SEGMENTING AND PROFILING CUSTOMERS OF A RETAIL STORE USING DATA MINING APPROACH

Dr Naveeta Mehta*

Ruchi Mittal**

ABSTRACT

The primary objective of this article is to demonstrate the impact of using data mining techniques in an attempt to bring the fields of management and IT much closer. This article is an empirical research on the application of descriptive data mining models in the field of shopping behavior in the context of apparel retailing. This paper uses Cluster Analysis and then explains how it can be applied to a survey based data for segmenting customers. In this study, customers have been clustered and then profiled in a grocery shopping environment based on their store loyalty, demographics and shopping behavior. The findings of this research provide some very significant insights into consumer behavior at the retail level which will help retailers and academicians design more customer centric marketing strategies.

Keywords: *Data Mining, Grocery Shopping, Cluster Analysis and Store Loyalty*

*Professor, MMICT & BM, Mullana (Ambala), Haryana, INDIA

**Associate Professor, MCA Deptt. MAIMT, Jagadhri, Haryana, India

INTRODUCTION

Data mining is an emerging field that focuses on access of information useful for high-level decisions. It is the confluence of multiple disciplines and enables business executives to manage their data to make relevant decisions [1]. Retail is the main field of application of data mining technology. It is India's largest industry accounting for over 10 per cent of the GDP and 8 per cent of employment. The industry is facing the new millennium, and the models of the past are not sufficient to ensure tomorrow's successes. The stereotype of a homogeneous market is a fiction that no longer exists. Firms are now employing the strategy of segmentation, viewing the market made up of small segments, each more homogeneous in important characteristics. [2]

LITERATURE REVIEW

Liu & Luo applied clustering data mining method to customers of store for the analysis of their characteristics and the relationship between customers and product categories [4]. Using data-mining techniques, Hockey Min profiled 301 supermarket consumers based on their loyalty to 10 supermarkets in U.S and found a significant relationship between valued consumers' shopping behavior and demographic variables [5]. In Asian context, a study of 400 supermarket consumers in Qatar, Jamal et al. used cluster analysis to identify six shopper typologies on specific reasons for shopping such as Socializing; Disloyal; Independent; Escapist; Apathetic; and Budget Conscious Shoppers [6]. In Indian context Goyal & Mittal [3] identified two shopper types i.e. Pro-Shopper and Anti-Shopper, based on the degree of enjoyment they felt from shopping. In another study of 300 shoppers, Sinha [7] clustered shoppers in two segments- (1) Fun Shoppers (2) Work Shoppers. Batra and Ahtola [8], also investigated consumers' attitude towards brands and behaviors, and argued that it has at least two distinct dimensions, hedonic and utilitarian.

OBJECTIVES OF THE STUDY

This study seeks to identify clusters of Indian grocery shoppers on the basis of their current store loyalty. The clusters will then be profiled to understand who represents a typical "loyal shopper" and a typical "non-loyal shopper". The research is conducted in the context of grocery retailing. The specific objectives are:

Objective 1: To use data mining techniques to identify the various customers types for grocery shopping scenario;

Objective 2: To find the characteristics and the behavioral patterns of various customer types identified;

RESEARCH METHODOLOGY

The data were collected from male and female adults above (or equal to) the age of 20 years comprising grocery shoppers. To obtain a profile of the respondents they were requested to complete questions regarding the following descriptive:

Demographics:

- Gender
- Age
- Occupation (Profession)
- Education
- Income

Shopping Behavior

- 'Shop Alone or Shop with someone';
- Expenditure on the shopping category (monthly) and
- Number of shopping trips (shopping frequency).

The final sample consisted of 352 respondents who volunteered to respond to our questionnaire.

STATISTICAL DATA MINING

The data obtained is classified using data mining technique, Two Step *Cluster Analysis* due to large sample size.

4.2.1 Two Step Cluster Analysis: The Two step cluster analysis procedure has been used when we have a large data set or we need a clustering procedure that can rapidly form clusters on the basis of either categorical or continuous data. It requires only one pass of data (which is important for very large data files), and it can produce solutions based on mixture of continuous and categorical variables and for varying number of clusters. Cluster analysis does not involve

any hypothesis testing and calculation of observed significance levels, other than for descriptive follow up, it's perfectly acceptable to cluster data that may not meet the assumptions for best performance but provide us the satisfactory results which is our major requirement.

