THE INTERNATIONAL JOURNAL OF BUSINESS & MANAGEMENT

Cluster Analysis for Market Segmentation: A Study Conducted in Aligarh Muslim University Malappuram Centre, Kerala

Waqaruddin Siddiqui
Student, Department of Business Administration, Aligarh Muslim University, Aligarh, UP, India

Abstract:
Regardless of them any kinds of procedures available for categorizing individuals into market segments on the basis of multivariate survey information, clustering is the most used procedure. A review of the application of such data-driven categorizing procedures reveals that standards have aroused that are questionable. For example, the investigative essence of categorizing procedure is typically not considered for, critical specifications of the algorithms used are neglected, therefore, it leads to a dangerous black-box approach, but the main reasons for these specific outcomes are not fully understood, pre-processing techniques are used uncritically resulting to segmentation solutions in needlessly altered data space, etc. Main aim of this research paper is to unearth typical patterns of data driven segmentation studies, giving a censorious analysis of resulted standards and recommending advancements.
For the study the information was gathered with the help of a questionnaire on the profile of target audience in terms of lifestyle, attitude, and perception.
The result was used in our analysis. We used the statistics computer programme SPSS to easier see the significance of results.

Keywords: *Segmentation, cluster, attitude, perception, market*

1. Introduction

It is recognized that the aimed population of a particular goods are not all alike. They are different in terms of demographics, attitudes, wants, and social affiliations. Mostly markets are consisting different individual customers, sub-markets or segments.
Segmentation and targeting of customers helps the marketer to provide a product within the target audience as per their needs and wants. It is a very important to institute the needs and values of the target customers in each segment, in order for companies to promote their products, brands or services properly.
Most of researches in this field have shown these facts about the audience. Hence, for marketers to frame the correct strategy it is important for the marketer to have a sound knowledge related to the needs and wants of the target audience. With the use the segmentation process the marketer will get the required information and then he will be able to frame a right and workable marketing strategy.
This paper will provide a base for exemplifying how the segmentation can identify the right aimed audience. It will be done by using many factors about the target audience i.e; their perception, attitude, and lifestyle etc.

1.1. Objective

The objective of this paper is to know the segmentation of target market audience and the concepts related to it. Also, by attaining knowledge of these concepts it should be possible to define the correct aimed customers in terms of attitude, perception and lifestyle by analysis.

1.2. Delineation

This paper will emphasize on the segmentation and elucidate a target audience which will be studied as per the survey donewith the help of questionnaire. Also, as we know this paper is not based upon the target audience for any specific brand therefore we will consider target audience in general. Due to the fact that a survey has been conducted with Aligarh Muslim University Centre, Malappuram, Kerala participants, the analysis in the report will not represent any other market but only the Aligarh Muslim University Centre, Malappuram, Kerala market. Hence, the theoretical part of the paper is aimed to give knowledge of the technical terms used in this paper. In addition to that, the results from the survey are purely based on the Aligarh Muslim University Centre, Malappuram, Kerala audience and have no relevance in other aspects beside this report.
For the analysis, the hierarchical and non-hierarchical method are to be the used.

1.3. Background of the Study

Cluster analysis is used to define a sample of subjects on the basis of a set of measured variables into various groups so that alike subjects are placed in the alike group. In marketing, it can be helpful to identify various groups of potential customers so that, for example, promotional programmes can be effectively targeted.

2. Methods of Cluster Analysis

Many methods are used to perform cluster analysis and some of these methods are defined below:

- Hierarchical methods– Here subjects start in their own different cluster. The two most similar clusters are combined and this is done till all subjects are in one cluster. At last, the optimum number of clusters is chosen from all cluster solution.
- Non-hierarchical methods is also used (generally known as k-means clustering methods).

2.1. Kinds of Data and Measures of Distance

Interval, ordinal or categorical data can be used. Though, if there is a combination of different kinds of variable then it will make the analysis complex because in cluster analysis must have technique of computing the distance between observations and the kind of measure employed will rely on the kind of data you are having. Many measures have been presented to measure distance for binary and categorical data.

3. Research Methodology

As we know that Cluster analysis is a multivariate method which defines a sample of subjects on the basis of a set of measured variables into various groups so that alike subjects are placed in the alike group.

3.1. Methods

Primarily, the techniques of clustering which we use in computer packages are of two kinds:

1- Hierarchical clustering or Linkage methods.
2- Non- hierarchical clustering or Nodal methods.

