Paper 447-2013

Multivariate Statistical Analysis in SAS: Segmentation and Classification of Behavioral Data

Rachel Poulsen, TiVo, Alviso, CA

ABSTRACT

An idiom in the customer service industry is “the customer is always right”. However, in many instances the customer will not speak up and another popular idiom must be used “Actions speak louder than words”. Customer actions can be measured to infer what they will not say. Once measured, segmentation analysis can be used to make sense of the large amount of behavioral data by placing customers into various segments. Classification models are then used to assign new customers to a segment. Statistical algorithms used to segment and classify observations include Collaborative Filtering and Machine Learning Models. This paper will illustrate how SAS® can be used to segment and classify observations using the FASTCLUS and DISCRIM procedures.

INTRODUCTION

Clustering and Classification analysis are at the heart of Multivariate statistics. Cluster analysis places observations into groups according to the natural association between the observations. Classification analysis will both describe the separation between groups and assign new observations to a group based off a list of variables. This paper will discuss various ways in which Cluster and Classification analysis can be used to better understand customer behavior. Viewership data collected from TiVo will be used as an example to illustrate how to perform Cluster and Classification analysis using procedures in SAS ®. (This paper does not describe any process or procedures actually used by TiVo. TiVo data is merely used for explanatory purposes). Viewership information was taken from a sample of 6,500 TiVo digital video recorders (DVRs) and was quantitatively measured in 10 different aspects of viewership behavior. The CLUSTER and FASTCLUS procedures in SAS ® were first used to place the observations into different groupings based on the 10 measured aspects. Using the DISCRIM procedure, the quality of each grouping was evaluated to understand the separation between groups. The DISCRIM procedure was then used to classify a list of new observations into groups.

DATA AND METHODOLOGY

Although SAS® is capable of handling a large number of variables and observations, the data have been simplified into 10 variables and a sample of 6,500 observations in order to better visualize and illustrate the methods used in this paper. The viewership of each observation across 10 television networks was selected for this analysis. Though there are thousands of networks, the 10 networks were selected because of their availability to subscribers and/or their targeted programming (ex: children’s programs). The 10 networks considered can be seen in Table 1.

Variable (Network)	Description	Variable type
ABC	One of the major Broadcast Networks	Positive continuous value representing how much the observation enjoyed watching the network
NBC	One of the major Broadcast Networks	Positive continuous value representing how much the observation enjoyed watching the network
CBS	One of the major Broadcast Networks	Positive continuous value representing how much the observation enjoyed watching the network
FOX	One of the major Broadcast Networks	Positive continuous value representing how much the observation enjoyed watching the network
CW	One of the less major Broadcast Networks	Positive continuous value representing how much the observation enjoyed watching the network
USA	One of the major cable drama networks available to most cable subscribers	Positive continuous value representing how much the observation enjoyed watching the network
TNT	One of the major cable drama networks available to most cable subscribers	Positive continuous value representing how much the observation enjoyed watching the network
NIK	Nickelodeon - One of the major cable children’s networks available to most cable subscribers	Positive continuous value representing how much the observation enjoyed watching the network

Variable (Network)	Description	Variable type
TOON	Cartoon Network - One of the major cable children's networks available to most cable subscribers	Positive continuous value representing how much the observation enjoyed watching the network
FNC	Fox News Network - One of the major cable Political News networks available to most cable subscribers	Positive continuous value representing how much the observation enjoyed watching the network

Table 1. Network Variable Definitions

An implicit rating of each Network was measured using the viewership information. The rating is an arbitrary number between 0 and 10 and represents how much time the user views each network. Smaller values indicate the user 'did not view' and larger values indicate the user 'viewed a lot'.

HEIRARCHICAL CLUSTERING

A cluster analysis was first used to group the observations according to the 10 network viewership aspects. In machine learning models, cluster analysis is referred to as "unsupervised learning" as it assumes no formal group membership and identifies natural groupings in the observations. The two types of cluster analysis methods used in grouping the 6,500 observations were hierarchical clustering and K-means clustering. Hierarchical clustering methods begin with one large group and then successively break down groups into dissimilar subgroups until each observation is its own group. K-means cluster methods find the optimal clustering for a given number, K, of groups. In Multivariate statistics, hierarchical clustering is useful for data exploration and K-means clustering is useful for modeling. Hierarchical clustering was first performed on the network viewership data to explore the natural groupings within the data. After natural groupings were identified, a K-means cluster analysis was performed to assign the 6,500 observations to groups.

Like all statistical methods, hierarchical clustering methods make assumptions about the data. It is necessary to understand and check these assumptions prior to running the analysis. Multivariate statistical assumptions are a little different than their one-dimensional counterparts. One of the one-dimensional modeling assumptions that can be relaxed in a multivariate cluster analysis is the independence between variables. On the contrary, redundancy between variables can be beneficial as it can identify similarities among observations. While collinearity can be relaxed in a cluster analysis, standardization among variables is necessary. All variables must be measured using the same unit of measure. If this assumption is not met, results will be largely influenced by the variable measured using the largest unit of measure. If there is some question as to whether the variables are standardized, procedures in SAS® such as PROC STANDARD and PROC ACECLUS can be used to standardize multivariate data. In the network viewership example, all variables are measured with the same unit so the standardization assumption was met.

