Chapter 7: Cluster Analysis Using SAS Enterprise Miner

Introduction

In Chapter 5, you explored the DMR Publishing customer data to familiarize yourself with the data and look for customer patterns and trends. In Chapter 6, you compared the performance of DMR Publishing customers to the performance of the entire publishing market. In this chapter, your goal is to gain a deeper understanding of your customer base by grouping your customers into segments that have similar characteristics. By unveiling the similarities and differences among your customers, you can design marketing programs and database enrichment strategies that align with your long-term strategic goals.

Project Overview

The leadership team at DMR Publishing Company has asked you to help it gain a better understanding of its customer base. The team wants to know whether all their customers look alike, or whether they naturally segment into different groups. The segmentation is performed with SAS Enterprise Miner.

The project has eight steps:

1. Initiate the project in SAS Enterprise Miner 13.1.
2. Input data and assign variable roles.

3. View variable distributions and transform if necessary.
4. Filter data.
5. Build clusters.
6. Build segment profiles.
7. Recommend business actions.

Cluster Analysis

Cluster analysis, or *clustering*, is a process that places observations into groups or segments that favor similarity within each group while favoring dissimilarity between groups. This ability to group or segment customers can be useful if you want to market to a group of your customers who look alike. You may have information about your customers, but you don't know what makes them similar or different. Cluster analysis performs this type of grouping.

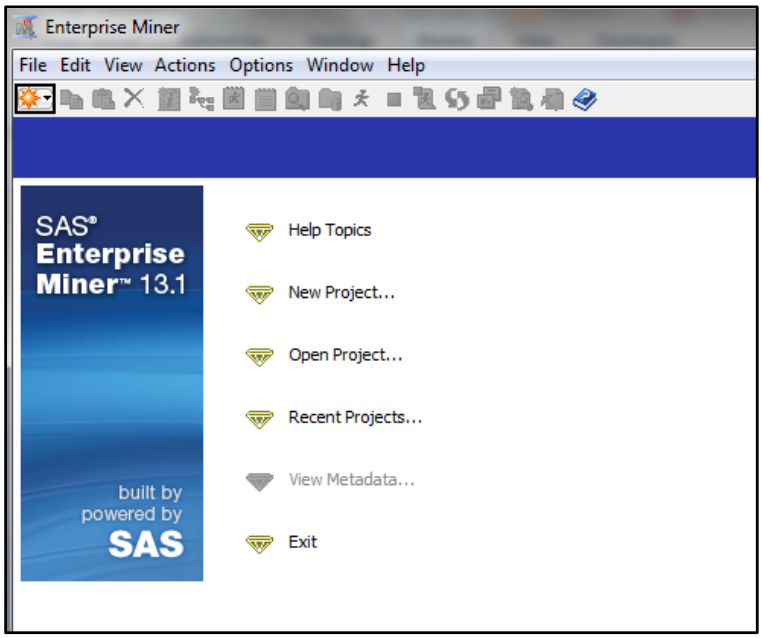
The clustering methods in the SAS Enterprise Miner cluster node perform disjoint cluster analysis by calculating the Euclidean distances between at least one quantitative variable and seeds. The *seeds* are the original centers of the clusters. The centers of the clusters change during the clustering process. You can control the clustering criterion that is used to measure the distance between data observations and seeds. The final clusters are mutually exclusive in that no observation populates more than one cluster. This feature of the cluster node makes it very useful for business purposes.

Initiate the Project

To open SAS Enterprise Miner, click the icon on your desktop or Start menu. Your first choice is to open an existing project or create a new project. Highlight and click **New Project** (Figure 7.1).

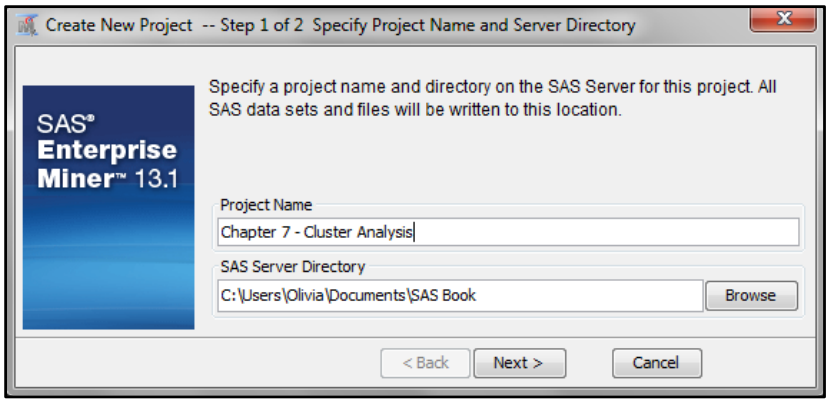
NOTE: It is possible to do basic cluster analysis in SAS Enterprise Guide. But SAS Enterprise Miner has automated and streamlined the process that is optimal for creating mutually exclusive clusters. This ability to create mutually exclusive clusters is essential for use in marketing and risk analysis.