The first type involves method like single linkage, complete linkage, and average linkage. Here, it is not required to specify in advance the number of clusters to be uprooted. An array of solutions is given by the computer system, from a 1-cluster solution to an n-cluster solution, where *n* is the total number of objects studied. The other type comprises the *K*-means approach, in which you have specify in advance the number of clusters to be uprooted from the data. Euclidean distance measure is one of the widely used to define distance between objects clustered.

3.2. Data/ Scales of Variables

Generally, interval-scaled variables are used for cluster analysis. Also, Continuous or ratio-scaled variables can be used. Sometimes standardization might be required if the units of measurement of various variables are alike.

4. Analysis of Data

4.1. Input Data:

A random sample of 100respondents of Aligarh Muslim University Malappuram Centre, Kerala is taken. The survey was done with the help of a questionnaire. The modelling and testing of market segmentation using clustering for estimation was based on the UG and PG class students of Aligarh Muslim University Centre, Malappuram, Kerala, India. It was necessary to depict the outline of the target audience in terms of lifestyle, attitudes and perceptions. The survey contained 15 different questions as given below in Table 1.

Var 1	I prefer to use email rather than writing a letter
Var 2	I feel that quality products are always priced high
Var 3	I think twice before I buy anything
Var 4	Television is a major source of entertainment
Var 5	A car is a necessary rather thana luxury
Var 6	I prefer fast food and ready-to-use products
Var 7	People are more health-conscious today
Var 8	Efficiency of Indian companies have increased due to the entry of foreign companies
Var 9	Women participates actively in purchase decisions
Var 10	I believe politicians can play a positive role
Var 11	I enjoy watching movies
Var 12	If I get a chance, I would like to settle abroad
Var 13	I always buy branded products
Var 14	I frequently go out on weekends
Var 15	I prefer to pay by credit card rather than in cash

Table1: variables chosen

Different variables used for marketing segmentation

4.2. Output and its Elucidation

4.2.1. Illustrative Analysis

A five-point rating scale was used to represent variables in segmentation. For this, the respondents were told to give response in categories of strongly agree as 5, agree as 4, No Opinion as 3, disagree as 2 and strongly disagree as 1. The Euclidean distance was used to measure the clustering analysis. Euclidean distance is perfectly appropriate for alike interval scaled variables. The input data matrix of 100 respondents with 15 variables is shown in Table 2. This could be explained as 35 respondents strongly agreed that they preferred emails to writing letters whereas 30 customers only agreed that they preferred email. Similarly, 10 customers strongly disagreed with the idea of emails, may be they didn't have access to internet or otherwise and so on.

Variable	Strongly agree (5)	Agree (4)	NAND (3)	Disagree (2)	Strongly disagree (1)
Var1	35	30	11	15	10
Var2	22	34	18	13	13
Var3	23	35	19	15	8
Var4	13	34	28	11	14
Var5	10	22	34	24	10
Var6	8	24	37	23	8
Var7	18	36	22	17	7
Var8	12	34	21	26	7
Var9	20	35	18	18	9
Var10	17	29	25	18	11
Var11	19	24	21	30	6
Var12	18	26	20	16	20
Var13	19	23	26	24	10
Var14	8	30	32	24	6
Var15	7	23	24	32	14

Table 2: Ron the scale of five points

4.2.2. Cluster Analysis

The output is attained by primarily applying a hierarchical cluster analysis to find out the total clusters existing in the data. The outputs figs 1, and 2 (agglomeration schedule, vertical icicle plot, and dendogram using average linkage, respectively).

The next step is a k-means output with a pre decided number of clusters to be identified. In this case, the output is for 4 clusters. We will see at both stage 1 and stage 2 outputs to realize the elucidation of both stages.

