

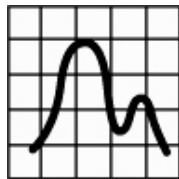


StatSoft®

data analysis • data mining • quality control • web-based analytics

Clustering Techniques and *STATISTICA*

Case Study: Defining Clusters of Shopping Center Patrons



STATISTICA
**Solutions for Business Intelligence,
Data Mining, Quality Control, and
Web-based Analytics**

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o
China: StatSoft China

France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.

Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS

Poland: StatSoft Polska Sp. z.o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda

S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Table of Contents

| | |
|---|-----------|
| CLUSTERING TECHNIQUES AND STATISTICA | 1 |
| Summary of Clustering Algorithms..... | 1 |
| Joining (Tree Clustering) | 2 |
| <i>k</i> -Means Clustering | 3 |
| EM (Expectation Maximization) Clustering..... | 3 |
| CASE STUDY: DEFINING CLUSTERS OF SHOPPING CENTER PATRONS | 5 |
| DATA ANALYSIS WITH STATISTICA..... | 7 |
| <i>k</i> -Means Cluster Analysis Results | 7 |
| EM Cluster Analysis Results | 8 |
| Feature Selection and Variable Screening Results | 9 |
| CONCLUSION..... | 10 |

Clustering Techniques and *STATISTICA*

The term *cluster analysis* (first used by Tryon, 1939) actually encompasses a number of different classification algorithms. A general question facing researchers in many areas of inquiry is how to *organize* observed data into meaningful structures, that is, to develop taxonomies. Clustering techniques have been applied to a wide variety of research problems. Hartigan (1975) provides an excellent summary of the many published studies reporting the results of cluster analyses. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc., is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc., by applying cluster analytic techniques. In general, whenever there is a need to classify a “mountain” of information into manageable meaningful piles, cluster analyses are of great utility.

Clustering algorithms work well with all kinds of data including categorical, numerical, and textual data. Usually the only decision the user must make is to ask for a specific number of candidate clusters. The advanced features in *STATISTICA* will help the user even with this aspect of the analyses (i.e., to determine the right number of clusters). The clustering algorithm will find the best partitioning of all the customer records (in our example) and will provide descriptions of the “means or centroids” (“average responses”) of each cluster in terms of the user’s input data. In many cases, these clusters have an obvious interpretation that provides insight into the “natural” segmentation of customers that visit the mall.

Clustering techniques generally belong to the group of undirected data mining tools; these techniques are also sometimes referred to as “unsupervised learning” because there is no particular dependent or outcome variable (such as *Amount Purchased*) to predict (and that would be used to “supervise” the learning process, i.e., to lead to the best predictive model). Instead, the goal of undirected data mining or unsupervised learning is to discover structure in the data as a whole. There is no target variable to be predicted, so no distinction is being made between independent and dependent variables.

Summary of Clustering Algorithms

Clustering techniques are used for combining observed cases or observations into clusters (groups) that satisfy two main criteria:

1. Each group or cluster is homogeneous; observations (cases) that belong to the same group are similar to each other.
2. Each group or cluster should be different from other clusters, that is, observations (cases) that belong to one cluster should be different from those contained in (grouped into) other clusters.

Let's start by understanding the different methods for clustering. Clustering methods have been widely distinguished into three categories.

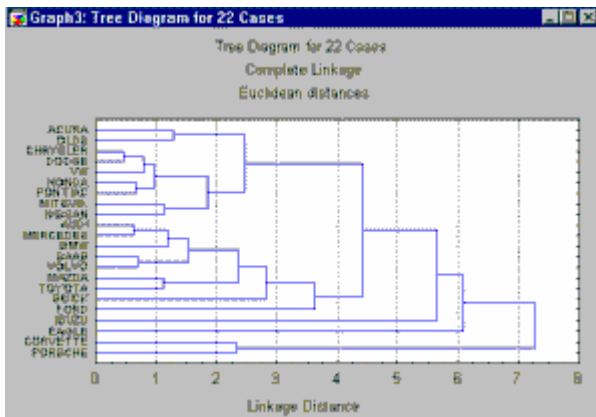
1. Joining (Tree Clustering)
2. k -Means Clustering
3. EM (Expectation Maximization) Clustering

Following are brief explanations of these methods.

Joining (Tree Clustering)

The purpose of this algorithm is to join together objects (e.g., animals) into successively larger clusters, using some measure of similarity or distance. A typical result of this type of clustering is the hierarchical tree.