4.2.2 Chi Square Analysis: When you cluster cases, you want to know how important the different variables are for the formation of the cluster. For categorical variables, SPSS calculates a chi-square value that compares the observed distribution of values of a variable within a cluster to the overall distribution of values. In this study Figure---- is a plot of the chi-square statistic for newspaper readership. Within each cluster, the observed distribution is compared to an expected distribution based on all cases. Large values of the statistic for a cluster indicate that the distribution of the variable in the cluster differs from the overall distribution. The **critical value line** that is drawn provides some notion of how dissimilar each cluster is from the average. If the absolute value of the statistic for a cluster is greater than the critical value, the variable is probably important in distinguishing that cluster from the others. This method can be followed for all variables used in the cluster analysis and the final result based on comparison of the test statistic with the critical value line can help us profile the clusters.

RESULTS and ANALYSIS

Objective 1: **To use data mining techniques to identify the various customers types for grocery shopping scenario;**

The data mining technique used here is two-step cluster analysis. The TwoStep Cluster Analysis procedure is an exploratory tool designed to reveal natural grouping within a data set that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- Handling of categorical and continuous variables. By assuming variables to be independent, a joint multinomial-normal distribution can be placed on categorical and continuous variables.

- Automatic selection of number of clusters by comparing the values of a model-choice criterion across different clustering solutions, the procedure can automatically determine the optimal number of clusters.
- Scalability. By constructing a cluster features (CF) tree that summarizes the records, the Two Step algorithm allows you to analyze large data files.

The two step clustering method is a scalable cluster analysis algorithm designed to handle very large data sets. It has two steps:

a) Pre cluster the cases into many small sub clusters and b) Cluster the sub clusters resulting from pre cluster step into the desired number of clusters.

Table 1: Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change(a)	Ratio of BIC Changes(b)	Ratio of Distance Measures (c)
1	5992.432			
2	4988.359	-1004.073	1.000	6.455
3	4955.988	-32.371	.032	1.209
4	4954.360	-1.628	.002	1.024
5	4956.176	1.816	-.002	1.086
6	4969.338	13.163	-.013	1.005
7	4983.208	13.870	-.014	1.106
8	5009.666	26.458	-.026	1.078
9	5044.718	35.052	-.035	1.059
10	5085.911	41.192	-.041	1.015
11	5128.659	42.748	-.043	1.005
12	5171.956	43.297	-.043	1.090
13	5223.693	51.737	-.052	1.122

14	5285.664	61.971	-.062	1.019
15	5349.182	63.518	-.063	1.075

a. The changes are from the previous number of clusters in the table 1.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Here, automated cluster selection has been used where number of clusters can be found using the Schwarz Bayesian Criterion (SBIC)- where 'I' stands for information. In statistics, the **Bayesian information criterion** (**BIC**) or **Schwarz criterion** (also **SBC**, **SBIC**) is a criterion for model among a class of parametric models with different numbers of parameters. Choosing a model to optimize BIC is a form of regularization. The number of clusters at which the **Schwarz criterion** BIC becomes small and the change in BIC between adjacent numbers of clusters is small, has been shown.

The algorithm selected two clusters which can also be determined from Table 2. As discussed from Table 2, Cluster-1 has 120 cases, while Cluster-2 has 152 cases. A total of 272 cases out of the total respondents 352 were clustered. 22.7% cases (80 nos. were excluded).

Table 2: Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	120	44.1%	34.1%
	2	152	55.9%	43.2%
	Combined	272	100.0%	77.3%
Excluded Cases		80		22.7%
Total		352		100.0%

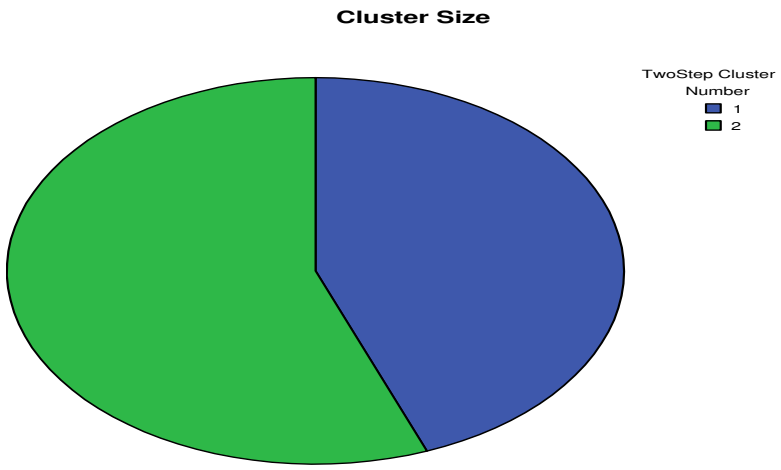

Figure 1: Graphical Representation of the Clusters.