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	50	51	.000	0	0	63
2	66	67	2.236	0	0	64
3	24	99	2.449	0	0	25
4	85	86	2.449	0	0	72
5	8	49	2.449	0	0	23
6	42	96	2.646	0	0	20
7	53	55	2.646	0	0	47
8	7	45	2.646	0	0	24
9	5	6	2.646	0	0	42
10	30	89	2.828	0	0	26
11	14	80	2.828	0	0	34
12	61	62	2.828	0	0	29
13	37	58	2.828	0	0	21
14	3	39	2.828	0	0	39
15	48	94	3.000	0	0	55
16	71	92	3.000	0	0	33
17	23	70	3.000	0	0	23
18	41	59	3.000	0	0	28
19	2	13	3.000	0	0	44
20	15	42	3.232	0	6	25
21	37	64	3.317	13	0	50
22	12	36	3.317	0	0	45
23	8	23	3.377	5	17	36
24	7	25	3.390	8	0	46
25	15	24	3.442	20	3	32
26	30	65	3.452	10	0	50
27	38	69	3.464	0	0	37
28	41	52	3.500	18	0	36

29	28	61	3.595	0	12	48
30	47	95	3.606	0	0	41
31	46	79	3.606	0	0	65
32	10	15	3.660	0	25	40
33	71	93	3.669	16	0	49
34	14	82	3.702	11	0	68
35	4	43	3.742	0	0	60
36	8	41	3.780	23	28	37
37	8	38	3.813	36	27	53
38	22	87	3.873	0	0	57
39	3	18	3.873	14	0	71
40	10	100	3.882	32	0	46
41	47	90	3.924	30	0	47
42	5	57	3.936	9	0	54
43	1	60	4.000	0	0	66
44	2	74	4.004	19	0	61
45	12	26	4.039	22	0	53
46	7	10	4.048	24	40	52
47	47	53	4.119	41	7	49
48	27	28	4.126	0	29	70
49	47	71	4.168	47	33	59
50	30	37	4.219	26	21	72
51	97	98	4.243	0	0	91
52	7	9	4.292	46	0	59
53	8	12	4.304	37	45	58
54	5	56	4.346	42	0	69
55	17	48	4.398	0	15	68
56	20	83	4.472	0	0	79
57	22	68	4.500	38	0	76
58	8	40	4.506	53	0	60
59	7	47	4.526	52	49	65
60	4	8	4.553	35	58	77
61	2	16	4.571	44	0	83
62	34	73	4.583	0	0	84
63	50	78	4.690	1	0	90
64	66	72	4.743	2	0	76
65	7	46	4.790	59	31	69
66	1	88	4.796	43	0	85
67	32	44	4.796	0	0	73
68	14	17	4.875	34	55	78
69	5	7	4.884	54	65	75
70	27	76	4.909	48	0	82
71	3	54	4.910	39	0	74
72	30	85	4.918	50	4	75
73	32	63	4.949	67	0	77
74	3	21	5.066	71	0	93
75	5	30	5.116	69	72	78
76	22	66	5.248	57	64	86
77	4	32	5.317	60	73	81
78	5	14	5.363	75	68	80
79	20	77	5.465	56	0	87
80	5	11	5.551	78	0	82
81	4	33	5.604	77	0	84
82	5	27	5.665	80	70	83
83	2	5	5.721	61	82	88
84	4	34	5.728	81	62	98
85	1	91	5.728	66	0	86
86	1	22	5.830	85	76	89
87	19	20	5.887	0	79	90
88	2	81	6.116	83	0	89
89	1	2	6.244	86	88	92
90	19	50	6.258	87	63	92
91	75	97	6.293	0	51	96
92	1	19	6.387	89	90	93
93	1	3	6.538	92	74	94
94	1	84	6.749	93	0	95
95	1	35	6.885	94	0	96
96	1	75	7.081	95	91	97
97	1	29	7.244	96	0	98
98	1	4	7.602	97	84	99
99	1	31	8.073	98	0	0

Table 3: (Agglomeration Schedule)