Consider a *Horizontal Hierarchical Tree Plot* (see graph below). On the left of the plot, we begin with each object in a class by itself. Now imagine that, in very small steps, we “relax” our criterion as to what is and is not unique. Put another way, we lower our threshold regarding the decision when to declare two or more objects to be members of the same cluster.



As a result, we *link* more and more objects together and aggregate larger and larger clusters of increasingly dissimilar elements. Finally, in the last step, all objects are joined together. In these plots, the horizontal axis denotes the linkage distance (in *Vertical Icicle Plots*, the vertical axis denotes the linkage distance). Thus, for each node in the graph (where a new cluster is formed) we can read off the criterion distance at which the respective elements were linked together into a new single cluster. When the data contain a clear structure in terms of clusters of objects that are similar to each other, then this structure will often be reflected in the hierarchical tree as distinct branches. As the result of successful analyses with the joining method, one is able to detect clusters (branches) and interpret those branches.

***k*-Means Clustering**

k-Means Clustering is very different from Joining (Tree Clustering) and is widely used in real-world scenarios. Suppose that you already have hypotheses concerning the number of clusters in your cases or variables. You may want to “tell” the computer to form, say, exactly 3 clusters that are to be as distinct as possible. This is the type of research question that can be addressed by *k*-Means Clustering. In general, the *k*-Means method will produce exactly *k* different clusters of greatest possible distinction.

Suppose a medical researcher has a “hunch” from clinical experience that her heart patients fall into three different categories with regard to physical fitness. She might wonder whether this intuition could be quantified, that is, whether a *k*-Means Cluster Analysis of the physical fitness measures would indeed produce the three clusters of patients as expected. If so, the means on the different measures of physical fitness for each cluster would represent a quantitative way of expressing the researcher’s hypothesis or intuition (i.e., patients in cluster 1 are high on measure 1, low on measure 2, etc.).

Computationally, you may think of this *k*-Means method as analyses of variance (ANOVA) “in reverse.” The program will start with *k* random clusters, and then move objects between those clusters with the goal to 1) minimize variability within clusters and (2) maximize variability between clusters. This is analogous to “ANOVA in reverse” in the sense that the significance test in ANOVA evaluates the between-group variability against the within-group variability when computing the significance test for the hypothesis that the means in the groups are different from each other. In *k*-Means Clustering, the program tries to move objects (e.g., cases) in and out of groups (clusters) to get the most significant ANOVA results.

Defining “distances” between and within clusters, and “correct” clustering. One important rule to note is that the classification produced with all clustering tools (*k*-Means as well as joining methods) is very dependent upon the particular metric that is used to determine the “distance” between objects (e.g., observations) and clusters. Remember that all clustering methods attempt to arrange observations so that those that are similar in some respect are assigned to the same cluster, and those that are dissimilar to different clusters. Because there are many ways in which “distance” can be defined (e.g., as Euclidean distance, squared Euclidean distance, percent disagreement, and so on), there is no such thing as a single correct classification, although there have been attempts to define concepts such as ‘optimal’ classifications.

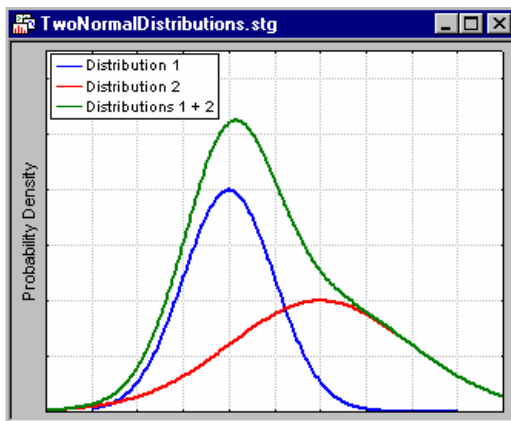
EM (Expectation Maximization) Clustering

The EM (Expectation Maximization) Clustering technique is another tool offered in *STATISTICA*. The general purpose of this technique is also to detect clusters in observations (or variables) and to assign those observations to the clusters. The EM (expectation maximization) algorithm extends the *k*-Means Clustering approach to clustering in two important ways:

1. Instead of assigning cases or observations to clusters to maximize the differences in means for continuous variables (or disagreement for categorical variables), the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters.

2. Unlike the classic implementation of k -Means Clustering, the general EM algorithm can be applied to both continuous and categorical variables (although in *STATISTICA* the classic k -Means algorithm was modified to accommodate categorical variables as well, by defining appropriate measures of “distance”).

