Practical #3: Cluster Analysis (1)

Objectives:

Learn how to use the Cluster and Segment profile tools of SAS Enterprise Miner

Submission:

- *What to submit*: a PDF report generated by the Reporter node/tool of SAS EM, containing the information (diagram, results etc.) about the exercise done by you for this practical.
- *Deadline of the submission*: 11:59PM (Adelaide Time), Tuesday of Week 5.
- *Submission link*: "**Submission Link of Prac #3"** in **Week 4 section** on Learnonline course site.
- *Marks*: Prac#3 (part of the ongoing assessment of the course) is worth 2% of the total marks of the course.

Instructions:

Cluster analysis with SAS EM

In this practical, you will learn to use the main features of the **Cluster tool** and **Segment Profile tool** of SAS Enterprise Miner, including

- Prepare for clustering
- Find clusters
- Analyse and interpret the results

Follow the steps below to compete the practical.

1. Create a project named **Prac#3**, and in the project, create a diagram named **myPrac3**.
2. Create a data source with the **CENSUS2000** data set, which is a postal code-level summary of the entire 2000 United States Census. This data set is in the same SAS folder as the PVA97NK data set we used before. It features seven variables:

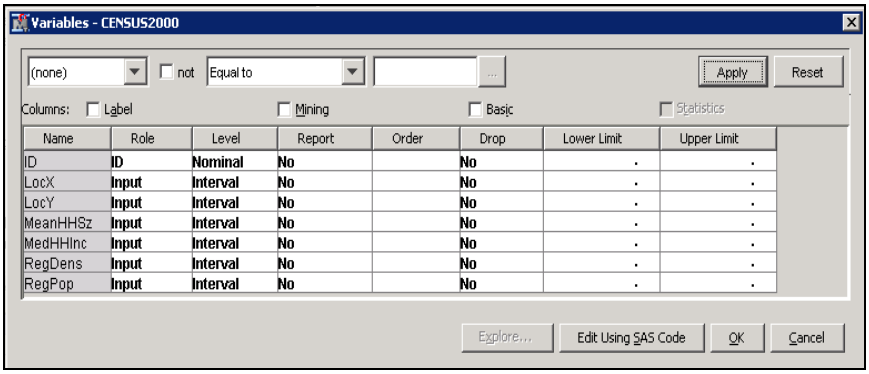
ID	postal code of the region
LOCX	region longitude
LOCY	region latitude
MEANHHSZ	average household size in the region
MEDHHINC	median household income in the region
REGDENS	region population density percentile (1=lowest density, 100=highest density)
REGPOP	number of people in the region

   Note: In the **Metadata Advisor Options** step of the Data Source Wizard, use the **Basic** setting.

3. Explore the data source.
   A worthwhile next step is to explore and validate the contents of the created data source. By assaying the prepared data, you substantially reduce the chances of erroneous results in your analysis, and you can gain insights graphically into associations between variables.

   Note: the following diagrams/screenshots are for illustration only, and in the windows you see, the values/diagrams might not be exactly the same as these shown below.

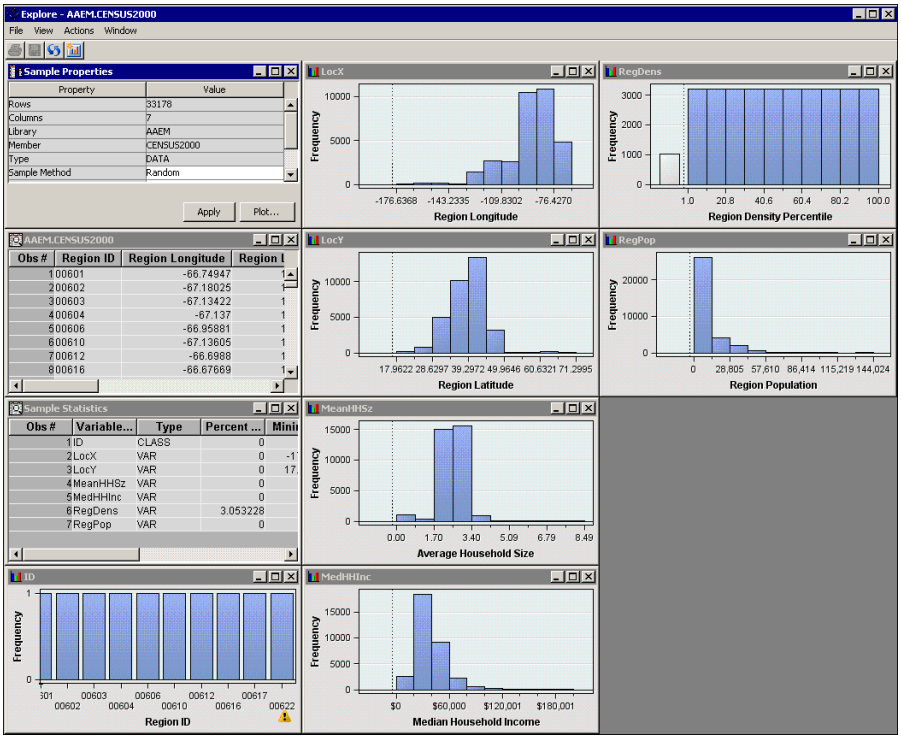
   a) Right-click the **CENSUS2000** data source and select **Edit Variables**. The **Variables - CENSUS2000** dialog box appears.



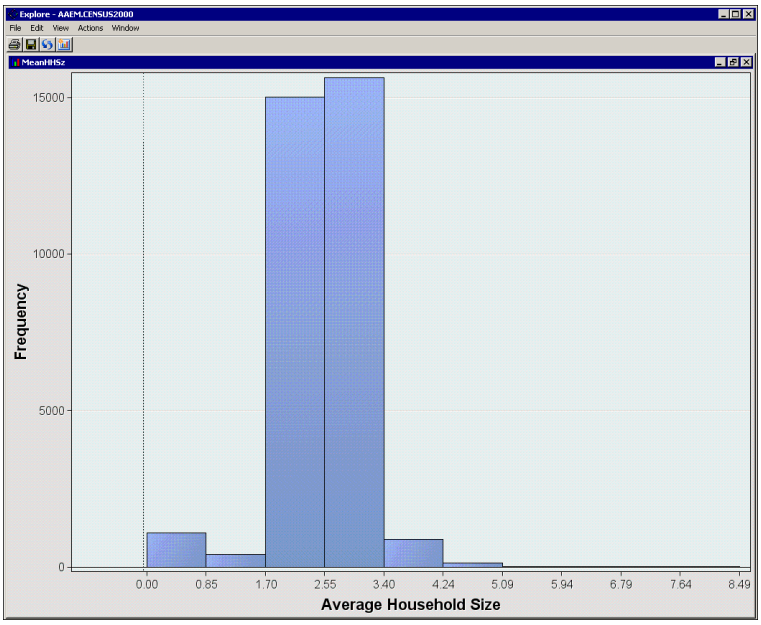
Inspect the role and level specified by SAS for each variable, to validate if they are consistent with the description of the variables in Step 2 above. Note that the Data Source Wizard assigns a default role to a variable based on the variable's name. For example, the variable ID was given the role ID based on its name. 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. ID variables are normally excluded when analyzing the data.

You can see that the role and level assigned by SAS are correct for this data source.