STORE LOYALTY		1		2		3		4		5		6		7	
		Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Cluster	1	8	72.7%	29	87.9%	43	100.0%	30	63.8%	8	12.7%	2	3.4%	0	.0%
	2	3	27.3%	4	12.1%	0	.0%	17	36.2%	55	87.3%	57	96.6%	16	100.0%
	Combined	11	100.0%	33	100.0%	43	100.0%	47	100.0%	63	100.0%	59	100.0%	16	100.0%

Table 3: Cluster wise Store Loyalty Ratings

As per the table3, it can be observed that Cluster 1 primarily consists of shoppers who are store non- loyal, whereas, Cluster 2 consists of store loyal shoppers. The Store loyalty rating of '7' denotes the highest degree of loyalty while '1' denotes the lowest degree of shopping. The same can also be observed from the figure 2. In this case the Test statistic is greater than the Critical value; these values are based on the statistical calculations of chi-square analysis. This confirms our observation from figure 2 that the variable Store Loyalty is a significant in differentiating the two clusters.

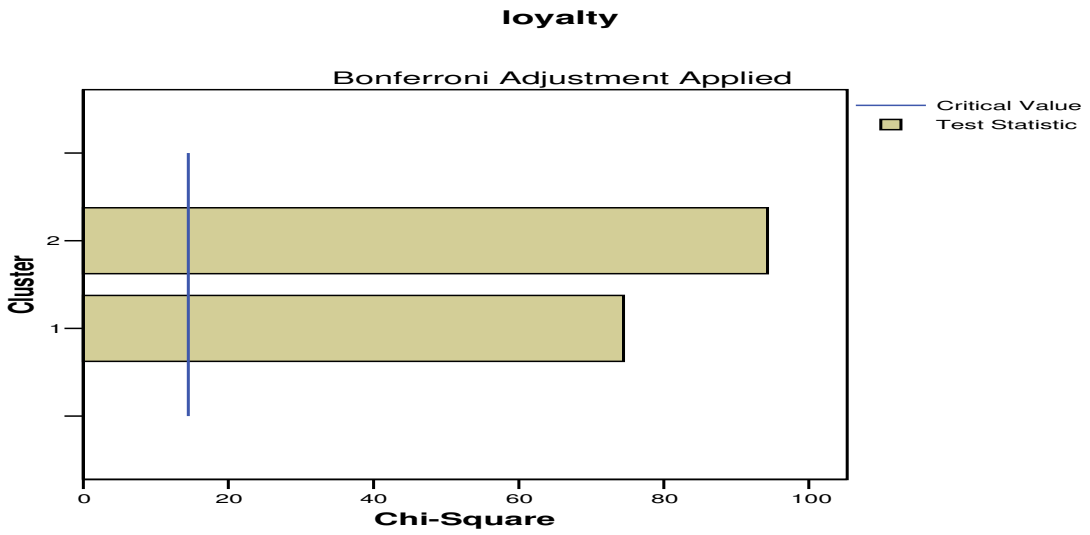

Figure 2: Chi Square analysis of Loyalty variable based on Bonferroni Adjustment.

Based on the Table 3 and figure 2; the two clusters are named as follows:

Cluster1: Store non-loyals

Cluster2: Store loyals

Objective 2: To find the characteristics and the behavioral patterns of various customer types identified;

Now that the clusters are formed, as per objective number 2, the profiling of each cluster needs to be done. The profiling has been done based on demographic and behavioral variables mentioned below:

Demographics:

- Gender
- Age
- Occupation (Profession)

- Education
- Income

Shopping Behavior

- 'Shop Alone or Shop with someone';
- Expenditure on the shopping category (monthly) and
- Number of shopping trips (shopping frequency).