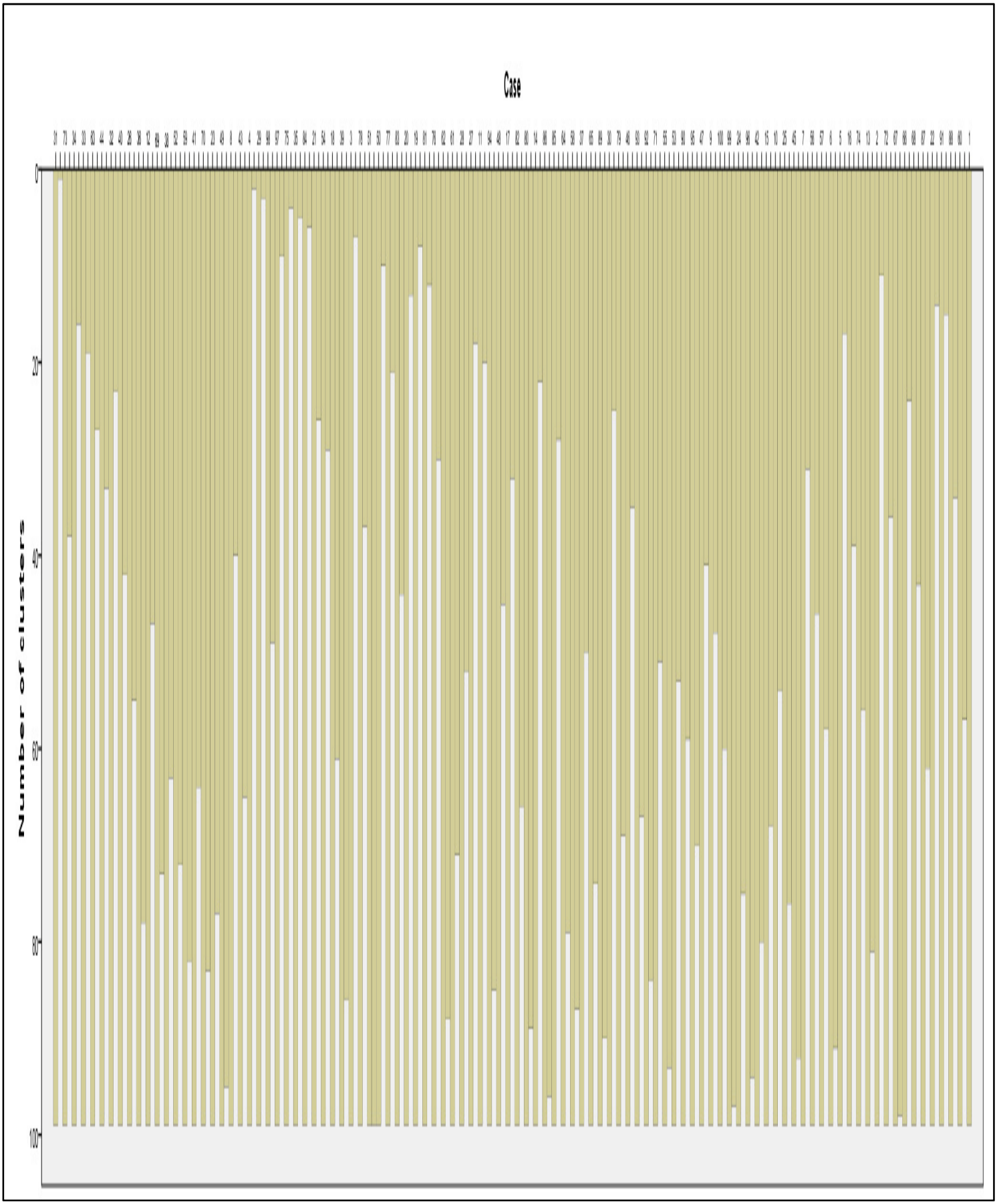

Figure 1: (vertical icicle Plot)