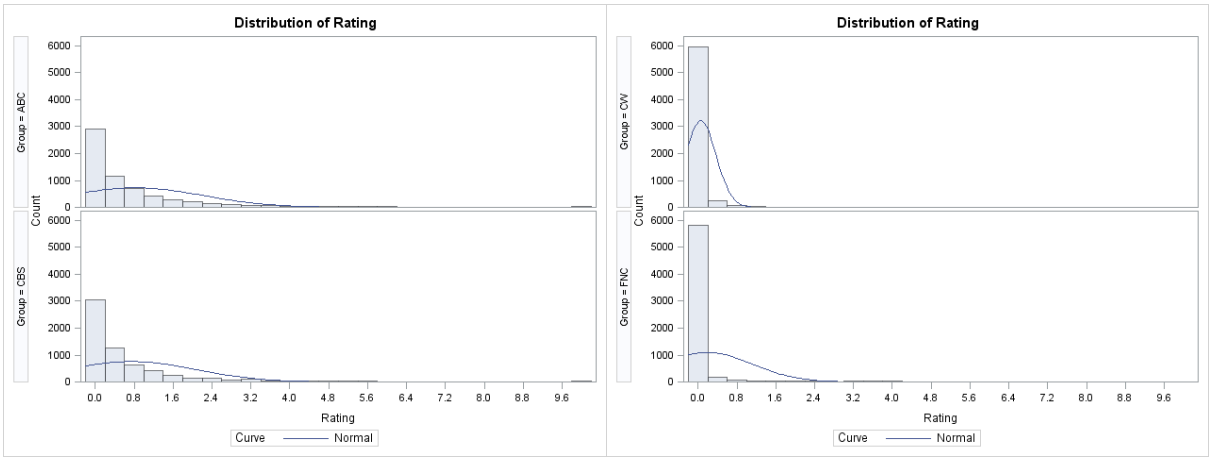
After assumptions have been checked, the method for creating clusters needs to be selected. There are various methods available for grouping observations. Understanding the methods available in SAS® is critical as the default method across statistical software may be different, yielding very different results. This is more true for multivariate statistical methods. The increase in dimensionality and complexity results in an increase in methodologies. Three of the most well-known methods for grouping observations in a cluster analysis are single linkage, complete linkage, and average linkage. The single linkage approach defines the similarity between two clusters by the similarity between the most similar items. The distance between two clusters is calculated as the minimum distance between any two observations in each cluster. At each step, the two clusters with the smallest minimum distance are merged into one cluster. This method has a tendency to form long chain-like clusters in the data. Intuitively, clusters should form compact, spherical shapes. The complete linkage approach defines the similarity between two clusters by the similarity between the most dissimilar observations. The distance between two clusters is calculated as the maximum distance between any two observations in each cluster. At each step, the maximum distance is found for every pair of clusters, and the two clusters with the smallest maximum distance are merged together. This method will result in spherical shaped clusters, but is heavily influenced by outliers. The average linkage approach combines the single linkage and complete linkage approaches, thereby compromising the sensitivity to outliers and the long chain-like clusters.

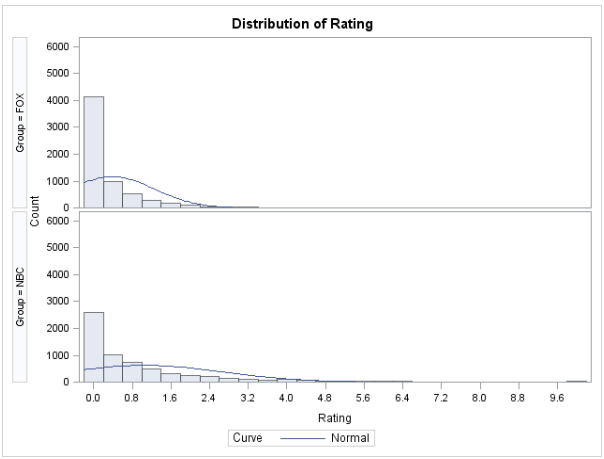
In addition to the three linkage approaches to grouping, Ward's method can be used to group observations by minimizing the variance in the distances between groups. Ward's method also creates small, compact clusters, but is also affected by outliers. In order to determine which grouping method to use in the network viewership example, some data exploration was initially completed. The distribution of each variable was considered and can be seen in Figure 1.

Visualizing the relationship between variables in multivariate statistics can be challenging because it requires viewing the data in more than 3 dimensions. One option for visualizing the relationship between all variables is to examine the

principal components. A principal component analysis will create uncorrelated linear combinations of the variables. The first two principal components can be thought of as the two dimensions among the variables that are the most un-related. Plotting the data against the first two principal components will give the most un-correlated view of the data, thereby allowing the separation between observations to be best seen in two dimensions. The relationships between the variables were visualized by plotting the data against the first two principal components and can be seen in Figure 2.

It can be seen in Figure 1 that all the variables in the network viewership data are zero-inflated, log-normal distributions with many outliers. These plots were created using PROC UNIVARIATE in SAS ®. This implies that complete linkage may not be the appropriate method for creating clusters.