CLUSTER PROFILING

A) Gender

		Male		Female	
		Frequency	Percent	Frequency	Percent
Cluster	Non-Loyals	80	76.2%	40	24.0%
	Loyals	25	23.8%	127	76.0%
	Combined	105	100.0%	167	100.0%

Table 4: Cluster-wise distribution of shoppers based on 'Gender'

Gender is one of the variables considered for profiling of the two clusters or segments of shoppers that have emerged. Table 4 shows that Men are more likely to be with-in the Store Non-loyals cluster (Cluster 1) as compared to women. Women seem more inclined to remain loyal (Cluster 2) to the same store especially in the context of grocery shopping.

The graphic representation of gender distribution as per Table 4 also shows that the distribution of males and females is different for the clusters vis-à-vis the overall distribution of gender.

Figure 3(a) validates the difference in distribution of the expected frequency from that of the observed frequency using the chi-square method (used for categorical variables).

For categorical variables, SPSS calculates a chi-square value that compares the observed distribution of values of a variable within a cluster to the overall distribution of values. Figure 3(b) is a plot of the chi-square statistic for gender. Within each cluster, the observed distribution is compared to an expected distribution based on all cases. Large values of the statistic for a cluster indicate that the distribution of the variable in the cluster differs from the overall distribution. The **critical value line** that is drawn provides some notion of how dissimilar each cluster is from the average. If the absolute value of the statistic for a cluster is greater than the critical value, the variable is probably important in distinguishing that cluster from the others. In this case, it can be observed that Gender is an important variable in differentiating both the clusters from the overall distribution.

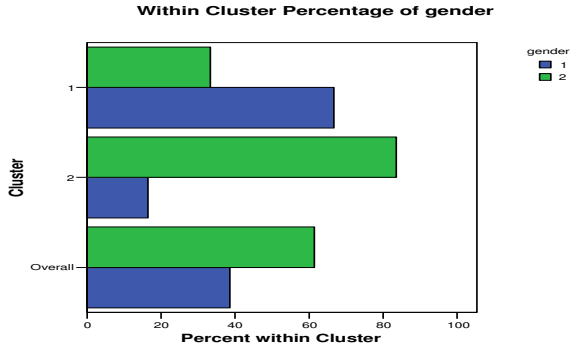

Figure 3(a)

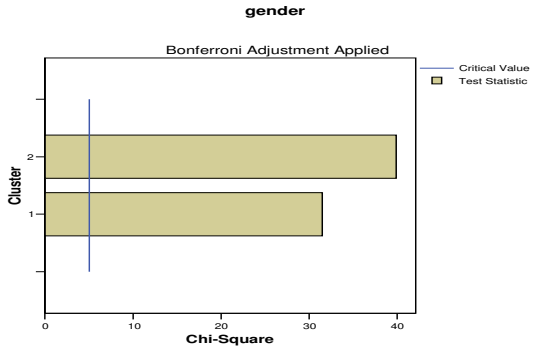

Figure 3(b)

B) AGE:

		20-29		30-39		>40	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	Loyals	39	36.1%	72	50.0%	9	45.0%
	Non-Loyals	69	63.9%	72	50.0%	11	55.0%
	Combined	108	100.0%	144	100.0%	20	100.0%

Table 5: Cluster-wise distribution of shoppers based on 'Age'

Age is one of the variables considered for profiling of the two clusters or segments of shoppers that have emerged. Across different age categories, it seems inconclusive to distinguish the clusters on the basis of age.

The chi-square statistic also indicates that age is an insignificant variable when it comes to profiling the clusters.

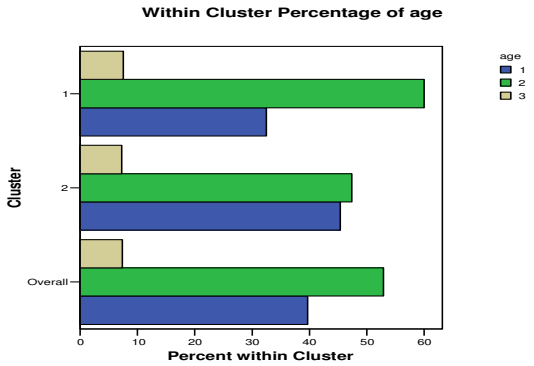

Figure 4(a)