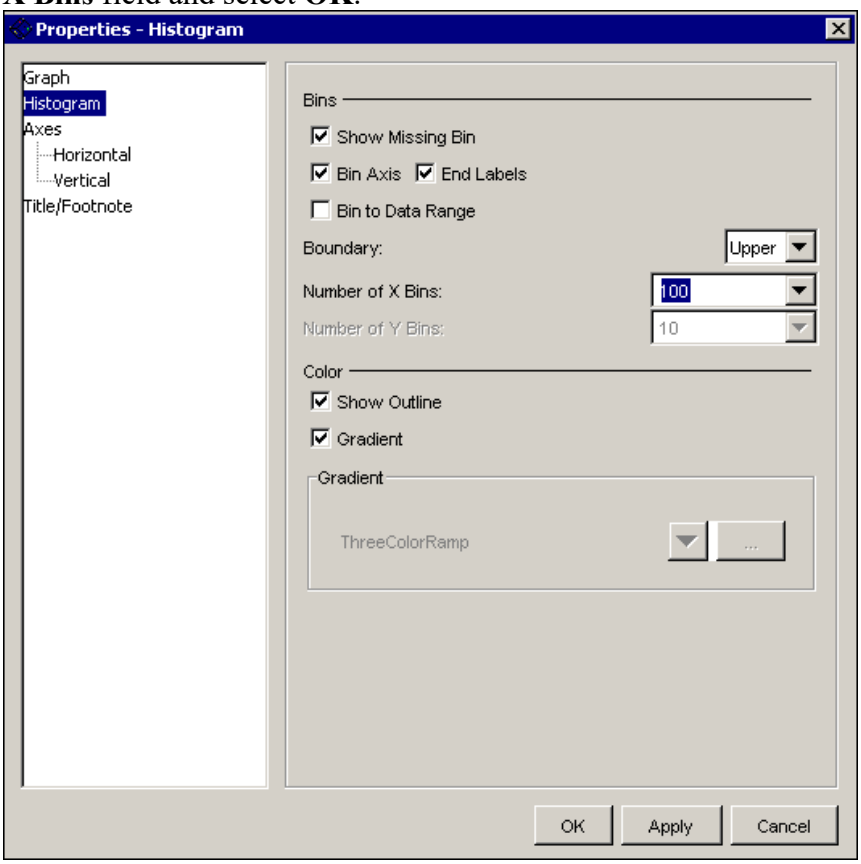
b) Select all listed inputs by dragging the cursor across all of the input names or by holding down the CTRL key and typing A. Select **Explore**. The Explore window appears, and displays histograms for all of the variables in the **CENSUS2000** data source.



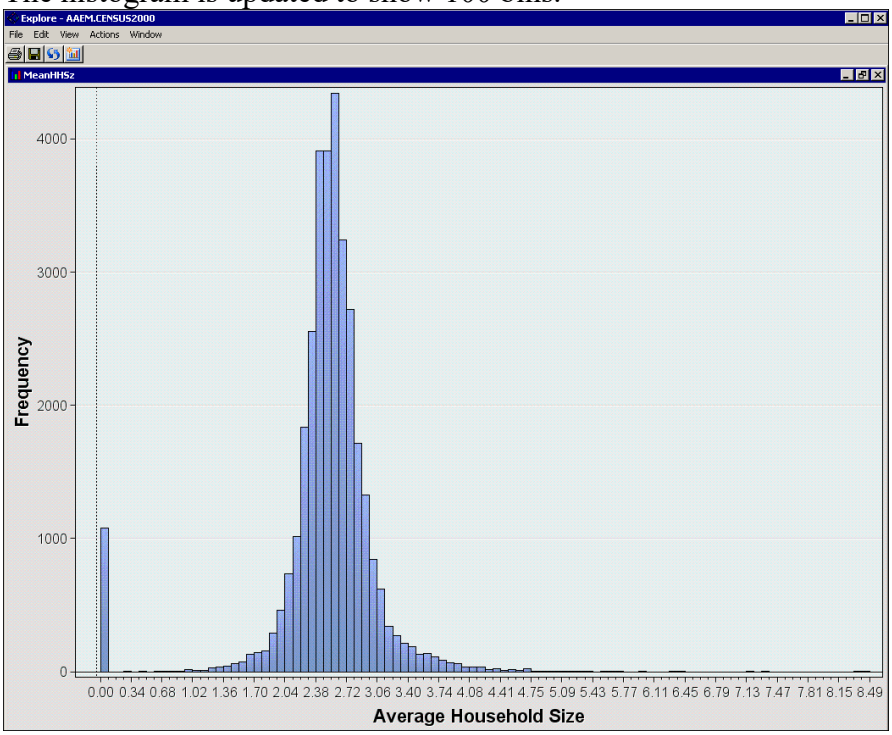
c) Maximize the **MeanHHSz** histogram by double-clicking its title bar. The histogram now fills the Explore window.



d) Increasing the number of histogram bins from the default of 10 increases your understanding of the data. You can use the **Properties - Histogram** dialog box to change the appearance of the corresponding histogram. Right-click in the histogram window and select **Graph Properties**. The **Properties - Histogram** dialog box appears. Type **100** in the **Number of X Bins** field and select **OK**.