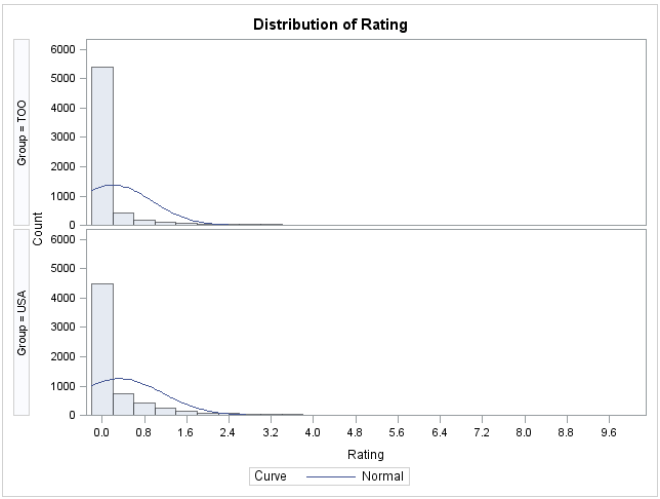


Figure 1. Variable Distribution Plots

Summary Plots

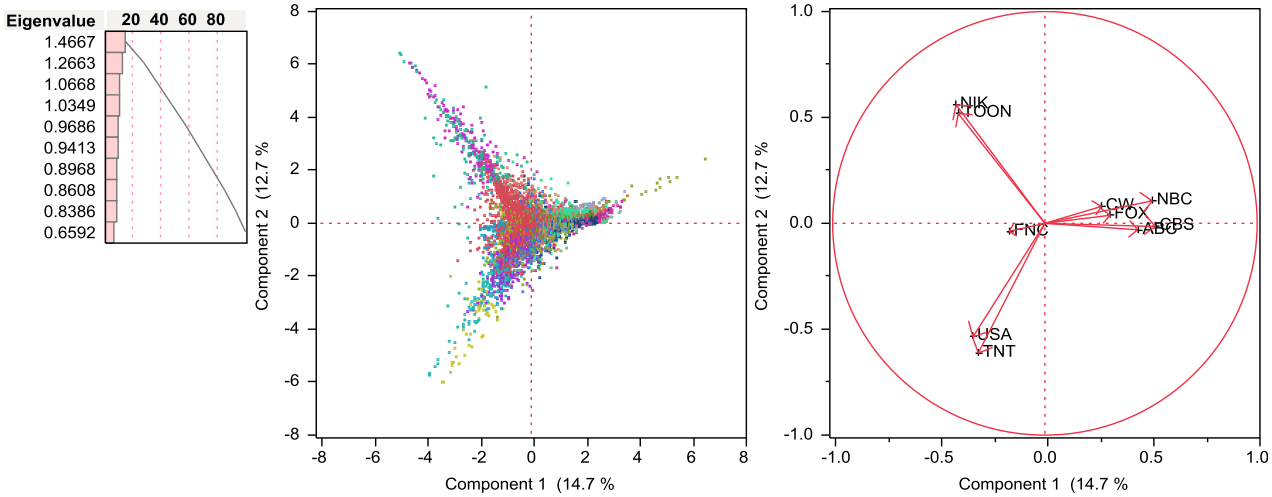


Figure 2. Principal Component Analysis in JMP

Three plots are illustrated in Figure 2. A horizontal Pareto chart of eigenvalues, a two-dimensional projection of the data plotted against the first two principal components, and a factor loading chart of the first two principal components. The eigenvalue bar chart shows the proportion of the total eigenvalue contribution for each eigenvalue. Six eigenvalues make up for about 80% of the total eigenvalue contribution. This number is also known as the "essential dimensionality" of the data. Though there are 10 variables in the network viewership example, only six dimensions are needed to visualize the relationship between the variables. The second chart shows the data plotted against the first two principal components. The data is centered in one place with three starbursts breaking off from the center. It can be seen in the third chart which variables contribute the most to each starburst. Interestingly the broadcast networks, children's networks, and cable drama networks contribute the most to the star-bursting behavior seen in the principal components plot. The political news network appears to slightly starburst off in its own direction. Adding another political news network may make this factor loading more prevalent. The logical and clear-cut groupings of these variables indicates that network viewership during the summer may be a good identifier of different viewer types.

In hierarchical cluster analysis, it is important to consider a few different linkage methods to identify which method appropriately clusters the data type. In this example, average linkage and Ward's method were considered. Complete linkage was not considered due to the many outliers in the data. Single linkage was initially considered but ruled out when the results gave very small R-squared values. The output for single linkage is not included.