The EM algorithm for clustering is described in detail in Witten and Frank (2001). The basic approach and logic of this clustering method is as follows. Suppose you measure a single continuous variable in a large sample of observations. Further, suppose that the sample consists of two clusters of observations with different means (and perhaps different standard deviations); within each sample, the distribution of values for the continuous variable follows the normal distribution. The resulting distribution of values (in the population) may look like this:



Mixtures of distributions. The illustration shows two normal distributions with different means and different standard deviations, and the sum of the two distributions. Only the mixture (sum) of the two normal distributions (with different means and standard deviations) would be observed. The goal of EM clustering is to estimate the means and standard deviations for each cluster to maximize the likelihood of the observed data (distribution). Put another way, the EM algorithm attempts to approximate the observed distributions of values based on mixtures of different distributions in different clusters.

Categorical variables. The EM algorithm can also accommodate categorical variables. The program will first randomly assign different probabilities (weights, to be precise) to each class or category, for each cluster. In successive iterations, these probabilities are refined (adjusted) to maximize the likelihood of the data given the specified number of clusters.

Classification probabilities instead of classifications. The results of EM clustering are different from those computed by k -Means Clustering. The latter will assign observations to clusters to maximize the distances between clusters. The EM algorithm does not compute actual assignments of observations to clusters, but classification *probabilities*. In other words, each observation belongs to each cluster with a certain probability. Of course, as a result you can usually review an actual assignment of observations to clusters, based on the (largest) classification probability.

Case Study: Defining Clusters of Shopping Center Patrons

This case study is based on a sub-sample from a survey containing 502 questions completed by patrons of a shopping mall in the San Francisco Bay area (see Hastie, Tibshirani, Friedman, 2001)). The general purpose of this study is to identify homogeneous “typical” groups of visitors to the shopping center. Once such “prototypical” groups of customers have been identified, special marketing campaigns, services, or recommendations for particular types of stores can be developed for each group in order to enhance the overall attractiveness and quality of the shopping experience.

This clustering example analyzes a large number of undifferentiated customers to see if they fall into natural groupings. This is a pure example of “undirected data mining” or “unsupervised learning” where the analyst has no clear prior knowledge of the “types” of customers that visit the shopping mall and is hoping that the data-mining tool will reveal some meaningful structure by identifying clusters or groups in these relatively homogeneous shoppers.

Specifically, the file used in this example contains 8995 observations with 14 variables regarding customer information. The information contained in this data set is summarized below:

1. ANNUAL INCOME OF HOUSEHOLD (PERSONAL INCOME IF SINGLE)

| | | |
|-------------------------|-------------------------|-------------------------|
| 1 -Less than \$10,000 | 4 -\$20,000 to \$24,999 | 7 -\$40,000 to \$49,999 |
| 2 -\$10,000 to \$14,999 | 5 -\$25,000 to \$29,999 | 8 -\$50,000 to \$74,999 |
| 3 -\$15,000 to \$19,999 | 6 -\$30,000 to \$39,999 | 9 -\$75,000 or more |

2. SEX

| | |
|---------|-----------|
| 1. Male | 2. Female |
|---------|-----------|

3. MARITAL STATUS

| | | |
|--------------------------------|--------------------------|--------------------------|
| 1. Married | 3. Divorced or separated | 5. Single, never married |
| 2. Living together not married | 4. Widowed | |

4. AGE

| | |
|---------------|----------------|
| 1. 14 thru 17 | 5. 45 thru 54 |
| 2. 18 thru 24 | 6. 55 thru 64 |
| 3. 25 thru 34 | 7. 65 and Over |
| 4. 35 thru 44 | |

5. EDUCATION

| | | |
|--------------------|----------------------------|---------------------|
| 1. Grade 8 or less | 3. Graduated high school | 5. College graduate |
| 2. Grades 9 to 11 | 4. 1 to 3 years of college | 6. Grad Study |

6. OCCUPATION

| | | |
|----------------------------------|----------------------------|---------------|
| 1. Professional/Managerial | 4. Clerical/Service Worker | 7. Military |
| 2. Sales Worker | 5. Homemaker | 8. Retired |
| 3. Factory Worker/Laborer/Driver | 6. Student, HS or College | 9. Unemployed |

7. LIVED HOW LONG IN THE SAN FRANCISCO /OAKLAND/SAN JOSE AREA

| | | |
|-----------------------|-----------------------|------------------------|
| 1. Less than one year | 3. Four to six years | 5. More than ten years |
| 2. One to three years | 4. Seven to ten years | |

8. DUAL INCOMES (IF MARRIED)

| | |
|----------------|-------|
| 1. Not Married | 3. No |
| 2. Yes | |

9. PERSONS IN YOUR HOUSEHOLD

| | | |
|----------|---------|-----------------|
| 1. One | 4. Four | 7. Seven |
| 2. Two | 5. Five | 8. Eight |
| 3. Three | 6. Six | 9. Nine or more |

10. PERSONS IN HOUSEHOLD UNDER 18

| | | |
|----------|---------|-----------------|
| 1. One | 4. Four | 7. Seven |
| 2. Two | 5. Five | 8. Eight |
| 3. Three | 6. Six | 9. Nine or more |

11. HOUSEHOLDER STATUS

| | |
|---------|-----------------------------|
| 1. Own | 3. Live with Parents/Family |
| 2. Rent | |

12. TYPE OF HOME

| | | |
|----------------|----------------|----------|
| 1. House | 3. Apartment | 5. Other |
| 2. Condominium | 4. Mobile Home | |

13. ETHNIC CLASSIFICATION

| | | |
|--------------------|---------------------|----------|
| 1. American Indian | 4. East Indian | 7. White |
| 2. Asian | 5. Hispanic | 8. Other |
| 3. Black | 6. Pacific Islander | |

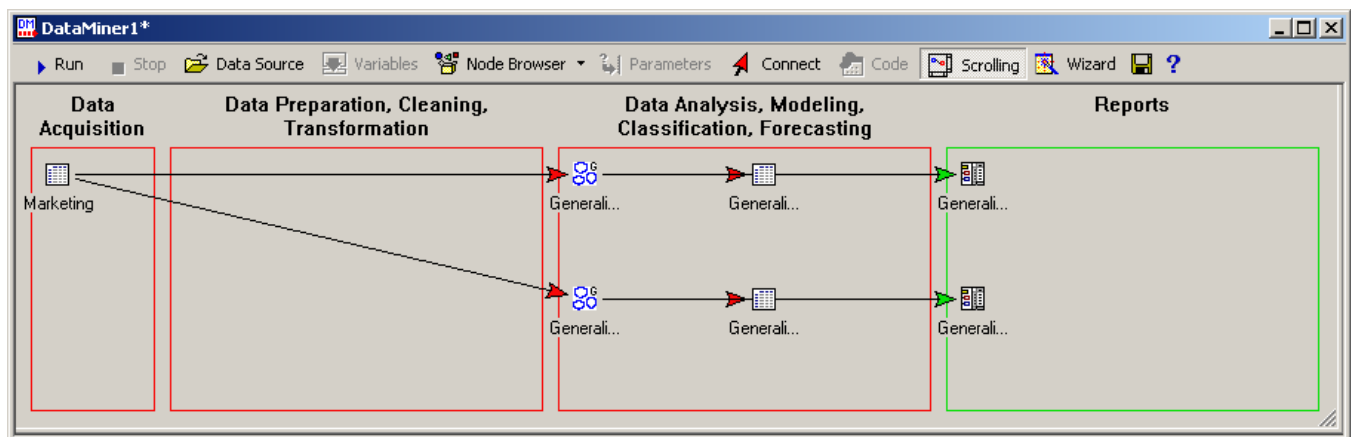
14 LANGUAGES SPOKEN MOST OFTEN IN YOUR HOME

| | | |
|------------|------------|----------|
| 1. English | 2. Spanish | 3. Other |
|------------|------------|----------|

Data Analysis with *STATISTICA*

Data preparation is a crucial step in any data mining project. In this case, however, the data preparation is complete. Data have been verified to be accurate and reasonable; there are no missing data, etc.

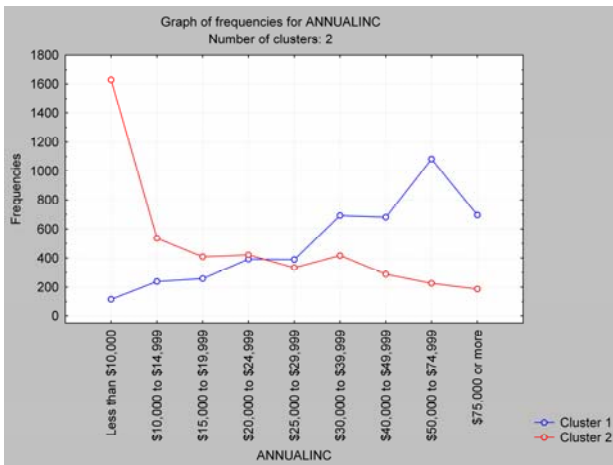
The following illustration shows the *STATISTICA Data Miner* workspace, where the Generalized *k*-Means Clustering tool and the Generalized EM Clustering tool are used to cluster data in the *Marketing* data set. Once the clusters are determined, the Feature Selection tool was used to determine the variables that most strongly influence cluster membership.