Figure 7.1: Initialize SAS Enterprise Miner



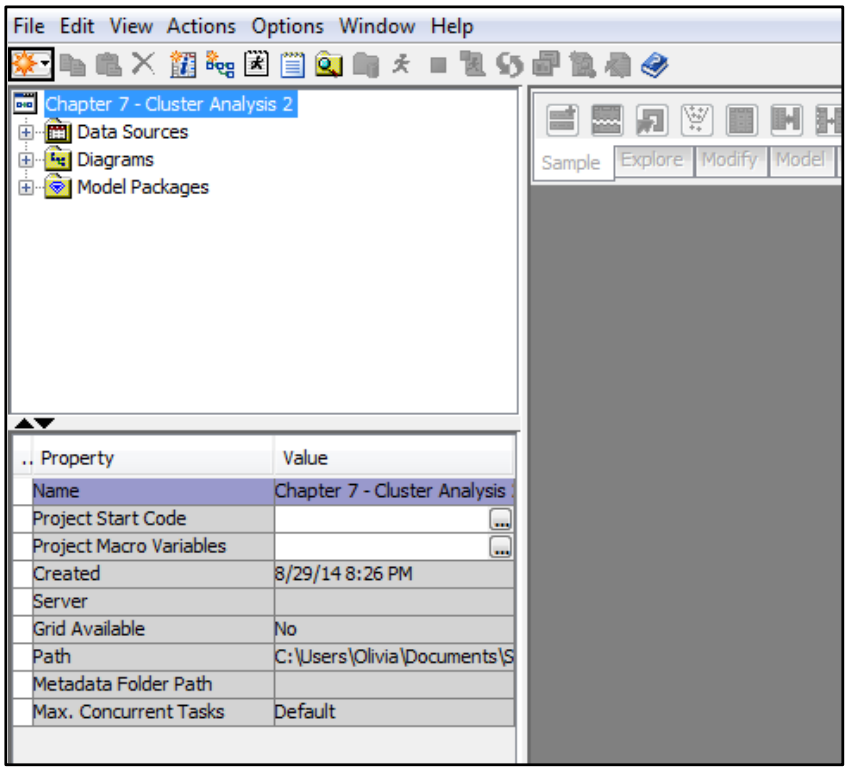
A window will open that asks you to name your project and select a SAS Server Directory (Figure 7.2). Depending on your setup, additional connections may be required. If this is the case or if you are not sure how to locate your data, contact your information technology department or other technical assistant. Otherwise, click browse and select a folder in which you would like to save your project files.

Figure 7.2: Create and Name New Project



When you are finished, click **Next** ▶and you will see window that summarizes your options. Click **Finish**. You are now in the EM workspace, as shown in Figure 7.3.

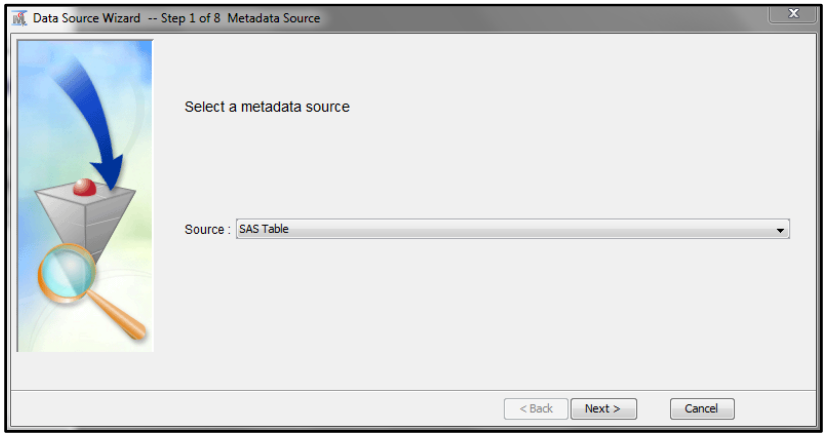
Figure 7.3: View SAS Enterprise Miner Workspace



Input the Data Source and Assign Variable Roles

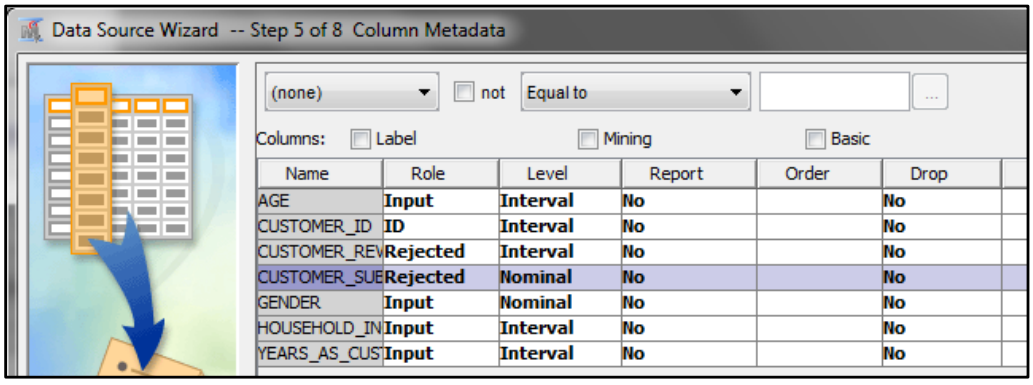
Next, double click on the Data Sources icon (upper left-hand menu, directly under the word ***Actions***). The Data Import Wizard will open, asking you to create a SAS Table (Figure 7.4).