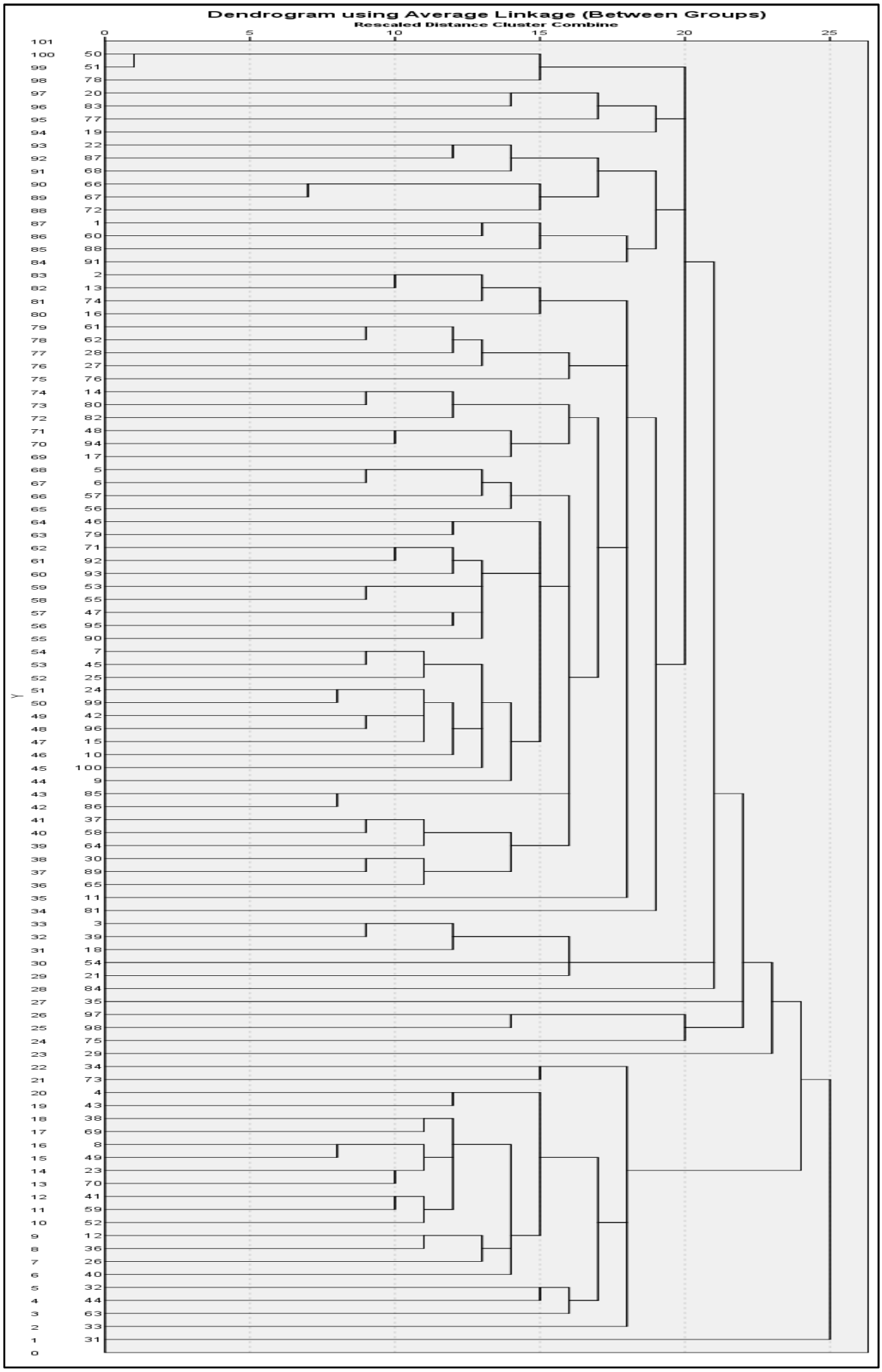

Figure 2: (Dendrogram)

4.3. Stage-1

See Table1, the agglomeration schedule, enables us to find out the large differences in the coefficient (4th column). The agglomeration schedule from upward to downward (stage 1 to 99) indicates the sequence in which cases get combined with others (or one cluster combines with another), until all 100 cases are combined together in one cluster at the last stage (stage 99). Therefore, stage 99 represents a 1-cluster solution, stage 98 represents a 2-cluster solution, and stage 97 represents a 3-cluster solution, and so on, going up from the final row to first row. We have to unearth that how many clusters are existing there in the data. We use the difference between rows in a measure called coefficient in column 4 to find out the total clusters in the data. Large difference of (8.077-7.602) can be seen in the coefficients between 1- cluster solution (stage 99) and the 2-cluster solution (stage 98). This is a difference of 0.475. Next it can be seen that the difference is of (7.602.79-7.244) which is equal to 0.358 (between stage 98, the 2-cluster solution and stage 97, the 3-cluster solution). The next one after that is (7.244-7.081), only 0.163, between stage 97 and stage 96. After this, we see that again there is a large difference between the 4- cluster and 5-cluster solutions, of (7.081-6.885) or 0.196.After that, the differences are smaller between consecutive rows of coefficients. A huge difference in the coefficient values between any two rows denotes a solution which is similar to the total number of clusters which the lower row represents.

Ignoring the first difference of 0.475 which would show only 1 cluster in the data, we see the next largest differences. 0.358 is the difference between row 2 from the bottom and row 3 from the bottom, denoting a 2-cluster solution. But almost the same is the difference between stage 96 and stage 95, indicating a 4-cluster solution. At this point of time, it is wisdom of the researcher, whether to go for a 2-cluster or a 4-cluster solution. Let us go with the 4-cluster solution in our case.