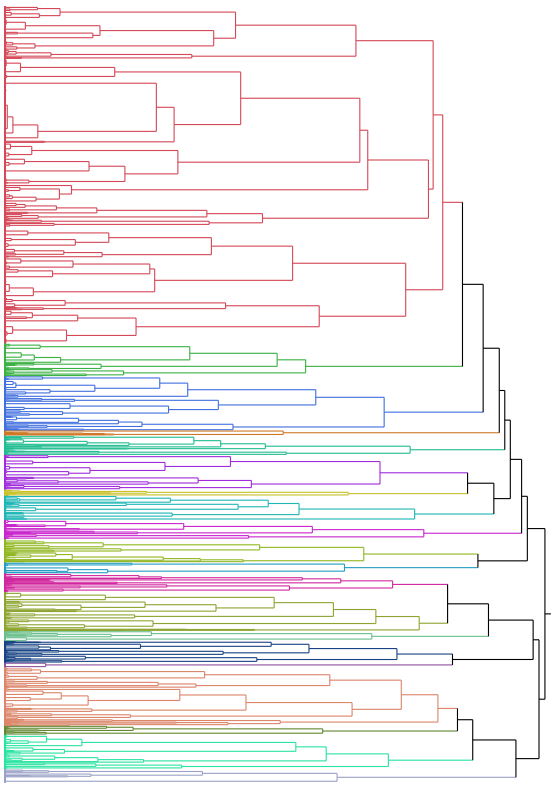


Figure 3. JMP Dendogram Using Ward's Method

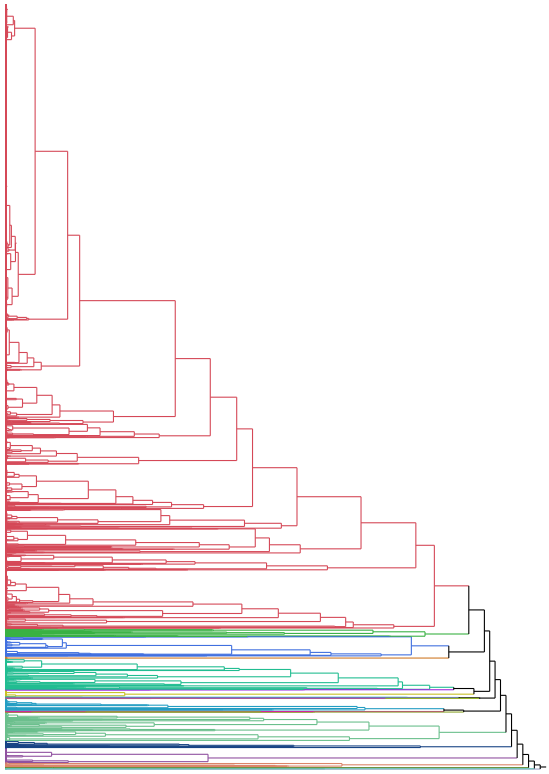


Figure 4. JMP Dendogram Using Average Linkage

Figures 3 and 4 are dendograms of the hierarchical cluster analysis. Figure 4 is a dendogram of the clusters using the average linkage method as the grouping criteria. In this case, it can be seen that average linkage creates one large cluster (colored in red) and a handful of very small clusters. Figure 3 is a dendogram of the clusters using Ward's method as the grouping criteria. In this case, it can be seen that Ward's method creates one larger cluster, and a handful of smaller clusters that appear to be more evenly distributed than the average linkage method. When classifying future observations into clusters it will be desirable to have larger clusters. The size of the clusters are similar to sample sizes in a linear regression model. If one cluster is very small, it will not be well represented, thereby resulting in a greater probability of misclassification during the classification analysis and potentially mis-repesenting the population. For this reason, it was decided to go with Ward's method as the grouping criteria.

SAS® provides multiple options for choosing an optimal number of groups. Recall that the goal of hierarchical clustering is to explore the data and understand how the data naturally forms into groups, providing potential clustering options. The final model will be selected using K-means clustering analysis. The Cubic Cluster Criterion (CCC) option is a metric developed by SAS®. It is a comparative measure of the deviation between the R^2 value if the data were obtained from a uniform distribution and the actual R^2 value for each number of clustering options. Large values greater than 1 indicate that the number of clusters is significant. If the data were obtained from a uniform distribution, all observations would belong to the same cluster and the R^2 value would be equal to the R^2 value of a uniform distribution. The pseudo option calculates a "Pseudo F-statistic" for each number of clustering options. It is intended to measure the 'compactness' of each cluster. Essentially, it is a measurement of the ratio of the between-group variation and the within-group variation. Larger values of the Pseudo F-statistic indicate better clustering.

Some other options to consider while using PROC CLUSTER are *trim*, *k=,* and *print*. The trim option is recommended when using grouping methods sensitive to outliers like the complete linkage method and Ward's method. The value given to the trim option is the percent of outlying observations desired for removal. The number chosen is arbitrary. However, large numbers should be avoided to maintain the integrity of the data. In this case, trim=10% was chosen. Trim cannot be specified without the 'k=' option as *k* represents the minimum number of observations needed to form a cluster. On large datasets, the print option will help suppress the output. Only output up to the number of clusters specified in the print option will be shown.

Finally, while running the CLUSTER procedure, ODS graphics can be used. More visualizations are created in the output when using ODS graphics, including plots of the criterion for the number of clusters and a dendogram of the hierarchical clusters. In this case, a dendogram could not be created due to the large number of observations. Dendograms were created in JMP and can be seen in Figures 3 and 4.

The CLUSTER procedure in SAS® was used to perform a Hierarchical cluster analysis on the network viewership data using Ward's method. Both the code and snapshots of the output can be seen below.

PROC CLUSTER source code (Ward's method):

```
ods graphics on;
proc cluster data=nets method=ward ccc pseudo trim=10 k=50 print=25;
      var ABC CW NBC CBS FOX TNT USA FNC NIK TOON;
id obs;
run;
ods graphics off;
```