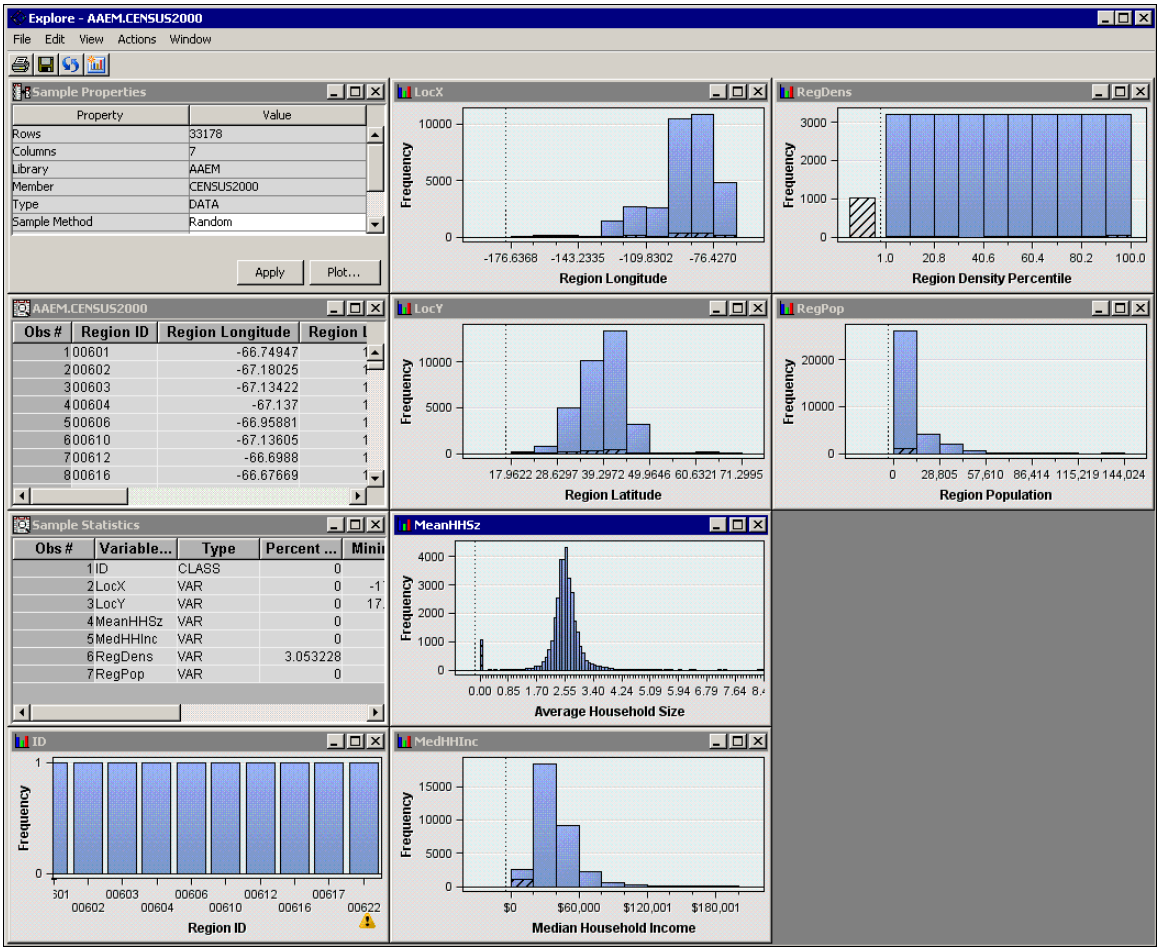


The histogram is updated to show 100 bins.



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. Select the bar near zero in the histogram. The shading of the bar changes to the pattern of parallel diagonal lines.

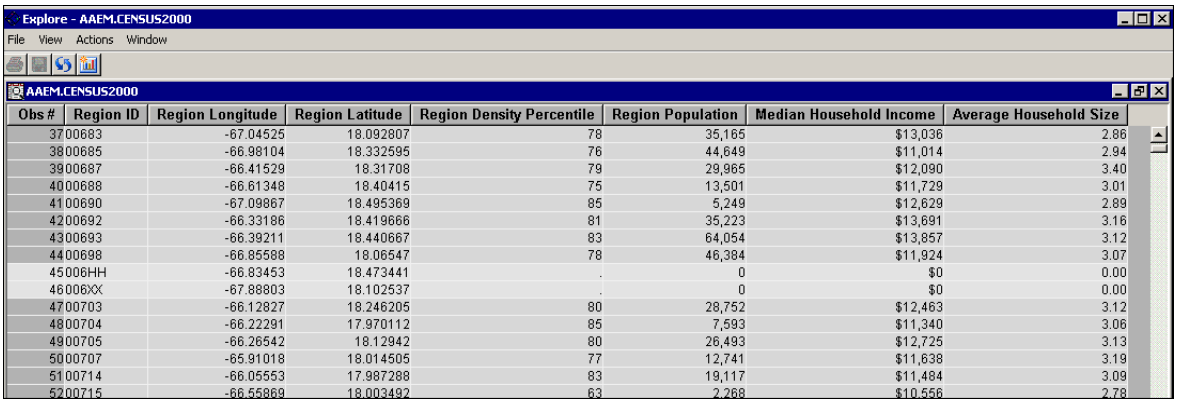
Restore the size of the window by double-clicking the title bar of the **MeanHHSz** window. The window returns to its original size.



The zero average household size (indicated by the diagonal line pattern) seems to be evenly distributed across the **longitude**, **latitude**, and **density percentile** variables. It seems concentrated on low **incomes** and **populations**, and also makes up the majority of the missing observations in the distribution of **Region Density**. It is worthwhile to look at the individual records of the explore sample.

e) Maximize the **CENSUS2000** data table (the second table in the left of the **Explore** window with the name **AAEM.CENSUS2000**).

f) Scroll in the data table until you see the first selected/highlighted row.


Explore - AAEM.CENSUS2000

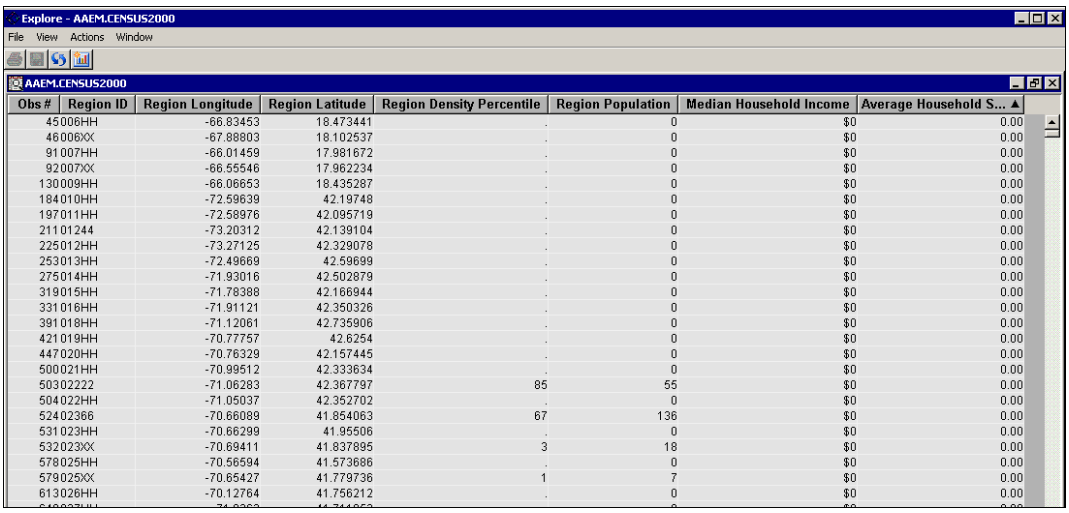
File View Actions Window

AAEM.CENSUS2000

Obs #	Region ID	Region Longitude	Region Latitude	Region Density Percentile	Region Population	Median Household Income	Average Household Size
3700683		-67.04525	18.092807	78	35,165	$13,036	2.86
3800685		-66.98104	18.332595	76	44,649	$11,014	2.94
3900687		-66.41529	18.31708	79	29,965	$12,090	3.40
4000688		-66.61348	18.40415	75	13,501	$11,729	3.01
4100690		-67.09867	18.495369	85	5,249	$12,629	2.89
4200692		-66.33186	18.419666	81	35,223	$13,691	3.16
4300693		-66.39211	18.440667	83	64,054	$13,857	3.12
4400698		-66.85588	18.06547	78	46,384	$11,924	3.07
45006HH		-66.83453	18.473441	.	0	$0	0.00
46006XX		-67.88803	18.102537	.	0	$0	0.00
4700703		-66.12827	18.246205	80	28,752	$12,463	3.12
4800704		-66.22291	17.970112	85	7,593	$11,340	3.06
4900705		-66.26542	18.12942	80	26,493	$12,725	3.13
5000707		-65.91018	18.014505	77	12,741	$11,638	3.19
5100714		-66.05553	17.987288	83	19,117	$11,484	3.09
5200715		-66.55869	18.003492	63	2,268	$10,556	2.78

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

g) Select the **Average Household Size** column heading twice to sort the table by ascending values in this field. Cases of interest are collected at the top of the data table.