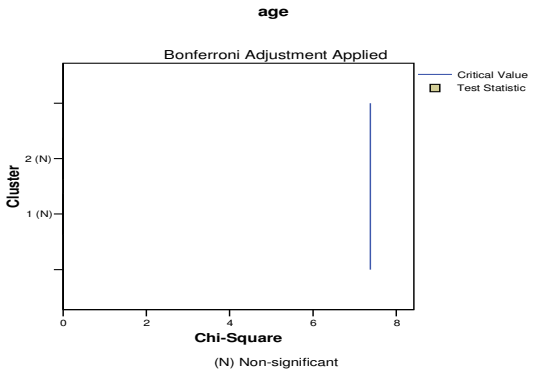

Figure 4(b)

C) Occupation

		Businessperson		Self-employed		Service		Homemaker		Student		Retired	
		Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Cluster	1	6	54.5	21	100.0	51	52.6	14	14.6	28	63.6	0	.0
	2	5	45.5	0	.0	46	47.4	82	85.4	16	36.4	3	100.
	Combined	11	100.0	21	100.0	97	100.0	96	100.0	44	100.0	3	100.0

Table 6: Cluster-wise distribution of shoppers based on 'Occupation'

Occupation is another considered for profiling of the two clusters or segments of shoppers that have emerged. Table 6 shows that Self-employed individuals, those in Service and Students are more likely to be with-in the store non-loyals cluster (Cluster 1). Homemakers and Retired individuals seem more inclined towards grocery store loyalty, as seen from their likelihood of being present in significant numbers in Cluster 2.

The graphic representation of Occupational distribution as per Figure 5(a) also shows that the distribution of various occupations is different for the clusters vis-à-vis the overall distribution for Occupation.

Figure 5(b) validates the difference in distribution of the expected frequency from that of the observed frequency using the chi-square method (used for categorical variables). By observing the **critical value line** it can be stated that occupation is an important variable and the variable is important in distinguishing both clusters from each other.

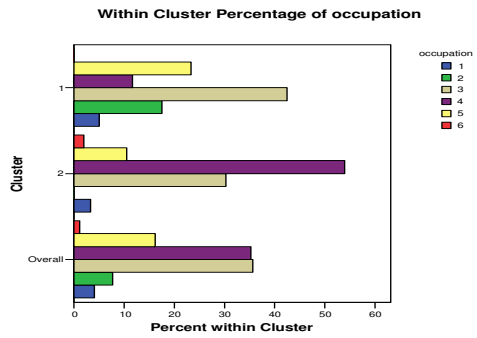

Figure 5(a)

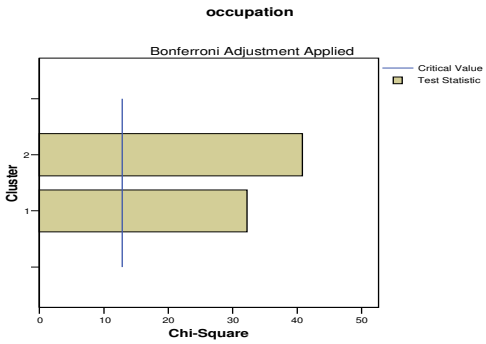

Figure 5(b)

D) Education:

		Undergraduates		Graduates		Post-Graduates	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	2	3.5%	63	39.9%	55	96.5%
	2	55	96.5%	95	60.1%	2	3.5%

Combined	57	100.0%	158	100.0%	57	100.0%

Table 7: Cluster-wise distribution of shoppers based on 'Education'

Education is another demographic variable considered for profiling of the two clusters or segments of shoppers that have emerged. Table 6 shows that higher educated individuals (Post Graduates and Graduates) are more likely to be with-in the store non-loyals cluster (Cluster 1). Less educated shoppers (Undergraduates) are inclined towards maintaining grocery store loyalty, as seen from their likelihood of being in the first cluster .i.e. Cluster 2.

The graphic representation of Education distribution as per Figure 6 also shows that the distribution of various occupations is different for the clusters vis-à-vis the overall distribution for Occupation.

Figure 6 validates the difference in distribution of the expected frequency from that of the observed frequency using the chi-square method (used for categorical variables). By observing the **critical value line** it can be stated that Education variable is probably important in distinguishing both clusters from each other.