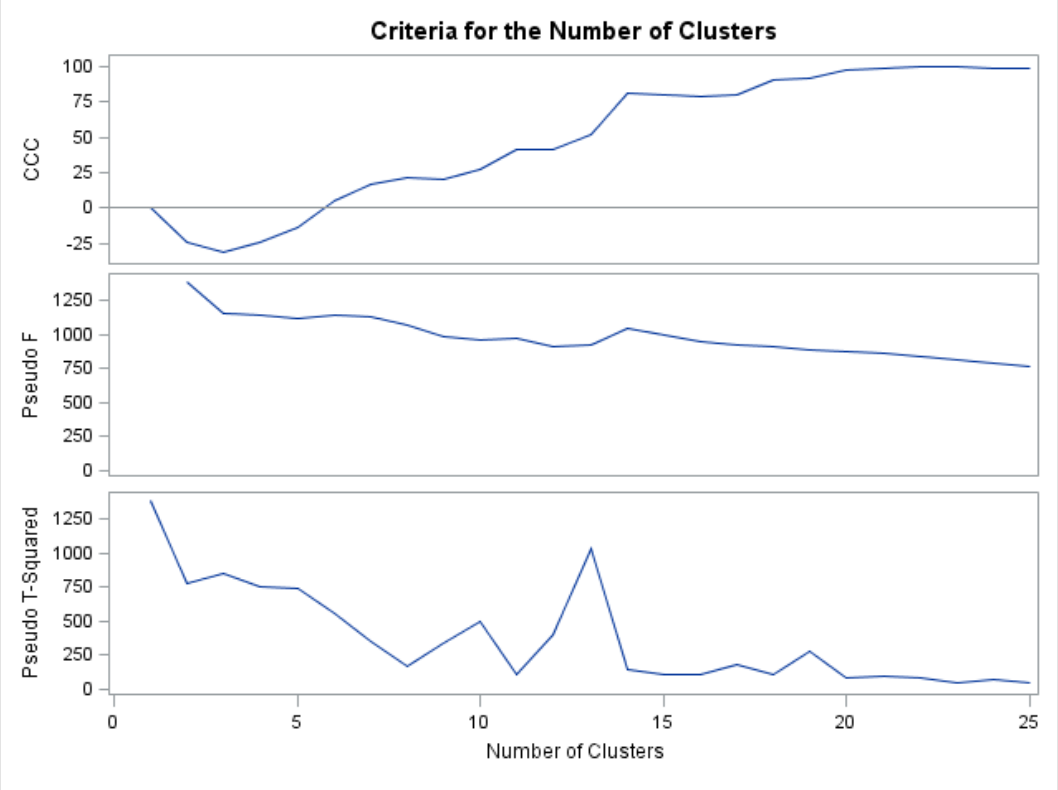


Figure 5. PROC CLUSTER Cluster Criteria Output

The PROC CLUSTER cluster criteria output can be seen in Figure 5. SAS® suggests that negative CCC values indicate outliers and local peaks in the CCC indicate good clusters in the data. In Figure 5, local peaks in the CCC graph occur at 8, 11, 14, and 16 clusters. SAS® also suggests that relative peaks in the pseudo F values indicate good clusters. Relative peaks in the pseudo F value chart occur at 8, 11, and 14 clusters. Finally, SAS® suggests that values immediately after relative peaks in the pseudo t-squared values indicate good clusters. Cluster values immediately after relative peaks in the pseudo t-squared chart are 4, 6, 11, 14, 18, and 20 clusters. Selecting the suggested clusters that overlap from the three criteria, clusters of 11 and 14 are potential cluster criteria for the network viewership data.

K-MEANS CLUSTERING

As discussed, hierarchical clustering is a good tool for exploratory analysis in Multivariate statistics. K-means clustering is a good tool for modeling. K-means clustering allows the observations to be moved from one group to another throughout the algorithm, a process that does not occur in hierarchical methods. K-means clustering will find

the optimal clustering for a given number (K) of groups. The results of the hierarchical cluster analysis are used to set *k* in the K-means cluster analysis. After assigning observations to clusters using K-means analysis, these assignments can then be used as the dependent variable in a discriminant analysis to create a classification model. This classification model can be used to assign future observations to a group. K-means clustering provides the "learning" component of a machine learning model as it allows the final classification model to evolve as the clusters evolve. The training data is used create the clusters and the new data can be assigned using knowledge about the clusters formed in the training data. In the example used in this paper, all 6,500 observations were considered the training data set.

The process for K-means clustering is as follows. After K is chosen, K items are selected to serve as seeds. These seeds are the initial values for the algorithm. Each item is evaluated by its distance to the nearest seed. As soon as a cluster has two observations, the cluster seed is replaced by the centroid or average value of the cluster. After all observations are assigned, each observation is again evaluated to see if it is closer to a different centroid than the centroid of its own cluster. This process will eventually converge until no more observations move clusters. There are various ways to choose the initial seeds. Since the algorithm converges, K random observations can be used as the initial seeds. However, the K-means method may be sensitive to the initial choice of seeds. If convergence is really slow or widely different results occur with different initial values, more strategic initial values may be needed. One option is to choose the K seeds by their maximum distance apart. The K observations with the furthest distance between them can serve as the initial seeds.

By default, the FASTCLUS procedure in SAS® uses a K-means clustering method. To establish convenient initial values for the seeds in the network viewership data, the FASTCLUS procedure was run with no iterations. The *outseed* option will save the final seeds in a SAS® data set. The *maxclusters* option refers to the maximum number of clusters to partition the data into, or the value for K in the K-means model. For establishing the initial seed values the maxcluster value must be greater than the greatest number of clusters desired for partitioning the data. In this example, a value of 20 was chosen as K-means models with 11 clusters and 14 clusters were created. The *maxiter* option specifies the number of iterations to run the K-means algorithm. SAS® suggests that the FASTCLUS procedure usually converges very quickly (within 3 iterations), so a large *maxiter* value is not needed. The iteration history in the FASTCLUS output will also specify whether or not convergence was achieved.

The SAS® code for the K-means cluster analysis for 11 and 14 clusters can be seen below.