We maysee the icicle plot (Fig. 1, for the graphically aligned reader) or we can look at the dendrogram (Fig. 2) for data as to which cases link up in what concatenation to form clusters. The numbers in column 2 and 3 of the agglomeration schedule also provide alike data. The case integration of each cluster is considerably clearer in the dendrogram. For example, for a 4-cluster solution, from the dendrogram, cluster 1 would consist of cases no. 4, 5, 3, 19, 20, 11, and 10. Cluster 2 would consist of cases no. 1, 2, 14, 15, and 9. Cluster 3 would consist of cases no. 13, 16, 7, and 17. Cluster 4 would consist of cases no. 6, 18, 8, and 12. Alternatively, it can be done by determining on the cut-off distance, on the scale at the top of the dendrogram and then viewing at the individualistic clusters at that distance. In this situation, a cutoff of 99 would capitulate the 4 clusters.

4.4. Stage 2:

Now we will do K-clustering because a K-means method normally provide seven more sturdy clusters. However, it requires a particular figure of the beginning points, to obtain an embryonic position. Thus, it is best used in combinations with stage 1.

In our case, Tables 1, 2, 3denoting the outputs of K-means clustering for a 4- cluster solution. These outputs provide us the embryonic cluster centres, the case listing of cluster integration (i.e., which case related to which of the respondent) and the solution.

The final cluster centres defines the average worth of all variable for all the 4 clusters. For example, cluster 1 is defined by the mean values of variable 1= 2.2, variable 2=2.2, variable 3=3.8, variable 4=3.2, etc. Likewise, cluster 3 is defined by variable 1=1.75, variable 2=2.0, variable 3=2.25, and variable 4=3.0 etc.

Now let's come again to the aboriginal variables (in this case the 15 statements in our questionnaire), and elucidate the clusters in terms of the 15 variables. E.g. cluster 3 consists of people who are inclined towards email than writing traditional letters (variable 1 value= 1.75 which is analogous to 'agree' on the scale of 1 to 5). Likewise, they are also people who think more than once before purchasing anything (variable 3 value 2.25)or we can say these people are careful shoppers. They agree (variable 2 value= 2.00) that high quality products are always priced high.

Taking the exact variables, 2nd cluster exhibits people prefer favouring mail over email, people who do not link excessive price with thesuperior quality, and tend to be impartial about care in shopping. In this manner, if we compare final cluster centre values on every variable, for single cluster at a time, a full image of the clusters appears.

Below, all 4 clusters are briefly elucidated:

4.5. Cluster-1

The persons which are in this cluster are email users, they think quality goods are always highly priced. They don't shop cautiously and they don't consider that car is essential. They are not sure that people are more health-conscious now or not, they are not agreed with the statement that women are active buying decision makers and they think efficiency of Indian companies have been somewhat boosted by foreign corporations. Persons in this cluster do not like T.V., or fast food etc. Moreover, they think that politicians can play vital role. They not at all enjoy movies and also going out on weekends, they are more inclined towards buying branded goods and they like to pay in cash. They may think to settle abroad.

Therefore, it is a cluster showing traditional values, except adopting the email use. They started to loosen their purse strings, and possibly they are in switching in some other elements like acceptance of females as decision makers and the emergence of credit cards.

4.6. Cluster-2

People in this group are regular e-mail writers, they are aggressive shoppers. They don't think much before spending money, don't give much value to TV and think car is a luxury. They are not much affectionate of fast food and convenience goods, these people do not consider that people are much conscious towards their health, they feel foreign companies are good for us and believe that females are involved actively in buying decision making. They do not think politicians can play active roles, don't like watching movies, they do not emphasize on branded products, do not go out on weekends, but prefer to pay by credit card and they think that quality goods can be priced low. Also, they do not want to settle outside their country.

4.7. Cluster-3

This group belongs to email users, and thy think quality is always estimated by the price. They think more before shopping, they are dispassionate about TV and for them car is a luxury. They don't like junk food much, agreed to the fact that people are health-conscious, and are not agree that foreign corporations have increased our efficiency, they don't have faith in female potential, don't like politicians and they always like seeing movies. They wanted settling abroad, always purchase branded goods, and go out on weekends and vaguely favour cash to credit cards.

