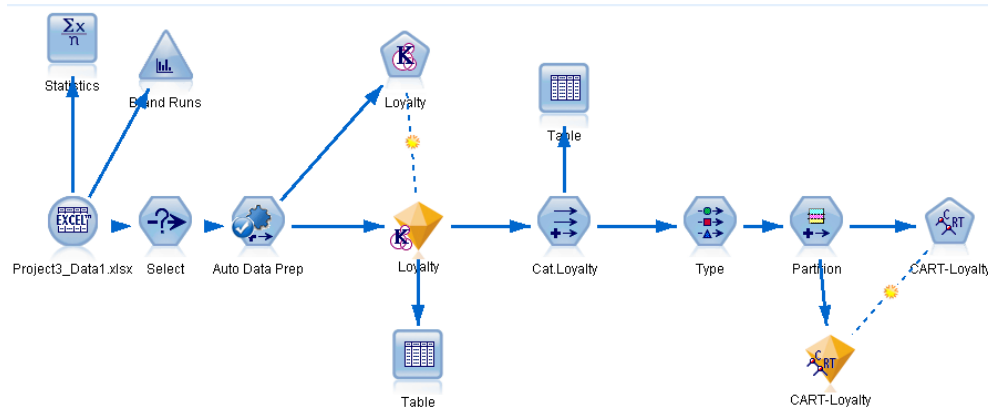
Brand Runs	
Statistics	
Count	600
Mean	15.752
Min	1.000
Max	74.000
Range	73.000
Standard Deviation	10.396

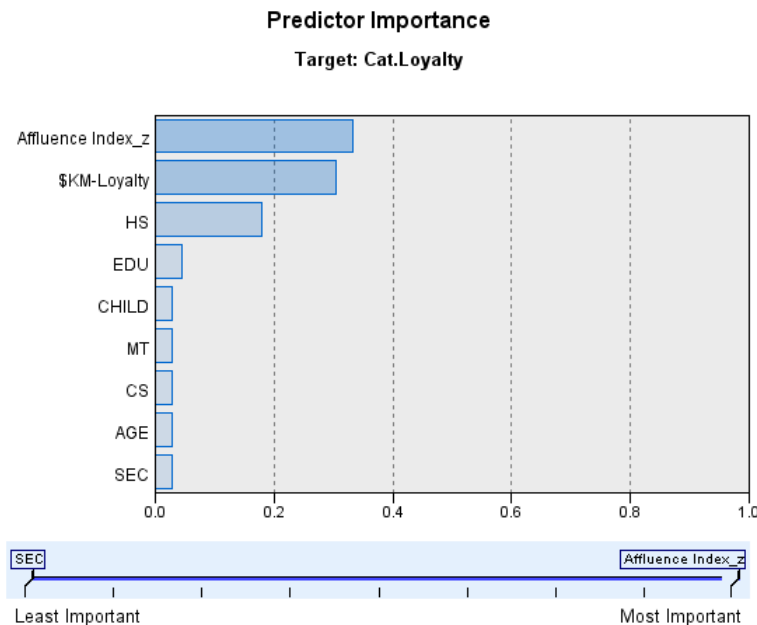
Brand Runs	Loyalty
< 3	Low
>= 3 and <= 25	Medium
>25	High

Brand Runs	Brand Runs_z
3	-1.227
25	0.8895

All the other relevant variables which influence purchase behavior and basis for purchase along with demographic variables are taken as inputs to CART.

The stream used is shown below.





The model shows that affluence index and the cluster grouping variable the most important predictor in the classification model. Demographic variables like size of household, education, number of children, native language, age and socio economic class play a less significant role in brand loyalty classification.

Model Summary

The CART classification recommendations are shown below along with support and confidence levels.

```

$KM-Loyalty in ["cluster-3"] [ Mode: Low ] (8)
├── Affluence Index_z <= -0.133 [ Mode: Low ] ⇒ Low (6; 1.0)
└── Affluence Index_z > -0.133 [ Mode: High ] ⇒ High (2; 1.0)
$KM-Loyalty in ["cluster-1" "cluster-4"] [ Mode: High ] (51)
├── Affluence Index_z <= -0.133 [ Mode: High ] (13)
│   ├── HS <= 4.500 [ Mode: High ] (5)
│   └── HS > 4.500 [ Mode: High ] ⇒ High (8; 1.0)
└── Affluence Index_z > -0.133 [ Mode: High ] ⇒ High (38; 1.0)
  
```

Since affluence index and cluster memberships are the only 2 important predictors and transforming back from z normalization for affluence index the recommendations from the CART decision tree are