Figure 7.4: Locate Data Source



Click **Next ▶**and browse for the DMR_CUSTOMER_BASE data set created in Chapter 3. Once you locate the data, select the data and click **OK**, then click **Next ▶**. The next window shows you a summary. Click **Next ▶**. For Meta Data Advisory options, click **Advanced** and **Next ▶**. A window appears that displays the variable characteristics and offers options for exploring the distributions (Figure 7.5).

Figure 7.5: Assign Variable Roles and Explore Variables

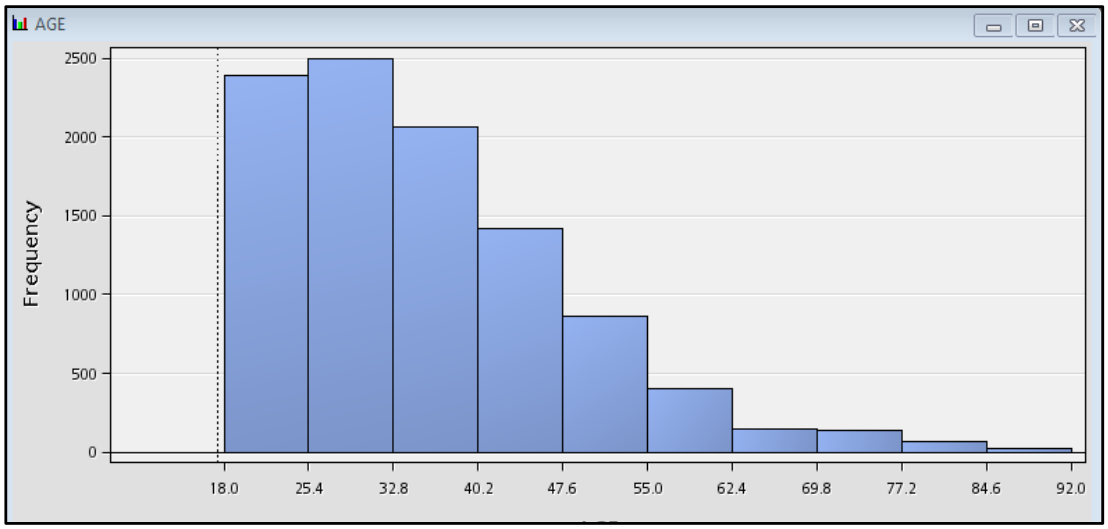


First, you need to change the role of CUSTOMER_REVENUE and CUSTOMER_SUBSCRIPTION_COUNT to "Rejected." These are outcome variables that will be used in future chapters. But for now, you do not want to include them in your analysis.

NOTE: It is necessary to have an ID variable for clustering in order to track the observations in each cluster.

Highlight AGE and click **Explore**. A large window opens and has four quadrants. Maximize the quadrant in the lower left to get Output 7.1.

Output 7.1: View Distribution of the Age Variable



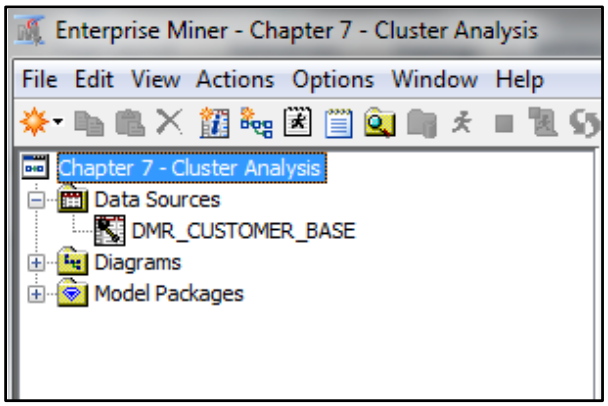
For clustering, you want variables to have a bell-shaped curve that represents a normal distribution. Because AGE is not normally distributed, you can use the **transform** function. Because AGE is skewed to the right, or positively skewed, you should use a log transformation after all the variables are explored.

NOTE: To close the **Results** window in SAS Enterprise Miner, click the **X** in the upper right-hand corner. It will close only the results window; the main project will remain open.

Explore the two remaining continuous variables, HOUSEHOLD_INCOME and YEARS_AS CUSTOMER, by following the same process. Because both variables are skewed to the right, these variables will also need to be transformed with use of the log transformation. But first, you must complete the import process. Click **Next** ▶ several times and then **Finish.**

Once the DMR_CUSTOMER_BASE data source is created, the data set name will appear under Data Sources in the upper diagram as shown in Figure 7.6.

Figure 7.6: Open Project View with Data Source in Clustering Diagram

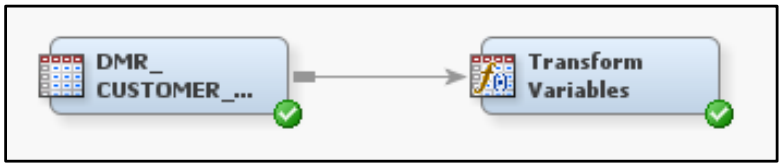


Next, you want to create a workspace for your cluster analysis. Go to the top menu and click **File** ▶ **New** ▶ **Diagram**. When the box opens, name the diagram Clustering and click **OK**. A work area will appear on the right. Place your cursor on the data set, DMR_CUSTOMER_BASE, and drag it into the clustering work area on the right.