PROC FASTCLUS source code (11 Clusters):

```
proc fastclus data=nets outseed=seeds maxclusters=20 maxiter=0;
      var ABC CW NBC CBS FOX TNT USA FNC NIK TOON;
run;
proc fastclus data=nets out=c11 seed=seeds maxclusters=11 maxiter=10;
      var ABC CW NBC CBS FOX TNT USA FNC NIK TOON;
run;
```

PROC FASTCLUS source code (14 Clusters):

```
proc fastclus data=nets out=c14 seed=seeds maxclusters=14 maxiter=10;
      var ABC CW NBC CBS FOX TNT USA FNC NIK TOON;
run;
```

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	178	0.7101	4.9834		6	5.0124
2	348	0.8738	5.0991		6	4.7257
3	564	0.8118	4.9230		6	4.6320
4	72	0.9544	4.2900		6	5.7895
5	450	0.8525	5.3383		6	4.3645
6	3857	0.4782	3.7611		11	2.8681

Cluster Summary						
Cluster	**Frequency**	**RMS Std Deviation**	**Maximum Distance from Seed to Observation**	**Radius Exceeded**	**Nearest Cluster**	**Distance Between Cluster Centroids**
7	12	1.0621	4.8705		6	5.8898
8	181	0.7707	5.7884		6	3.9042
9	199	0.7262	5.6396		6	4.2804
10	168	0.7068	6.2012		6	3.8348
11	323	0.6886	7.0449		6	2.8681

Figure 6. Select PROC FASTCLUS Output for 11 Clusters

Cluster Summary						
Cluster	**Frequency**	**RMS Std Deviation**	**Maximum Distance from Seed to Observation**	**Radius Exceeded**	**Nearest Cluster**	**Distance Between Cluster Centroids**
1	174	0.7036	4.9297		8	5.0516
2	171	0.8145	4.6430		4	4.0727
3	231	0.7422	4.4278		6	4.0051
4	485	0.6020	3.9515		8	2.4790
5	275	0.6894	6.7983		8	3.0679
6	685	0.5723	4.0209		8	2.7901
7	11	1.0713	4.6261		4	5.9434
8	2894	0.3943	3.2423		12	2.1006
9	186	0.7243	5.4953		8	4.3832
10	58	0.8972	4.6628		8	6.3849
11	165	0.7671	5.6092		8	4.0790
12	629	0.5283	3.9909		8	2.1006
13	230	0.8218	4.4033		12	3.9241
14	158	0.6985	6.0790		8	3.8988

Figure 7. Select PROC FASTCLUS Output for 14 Clusters

Three goodness-of-fit metrics are given in the PRCO FASTCLUS output and include an approximate R-squared value, a pseudo F-statistic, and the CCC value. The goodness-of-fit metrics for the 11-Cluster model resulted in an approximate R-squared value 0.49, a pseudo F-statistic of 1146, and a CCC value of 134. The goodness-of-fit metrics for the 14-Cluster model resulted in an approximate R-squared value of 0.52, a pseudo F-statistic of 1227, and a CCC value of 186. Higher values for all three metrics are considered better goodness-of-fit.

Figure 6 shows the Cluster Summary output from the FASTCLUS procedcure for the K-means method using 11 clusters. Notice how cluster 6 is the nearest cluster to all other clusters. Cluster 6 also has the largest frequency of observations. This suggests that cluster 6 is not well defined. For an algorithm like the K-means algorithm where the

number of final clusters is specified, it is not unusual to have a miscellaneous group. This group is usually defined as the group in which "observations do not belong to any other group". Also notice how group 7 has a very low frequency of observations in the 11-Cluster model. A small group such as group 7 is not ideal for classification because future observations assigned in group 7 will be classified based off the behavior of only 12 observations.

As can be seen in Figure 7, adding more clusters can help the miscellaneous group by pulling more observations away from the group and placing them into new groups. However, adding too many clusters can sometimes hurt the smaller groups by pulling observations away from them. The 14-Cluster model still contains a miscellaneous group, but it is not quite as influential as the miscellaneous group in the 11-Cluster model. While half of the clusters are located close to Group 8, some are close to different groups indicating better separation between groups. Large miscellaneous groups and small unique clusters suggest that more variables are needed to build a strong model for classifying future observations into groups. The low R-squared for these models suggest the same – more variables are needed to understand the separation between groups. Summer network viewership behavior may be influential in grouping observations as seen in the factor loadings and principal component analysis in Figure 2, but does not appear to tell the whole story of Summer viewership behavior. More variables will be needed to create a strong model for classifying future observations. In this example the 14-Cluster model was selected for the better goodness-of-fit metrics and separation between groups.

CLASSIFICATION

Discriminant analysis includes both a descriptive and a predictive component of clustering. Classification analysis is the predictive aspect of Discriminant analysis. Classification analysis differs from Cluster analysis in that the Clusters and group assignments are known. In Classification analysis, observations are allocated into groups using a set of variables. In the example used in this paper, the variables used to classify a new observation into a group can be seen in Table 1. The predictive component of classification analysis comes when observations are placed into groups according to their variable measures. In order to classify an observation, a training data set (with already pre-defined clusters) must be used to create the classification rules. The output from the 14-Cluster K-means model was used as the training data set to create the final classification model.

The DISCRIM procedure takes a training data set containing a list of quantitative variables and a classification variable and produces a model to classify observations into groups based solely on the quantitative variables. The default method for the DISCRIM procedure assumes multivariate normality. The distribution plots in Figure 1 show the network viewership data is far from multivariate normal. For this reason, a non-parametric method was used to classify new observations into groups. The classification criterion can be applied to a second dataset of new observations during the same execution the criterion is established. In this example the dataset *testdata* contains a list of 15 new observations to be classified into groups. The dataset *newgroups* contains the group assignments for each of the new observations. The method used to classify observations into groups is the non-parametric method known as the *Nearest Neighbor Classification Rule.* The rule classifies a new observation into a group by examining the behavior of its nearest *k* observations. The new observation is then classified into the group to which the majority of the *k* observations belong. *K* is an arbitrary value and is not the same *k* as selected in the K-means procedure. The source code for the DISCRIM procedure can be seen below.