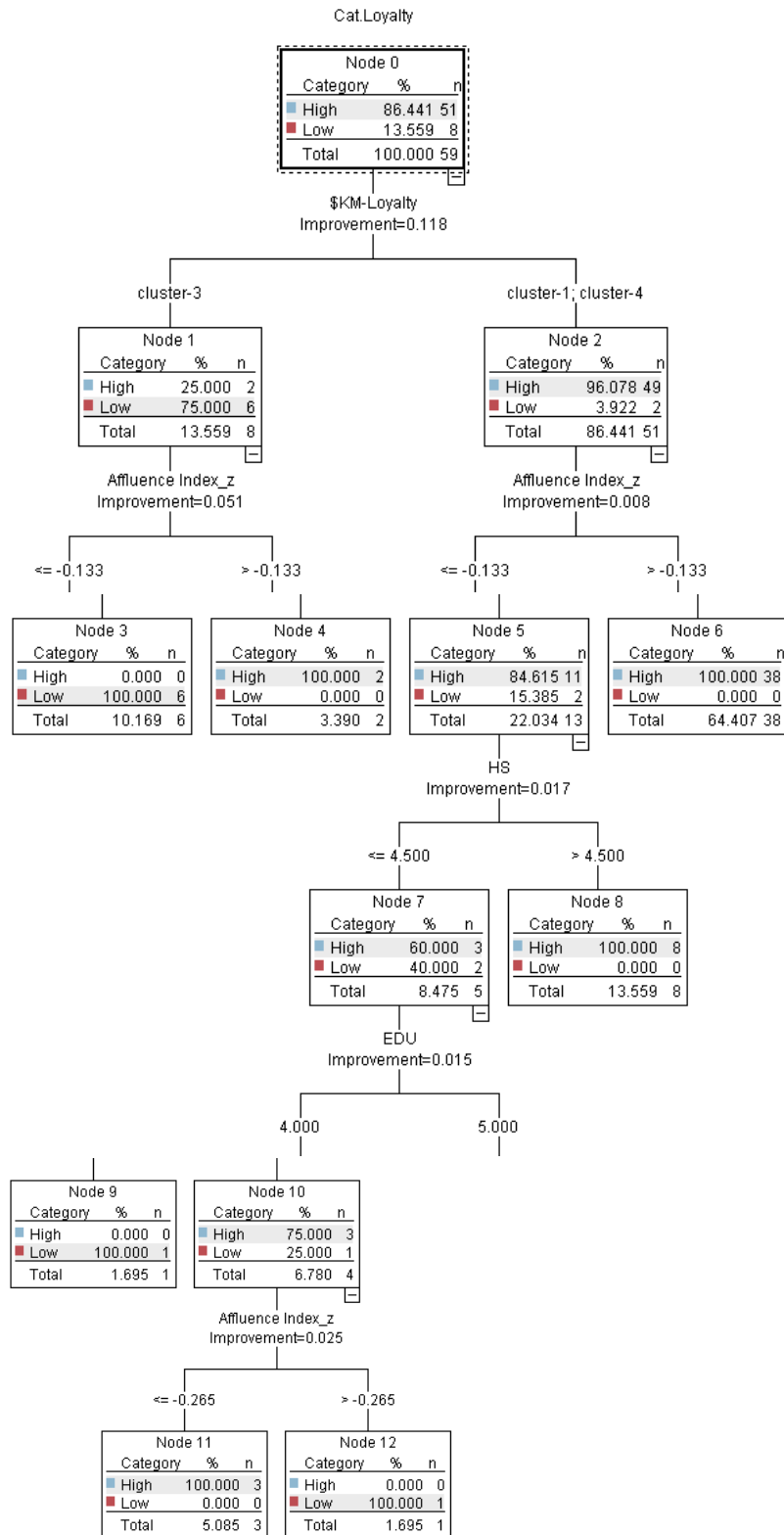
Consumers who are members of cluster 3 (Mainstream) and affluence index ≤ 15.5 ($z = -0.133$) have low brand loyalty.

Consumers who are members of cluster 3 (Mainstream) and affluence index > 18.54 ($z = 0.133$) have high brand loyalty.

Consumers who are members of cluster 1 (Premium Brand) or Cluster 4 (Value Beauty) irrespective of their affluence index and household size have high brand loyalty.

Cluster 2 customers (Price) have no brand loyalty. They are shopping for the lowest price regardless of brand and will buy the product with a price promotion over brand preference.

The CART decision tree



Conclusions

Three K- Means cluster models – Purchase Behavior Cluster Model, Basis for Purchase Cluster Model and All Variable Cluster Model were built in order to segment the IMRB household data to guide in development of advertising and promotional campaigns based on brand loyalty.

The Purchase Behavior Cluster Model was developed with 8 variables that describe the purchasing habits of consumers, shows that there are 4 clusters of consumers (Premium Brand, Price Only, Mainstream/Popular/Beauty and Value/Beauty) with different levels of brand loyalty given by number of brands bought, max to one main brand, volume of purchase etc.

The Basis for Purchase Cluster Models developed with 9 variables that describe the purchasing habits of consumers in response to discount promotions, selling propositions and price categories shows that there are 4 clusters of consumers with different levels of these variables.

The cluster models reduce the large number of variables in the dataset pertaining to the purchasing habits of consumers

CART classification decision tree, developed with Brand Runs as target and cluster field and demographic variable fields as input, gives a good picture of likely characteristics loyal consumers.

Brand loyal consumers who have bought same brand consequently for more than 25 times are those who are described as follows:

1. Members of the largest cluster, Cluster 3 (Mainstream), with affluence index > 18.54 .
2. Members of Cluster 1 (Premium Brand) and Cluster 4 (Value Beauty) irrespective of their affluence index and household size.

These are the customers to whom the advertising and promotional campaigns materials should be directed to. Promotional offers could include offers for premium brands, beauty and popular soap brands to attract these consumers. These consumers are less likely to purchase sub-popular and carbolic soaps even with promotions.

Customers from Cluster 2 (Price Only) are not brand loyal and will purchase if the promotion provides the lowest price. They are not likely to purchase again if there is no promotion offered.