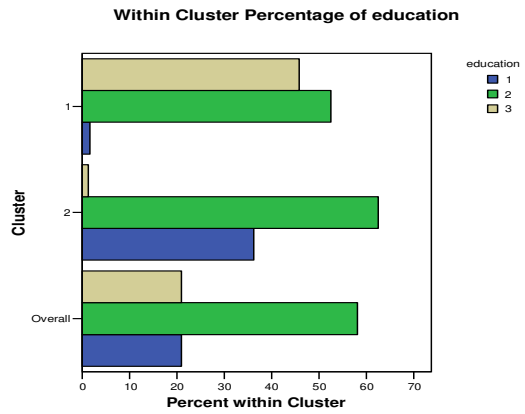

Figure 6(a)

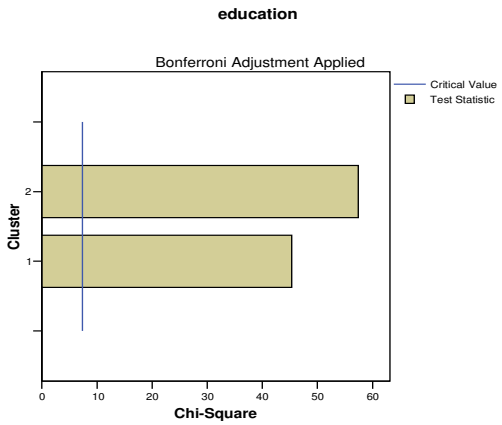

Figure 6(b)

E) INCOME:

		<10,000		10,000 – 20,000		20,001-30,000		>30,000	
		Freq	%	Freq	%	Freq	%	Freq	%
Cluster	1	0	.0%	13	15.3%	49	53.3%	58	77.3%
	2	20	100.0%	72	84.7%	43	46.7%	17	22.7%
	Combined	20	100.0%	85	100.0%	92	100.0%	75	100.0%

Table 8: Cluster-wise distribution of shoppers based on 'Income'

Monthly Household Income (MHI) is another demographic variable considered for profiling of the two clusters or segments of shoppers that have emerged. Table 8 shows that higher income individuals (MHI more than Rs 20,000) are more likely to be with-in the store non-loyalty cluster (Cluster 1). Shoppers with lesser incomes (MHI Less than Rs 20,000) are inclined towards maintaining store loyalty, as seen from their likelihood of being in Cluster 2.

The graphic representation of Income distribution as per Figure 7(a) also shows that the distribution of various Income categories is different for the clusters vis-à-vis the overall distribution for MHI.

Figure 7(b) validates the difference in distribution of the expected frequency from that of the observed frequency using the chi-square method (used for categorical variables). By observing the **critical value line** it can be stated that the MHI variable is important in distinguishing both clusters from each other.

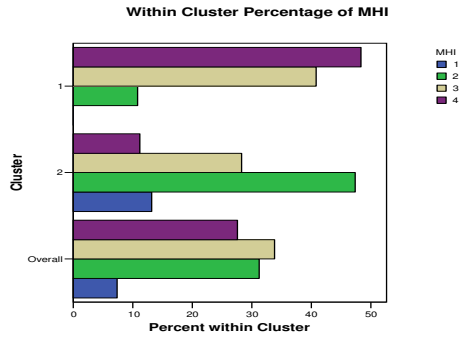

Figure 7(a)

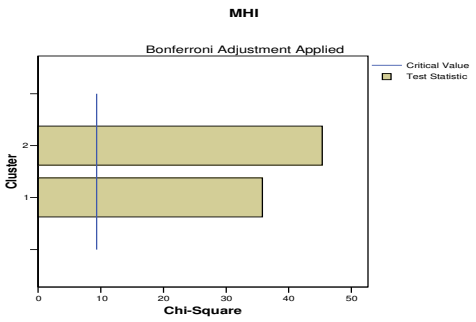

Figure 7(b)

F) Shop-with

		1		2	
		Frequency	Percent	Frequency	Percent
Cluster	1	52	50.0%	68	40.5%
	2	52	50.0%	100	59.5%
	Combined	104	100.0%	168	100.0%

Table 9: Cluster-wise distribution of shoppers based on 'Shop-with'

'Shop-with someone or alone' is one of the three behavioral variables considered for profiling of the two clusters or segments of shoppers that have emerged. Table 9 and Figure show that "Shopping with someone or alone" is not a significant variable to distinguish between Cluster 1 and Cluster 2.