Explore - AAEM.CENSUS2000

File View Actions Window

AAEM.CENSUS2000

Obs #	Region ID	Region Longitude	Region Latitude	Region Density Percentile	Region Population	Median Household Income	Average Household S... ▲
45006HH		-66.83453	18.473441	.	0	$0	0.00
46006XX		-67.88803	18.102537	.	0	$0	0.00
91007HH		-66.01459	17.981672	.	0	$0	0.00
92007XX		-66.55546	17.962234	.	0	$0	0.00
130009HH		-66.06653	18.435287	.	0	$0	0.00
184010HH		-72.59639	42.19748	.	0	$0	0.00
197011HH		-72.58976	42.095719	.	0	$0	0.00
21101244		-73.20312	42.139104	.	0	$0	0.00
225012HH		-73.27125	42.329078	.	0	$0	0.00
253013HH		-72.49669	42.59699	.	0	$0	0.00
275014HH		-71.93016	42.502879	.	0	$0	0.00
319015HH		-71.78388	42.166944	.	0	$0	0.00
331016HH		-71.91121	42.350326	.	0	$0	0.00
391018HH		-71.12061	42.735906	.	0	$0	0.00
421019HH		-70.77757	42.6254	.	0	$0	0.00
447020HH		-70.76329	42.157445	.	0	$0	0.00
500021HH		-70.99512	42.333634	.	0	$0	0.00
50302222		-71.06283	42.367797	85	55	$0	0.00
504022HH		-71.05037	42.352702	.	0	$0	0.00
52402366		-70.66089	41.854063	67	136	$0	0.00
531023HH		-70.66299	41.95506	.	0	$0	0.00
532023XX		-70.69411	41.837895	3	18	$0	0.00
578025HH		-70.56594	41.573686	.	0	$0	0.00
579025XX		-70.65427	41.779736	1	7	$0	0.00
613026HH		-70.12764	41.756212	.	0	$0	0.00

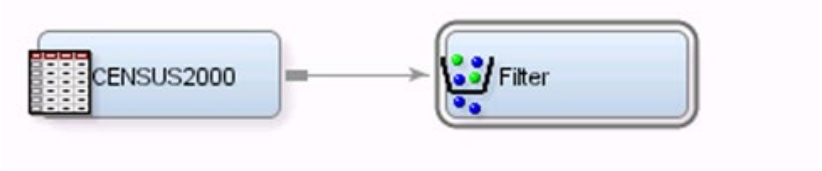
Most of the cases with zero Average Household Size have zero or missing on the remaining non-geographic attributes. There are some exceptions, but it could be argued that cases such as this are not of interest for analyzing household demographics. The next step shows how to remove cases such as this from the subsequent analyses.

Close the **Explore** and **Variables** windows.

4. Filter cases.

a) Drag the **CENSUS2000** data source to the diagram workspace to create an Input Data node for a process flow.

b) Select the **Sample** tab to access the Sample tool group and drag the **Filter** tool into the workspace and connect it to the **CENSUS2000** Input Data node.

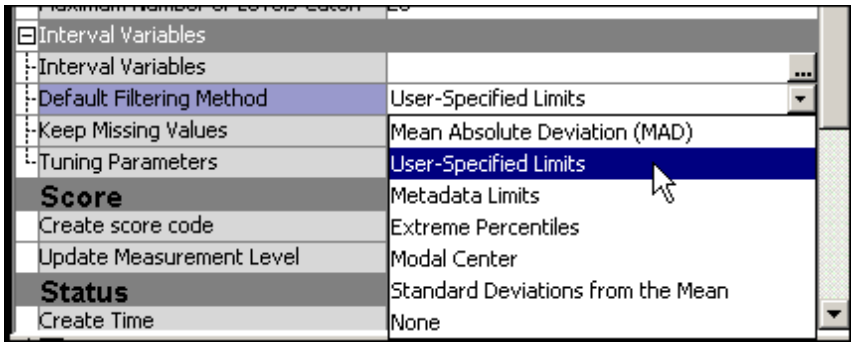


c) Select the **Filter** node and examine the Properties panel.

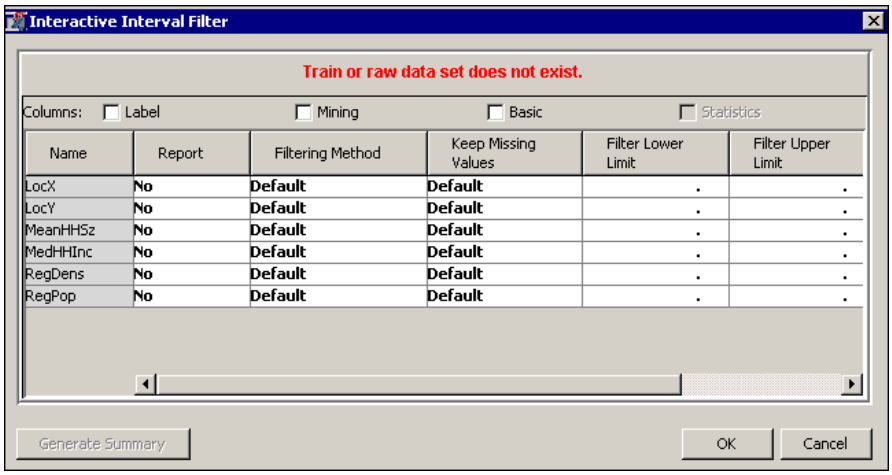
Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cutoff	25
Interval Variables	
Interval Variables	
Default Filtering Method	Standard Deviations from the Mean
Keep Missing Values	Yes
Tuning Parameters	
Score	
Create score code	Yes
Update Measurement Level	No

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 criterion in the **Interval Variables** section is considered.

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

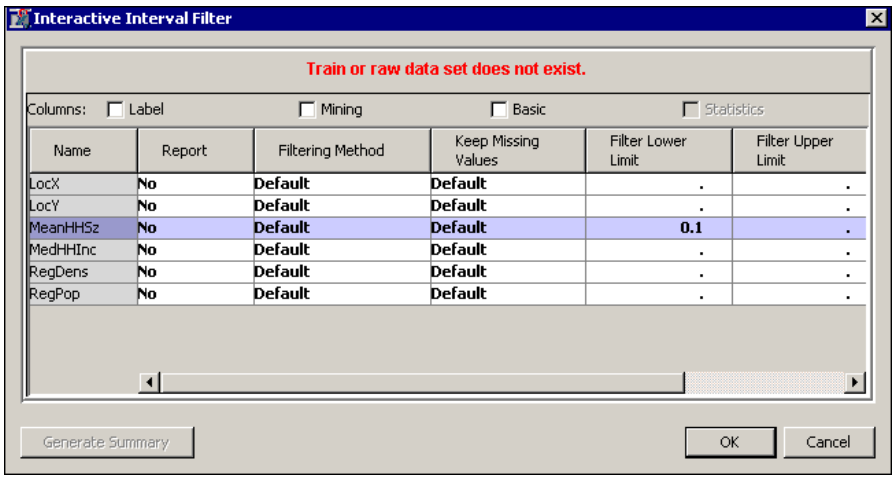


Then Select the ellipsis (**…**) next to the **Interval Variables** field. The **Interactive Interval Filter** window appears.



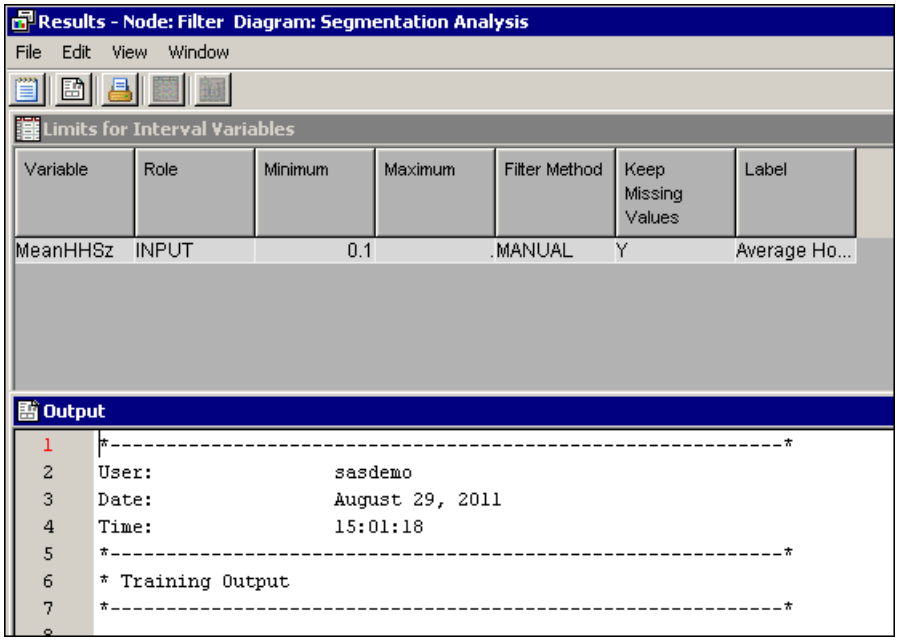
You are warned at the top of the dialog box that the Train or raw data set does not exist. This indicates that you are restricted from the interactive filtering elements of the node, which are available after a node is run. You can, nevertheless, enter filtering information.

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

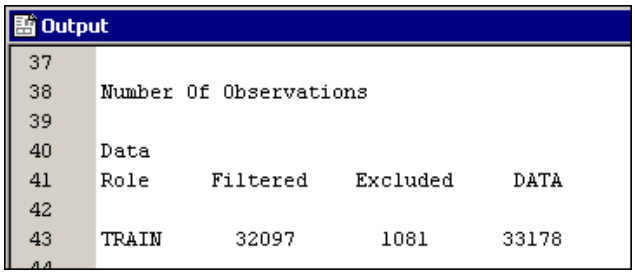


Select **OK** to close the Interactive Interval Filter dialog box. You are returned to the SAS Enterprise Miner interface window. All cases with an average household size less than 0.1 will be filtered from subsequent analysis steps.

f) Run the Filter node and view the results. The Results window appears.



g) Go to line 38 in the Output window.



The Filter node removed 1081 cases with a household size of zero.

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

5. Set Cluster Tool Options.

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

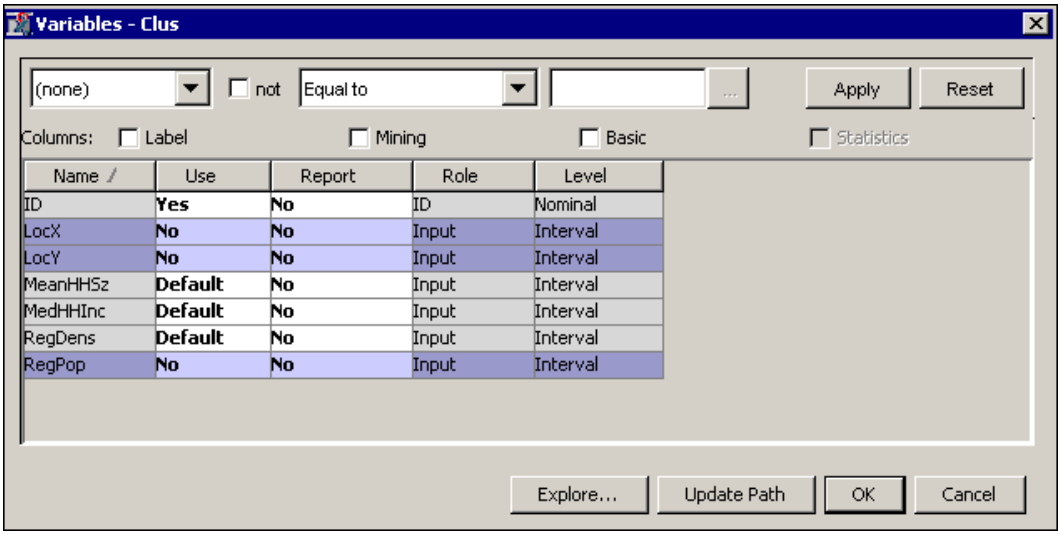
a) Select the **Explore** tab, locate and drag a **Cluster** tool into the diagram workspace, and connect the **Cluster** node to the **Filter** node

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

- ignore irrelevant inputs
- standardize the inputs to have a similar range

b) Select the **Variables** property in the Train section for the Cluster node, click on ellipsis, the Variables window appears.

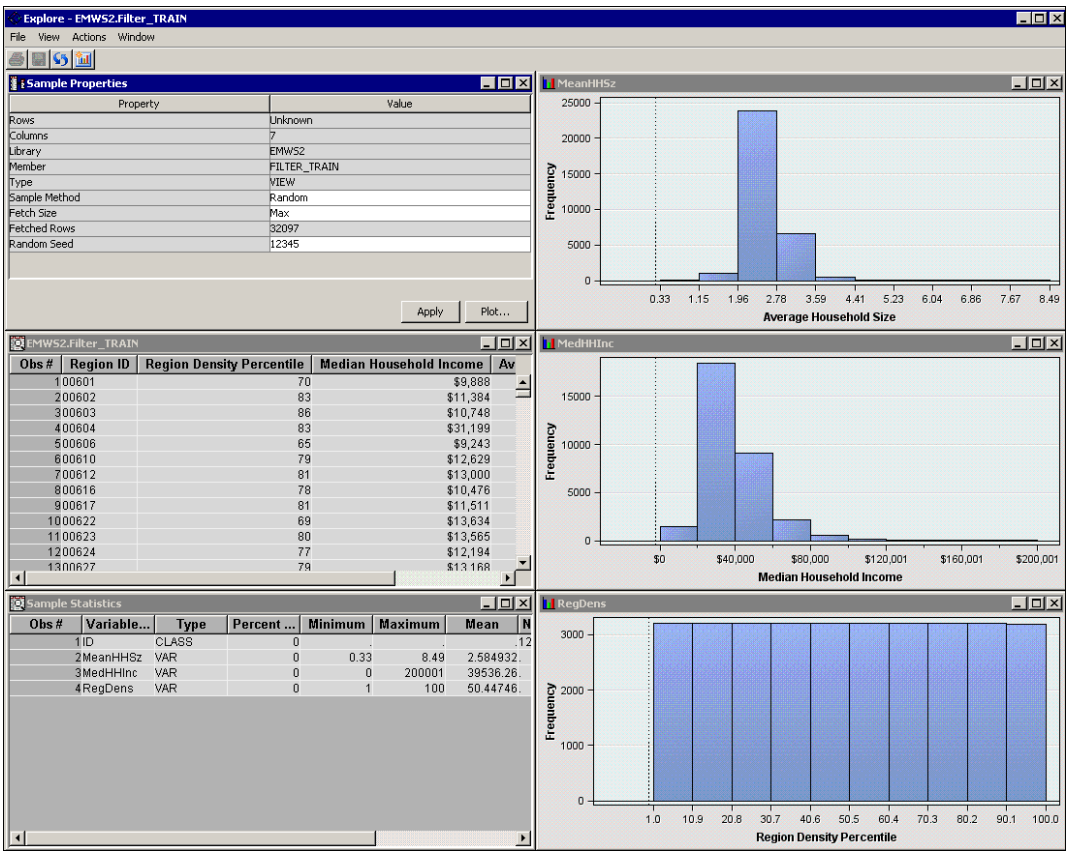
c) Select **Use → No** for **LocX**, **LocY**, and **RegPop**.



Name	Use	Report	Role	Level
ID	**Yes**	**No**	ID	Nominal
LocX	**No**	**No**	Input	Interval
LocY	**No**	**No**	Input	Interval
MeanHHSz	**Default**	**No**	Input	Interval
MedHHInc	**Default**	**No**	Input	Interval
RegDens	**Default**	**No**	Input	Interval
RegPop	**No**	**No**	Input	Interval

The Cluster node creates segments using the inputs **MedHHInc**, **MeanHHSz**, and **RegDens**. Note that variable ID is not used as mentioned before. Segments are created based on the (Euclidean) distance between each case in the space of selected inputs. If you want to use all the inputs to create clusters, these inputs 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 standardize the input measurements using the Transform Variables node. However, it is easier to use the built-in property in the Cluster node.

d) Select the inputs **MedHHInc**, **MeanHHSz**, and **RegDens** and select **Explore**. The Explore window appears.



The inputs that are selected for use in the cluster are on three entirely different measurement scales. They need to be standardized if you want a meaningful clustering.

e) Close the **Explore** window. Select **OK** to close the Variables window.

f) Note the default setting for Internal Standardization: **Internal Standardization → Standardization**. No change is required because standardization will be performed on input variables. Distances between points are calculated based on standardized measurements.

Property	Value
General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	Automatic

6. Create Clusters with the Cluster Tool, using the automatic setting to select cluster number.

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

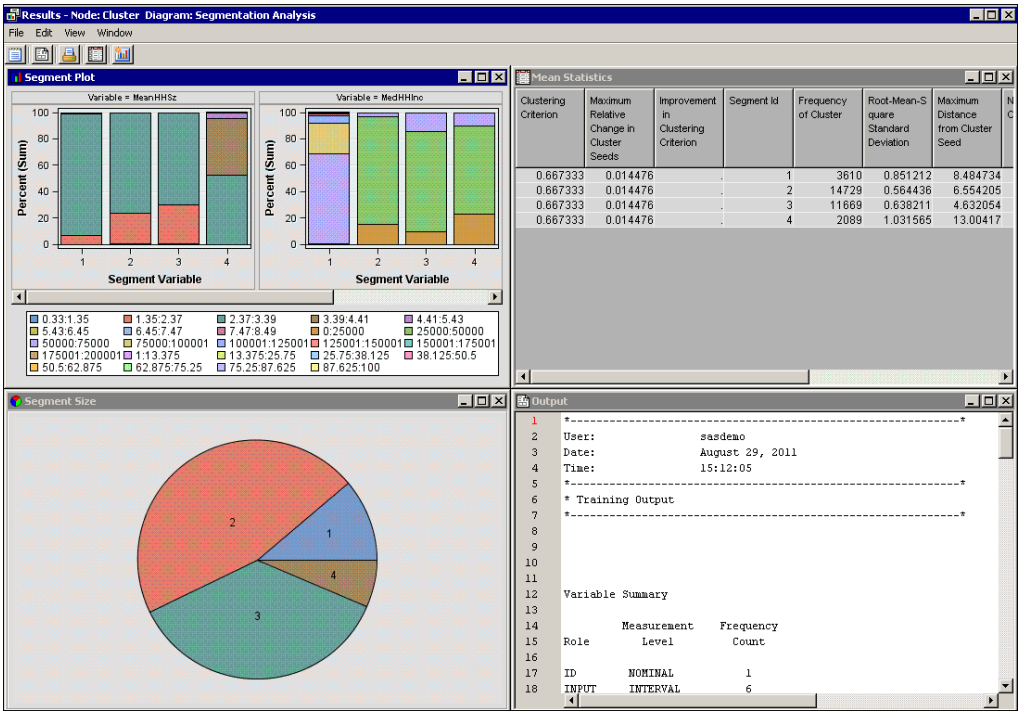
Step 1 A large number of cluster seeds are chosen (50 by default) and placed in the input space. Cases in the data set are assigned to the closest seed, and an initial clustering of the data is completed. The means of the input variables in each of these preliminary clusters are substituted for the original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

Step 2 A hierarchical clustering algorithm (Ward's method by default) 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) 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 data cases.

Run the Cluster node and select **Results**. The **Results - Cluster** window appears.



The Results - Cluster window contains four embedded windows.

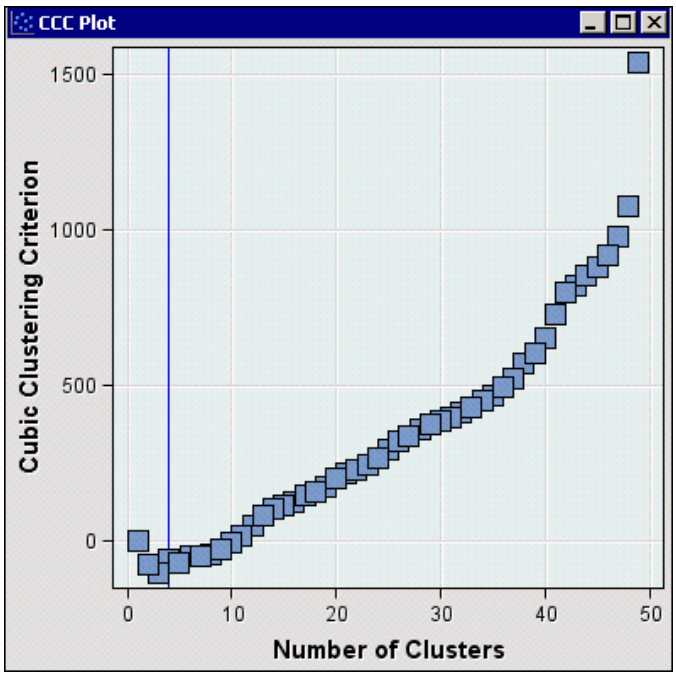
- The **Segment Plot** window attempts to show the distribution of each input variable by cluster.
- The **Mean Statistics** window lists various descriptive statistics by cluster.
- The **Segment Size** window shows a pie chart describing the size of each cluster formed.
- The **Output** window shows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

Examine and try to understand the information in each of the four embedded windows. Use the Help reference of the Cluster node to help you learn and understand the results shown in these windows. Note that the Mean Statistics contains some information which is helpful for comparing the quality of different clustering results, e.g. Root-Mean-Square Standard Deviation (indicating compactness of the clusters) and Distance to Nearest Cluster (indicating whether the clusters are well-separated).

Also explore the menu items of the Results window. Lots of useful functions can be found there, e.g., printing an embedded result window, plotting cluster distance etc.

Apparently, the Cluster node found four clusters in the **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.

Select **View → Summary Statistics → CCC Plot**. The CCC Plot window appears.



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 utility 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.

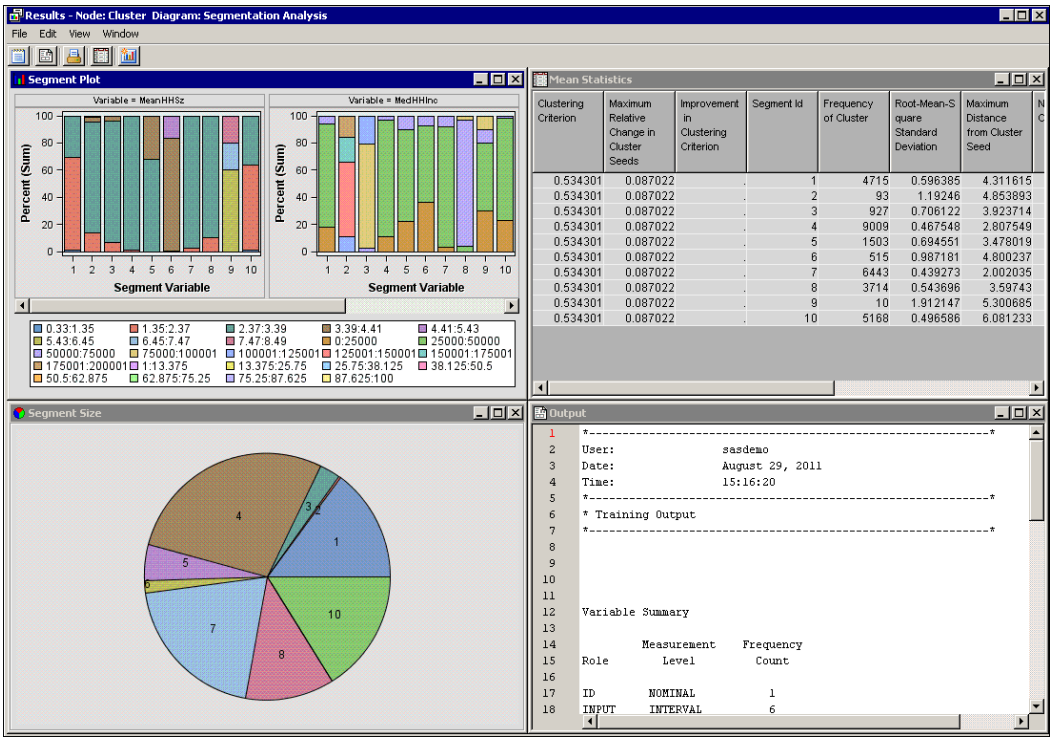
7. Specify the cluster number by yourself.
You might want to increase the number of clusters created by the Cluster node. You can do this by specifying the desired number of clusters by yourself.

a) In the Properties panel for the Cluster node, select **Specification Method → User Specify**.

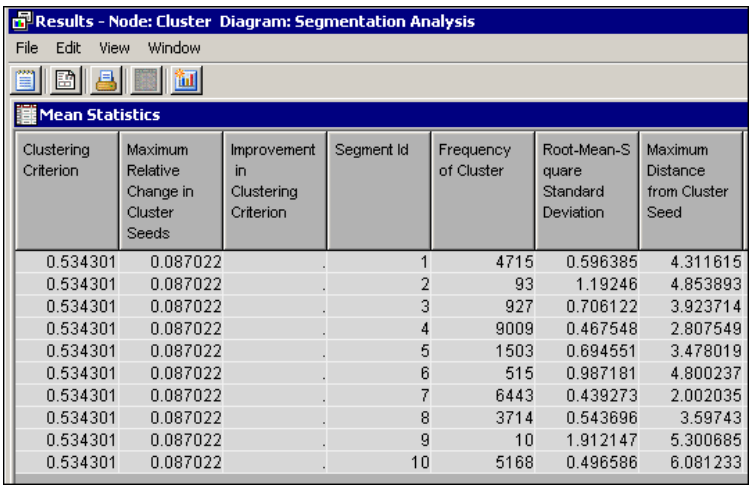
Property	Value
General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	10
Selection Criterion	

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

b) Run the Cluster node and select **Results**. The Results - Node: Cluster Diagram window appears, 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.



Results - Node: Cluster Diagram: Segmentation Analysis

File Edit View Window

Mean Statistics

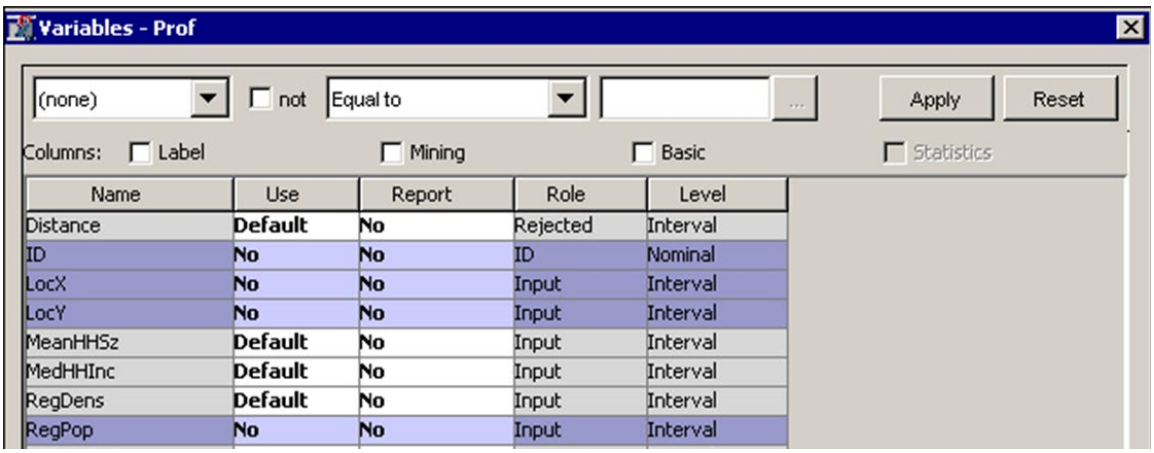
Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed
0.534301	0.087022	.	1	4715	0.596385	4.311615
0.534301	0.087022	.	2	93	1.19246	4.853893
0.534301	0.087022	.	3	927	0.706122	3.923714
0.534301	0.087022	.	4	9009	0.467548	2.807549
0.534301	0.087022	.	5	1503	0.694551	3.478019
0.534301	0.087022	.	6	515	0.987181	4.800237
0.534301	0.087022	.	7	6443	0.439273	2.002035
0.534301	0.087022	.	8	3714	0.543696	3.59743
0.534301	0.087022	.	9	10	1.912147	5.300685
0.534301	0.087022	.	10	5168	0.496586	6.081233

8. Profile clusters.
There is a 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. (Note that using SAS terminology, segment refers to cluster.)

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

b) To best describe the segments, you should pick a reasonable subset of the available input variables. Select the **Variables** property for the Segment Profile node.

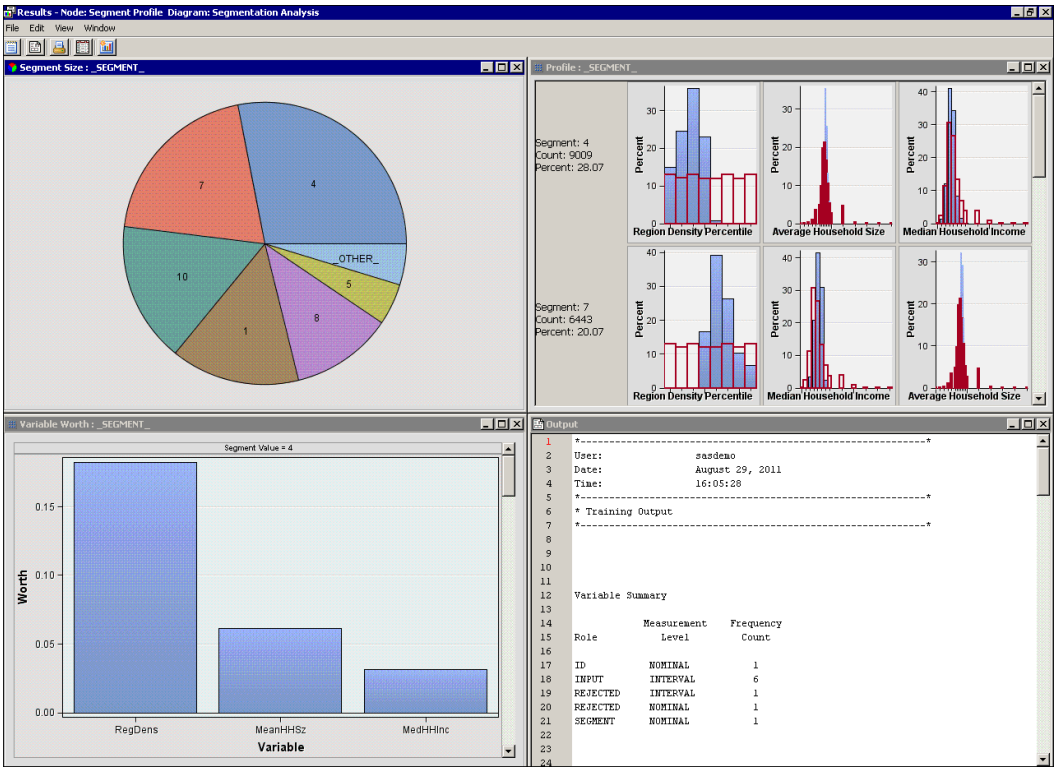
Select **Use → No** for **ID**, **LocX**, **LocY**, and **RegPop**.



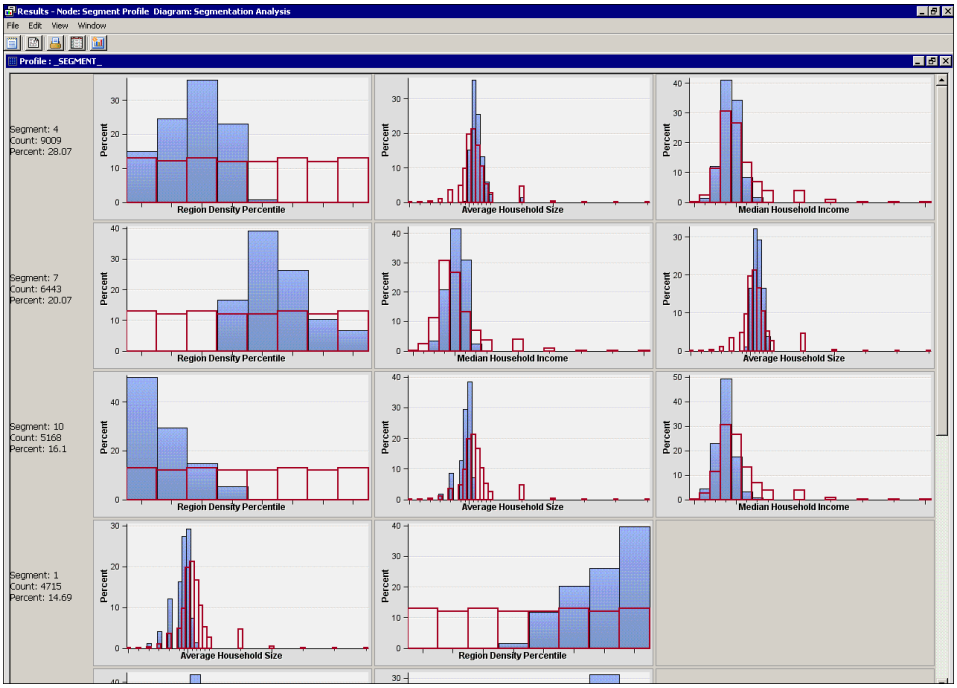
Name	Use	Report	Role	Level
Distance	Default	No	Rejected	Interval
ID	No	No	ID	Nominal
LocX	No	No	Input	Interval
LocY	No	No	Input	Interval
MeanHHSz	Default	No	Input	Interval
MedHHInc	Default	No	Input	Interval
RegDens	Default	No	Input	Interval
RegPop	No	No	Input	Interval

Select **OK** to close the Variables dialog box.

c) Run the Segment Profile node and select **Results**. The Results - Node: Segment Profile Diagram window appears.

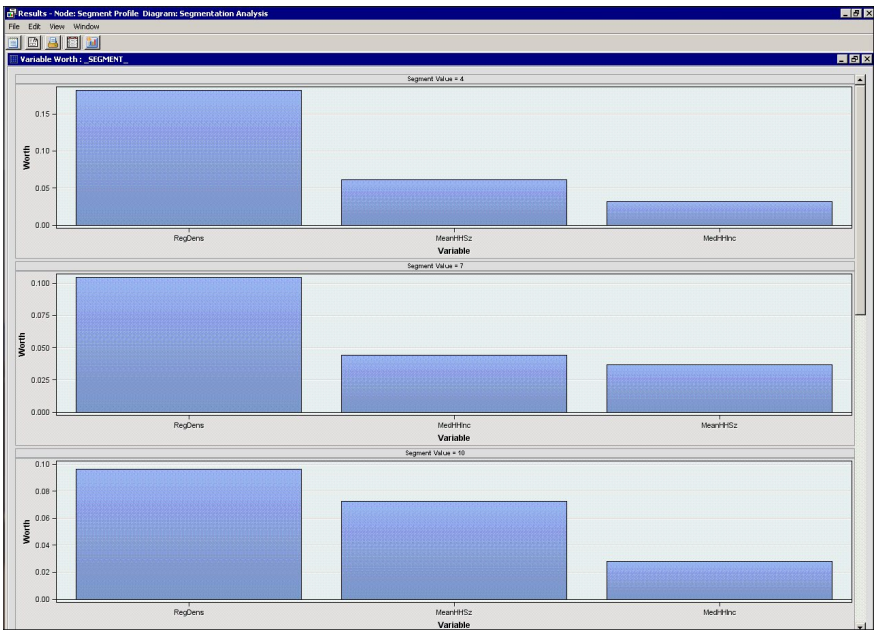


d) Maximize the **Profile** window.



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

e) 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.

Similar analyses can be used to describe the other segments.

Refer to the Help reference of the Segment Profile tool to learn more about this tool, particularly how to interpret the results generated by this tool.

9. Use the Reporter node to produce a report for Practical #3 submission.
Save the PDF report, and submit it by following the instruction given on page 1 of the practical document.