Transform Variables

Next, look at the menu above the diagram and click the **Modify** tab. The last icon in the row above is **Transform Variables**. Drag the **Transform Variables** icon to the **Diagram** and connect it to **DMR _CUSTOMER** data with an arrow (Figure 7.7).

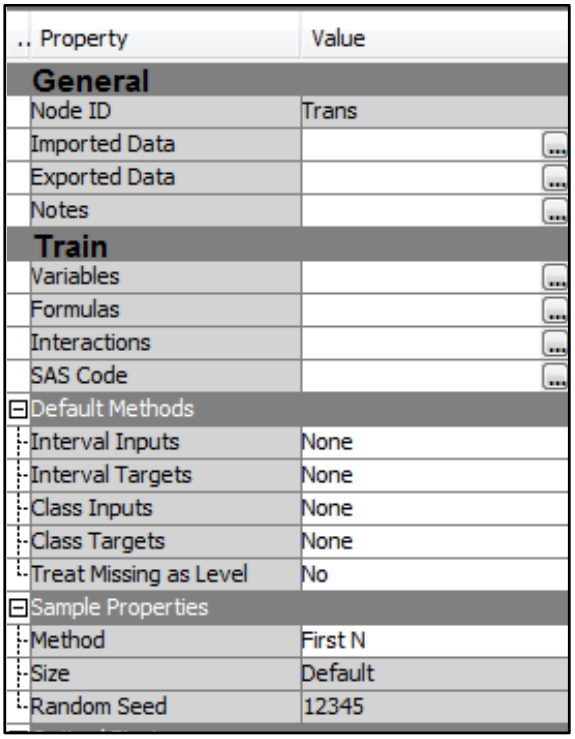
Figure 7.7: Use the Transform Variable Icon



To connect the DMR Customer data to the Transform Variables node, right-click on the **Transform Variables** node and select **Run.** When offered to view the results, click **OK**.

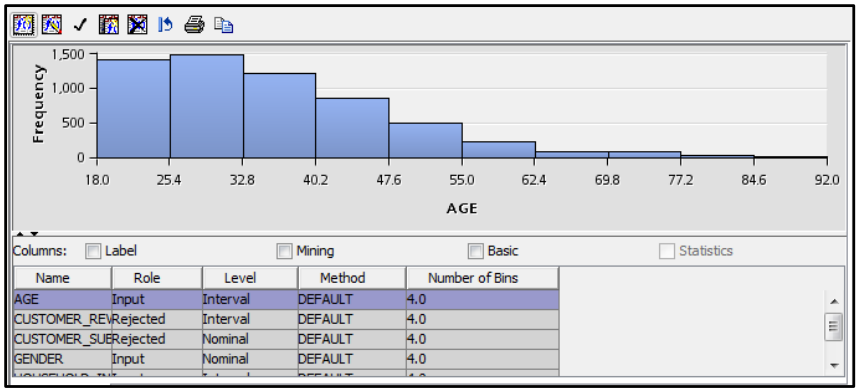
With the Transform Variables node still highlighted, look to the lower left of the work area shown in Figure 7.8.

Figure 7.8: Locate the Transform Variable Formula Menu



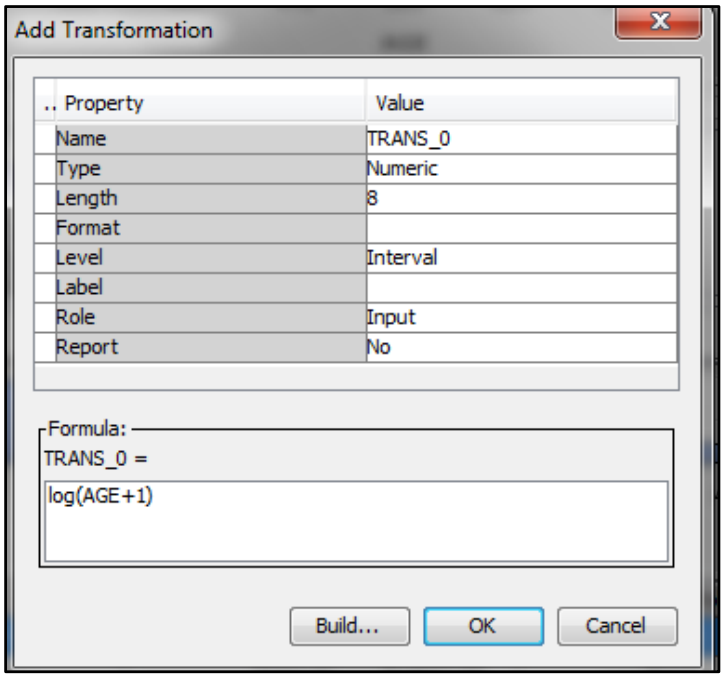
Under **Train**, next to **Formulas**, click the three dots to the right. A window will open that shows all the variables and each distribution, depending on which variable is selected (Figure 7.9).

Figure 7.9: View the Transformation Overview



Highlight the variable AGE. In the upper left-hand corner, click the **Create** icon. The word 'Create' will appear when you hover over the icon. The box in Figure 7.10 will appear.