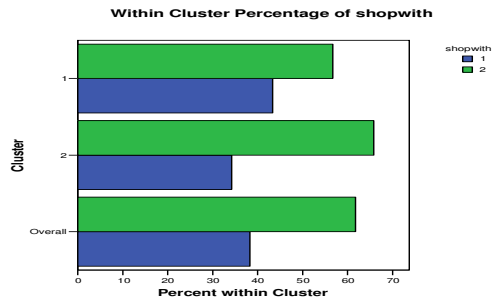

Figure 8(a)

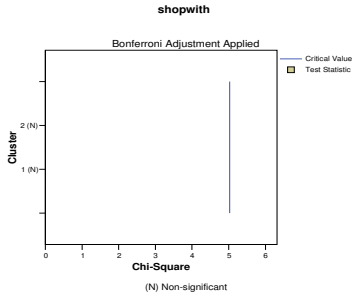

Figure 8(b)

G) Monthly Expenditure on Groceries:

		<500		501-1000		>1000	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	46	82.1%	39	37.1%	35	31.5%
	2	10	17.9%	66	62.9%	76	68.5%
	Combined	56	100.0%	105	100.0%	111	100.0%

Table 10: Cluster-wise distribution of shoppers based on 'Monthly Expenditure'

The spend per month is the last of the three behavioral variables considered for profiling of the two clusters or segments of shoppers that have emerged. Table 10 shows that individuals in the non-loyalty cluster (Cluster 1) spend less compared to shoppers represented in the store loyal cluster (Cluster 2).

The graphic representation of Shopper Spend distribution as per Figure 9(a) also shows that the distribution of various Shopping categories is different for the clusters vis-à-vis the overall distribution for 'Spend'.

Figure 9(b) validates the difference in distribution of the expected frequency from that of the observed frequency using the chi-square method (used for categorical variables). By observing the **critical value line** it can be stated that the 'Spend' variable is important in distinguishing both clusters from each other.

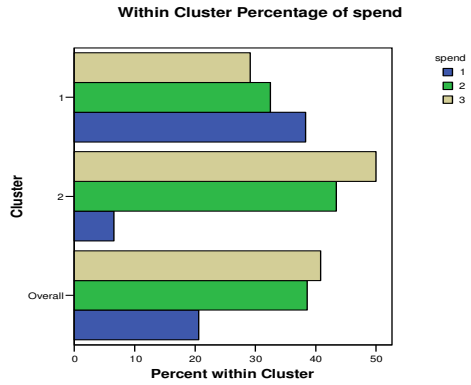

Figure 9(a)

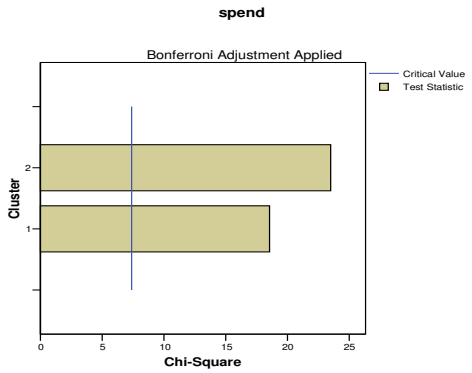

Figure 9(b)

H) No. of Shopping Trips:

		1-4 trips / month		5-9 trips / month		10-15 trips / month		>15 trips / month	
		Freq	%	Freq	%	Freq	%	Freq	%
Cluster	1	48	98.0%	68	60.2%	4	5.6%	0	.0%
	2	1	2.0%	45	39.8%	67	94.4%	39	100.0%
	Combined	49	100.0%	113	100.0%	71	100.0%	39	100.0%

Table 11: Cluster-wise distribution of shoppers based on 'Gender'

The number of 'Shopping Trips' is another of the three behavioral variables considered for profiling of the two clusters or segments of shoppers that have emerged. Table 11 shows that individuals in the store non-loyal cluster (Cluster 1) make fewer shopping trips compared to shoppers represented in the store loyal cluster (Cluster 2).

The graphic representation of Shopping Trips distribution as per Figure 10(a) also shows that the distribution of various Shopping categories is different for the clusters vis-à-vis the overall distribution for Occupation.

Figure 10(b) validates the difference in distribution of the expected frequency from that of the observed frequency using the chi-square method (used for categorical variables). By observing

the **critical value line** it can be stated that the 'Shop-with…' variable is probably important in distinguishing both clusters from each other.