PROC DISCRIM source code:

```
PROC DISCRIM data=c14 TESTDATA=testdata testout=newgroups method=npar k=5;
   var ABC CW NBC CBS FOX TNT USA FNC NIK TOON;
   CLASS CLUSTER;
   title '5 NEAREST NEIGHBORS Classification Analysis of Network Viewership';
run;
```

RESULTS

The results of the DISCRIM procedure using a non-parametric method do not include the criterion for classifying observations into groups. The criterion used to classify observations into groups using a parametric approach is known as the discriminant functions. These functions are useful in understanding which variables contribute to the separation between groups and how influential each variable is in determining that separation. However, some understanding into the separation between groups in a non-parametric approach can be found by inputting some clever test data. 15 new observations were created as a test data set in the network viewership example. This data can be seen in Table 2.

Observation	ABC	CW	NBC	CBS	FOX	TNT	USA	FNC	NIK	TOON
1	10	0	0	0	0	0	0	0	0	0
2	0	10	0	0	0	0	0	0	0	0

3	0	0	10	0	0	0	0	0	0	0
4	0	0	0	10	0	0	0	0	0	0
5	0	0	0	0	10	0	0	0	0	0
6	0	0	0	0	0	10	0	0	0	0
7	0	0	0	0	0	0	10	0	0	0
8	0	0	0	0	0	0	0	10	0	0
9	0	0	0	0	0	0	0	0	10	0
10	0	0	0	0	0	0	0	0	0	10
11	2	2	2	2	2	0	0	0	0	0
12	0	0	0	0	0	5	5	0	0	0
13	0	0	0	0	0	0	0	0	5	5
14	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0

Table 2. Test Data

Recall that a high score for a variable is 10. The sum of the variables for an observation also does not usually exceed 10. In the test data, the first 10 observations give a high weight to only one of the variables. Observations 11, 12, and 13 are given weight based off the results of the principal component analysis output in Figure 2 where networks were weighted according to broadcast networks, cable drama networks, and cable children's networks. Observation 14 gives equal weight to all variables and observation 15 gives no weight to any variables. The final group assignments for these 15 observations will give insight into which variables influence each group. The source code for viewing the final group assignments is a simple PROC PRINT and can be seen below, followed by the results.

PROC PRINT source code:

```
proc print data=newgroups;
run;
```

Obs	ABC	CW	NBC	CBS	FOX	TNT	USA	FNC	NIK	TOON	_INTO_
1	10	0	0	0	0	0	0	0	0	0	13
2	0	10	0	0	0	0	0	0	0	0	7
3	0	0	10	0	0	0	0	0	0	0	3
4	0	0	0	10	0	0	0	0	0	0	2
5	0	0	0	0	10	0	0	0	0	0	10
6	0	0	0	0	0	10	0	0	0	0	11
7	0	0	0	0	0	0	10	0	0	0	5
8	0	0	0	0	0	0	0	10	0	0	1
9	0	0	0	0	0	0	0	0	10	0	9
10	0	0	0	0	0	0	0	0	0	10	14
11	2	2	2	2	2	0	0	0	0	0	4
12	0	0	0	0	0	5	5	0	0	0	5
13	0	0	0	0	0	0	0	0	5	5	14
14	1	1	1	1	1	1	1	1	1	1	5
15	0	0	0	0	0	0	0	0	0	0	8

Figure 8. Select PROC DISCRIM Output

Figure 8 shows the group assignments for each observation. All groups are represented except group 6 and 12. Using the group assignments from Figure 8, each group and the separation between groups can be better understood. For example, the more time an individual spends watching the NBC network, the more likely they are to

be classified into group number 7. Table 3 shows how each group can be defined using the information from Figure 8.

Group	Network Viewer Type
1	Loyal FNC Network Viewers
2	Loyal CBS Network Viewers
3	Loyal NBC Network Viewers
4	Loyal Broadcast Network Viewers (Maybe non-Cable Subscribers)
5	More Cable Drama Network Viewers
6	
7	Loyal CW Network Viewers
8	Loyal Other Network Viewers (Maybe non-tv-watchers)
9	Loyal NIK Network Viewers
10	Loyal FOX Network Viewers
11	Loyal TNT Network Viewers
12	
13	Loyal ABC Network Viewers
14	More Children's Network Viewers

Table 3. Variable Impact on Group Assignment

In order to understand groups 6 and 12 and gain a greater understanding of the classification of all groups, the classified members of each group can be explored individually. Using cross-validation, observations from the original training dataset can be classified into a group as if it had no prior classification. Cross-validation removes one observation, calculates the classification criterion without that observation, and then uses the established classification rule to assign the observation to a group. The additional statement *crosslist* will include cross-validation output from the DISCRIM procedure.

PROC DISCRIM source code (with Cross-Validation):

```
PROC DISCRIM data=c14 method=npar k=5 crosslist outcross=cvresults;
   var ABC CW NBC CBS FOX TNT USA FNC NIK TOON;
   CLASS CLUSTER;
   title '5 NEAREST NEIGHBORS Classification Analysis of Network Viewership';
run;
```