Figure 7.10: Transform the Age Variable

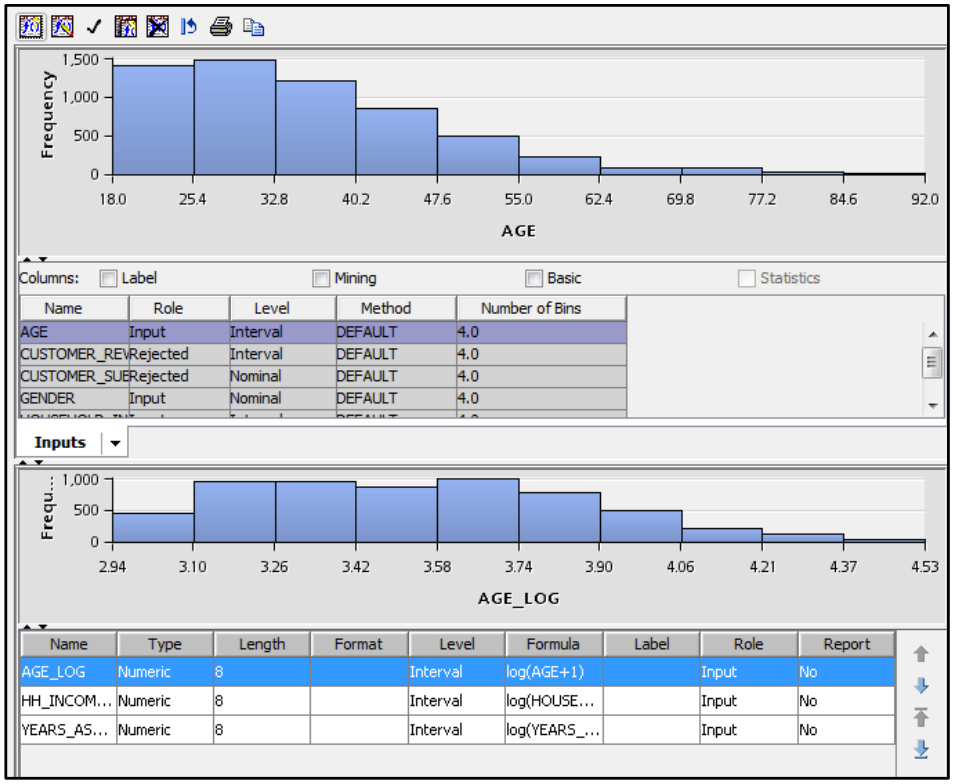


To the right of **Property** under **Value**, change the name from **TRANS_01** to **AGE_LOG**. Repeat the process for HOUSEHOLD_INCOME and YEARS_AS_CUSTOMER, using similar **Names** and log transformations based on the formulas in Figure 7.10. After each formula is typed in, click **OK**. Note

that the log of HOUSEHOLD_INCOME becomes HH_INCOME_LOG and equals log (HOUSEHOLD_INCOME+100). The name for log of YEARS_AS_CUSTOMER is YEARS_AS_CUST_LOG.

The transformation formulas appear at the bottom of the Formulas window as shown in Figure 7.11.

Figure 7.11: Create the Transformation Formulas for Age, Household Income, and Years as Customer



Click **OK**. Then, right-click on the **Transform Variables** icon and hit **Run**. After the run is complete, click **OK**.

Filter Data

Because clustering is sensitive to outliers, you will get better results if you run your variable through a filtering process. Above your diagram, click on the **Sample** tab and go to the fourth icon, **Filter**. Drag it onto the diagram and connect it with an arrow (Figure 7.12).

Figure 7.12: Filter the Data

Right click on the **Filter** node and select **Run**.

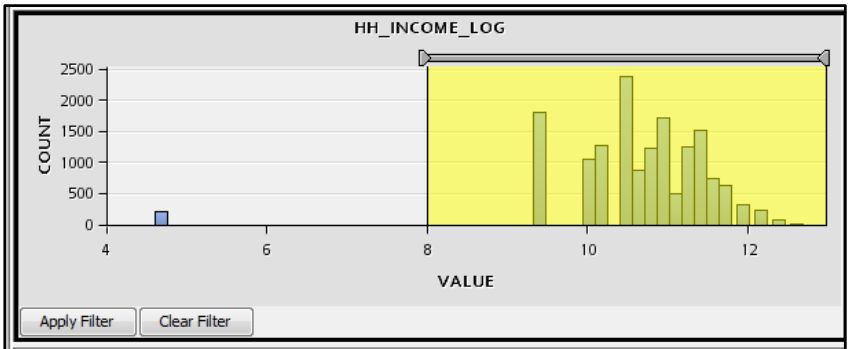
To look for potential outliers, highlight the filter icon, go to the lower left menu, and click the three dots to the right of **Interval Variables** (Figure 7.13).

Figure 7.13: Locate the Filter Menu

Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percent	0.01
Maximum Number of Levels	25
Interval Variables	
Interval Variables	
Default Filtering Method	Standard Deviations from t
Keep Missing Values	Yes

This dropdown option opens a window that allows you to view each distribution (Figure 7.14). For HH_INCOME_LOG, set the **Filter Lower Limit** to 8 and click **Apply Filter**.

Figure 7.14: Specify Filter Variables



Columns: Label Mining Basic Statistics

Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit
AGE_LOG	No	Default	Default	.	.
CUSTOMER_REVENUE	No	Default	Default	.	.
HH_INCOME_LOG	No	User Specified	Default	8	12.96682
YEARS_AS_CUST_LOG	No	Default	Default	.	.

Click **OK.** Then right-click the **Filter** icon and select **Run**. When it completes, click **OK**.

Build Clusters

Above the diagram window, go to the **Explore** tab and pick the second icon, **Cluster**. Drag the **Cluster** icon onto the diagram and connect it with the **Filter** node (Figure 7.15).

Figure 7.15: Add the Cluster Process

When you highlight the Cluster icon, the Property menu appears on the lower left (Figure 7.16).

Figure 7.16: View the Property Menu for the Add-Cluster Process

Property	Value
General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Range
Number of Clusters	
Specification Method	Automatic
Maximum Number of Cluste	10
Selection Criterion	
Clustering Method	Centroid
Preliminary Maximum	50
Minimum	2
Final Maximum	20
CCC Cutoff	3
Encoding of Class Variables	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Full Replacement
Minimum Radius	0.0
Drift During Training	No

Continuous variables come in different scales, such as counts, minutes, and dollars. You will want to standardize these variables for clustering. Otherwise, the variables with the higher scale will have an advantage. Under **Train**, change **Internal Standardization** to **Range**. This option standardizes the values for each variable to a value between 0 and 1. For **Clustering Method**, select **Centroid**. This method is usually better than the Ward method for handling contrasting data. Change the **Seed Initialization Method** to **Full Replacement** to select seeds that are well-separated. Right click, then click **Run**. When the run is complete, click on **Results**.

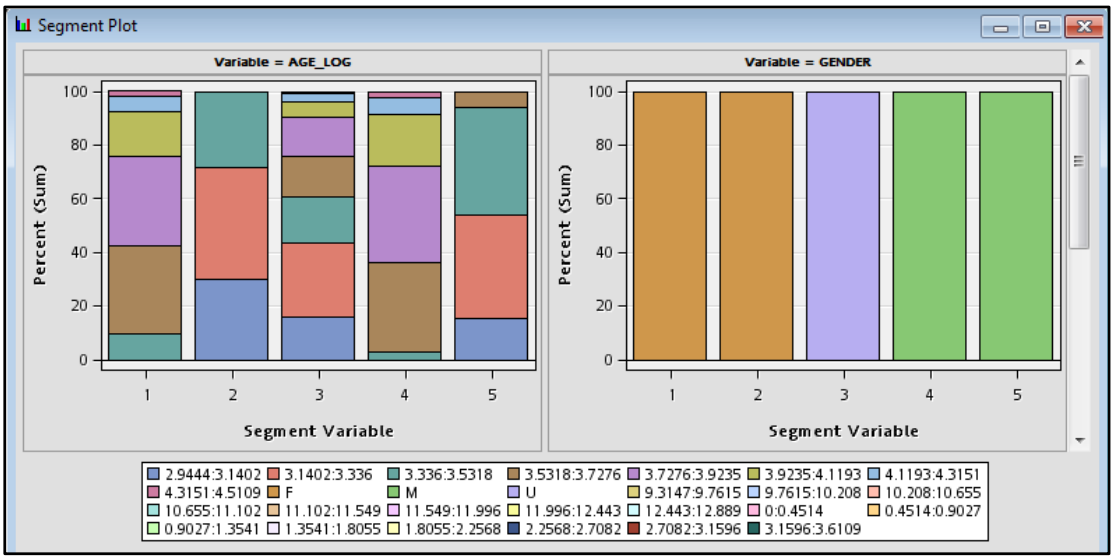
Output 7.2 shows the upper left and lower left quadrants of in the Results windows four quadrants. Each quadrant offers insights into the results of the process as defined:

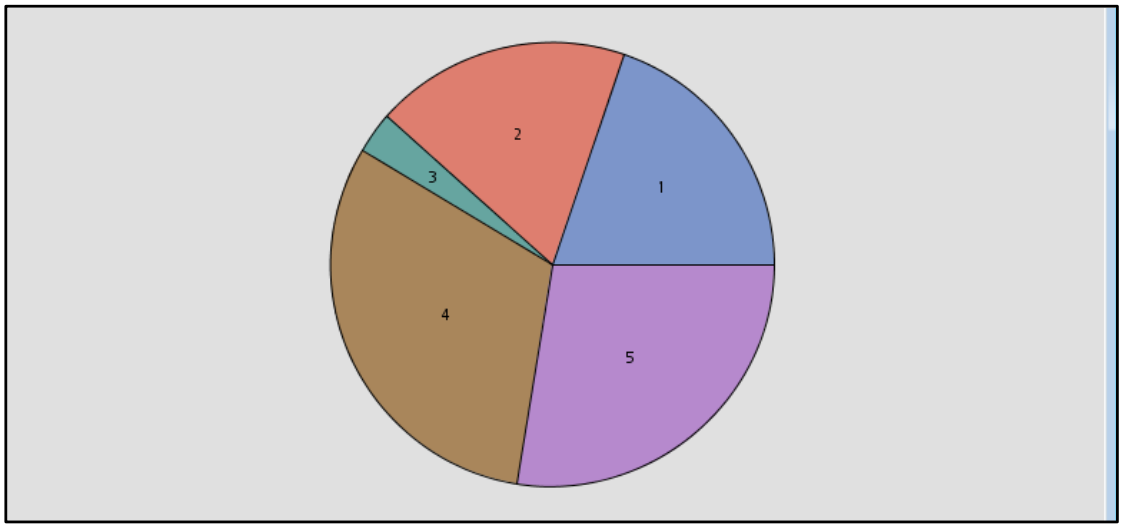
- *Segment Plot*—the upper left quadrant displays a segment plot of the clustering variables with the highest importance. The results show how the values of age, log, and gender are distributed among the clusters. To see the value of the segment and other statistics, place your cursor on the color within each segment.

- *Segment Size*—the lower left quadrant offers a visual display of the size of each cluster in a pie chart. To view the number of customers in each cluster, hover your cursor over each segment of the pie chart.

Because the next step is to build segment profiles, the main interest in these four quadrants is the number of clusters displayed in the lower left quadrant. The pie chart shows that there are five total clusters. Four to eight clusters is a good amount when building segment profiles.

Output 7.2: View the Cluster Results



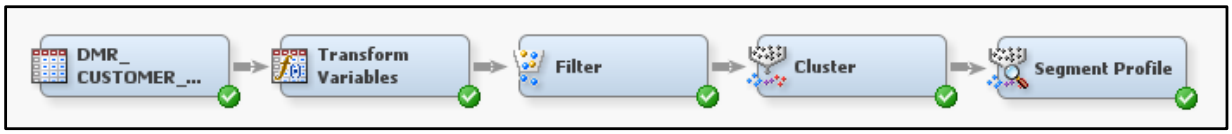


When you are finished viewing the cluster results, close the Results window and return to the diagram.

Build Segment Profiles

In the menu above the diagram, click **Assess** and **Segment Profile**. Drag the Segment Profile icon into the diagram and connect with an arrow (Figure 7.17).

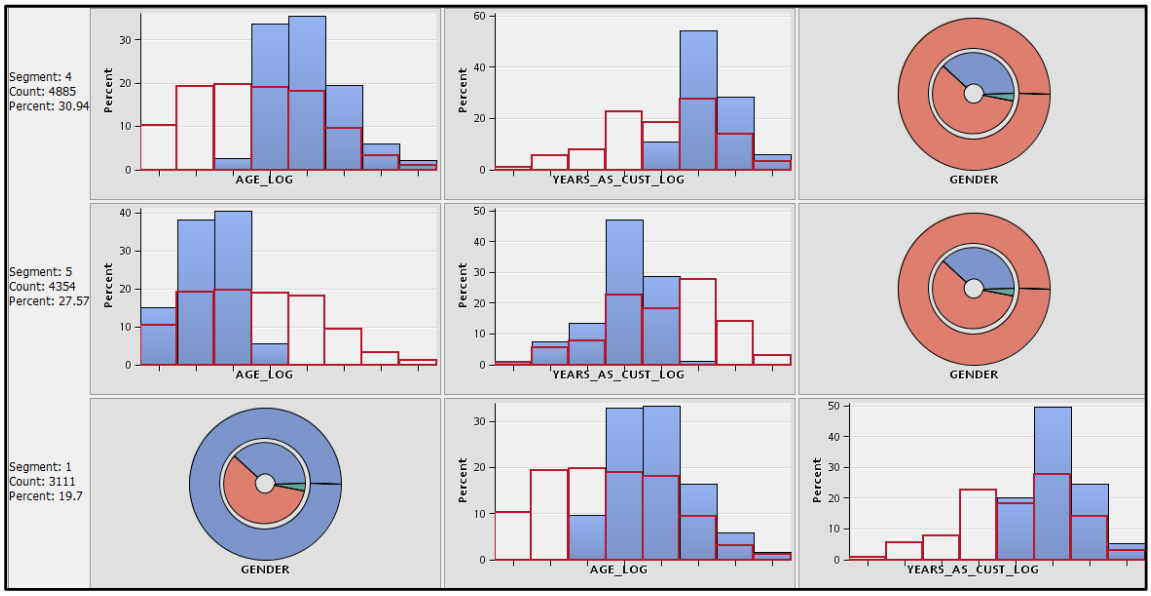
Figure 7.17: Add the Segment Profile Process



Right-click and **Run**. When the results window appears, click **Results**.

The initial Results window displays four quadrants. Focus on the upper right quadrant, which displays the profiles of each segment. Maximize the upper right quadrant for a closer look, as partially seen in Output 7.3.

Output 7.3: View the Segment Profile Results



Maximize the upper right quadrant for a closer look, as seen in Output 7.4.

This output displays the characteristics of the customers in each segment. The segments are arranged in order of the size of the segment. The segment with the greatest number of customers is listed first. The segment with the least number of customers is listed last:

Segment 4 has 4,885 customers. AGE_LOG is the strongest predictor. The solid bars represent the distribution within the segment. The dark red outline represents the overall population. So this cluster

contains customers that are older than average. They have also been customers longer than average. The third characteristic is gender. The customers in this cluster are all male. If you want to see the values, right-click next to the circle and select **Expand**. Then, hover over the edge of the outside circle, and a window will appear that gives you values for each area. Finally, the household income log shows the same trend toward higher than average.

Segment 5 has 4,354 customers. AGE_LOG is the strongest predictor. This cluster contains customers that are younger than average. They have also been customers for less time than average. The customers in this cluster are all male. Finally, household income shows the same trend towards lower than average.

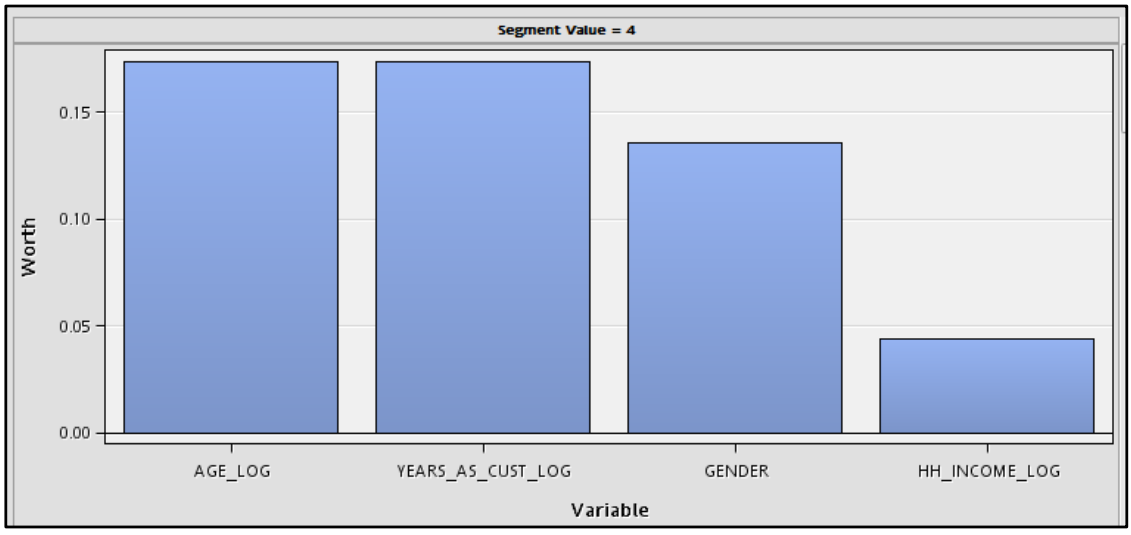
Segment 1 has 3,111 customers. GENDER is the strongest predictor. This cluster is all female. AGE is next. This cluster contains customers that are older than average. They have also been customers longer than average. Finally, household income shows the same trend towards slightly higher than average.

Segment 2 has 2,941 customers. GENDER is the strongest predictor. This cluster is all female. This cluster contains customers that are younger than average. They have also been customers less time than average. Finally, household income shows the same trend toward lower than average.

Segment 3 has only 499 customers. Its only characteristic is GENDER, which is all unknown. This result doesn't show in Output 7.4, but is visible when the process is run in SAS Enterprise Miner.

Another useful result is the **Variable Worth** in the lower left quadrant of the **Results** window, as shown in Output 7.4.

Output 7.4: View Variable Worth



AGE_LOG is the strongest contributor to the cluster segments, followed by YEARS_AS_CUST_LOG, which is very close. GENDER is third in importance as a contributor. And, finally, HH_INCOME_LOG is the least powerful contributor to the cluster segments.

Analyze Clusters and Recommend Marketing or Product Development Actions

This set of results informs DMR Publishing that it can focus its marketing and product development for different age groups with consideration for loyalty (years as customer), gender, and household income.

Because Segment 4 is the largest, it is a good place to focus your analysis. However, one thing to notice is the similarity between Segment 4 and the third largest segment, Segment 1. They both have loyal customers (based on years as customer) who are older than average and have slightly higher than average income. The main difference is their gender. Together, Segment 4 and Segment 1 represent 50% of the DMR Publishing customer base. Therefore, you may want to consider a two-level approach:

Consider separate marketing actions or product development for gender-specific publications, such as men's health or women's fashion magazines. Or, your approach might be as simple as different magazine covers or advertisements aimed at each gender.

Consider combining these two segments for publications that are not gender-specific, such as cooking and travel magazines or business journals.

Segment 5, the second largest cluster, offers another opportunity. This group of all male, younger-than-average customers is a prime audience for publications that appeal to that demographic group. If DMR Publishing doesn't already offer some sports or technology magazines, you might suggest that they add some to their list of publications.

Notes from the Field

Segment profiles are used to describe the clusters and make them actionable. As you have seen, the patterns revealed and the insights that emerge can guide creative marketing decisions and well as product or service development. Once you understand your customers based on your existing data, consider purchasing additional data to enrich your clusters, enhance your analysis, and grow your customer base. As discussed in Chapter 2, there are many good sources of external data. You can purchase characteristics such as hobbies and interests, buying patterns, and social or online behavior and append them to your existing customer base. Once your customer data is enriched, you can rerun your cluster segments and refine your marketing and product development strategies.

When sharing your results or proposing strategic initiatives with your stakeholders and end-users, speak in terms of their business objectives and relate your recommendations to the strategic goals of the company.