***k*-Means Cluster Analysis Results**

In the results below, the *Final Classification* and *Distance to Centroid* are given as output for the *k*-Means Cluster Analysis. This output can be used to compare cases in the various predicted clusters and their strength of cluster membership.

| Data: Generalized K-Means cluster analysis (16v by 8993c) | | | | | | |
|---|--------------------------|----------------|-------------------|-------------------|----------------------------|----------------------------|
| Marketing | | | | | | |
| | 11 HSEHLSTATUS | 12 HOMETYPE | 13 ETHNICCLASS | 14 LANGUAGEHME | 15 Final classification | 16 Distance to centroid |
| 1 | Own | House | White | 101 | 1 | 3.16227766 |
| 2 | Own | House | White | English | 1 | 3.46410162 |
| 3 | Rent | Apartment | White | English | 1 | 3.46410162 |
| 4 | Live with Parents/Family | House | White | English | 1 | 3.60555128 |
| 5 | Live with Parents/Family | House | White | English | 1 | 3.60555128 |
| 6 | Own | House | White | English | 1 | 3.60555128 |
| 7 | Rent | Apartment | White | English | 2 | 3.60555128 |
| 8 | Rent | Apartment | White | English | 2 | 3.60555128 |
| 9 | Rent | Apartment | White | English | 1 | 3.60555128 |
| 10 | Rent | Apartment | White | English | 1 | 3.46410162 |

An important variable in determining clusters for the *k*-Means algorithm is Annual Income. In the plot below, the frequencies of each income category for the two clusters are available. Lower income individuals are more often classified in Cluster 2, and higher income individuals are classified in Cluster 1.

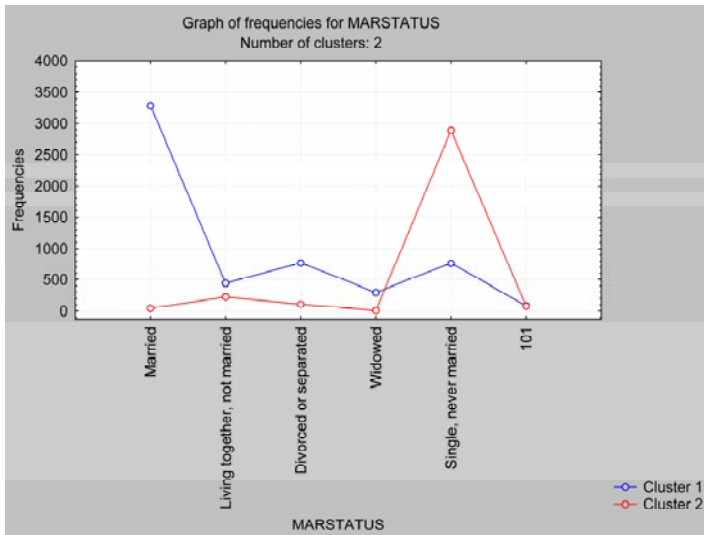


EM Cluster Analysis Results

The output for EM Cluster Analysis will be quite similar to that of *k*-Means Clustering. Along with the final cluster classification output, EM Cluster Analysis provides probabilities of cluster membership. In the output spreadsheet below, these probabilities are shown for the first several cases in the data set.

| Data: Classification probabilities (weights) for EM clustering (Marketing)* | | | |
|---|-----------|-----------|----------------------|
| Classification probabilities (weights) for EM clustering (Marketing) | | | |
| Number of clusters: 2 | | | |
| Total number of training cases: 8993 | | | |
| | Cluster 1 | Cluster 2 | Final classification |
| 1 | 1.000000 | 0.000000 | 1 |
| 2 | 1.000000 | 0.000000 | 1 |
| 3 | 0.999993 | 0.000007 | 1 |
| 4 | 0.000000 | 1.000000 | 2 |
| 5 | 0.000000 | 1.000000 | 2 |
| 6 | 1.000000 | 0.000000 | 1 |
| 7 | 0.000134 | 0.999866 | 2 |
| 8 | 0.953011 | 0.046989 | 1 |
| 9 | 0.999997 | 0.000003 | 1 |
| 10 | 1.000000 | 0.000000 | 1 |

Similar to the *k*-Means results, the frequencies of observations classified in the 2 clusters can be seen for Marital Status. Married people are classified in Cluster 1. Single, never married people are most likely classified in Cluster 2. From this plot, it is easy to tell that Marital Status is an important variable to determine cluster membership.