Cluster	Average ABC Score	Average CW Score	Average NBC Score	Average CBS Score	Average FOX Score	Average TNT Score	Average USA Score	Average FNC Score	Average NIK Score	Average TOON Score	Network Viewer Type
1	0.32	0.01	0.46	0.34	0.15	0.34	0.25	4.81	0.10	0.05	Mostly FNC
2	0.84	0.12	1.07	6.47	0.56	0.06	0.05	0.01	0.01	0.01	Little Cable, mostly CBS
3	0.74	0.05	6.95	0.88	0.30	0.09	0.05	0.02	0.04	0.01	Little Cable, mostly NBC
4	0.69	0.10	0.90	2.38	0.40	0.32	0.25	0.05	0.06	0.07	Some CBS
5	0.44	0.03	0.38	0.43	0.26	0.83	3.04	0.04	0.13	0.08	Mostly USA
6	0.65	0.04	2.91	0.61	0.36	0.28	0.19	0.05	0.12	0.08	Mostly NBC
7	0.48	5.95	0.72	1.91	0.79	0.00	0.00	0.00	0.00	0.00	No Cable, mostly CW
8	0.25	0.03	0.35	0.25	0.29	0.33	0.22	0.06	0.14	0.13	Little of everything
9	0.27	0.01	0.43	0.24	0.16	0.12	0.10	0.05	4.17	0.48	Mostly NIK
10	0.44	0.11	0.71	0.52	6.35	0.05	0.20	0.03	0.05	0.01	Little Cable, mostly FOX
11	0.39	0.04	0.43	0.47	0.22	4.11	0.37	0.03	0.10	0.05	Mostly TNT

12	2.11	0.04	0.71	0.46	0.46	0.46	0.21	0.05	0.12	0.08	Mostly ABC
13	5.98	0.06	1.08	0.88	0.38	0.15	0.04	0.02	0.04	0.02	Mostly ABC, some NBC
14	0.24	0.04	0.41	0.20	0.19	0.16	0.20	0.06	0.74	3.78	Mostly TOON

Table 4. Average Variable Scores per Group

Table 4 shows the average scores of each variable for each group. Highlighted values are those with larger scores for each group and indicate a greater variable influence on the corresponding group. For example, the more an individual spends time watching the FNC network, the more likely they are to be classified into group 1. Group 6 appears to be influenced mostly by its viewership preference for the NBC network, but is different from group 3 in its lower overall television viewership as the scores for each variable are small. Group 12 appears to be influenced mostly by its viewership preference for the ABC network, but is different from group 13 in its lower overall television viewership. Group 8 is the clear miscellaneous group as there is no clear network viewership preference for this group.

One final method for understand the final classification assignments is to view the misclassification rates for each group. The misclassification rates are calculated from the cross-validation results. Using the training data set, the cross-validation method calculates the new group assignment and knows the original group assignment. The misclassification rates for each group are included in the output of the DISCRIM procedure.

Error Count Estimates for CLUSTER

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
Rate	0.0057	0.0234	0.0173	0.1093	0.0327	0.1066	0.0000	0.1299	0.0161	0.0345	0.0121	0.1129	0.0304	0.0127	0.0460
Priors	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	

Figure 9. PROC DISCRIM Cross-Validation Error Counts Output

Figure 9 shows that the total percent of misclassified observations in the network viewership data was 4.6%. The group with the largest misclassification rate was group 8, with 13% of its original members being classified into a different group. Group 8 is the miscellaneous group, so a larger misclassification rate is expected. Lower viewership groups, such as group 12 and group 6 have large misclassification rates with 11% of their original members being classified into different groups. Using the cross-validation output, it can be seen which group these misclassified members most likely fall to. In this model, group 12 members falling into group 13 is not bad for the model as both groups prefer ABC. Table 5 below shows the percent of misclassified observations falling into each group.

		Classified Into													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Original Group	1	99%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
	2	0%	98%	0%	1%	0%	0%	0%	0%	1%	0%	1%	0%	1%	0%
	3	0%	0%	98%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	4	0%	4%	0%	89%	0%	2%	0%	1%	0%	0%	0%	2%	1%	0%
	5	1%	0%	0%	0%	97%	0%	0%	1%	0%	0%	0%	0%	0%	1%
	6	0%	0%	3%	1%	0%	89%	0%	1%	1%	0%	0%	2%	1%	0%
	7	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%
	8	0%	0%	0%	2%	1%	3%	0%	87%	1%	0%	1%	4%	0%	0%
	9	0%	0%	0%	0%	1%	0%	0%	1%	98%	0%	0%	0%	0%	0%
	10	0%	2%	0%	2%	0%	0%	0%	0%	0%	97%	0%	0%	0%	0%
	11	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%	99%	0%	0%	0%
	12	1%	0%	0%	2%	1%	2%	0%	1%	0%	0%	1%	89%	3%	0%
	13	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	97%	0%
	14	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	0%	0%	0%	99%

Table 4. Classification Rates

CONCLUSION

It has been determined that the 6,500 observations in the network viewership data are best clustered into 14 groups separated by network loyalty. Surprisingly the separation between groups was largely determined by network loyalty, and not network genre like cable drama networks or children's networks. Observations proved to prefer one network over the others. Some useful next steps to this type of analysis include adding more variables. The final K-means models gave weak goodness-of-fit statistics. These models can be made stronger by finding more variables that can classify the behavior of the groups. A larger sample size may also assist in providing more information about the final groups. The groups with small group memberships can benefit by a larger sample.

In a world where technology and information thrive, data can do a lot of talking. Clustering and Classification are at the heart of Multivariate statistical analysis and SAS® procedures such as PROC CLUSTER, PROC FASTCLUS, and PROC DISCRIM are efficient and simple methods for clustering and classifying multivariate data. Once understood, these tools can be very powerful and can make multivariate data less overwhelming and more effective.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Rachel Poulsen
Enterprise: TiVo
Address:
City, State ZIP: Alviso, CA
Work Phone:
Fax:
E-mail: rlpoulsen@gmail.com
Web: