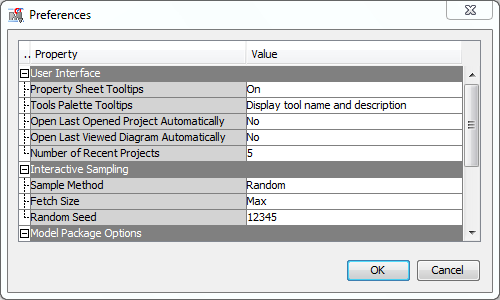
Clustering and Segmenting of Census Data

(adapted from Applied Analytics using SAS Enterprise Miner, SAS Institute, Cary, NC. 2010)

This demonstration introduces SAS Enterprise Miner tools and techniques for cluster and segmentation analysis. There are five parts:

* define the diagram and data source
* explore and filter the training data
* integrate the Cluster tool into the process flow and select the number of segments to create
* run a segmentation analysis
* use the Segment Profile tool to interpret the analysis results

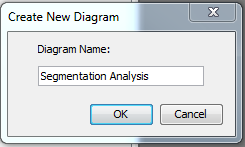
**Before you start**, make sure that the defaults for sampling are set up properly. Select **Options** ⇨ **Preferences** from the mainmenu and make sure that Sample Method is set to “Random” and Fetch Size is set to “Max.” It should look like this:



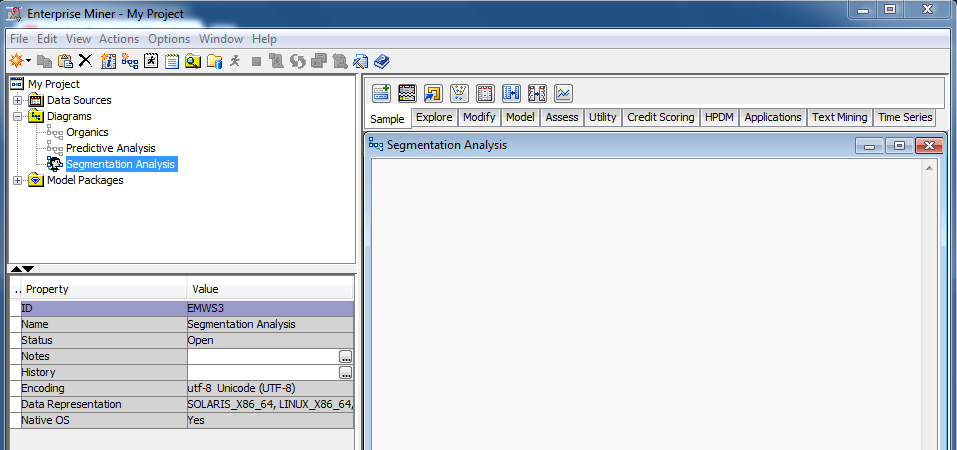
#### Diagram Definition

Use the following steps to define the diagram for the segmentation analysis. You can use the project you’ve already been using.

1. Right-click **Diagrams** in the Project panel and select **Create Diagram**. The Create New Diagram window opens and requests a diagram name.



1. Type **Segmentation Analysis** in the Diagram Name field and select **OK**. SAS Enterprise Miner creates an analysis workspace window named Segmentation Analysis.



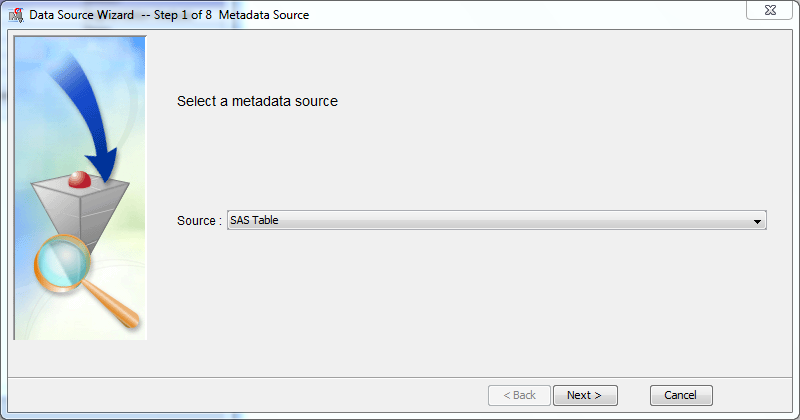
You use the Segmentation Analysis window to create process flow diagrams.

#### Data Source Definition

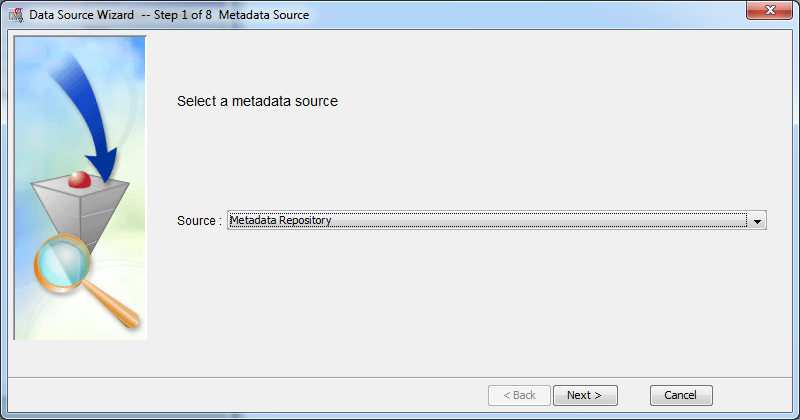
Follow these steps to create the segmentation analysis data source.

Follow these steps to specify a data source.

1. Select **File** ⇨ **New** ⇨ **Data Source…** from the main menu. The Data Source Wizard – Step 1of 8 Metadata Source opens.

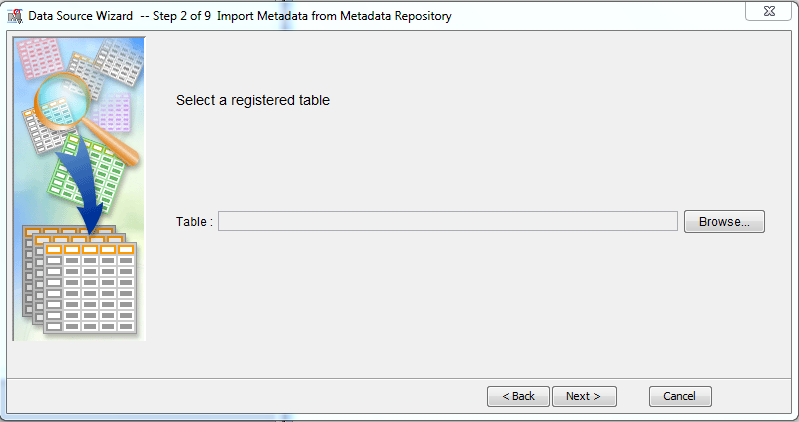


Click on **Source:** and select **Metadata Repository**

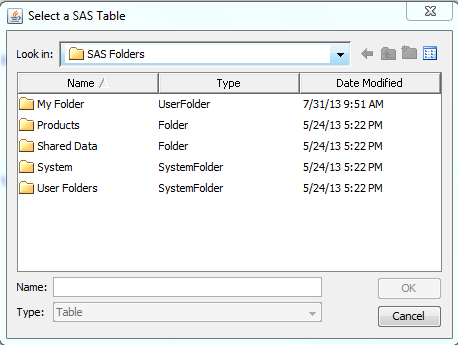


1. Select **Next >**

The Data Source Wizard continues to Step 2 of 8 Select a SAS Table.



1. In this step, select the SAS table that you want to make available to SAS Enterprise Miner. Click **Browse** on the right hand side.



1. Browse through the folders: **Shared Data**/**Libraries**/**AAEM**. Then select **CENSUS2000** and click **OK**.

The **Census2000** data is a postal code-level summary of the entire 2000 United States Census.   
It has seven variables:

**ID** postal code of the region

**LOCX** region longitude

**LOCY** region latitude

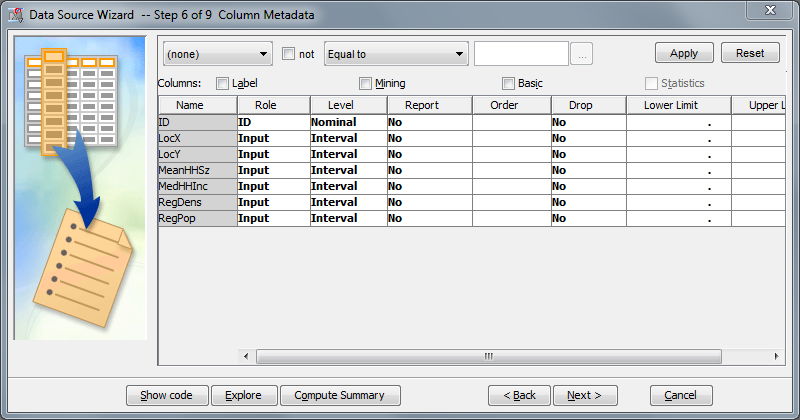
**MEANHHSZ** average household size in the region

**MEDHHINC** median household income in the region

**REGDENS** region population density percentile (1=lowest density, 100=highest density)

**REGPOP** number of people in the region

1. Select **Next >** until you reach Step 6.



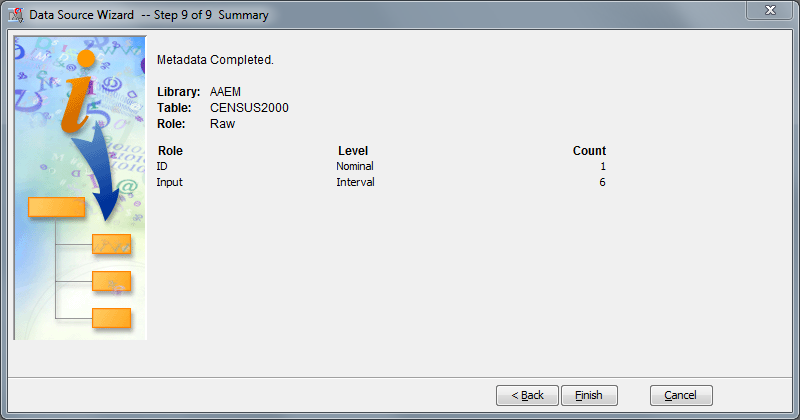
Step 6 lets you to specify the role and level for each variable in the table. A default role is assigned based on the name of a variable. For example, the variable **ID** was given the role ID.

**ID is an important designation – it means that this is the identifier for the data element (sort of like a primary key in a database table). Having a variable with an ID role is important so SAS can know how to tell the cases (rows of input data) apart.**

When a variable does not have a name corresponding to one of the possible variable roles, it will, using the Basic setting, be given the default role of **input**. An input variable is used for various types of analysis to describe a characteristic, measurement, or attribute of a record, or *case*, in a SAS table.

The metadata settings are correct for the upcoming analysis.

1. Keep selecting **Next>** until you get to Step 9. Step 9 provides summary information about the created data set.

s

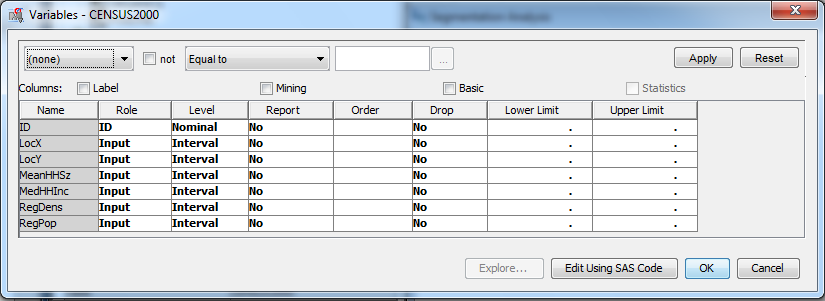
1. Select **Finish** to complete the data source definition. The **CENSUS2000** table is added to the   
   Data Sources entry in the Project panel.

Exploring and Filtering Analysis Data

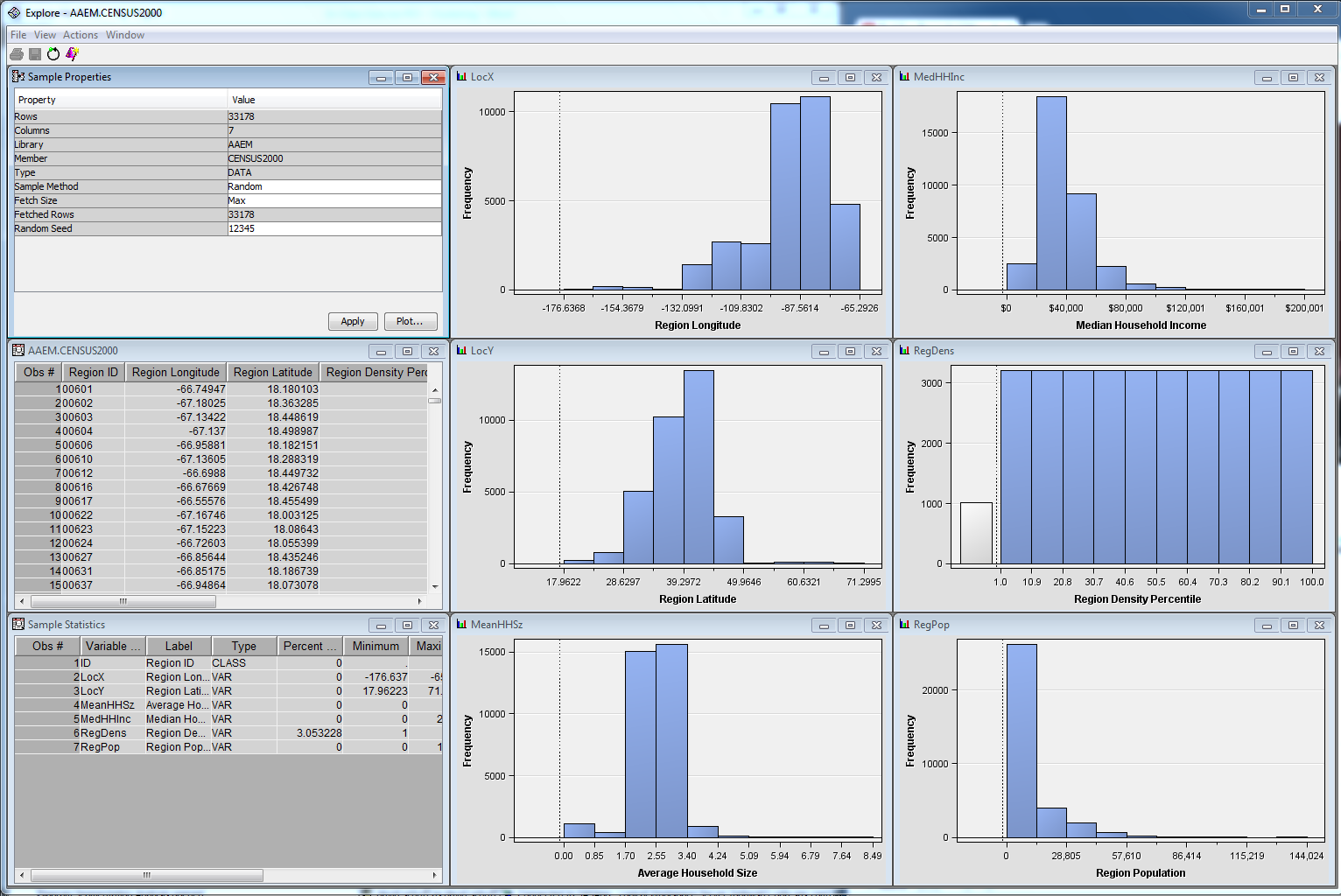
The next step in defining the data source is to explore and validate its contents. Doing this substantially reduces the chances of errors in your analysis. You can also gain insights graphically into associations between variables.

#### Data Source Exploration

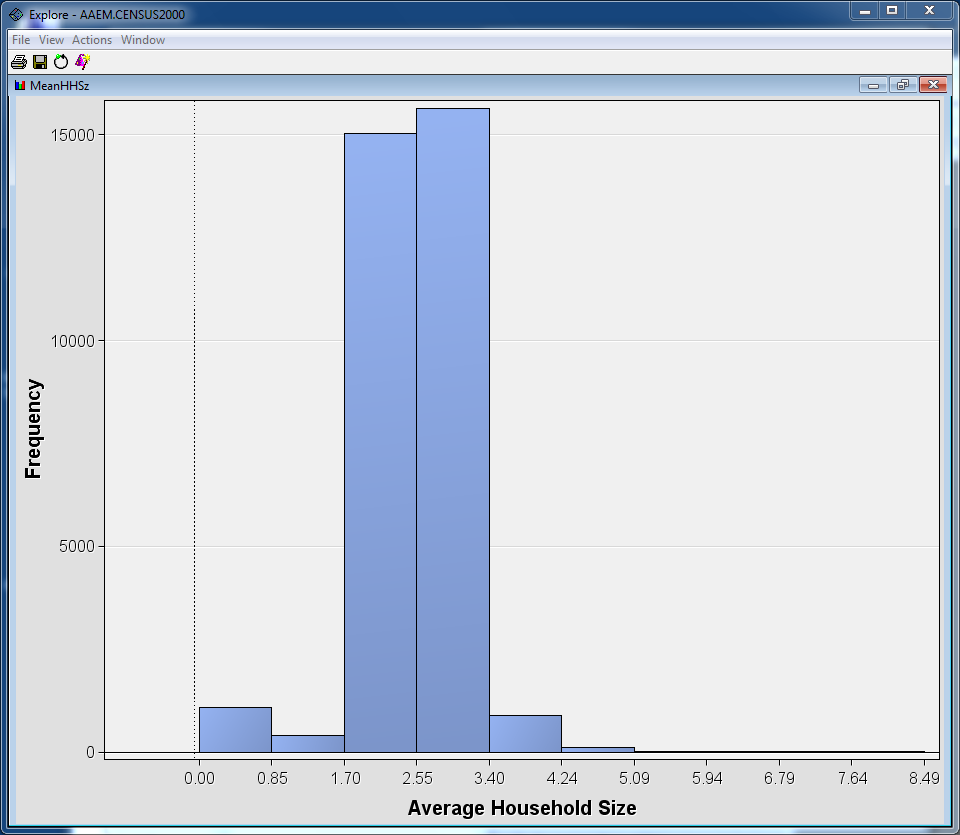
1. Right-click the **CENSUS2000** data source and select **Edit Variables…** from the shortcut menu.   
   The Variables - CENSUS2000 dialog box opens.



1. Examine histograms for the variables by selecting all listed input variables by dragging the cursor across all of the variable names or by pressing CTRL+**A**.
2. Select **Explore…**. The Explore window opens, and this time displays histograms for all of the variables in the **CENSUS2000** data source.

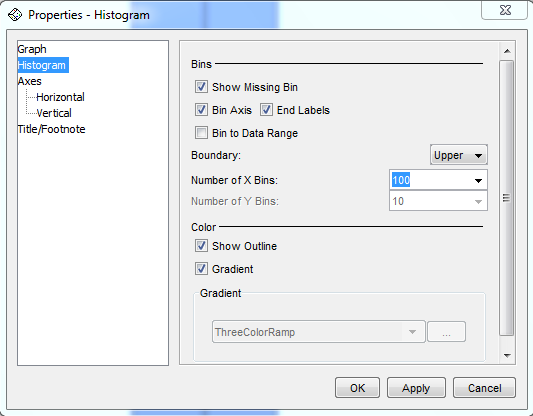
s

1. Maximize the MeanHHSz histogram by double-clicking its title bar. The histogram now fills the Explore window.

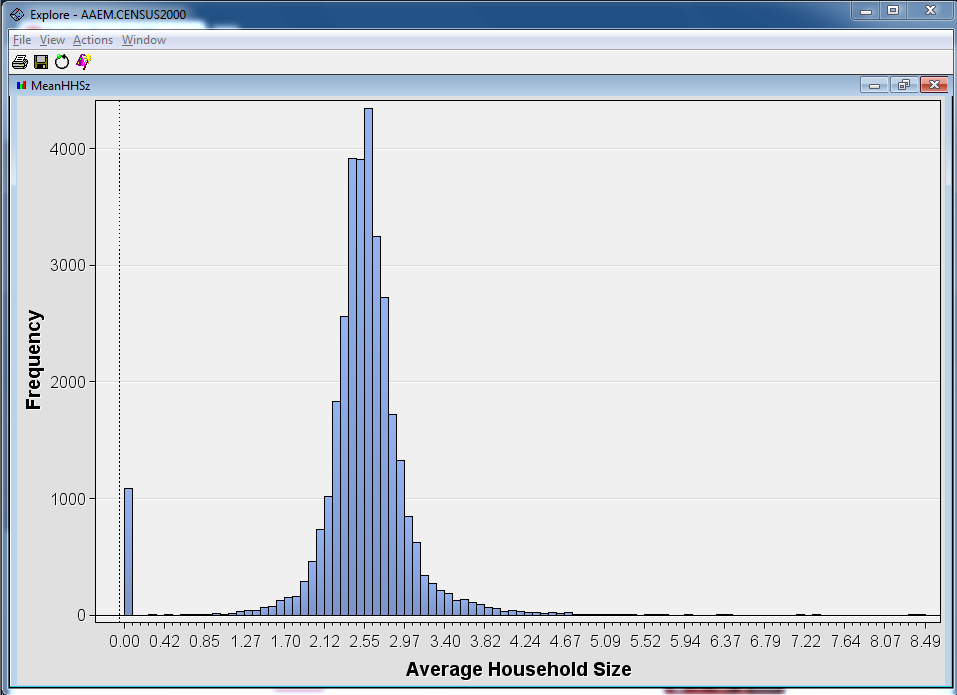


As before, increasing the number of histogram bins from the default of 10 increases your understanding of the data.

1. Right-click in the histogram window and select **Graph Properties…** from the shortcut menu.   
   The Properties - Histogram dialog box opens.  
     
   Type **100** in the Number of X Bins field and select **OK**.

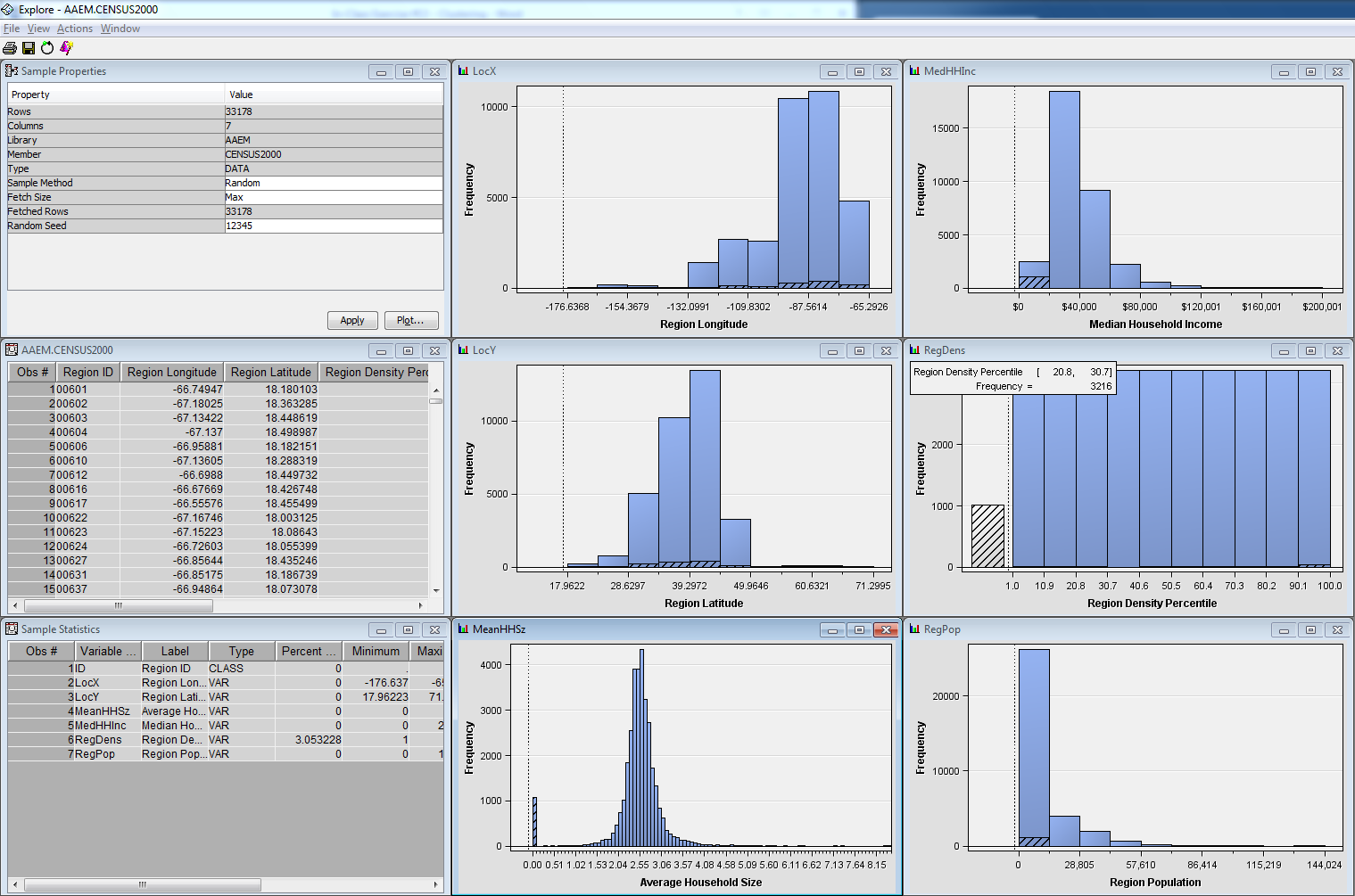


1. The histogram will reflect the change.

s

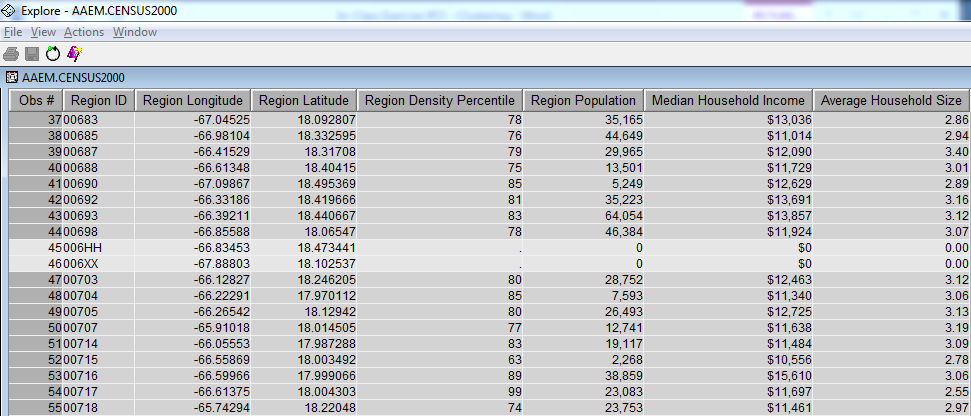
There is a curious spike in the histogram at (or near) zero. A zero household size does not make sense in the context of census data.

1. Select the bar near zero in the histogram.
2. Restore the size of the window by double-clicking the title bar of the MeanHHSize window.



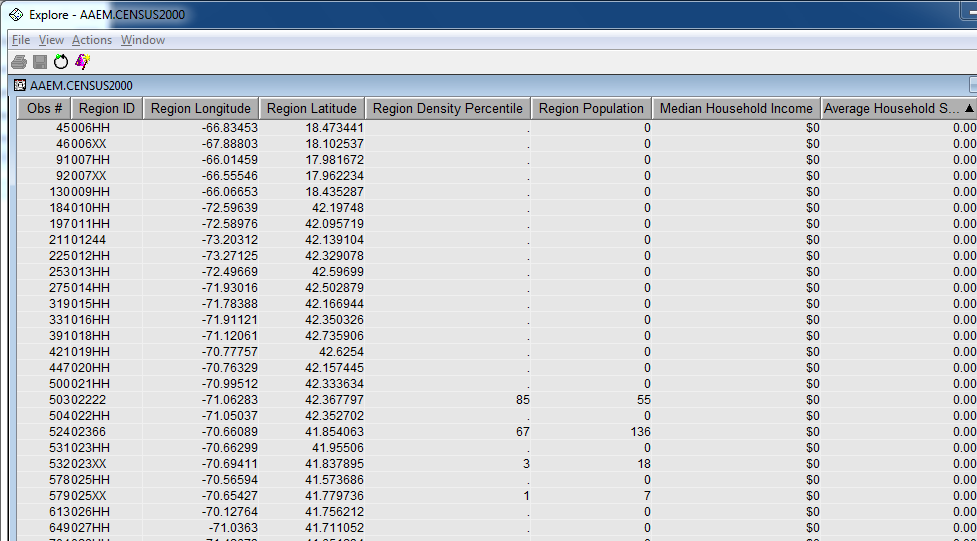
The zero values seem to be evenly distributed across the longitude, latitude, and density variables. It also seems concentrated on low incomes and populations, and makes up the majority of the missing observations in the distribution of Region Density. Now let’s look at the individual records.

1. Maximize the **CENSUS2000** data table. Scroll in the data table until you see the first highlighted row.



Records 45 and 46 (among others) have the zero Average Household Size characteristic. Other fields in these records also have unusual values.

1. Click the **Average Household Size** column heading (the last column) twice to sort the table by descending values in this field (the arrow should point up in the column heading). Cases of interest are collected at the top of the data table.



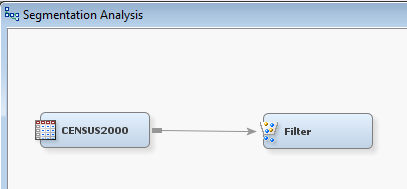
Most of the cases with zero Average Household Size have zero or missing values for the other non-geographic attributes. There are some exceptions, but it could be argued that these cases are not useful for analyzing household demographics. So we should remove these cases from the rest of the analysis.

1. Close the Explore and Variables windows.

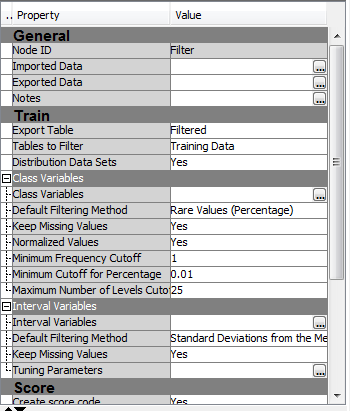
#### Case Filtering

The SAS Enterprise Miner Filter tool enables you to remove unwanted records from an analysis.   
Use these steps to build a diagram to read a data source and to filter records.

1. Drag the **CENSUS2000** data source to the Segmentation Analysis workspace window.
2. Select the **Sample** tab to access the Sample tool group.
3. Drag the **Filter** tool (fourth from the left) from the tools pallet into the Segmentation Analysis workspace window and connect it to the **CENSUS2000** data source.

s

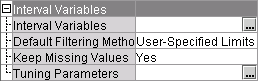
1. Select the **Filter** node and examine the Properties panel.



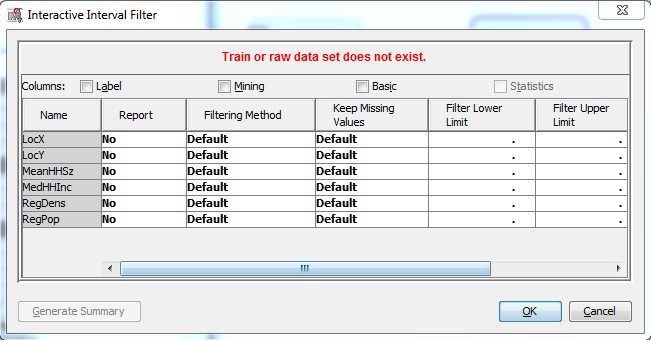
Based on the values of the properties panel, the node will, by default, filter cases in rare levels in any class input variable and cases exceeding three standard deviations from the mean on any interval input variable.

Because the **CENSUS2000** data source only contains interval inputs, only the Interval Variables criterion is considered.

1. Change the Default Filtering Method property in the Interval Variables section to **User-Specified Limits**.

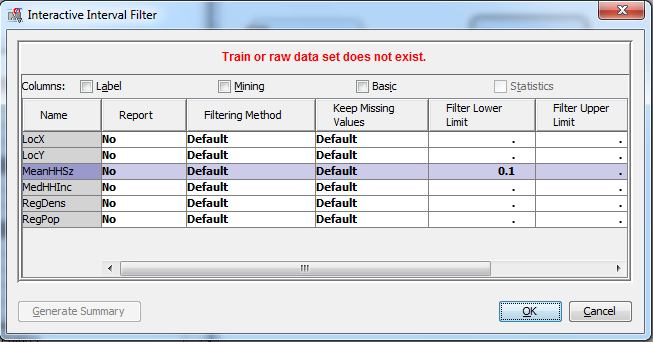


1. Select the Interval Variables ellipsis (**…**). The Interactive Interval Filter window opens.



You are warned at the top of the dialog box that the **Train** or raw data set does not exist.   
This indicates that you are restricted from the interactive filtering elements of the node, which   
are available after a node is run. You can, nevertheless, enter filtering information.

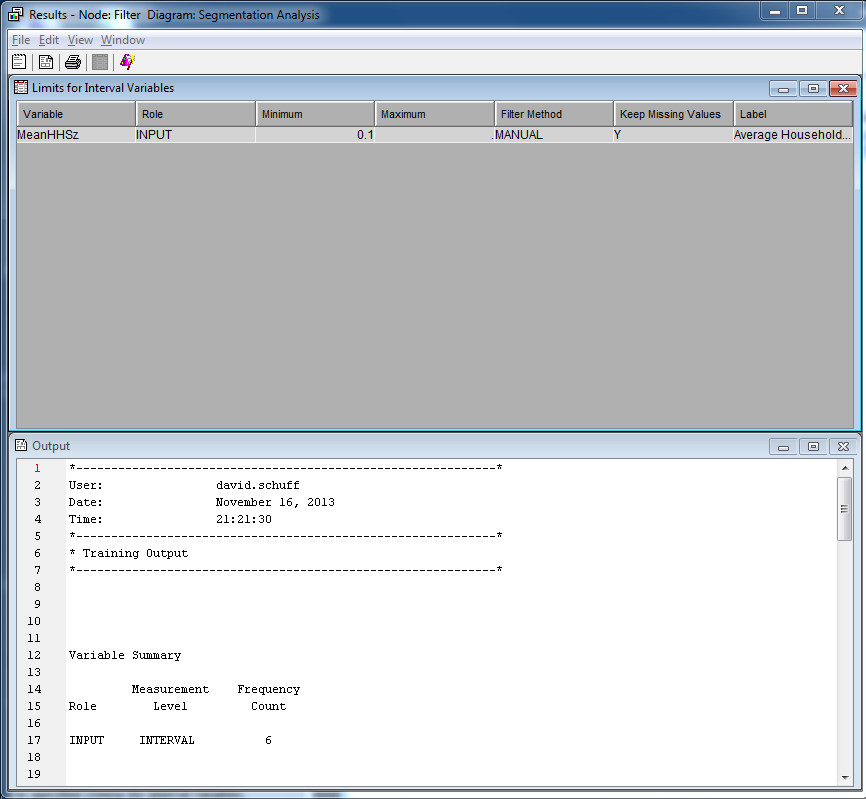
1. Type **0.1** as the Filter Lower Limit value for the input variable **MeanHHSz**.



1. Select **OK** to close the Interactive Interval Filter dialog box. You are returned to the SAS Enterprise Miner interface window.

All cases with an average household size less than 0.1 will be filtered from the rest of the analysis.

1. Run the Filter node and view the results. The Results window opens.



1. Go to line 38 in the Output window (find the section titled “Number Of Observations”).

Number Of Observations

Data

Role Filtered Excluded DATA

TRAIN 32097 1081 33178

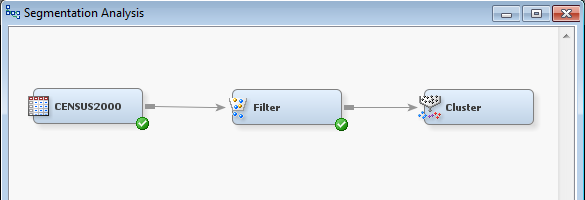
The Filter node removed 1081 cases with zero household size.

1. Close the Results window. The **CENSUS2000** data is ready for segmentation.

Setting Cluster Tool Options

The Cluster tool performs *k*-means cluster analyses, a widely used method for cluster and segmentation analysis. This demonstration shows you how to use the tool to segment the cases in the **CENSUS2000** data set.

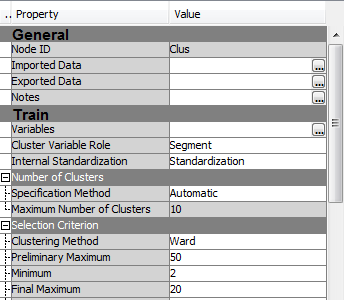
1. Select the **Explore** tab.
2. Locate and drag a **Cluster** tool into the diagram workspace.
3. Connect the **Filter** node to the **Cluster** node.



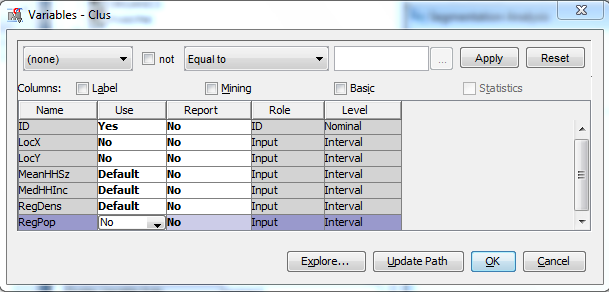
To create meaningful segments, you need to set the Cluster node to do the following:

* ignore irrelevant inputs (variables) – we only want to form clusters based on variables that matter
* standardize the inputs to have a similar range – this makes the variables directly comparable, even if they originally have different units

1. Select the **Variables** from the Train section of the Property panel for the Cluster node by clicking on the ellipsis (…). The Variables window opens.



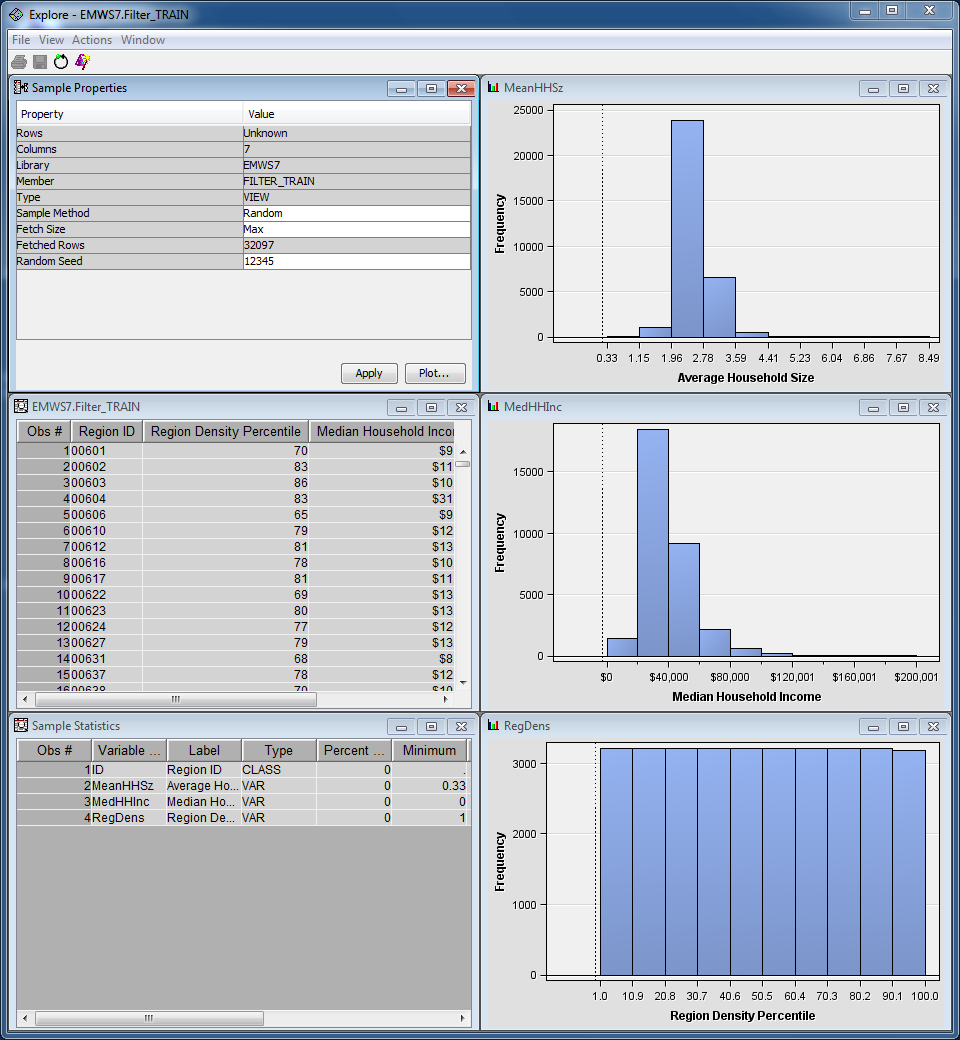
1. Select **Use** ⇨ **No** for **LocX**, **LocY**, and **RegPop**.



The Cluster node creates segments using the inputs **MedHHInc**, **MeanHHSz**, and **RegDens**.

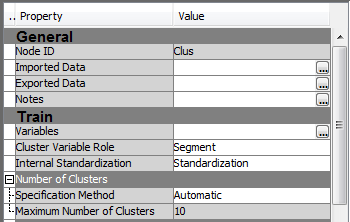
Segments are created based the distance between each case. The inputs used to create the clusters should have similar measurement scales. Calculating distances using standardized distance measurements (subtracting the mean and dividing by the standard deviation of the input values) is one way to ensure this. You can also standardize the input measurements using the Transform Variables node or by using the built-in property in the Cluster node.

1. Select the inputs **MedHHInc**, **MeanHHSz**, and **RegDens** and select **Explore…**. The Explore window opens.



You’ll notice that the inputs selected for use in the cluster are on three entirely different measurement scales (look at the y-axis for each histogram).   
  
They need to be standardized if you want to create meaningful clusters (and you do want that!).

1. Close the Explore window.
2. Select **OK** to close the Variables window. DO NOT CLICK CANCEL or it won’t exclude those three variables you marked as “No.”
3. Make sure that **Internal Standardization** ⇨ **Standardization** is selected from the Train section of the Properties pane. Distances between points are calculated based on standardized measurements.



**🖉** Another way to standardize an input is by subtracting the input’s minimum value and dividing by the input’s range. This is called *range standardization*. Range standardization rescales the distribution of each input to the unit interval, [0,1].

The Cluster node is ready to run.

Creating Clusters with the Cluster Tool

By default, the Cluster tool attempts to automatically determine the number of clusters in the data. A three-step process is used.

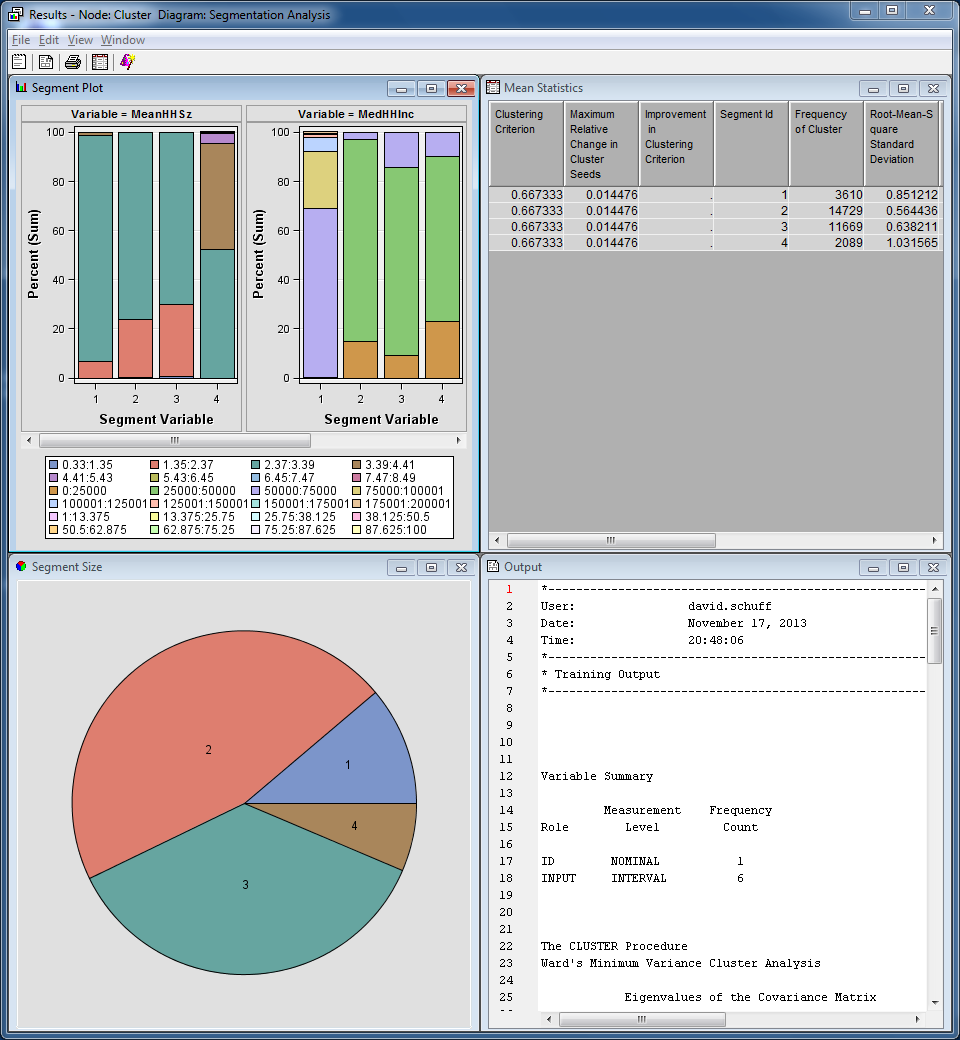
**Step 1** A large number of cluster seeds are chosen (50 by default) and placed in the input space. Cases   
in the training data are assigned to the closest seed, and an initial clustering of the data is completed. The means of the input variables in each of these preliminary clusters are substituted for the original training data cases in the second step of the process.

**Step 2** A hierarchical clustering algorithm (Ward’s method) is used to sequentially consolidate the clusters that were formed in the first step. At each step of the consolidation, a statistic named the *cubic clustering criterion* (CCC) (Sarle 1983) is calculated. Then, the smallest number of clusters that meets both of the following criteria is selected:

* The number of clusters must be greater than or equal to the number that is specified as the Minimum value in the Selection Criterion properties.
* The number of clusters must have cubic clustering criterion statistic values that are greater than the CCC threshold that is specified in the Selection Criterion properties.

**Step 3** The number of clusters determined by the second step provides the value for *k* in a *k*-means clustering of the original training data cases.

1. Run the Cluster node and select **Results…**. The Results - Cluster window opens.

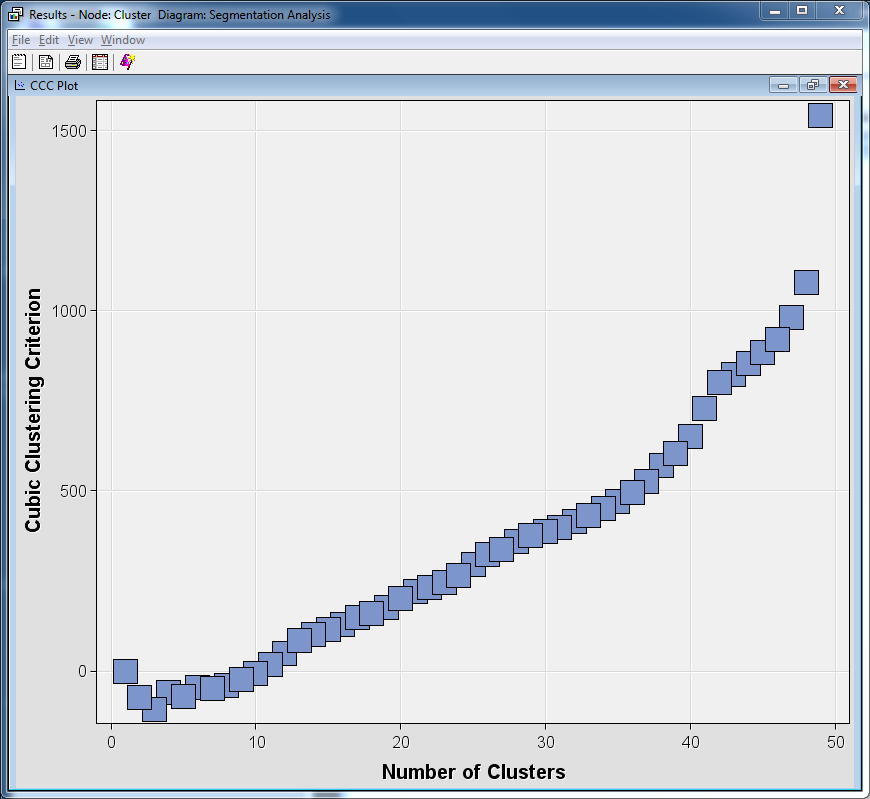


The Results - Cluster window contains four embedded windows.

* The **Segment Plot** window attempts to show the distribution of each input variable by cluster.
* The **Mean Statistics** window lists various descriptive statistics by cluster.
* The **Segment Size** window shows a pie chart describing the size of each cluster formed.
* The **Output** window shows the output of various SAS procedures run by the Cluster node.

Apparently, the Cluster node found four clusters in **CENSUS2000** data. Because the number of clusters is based on the cubic clustering criterion, it might be interesting to examine the values of this statistic for various cluster counts.

1. Select **View** ⇨ **Summary Statistics** ⇨ **CCC Plot**. The CCC Plot window opens.

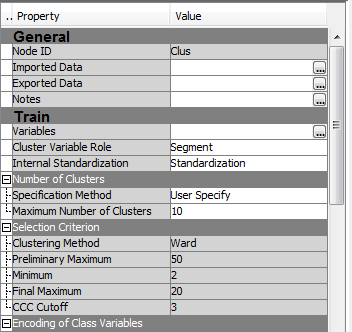


In theory, the number of clusters in a data set is revealed by the peak of the CCC versus Number of Clusters plot. However, when no distinct concentrations of data exist, the usefulness of the CCC statistic is somewhat suspect. SAS Enterprise Miner attempts to establish reasonable defaults for its analysis tools. The appropriateness of these defaults, however, strongly depends on the analysis objective and the nature of the data.  
  
Close the plot window.

Specifying the Segment Count

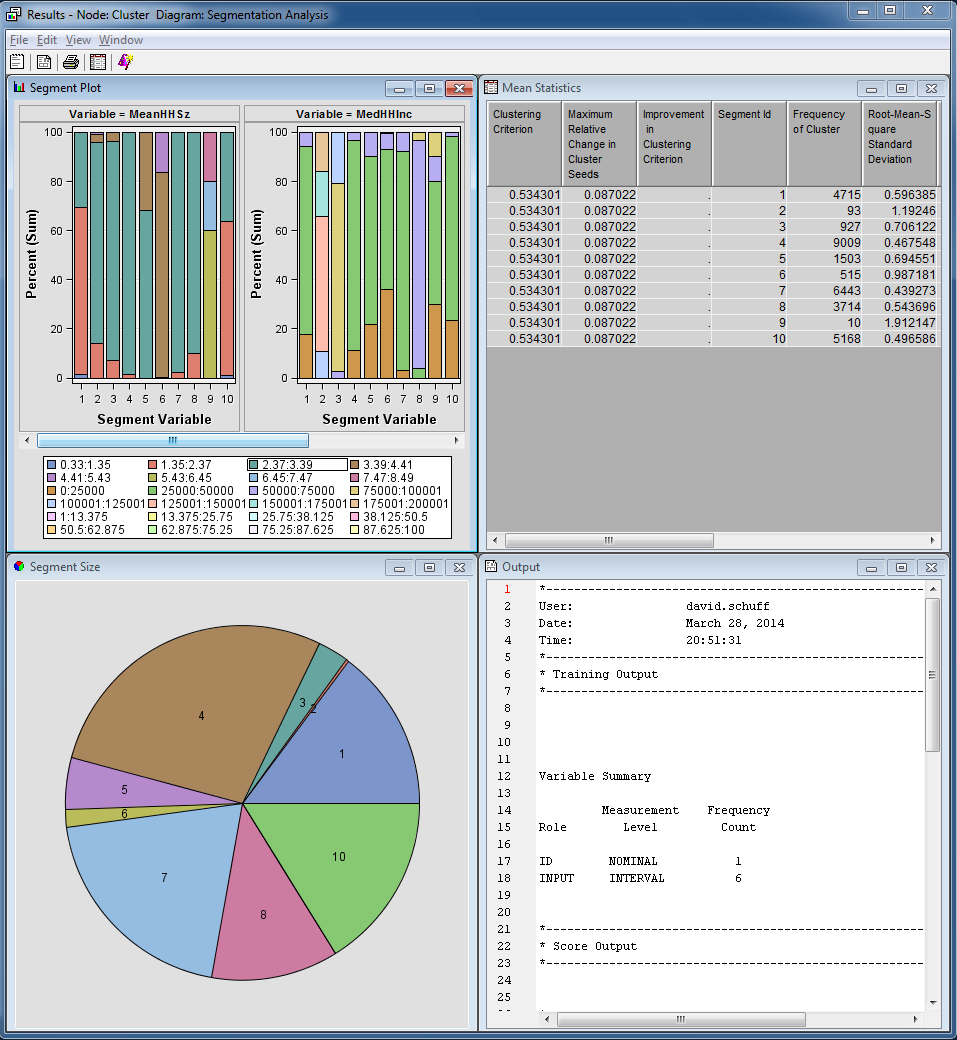
You might want to increase the number of clusters created by the Cluster node. You can do this by changing the CCC cutoff property or by specifying the desired number of clusters.

1. In the Properties panel for the Cluster node, select **Specification Method** ⇨ **User Specify**.



The User Specify setting creates a number of segments indicated by the Maximum Number   
of Clusters property listed above it (in this case, 10).

1. Run the Cluster node again and select **Results…**. The Results - Node: Cluster Diagram window opens, and shows a total of 10 generated segments.

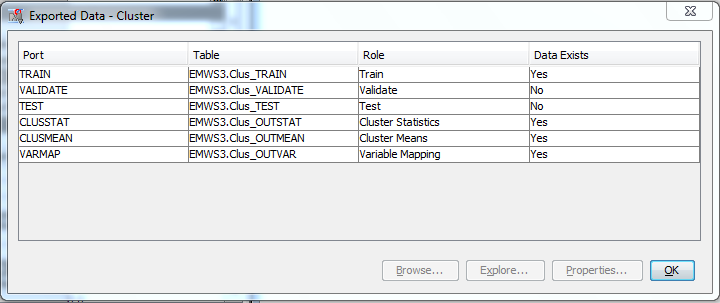


As seen in the Mean Statistics window, segment frequency counts vary from 10 cases to more than 9,000 cases.

Exploring Segments

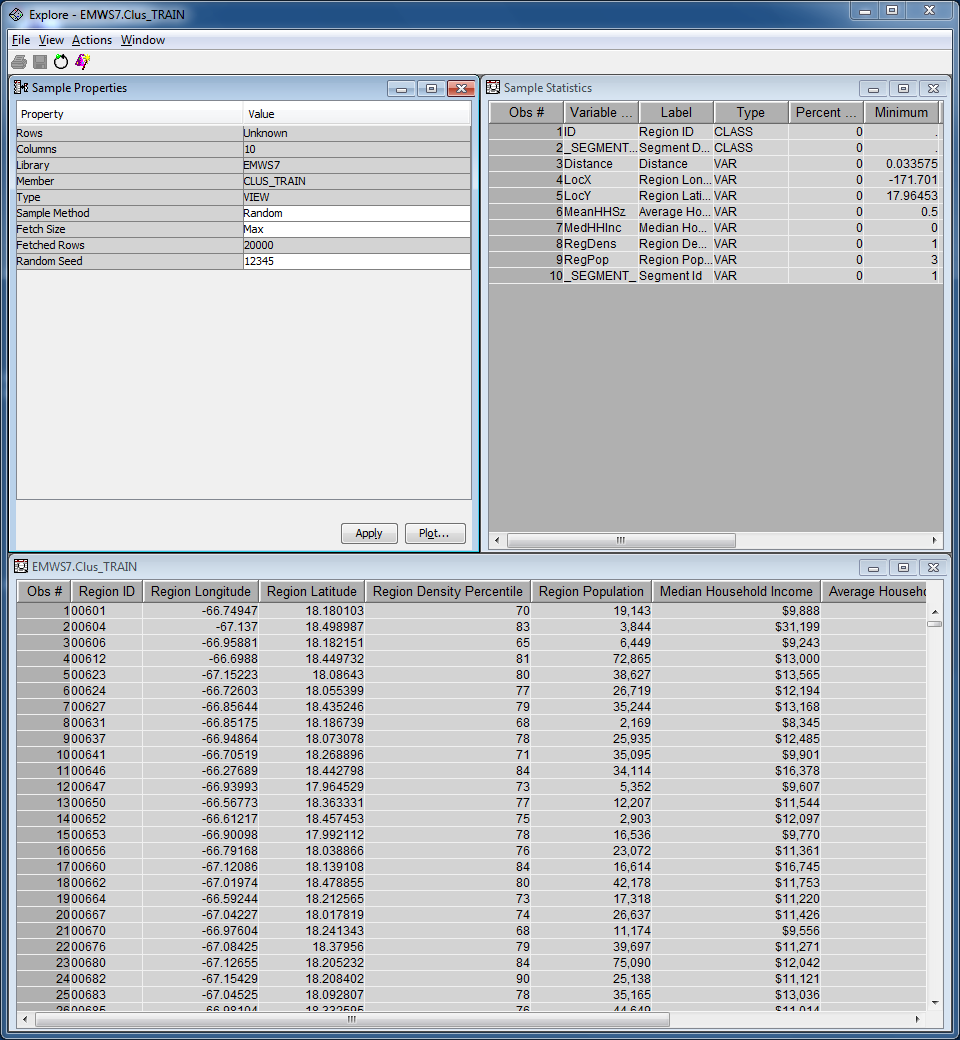
While the Results window shows a variety of data summarizing the analysis, it is difficult to understand the composition of the generated clusters. If the number of cluster inputs is small, the Graph wizard can aid in interpreting the cluster analysis.

1. Close the Results – Node Cluster window.
2. Select **Exported Data** from the Properties panel for the Cluster node by clicking on the ellipsis (…). The Exported Data - Cluster window opens.



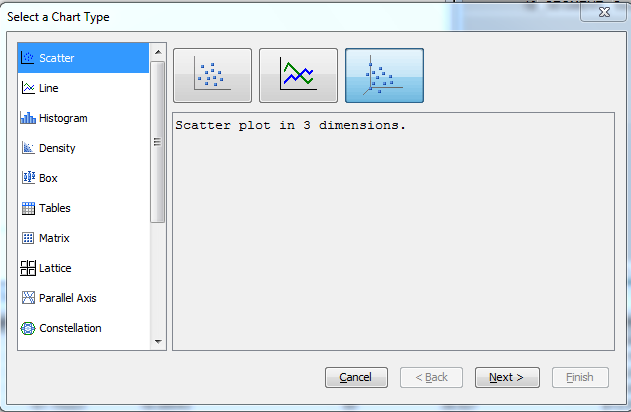
This window shows the data sets that are generated and exported by the Cluster node.

1. Select the **Train** data set and select **Explore…**. The Explore window opens.

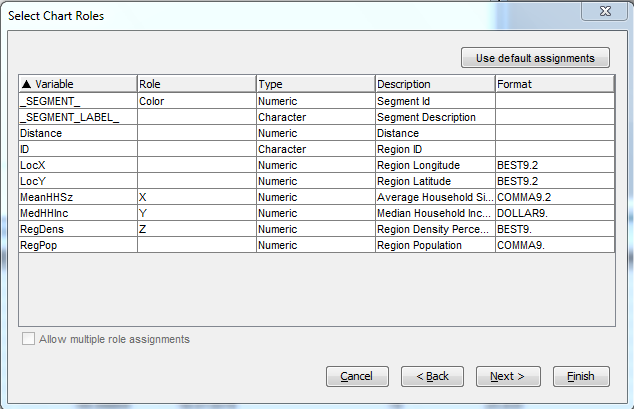


You can use the Graph Wizard to generate a three-dimensional plot of the **CENSUS2000** data.

1. Select **Actions** ⇨ **Plot**. The Select a Chart Type window opens. Select the icon for a three-dimensional scatter plot.

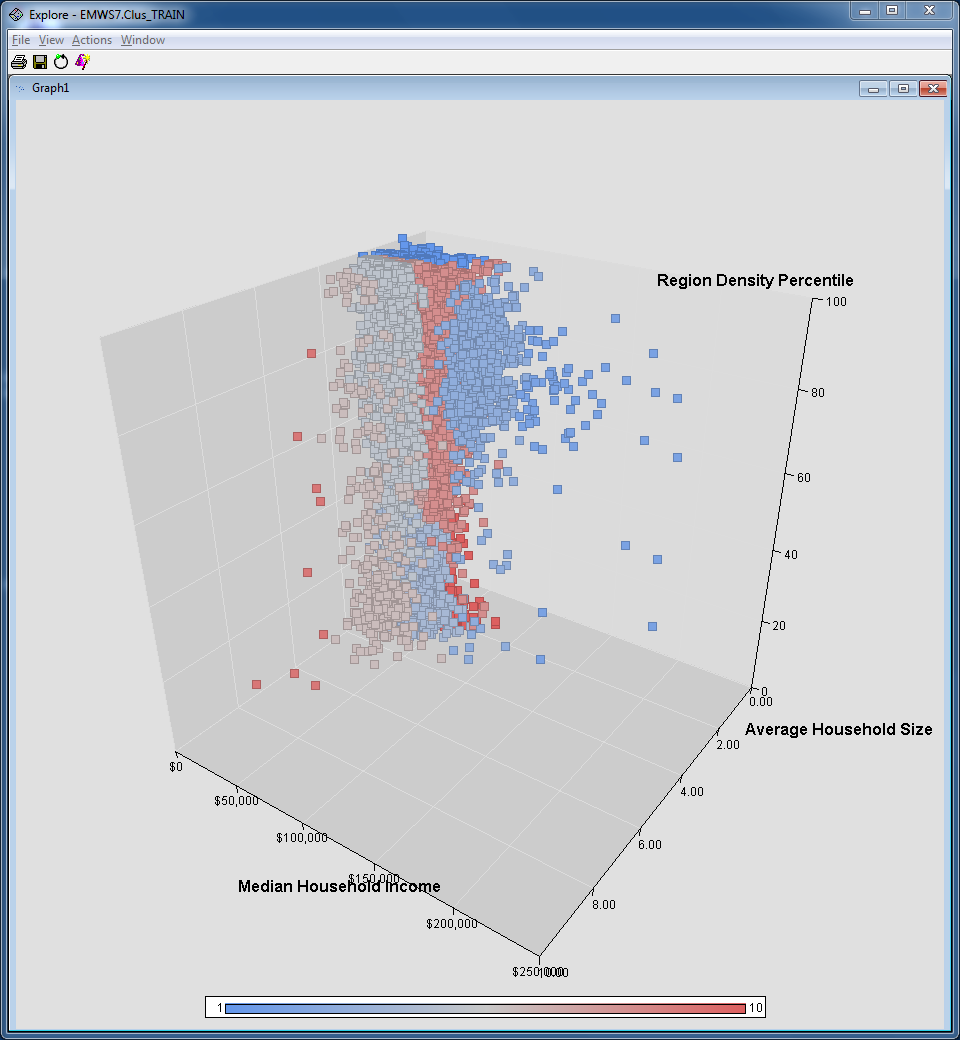


1. Select **Next >**. The Graph Wizard proceeds to the next step. Select **Chart Roles**.
2. Select roles of **X**, **Y**, and **Z** for **MeanHHSz**, **MedHHInc**, and **RegDens**, respectively.
3. Select **Role** ⇨ **Color** for **\_SEGMENT\_**.



1. Select **Finish**.

The Explore window opens with a three-dimensional plot of the **CENSUS2000** data. The graph will look basically like this:

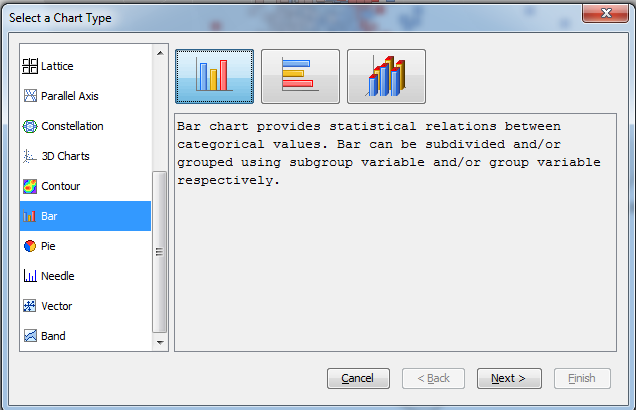


1. Rotate the plot by holding down the CTRL key and dragging the mouse.

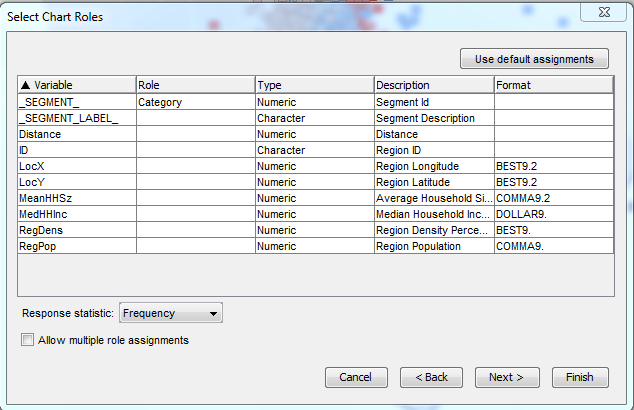
Each square in the plot represents a unique postal code. The squares are color-coded by cluster segment.

To further aid interpretability, add a distribution plot of the segment number.

1. Select **Action** ⇨ **Plot…**. The Select a Chart Type window opens.
2. Select a **Bar** chart (you may need to scroll down).

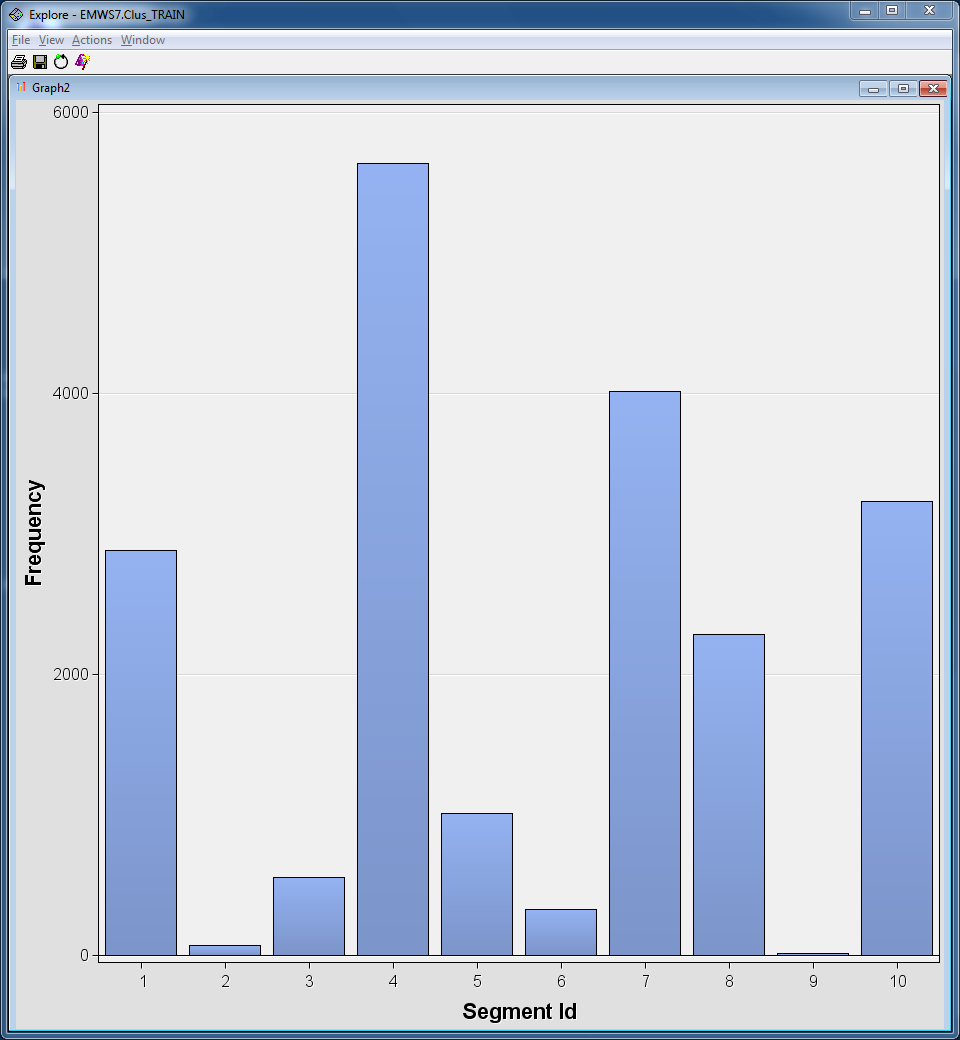


1. Select **Next >**.
2. Select **Role** ⇨ **Category** for the variable **\_SEGMENT\_**.



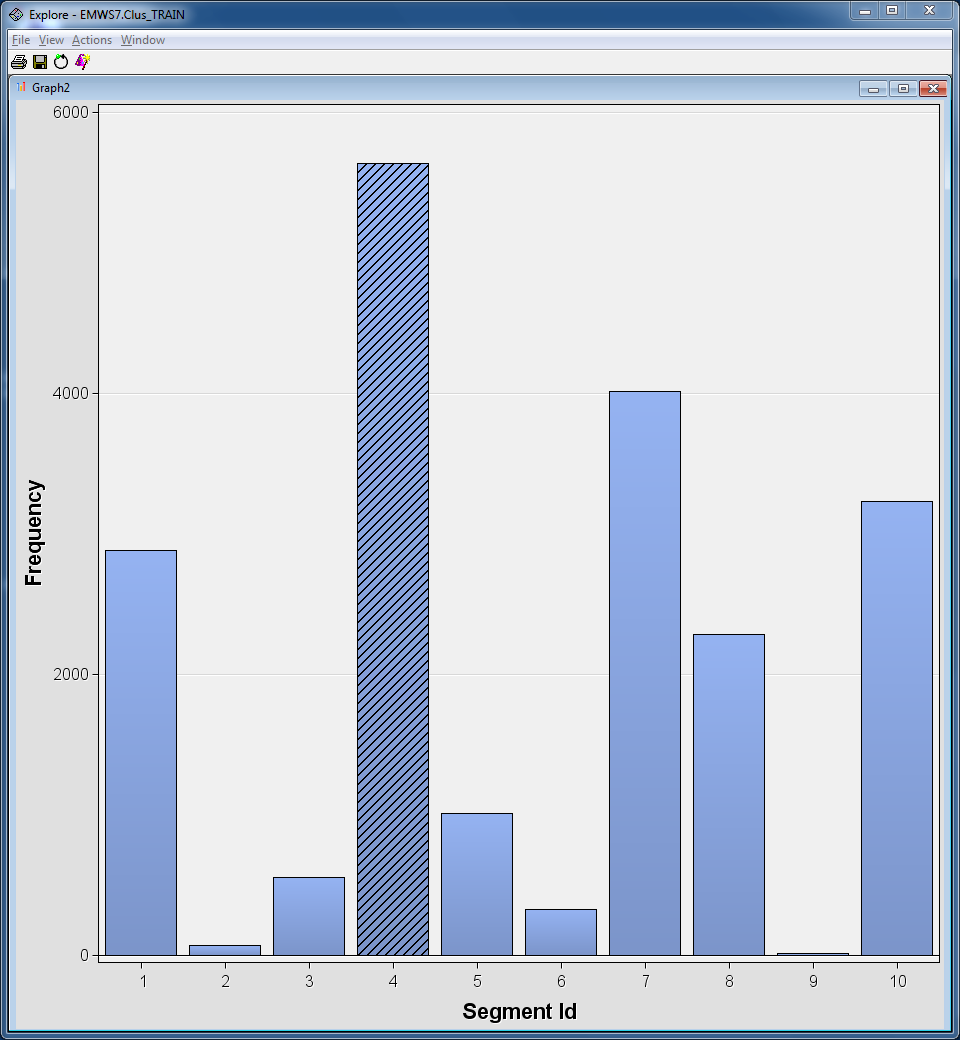
1. Select **Finish**.

A histogram of **\_SEGMENT\_** opens.

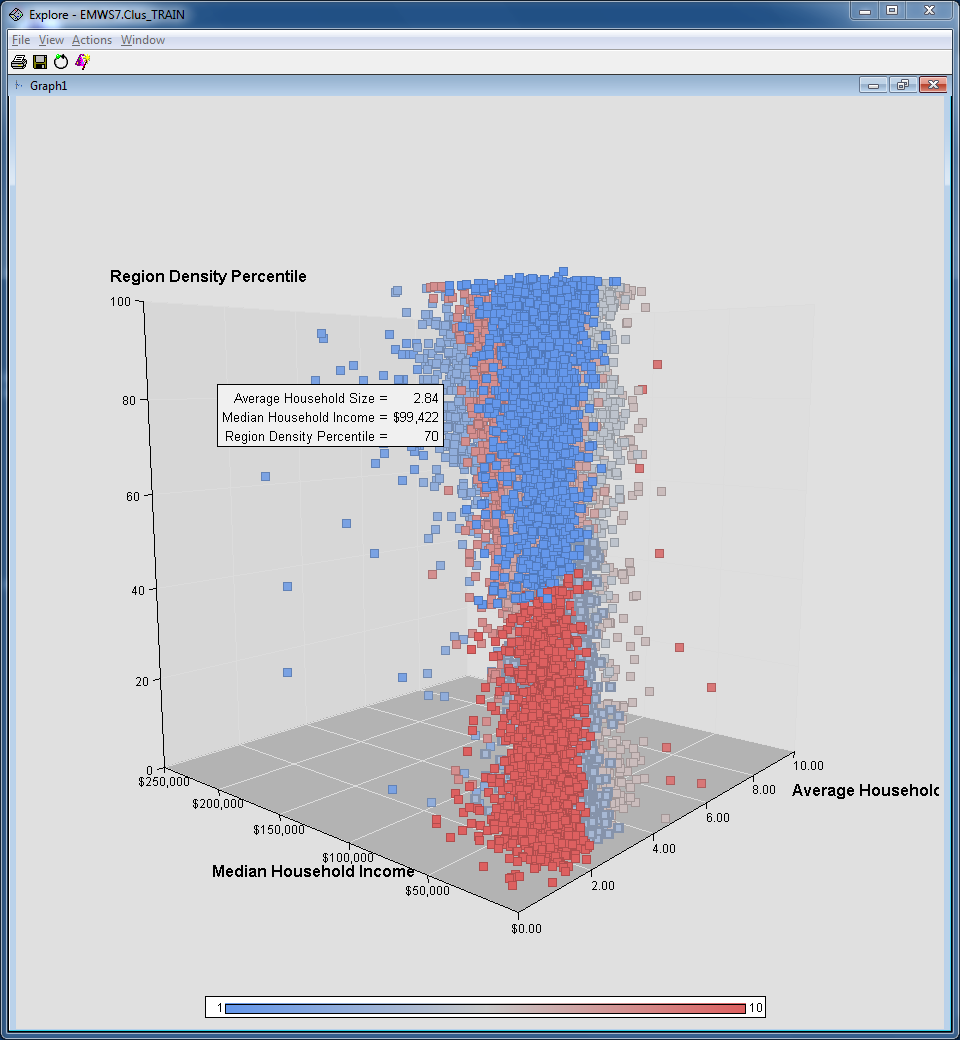


By itself, this plot is of limited use. However, when the plot is combined with the three-dimensional plot, you can easily interpret the generated segments.

1. Select the tallest segment bar in the histogram, segment 4.



1. Select the three-dimensional plot. Cases corresponding to segment 4 are highlighted. You can find the three-dimensional plot by selecting Windows/Tile and then double-clicking on the plot to maximize it.
2. Rotate the three-dimensional plot to get a better look at the highlighted cases.



Cases in this largest segment correspond to households averaging between two and three members, lower population density, and median household incomes between $20,000 and $50,000.

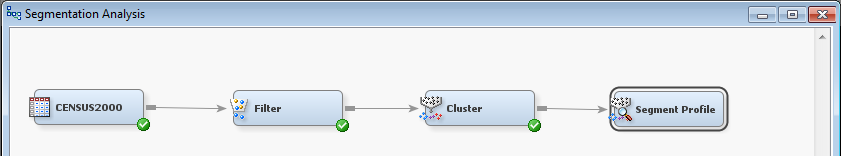
1. Click some of the other segments in the histogram to see the observations in those clusters.
2. Close the Explore, Exported Data, and Results windows.

Profiling Segments

You can gain a great deal of insight by creating plots as in the previous demonstration. Unfortunately, if more than three variables are used to generate the segments, the interpretation of such plots becomes difficult.

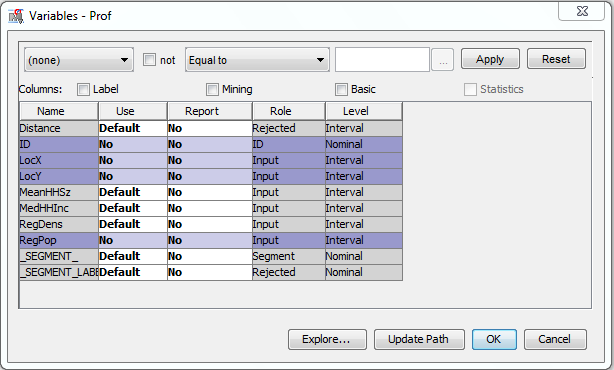
Fortunately, there is another useful tool in SAS Enterprise Miner for interpreting the composition of clusters: the Segment Profile tool. This tool enables you to compare the distribution of a variable in an individual segment to the distribution of the variable overall. As a bonus, the variables are sorted by how well they characterize the segment.

1. Drag a **Segment Profile** tool from the Assess tool palette into the diagram workspace.
2. Connect the **Cluster** node to the **Segment Profile** node.

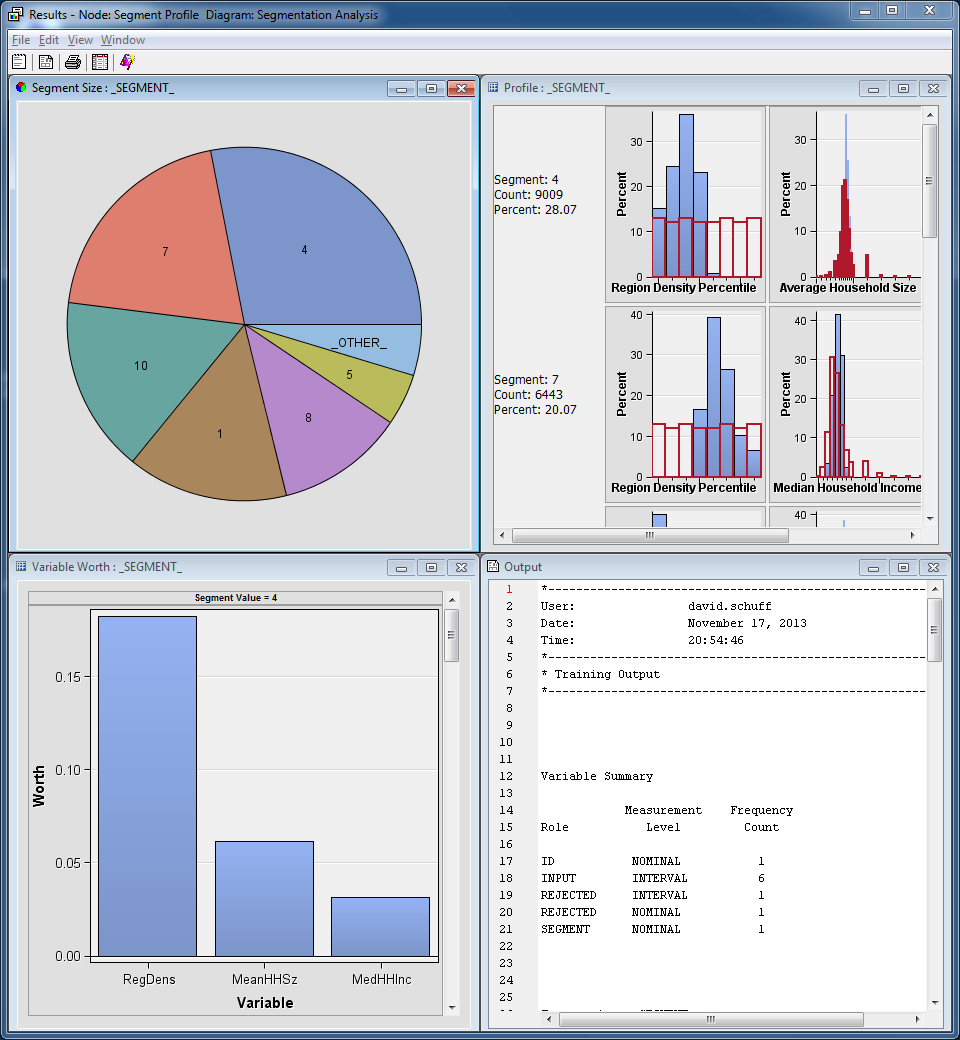


To best describe the segments, you should pick a reasonable subset of the available input variables.

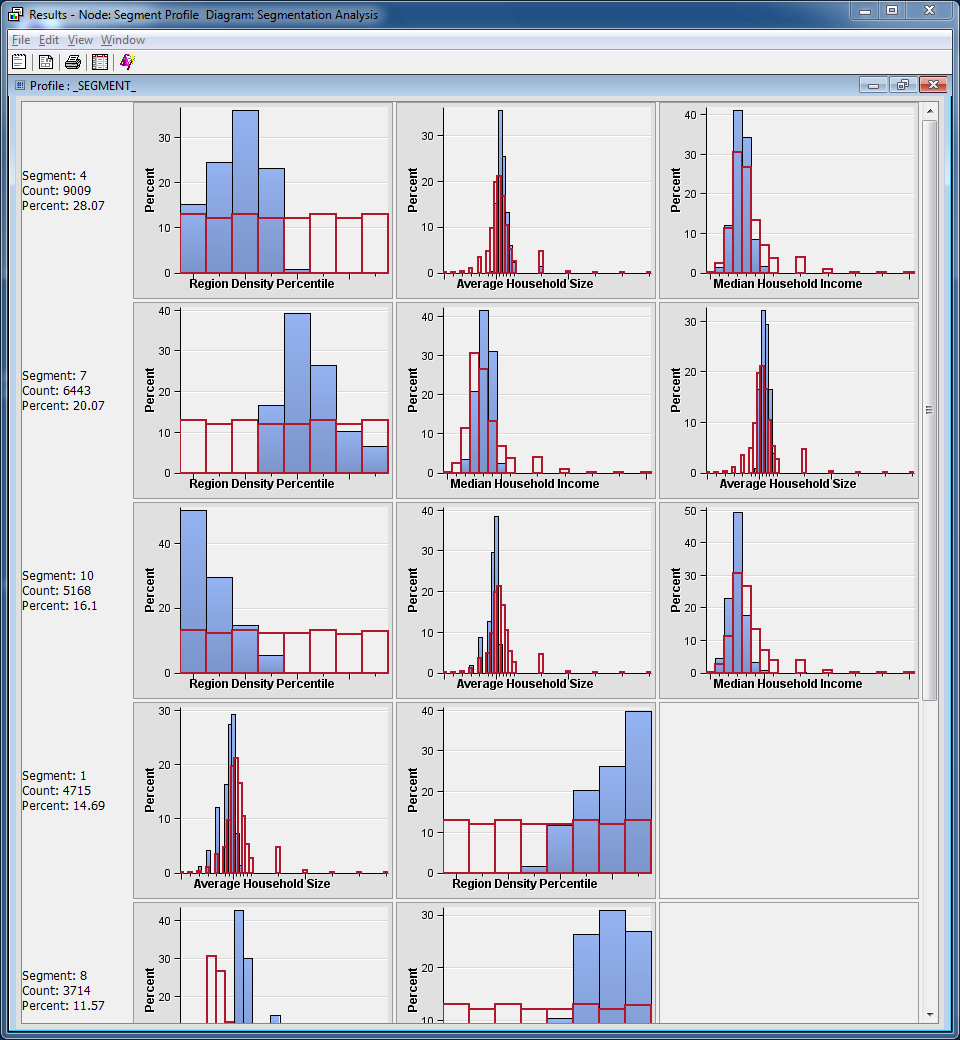
1. Select the **Edit** **Variables** property for the Segment Profile node.
2. Select **Use** ⇨ **No** for **ID**, **LocX**, **LocY**, and **RegPop** (you can CTRL+click to select multiple rows). You’ve excluded them because they aren’t meaningful for this analysis.



1. Select **OK** to close the Variables dialog box.
2. Run the Segment Profile node and select **Results…**. The Results - Node: Segment Profile Diagram window opens.



1. Maximize the Profile:\_SEGMENT\_ window.



Features of each segment become apparent. For example, segment 4, when compared to the overall distributions, has a lower Region Density Percentile, more central Median Household Income, and slightly higher Average Household Size.

You can figure this out by comparing the segment distribution (the solid bars) with the overall population (the “hollow” bars that appear as an outline).

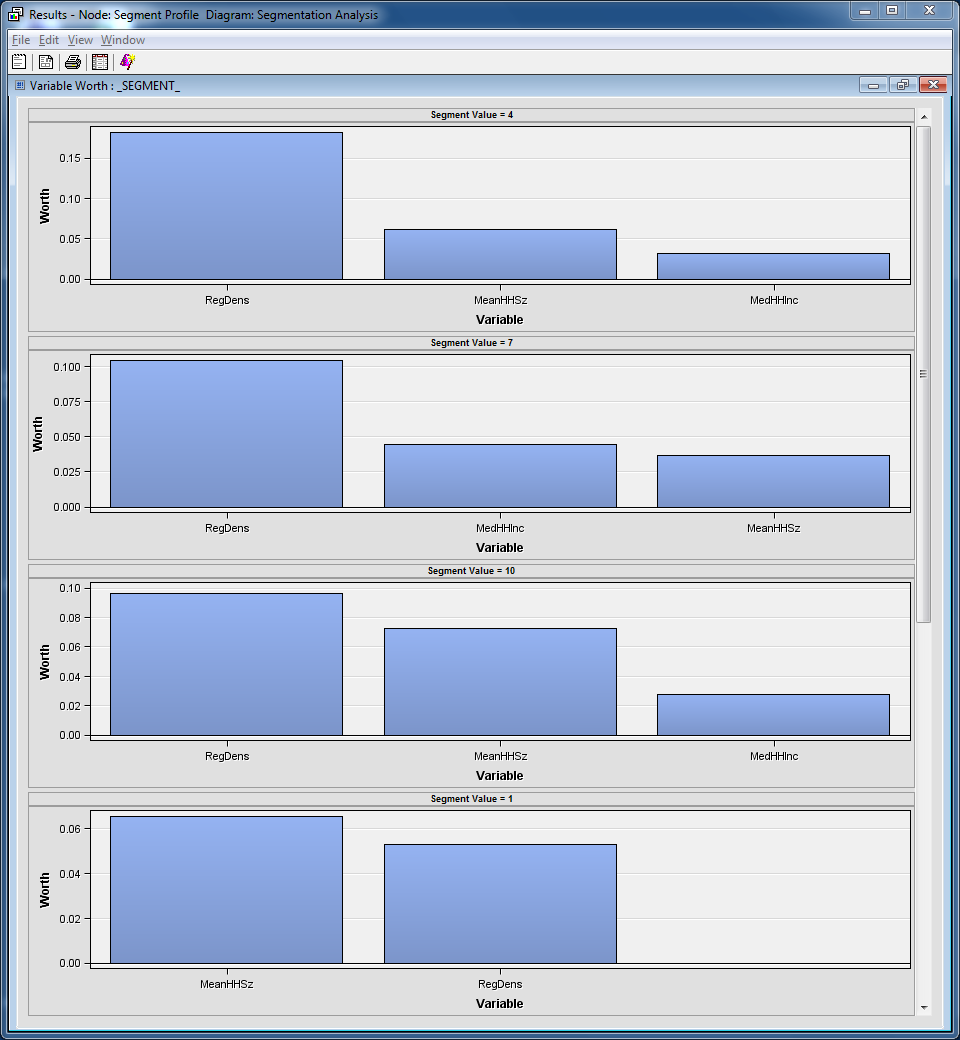
**What this means in practical terms:**

The **regions** in segment 4 are **less dense than average**. You know this because lower densities have higher frequencies – there are a lot of low density regions and very few high density regions.

**Median household income** in segment 4 are **more “central” than average**. We call it more “central” because the middle of the distribution is higher than the average, and the ends are slightly lower. So there are more “average” income households in this segment than in the rest.

**Household size** in segment 4 is **slightly higher than average**. This is a little more difficult one to see, but notice that the histogram skews more to the right than the average. So the number of households with more people is larger than the rest.

1. Maximize the Variable Worth: \_SEGMENT\_ window.



The window shows the relative worth of each variable in characterizing each segment. For example, segment 4 is largely characterized by the **RegDens** input, but the other two inputs also play a role.

Again, similar analyses can be used to describe the other segments. The advantage of the Segment Profile window (compared to direct viewing of the segmentation) is that the descriptions can be more than three-dimensional.