4.8. Cluster-4

This group is not much particular to email, compute quality by the price, and spend much. They enjoy watching TV, they realize the necessity of a car, are not much aligned towards junk food, agreed to the fact that people are health-conscious, and are not agree that foreign corporations have increased our efficiency, they don't have faith in female potential, little bit they think politicians can play active roles, don't watch movies at all, don't want to settle outside their country, unconcerned to brands, somewhat outgoing and somewhat in lean towards credit cards than cash. This group may desire worth for money, but if they find worth, they might spend much.

In short we can say the cluster analysis of this sample briefs us about the possible segments which prevail in targeted population.

5. Conclusions

In short cluster analysis of this sample briefs us about the possible segments which prevail in targeted population. As the total clusters known, *k*-means clustering method was applied. For determining *k*-means clustering, the primary cluster centers were taken and then last stable cluster centers were determined by continuing number of iterations until means had stopped changing with next iterations. This convergent situation was also attained by fixing a starting value for change in mean. The last cluster centers hold the mean values for all variable in each group. These studies in the last given us idea of using this method for market segmentation for approximations. A brilliant computing based set up was instituted and it presented outcomes automatically to the mangers to understand for fast decision making procedure. In coming time there will be more work which will include even more trials and automation of the market forecasting and planning.

In our paper we found there are 4 clusters of customers (Students of Aligarh Muslim University Centre, Malappuram) on the basis of their alike attributes which we have already discussed. They were divided into these groups so that an appropriate segmentation can be done.

6. References

i. I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.

ii. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.

iii. MacKay and David, "An Example Inference Task: Clustering," Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.

iv. M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in Proc.10 th ACM Symposium on Computational Geometry, 1994, pp. 332-339.

v. D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hard Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

vi. S. Dasgupta and Y. Freund, "Random Trees for Vector Quantization," IEEE Trans.on Information Theory, vol. 55, pp. 3229-3242, 2009.

vii. M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The Planar K-Means Problem is NP-Hard," LNCS, Springer, vol. 5431, pp. 274-285, 2009.

viii. A. Vattani, "K-means exponential iterations even in the plane," Discrete and Computational Geometry, vol. 45, no. 4, pp. 596-616, 2011.

ix. C. Elkan, "Using the triangle inequality to accelerate K-means," in Proc. the 12th International Conference on Machine Learning (ICML), 2003.

x. H. Zha, C. Ding, M. Gu, X. He, and H. D. Simon, "Spectral Relaxation for K-means Clustering," Neural Information Processing Systems, Vancouver, Canada, vol.14, pp. 1057-1064, 2001.

xi. C. Ding and X.-F. He, "K-means Clustering via Principal Component Analysis," in Proc. Int'l Conf. Machine Learning (ICML), 2004, pp. 225-232.

xii. "Chapter 6. Introduction to Clustering Procedures" in SAS Institute Inc., SAS/STAT®User's Guide, Version 6, Fourth Edition, Volume 1, Cary, NC: SAS Institute Inc., 1989. pp.53102.

xiii. Market Segmentation Using K-Means Cluster Analysis Harry B. Rowe March 26, 2012

xiv. Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.

xv. Manly, B.F.J. (2005), Multivariate Statistical Methods: A primer, Third edition, Chapman and Hall.

xvi. Rencher, A.C. (2002), Methods of Multivariate Analysis, Second edition, Wiley.

APPENDIX

Questionnaire

Section-1: General Information

Gender_____ (M/F),Age_____ (yrs.)
Current Education_____Occupation_____

Section-2: Survey Questionnaire

SN.	STATEMENTS	1	2	3	4	5
1.	I prefer to use email rather than writing a letter					
2.	I feel that quality products are always priced high					
3.	I think twice before I buy anything					
4.	Television is a major source of entertainment					
5.	A car is a necessary rather thana luxury					
6.	I prefer fast food and ready-to-use products					
7.	People are more health-conscious today					
8.	Entry of foreign companies has increased the efficiency of Indian companies					
9.	Women are active participants in purchase decisions					
10.	I believe politicians can play a positive role					
11.	I enjoy watching movies					
12.	If I get a chance, I would like to settle abroad					
13.	I always buy branded products					
14.	I frequently go out on weekends					
15.	I prefer to pay by credit card rather than in cash					

1= Strongly Agree, 2= Agree, 3= Neither Agree nor Disagree, 4= Disagree, and 5= Strongly Disagree