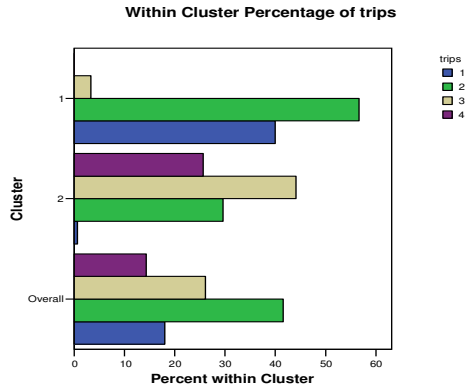

Figure 10(a)

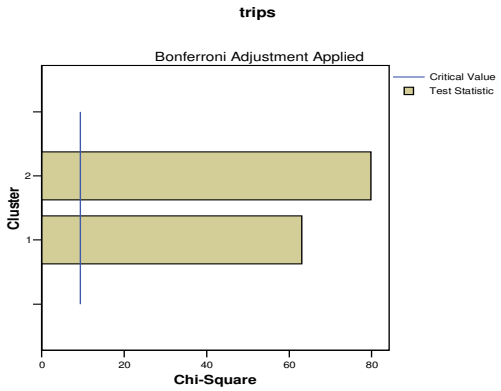

Figure 10(b)

DISCUSSIONS AND CONCLUSION

Marketers are often interested in attracting not just brand users, but perhaps more importantly, those who consistently purchase the company's brand. In the context of retailing, this would mean identifying and attracting regular visitors and buyers at a retail store. Using the Data-mining technique of 2-step clustering this study, in the context of grocery shopping, has identified two clusters from customer data based on loyalty behavior.

The two clusters: Store Loyals and Store Non-loyals have further been profiled on the basis of demographics and shopping behavior. Interestingly loyal shoppers have been profiled as females across all age groups. They are likely to be housewives (homemakers) or retired individuals. Shoppers with lesser incomes (less than Rs 20000 per month) and lesser education levels are also likely to be store loyals but they end up spending more and also make several shopping trips.

On the other hand, store non-loyals are most likely to be males across all age groups. They are generally self-employed or in service or students. Store non-loyals are also comparatively better educated and have higher incomes (more than Rs 20,000). They also spend less on groceries and make fewer trips to the market-place.

REFERENCES

- Kapil, A., Mittal, R. and Mittal, A., Descriptive Modeling of Customers in a Retail Store- A Data Mining Approach, PCTE Journal of Computer Science, Vol. 5, No. 1, 2008, pp. 82-91.
- Loudon, D.L. & Della Bitta, A.J., Consumer Behavior, Tata McGraw-Hill, 2002.
- Goyal, B.B. & Mittal, Amit, Gender Influence on Shopping Enjoyment: An Empirical Study, Indian Management Studies Journal, Vol. 11, No. 2, 2007, pp. 103-116.
- W. Liu & Y. Luo., Applications of clustering data mining in customer analysis in department store, In *Proc*. IEEE *Int. Conf*. Services Systems and Services Management, Vol. 2, 2005, pp. 1042-1046.
- H. Min., Developing the profiles of supermarket customers through data mining, The Service Industries Journal, Vol. 26, No. 7, 2006, pp. 747-763.
- Jamal, A; Davies, F; Chudry, F and Al-Marri M, Profiling Consumers: A Study of Qatari Consumers' Shopping Motivations, Journal of Retailing and Consumer Services, Vol. 13, 2006, pp. 67-80.
- Sinha, P.K, Shopping orientation in the evolving Indian market, Vikalpa, Vol. 28, No. 2, 2003, pp. 13-22.
- Batra, R. and Ahtola, O.T, Measuring the hedonic and utilitarian sources of consumer attitudes, Marketing Letters, Vol. 2, 1991, pp. 159-170.
- O'Guinn, T.C. and Faber, R.J., Compulsive Buying: A Phenomenological Exploration, Journal of Consumer Research, Vol. 16, 1989 pp 147-157.
- De Wulf, K., Odekerken-Schroder, G. and Iacobucci, D., Investments in Consumer Relationships: A Cross-Country and Cross-Industry Exploration, Journal of Marketing, Vol. 65, 2001, pp 33-50.