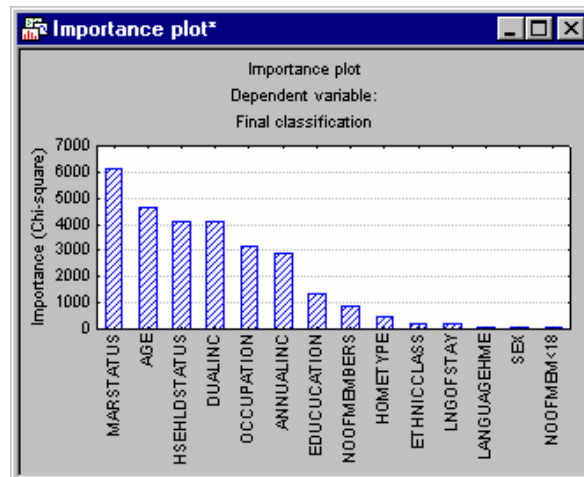


Feature Selection and Variable Screening Results

Now that we have detected cluster assignments in our data set, it's of great importance to find the predictor variables (typically with data set having numerous variables) that show the strongest relationship to the final cluster assignment variable, thereby identifying the ones that best discriminate between the clusters.

The resulting spreadsheet clearly arranges the “Best Categorical Predictor” variables from top to bottom based on the chi-square test in order of their relative importance. The histogram, showing the same information graphically, clearly indicates that MARITALSTAT was the most influential predictor in the clustering process followed by AGE, HSEHLDSTATUS etc.

| Variable | Chi-square | p-value |
|--------------|------------|----------|
| MARSTATUS | 6157.189 | 0.000000 |
| AGE | 4667.922 | 0.000000 |
| HSEHLDSTATUS | 4104.604 | 0.000000 |
| DUALINC | 4089.287 | 0.000000 |
| OCCUPATION | 3158.320 | 0.000000 |
| ANNUALINC | 2891.639 | 0.000000 |
| EDUCUCATION | 1357.628 | 0.000000 |
| NOOFMEMBERS | 845.233 | 0.000000 |
| HOMETYPE | 495.968 | 0.000000 |
| ETHNICCLASS | 222.661 | 0.000000 |
| LNGOFSTAY | 217.168 | 0.000000 |
| LANGUAGEHME | 82.725 | 0.000000 |
| SEX | 78.822 | 0.000000 |
| NOOFMEM<18 | 46.056 | 0.000000 |



Conclusion

Summarizing the analyses of the best predictors into a short table, we get the following results. Note that this table indicates the respective “dominant” class or category for each of the variables listed in the rows, i.e., the class or category that occurred with the greatest frequency within the respective cluster.

| | Cluster 1 | Cluster 2 |
|-------------------------|------------------------------|-------------------------|
| Marital Status | Married | Single, Never Married |
| Age | 25 through 34, 35 through 44 | 18 through 24 |
| Household Status | Own | Rent |
| Dual Income | Yes, Dual Income | Not Married |
| Occupation | Professional/Managerial | Students, HS or College |
| Annual Income | \$50,000 to \$74,999 | Less than 10,000 |

We can conclude from this table that Cluster 1 includes Married Individuals from Dual income households, between the Ages of 25 and 44, who Own houses, hold Professional/ Managerial positions, and have Annual Incomes between \$50,000 and \$75,000. Cluster 2 members are Single, between the Ages of 18-24, Rent houses, are in High School or College, and earn less than \$10,000. The above information clearly indicates that the clustering technique has helped us identify two meaningful distinct groups of shoppers from our marketing data set.

This information could be exploited to better serve the needs of the customers visiting the mall, which can improve sales, paving the way for higher profitability. For example, special marketing campaigns could be designed based on this better understanding of who visits the mall (e.g., special promotions for college students); this information could also be used to market store locations in the mall to prospective retailers who specifically cater to the groups that we identified. In general, the better we know and “understand” our customers, the better we are prepared to serve them and, hence, ensure a successful retail enterprise.