

Chapter 8 Introduction to Pattern Discovery

0.1	Introduction.....	Error! Bookmark not defined.
0.2	A Section Title.....	Error! Bookmark not defined.
	Demonstration: <Type title of demo here.>.....	Error! Bookmark not defined.
	Exercises	Error! Bookmark not defined.
0.3	Chapter Summary.....	Error! Bookmark not defined.
0.4	Solutions	Error! Bookmark not defined.
	Solutions to Exercises	Error! Bookmark not defined.
	Solutions to Student Activities (Polls/Quizzes)	Error! Bookmark not defined.

8.1 Introduction

Pattern Discovery



The Essence of Data Mining?

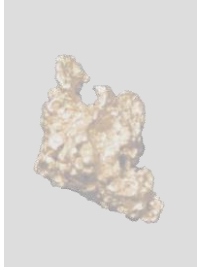
“...the discovery of interesting, unexpected, or valuable structures in large data sets.”

– David Hand

3 ...

There are a multitude of definitions for the field of data mining and knowledge discovery. Most center on the concept of pattern discovery. For example, David Hand, Professor of Statistics at Imperial College, London and a noted data mining authority, defines the field as “...*the discovery of interesting, unexpected, or valuable structures in large data sets.*” (Hand 2005) This is made possible by the ever-increasing data stores brought about by the era’s information technology.

Pattern Discovery



The Essence of Data Mining?

“...the discovery of interesting, unexpected, or valuable structures in large data sets.”

– David Hand

“If you’ve got terabytes of data, and you’re relying on data mining to find interesting things in there for you, you’ve lost before you’ve even begun.”

– Herb Edelstein

4

While Hand’s pronouncement is grandly promising, experience has shown it to be overly optimistic. Herb Edelstein, President of Two Crows Corporation and an internationally recognized expert in data mining, data warehousing, and CRM, counters with the following (Beck (Editor) 1997):

“If you’ve got terabytes of data, and you’re relying on data mining to find interesting things in there for you, you’ve lost before you’ve even begun. You really need people who understand what it is they are looking for – and what they can do with it once they find it.”

Many people think data mining (in particular, pattern discovery) means magically finding hidden nuggets of information without having to formulate the problem and without regard to the structure or content of the data. This is an unfortunate misconception.

Pattern Discovery Caution




- Poor data quality
- Opportunity
- Interventions
- Separability
- Obviousness
- Non-stationarity

5


In his defense, David Hand is well aware of the limitations of pattern discovery and provides guidance on how these analyses can fail (Hand 2005). These failings often fall into one of six categories:

- **Poor data quality** assumes many guises: inaccuracies (measurement or recording errors), missing, incomplete or outdated values, and inconsistencies (changes of definition). Patterns found in false data are fantasies.
- **Opportunity** transforms the possible to the perceived. Hand refers to this as the problem of multiplicity, or the law of truly large numbers. Examples of this abound. Hand notes the odds of a person winning the lottery in the United States are extremely small and the odds of that person winning it twice are fantastically so. However, the odds of **someone in the United States** winning it twice (in a given year) are actually better than even. As another example, you can search the digits of π for “prophetic” strings such as your birthday or significant dates in history and usually find them, given enough digits (www.angio.net/pi/piquery).
- **Intervention**, that is, taking action on the process that generates a set of data, can destroy or distort detected patterns. For example, fraud detection techniques lead to preventative measures, but the fraudulent behavior often evolves in response to this intervention.
- **Separability** of the interesting from the mundane is not always possible, given the information found in a data set. Despite the many safeguards in place, it is estimated that credit card companies lose \$0.18 to \$0.24 per \$100 in online transactions (Rubinkam 2006).
- **Obviousness** in discovered patterns reduces the perceived utility of an analysis. Among the patterns discovered through automatic detection algorithms, you find that there is an almost equal number of married men as married women, and you learn that ovarian cancer occurs primarily in women and that check fraud occurs most often for customers with checking accounts.
- **Non-stationarity** occurs when the process that generates a data set changes of its own accord. In such circumstances, patterns detected from historic data can simply cease. As Eric Hoffer states, “*In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists.*”


Pattern Discovery Applications




Data reduction




Novelty detection



Profiling



Market basket analysis






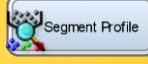




Sequence analysis

6
...

Despite the potential pitfalls, there are many successful applications of pattern discovery:

- **Data reduction** is the most ubiquitous application, that is, exploiting patterns in data to create a more compact representation of the original. Though vastly broader in scope, data reduction includes analytic methods such as cluster analysis.
- **Novelty detection** methods seek unique or previously unobserved data patterns. The methods find application in business, science, and engineering. Business applications include fraud detection, warranty claims analysis, and general business process monitoring.
- **Profiling** is a by-product of reduction methods such as cluster analysis. The idea is to create rules that isolate clusters or segments, often based on demographic or behavioral measurements. A marketing analyst might develop profiles of a customer database to describe the consumers of a company's products.
- **Market basket analysis**, or *association rule discovery*, is used to analyze streams of transaction data (for example, market baskets) for **combinations** of items that occur (or do not occur) more (or less) commonly than expected. Retailers can use this as a way to identify interesting combinations of purchases or as predictors of customer segments.
- **Sequence analysis** is an extension of market basket analysis to include a time dimension to the analysis. In this way, transactions data is examined for **sequences** of items that occur (or do not occur) more (or less) commonly than expected. A Webmaster might use sequence analysis to identify patterns or problems of navigation through a Web site.





Pattern Discovery Tools

	Data reduction	 Cluster
	Novelty detection	 Segment Profile
	Profiling	 SOM/Kohonen
	Market basket analysis	
	Sequence analysis	

7 ...

The first three pattern discovery applications are primarily served (in no particular order) by three tools in SAS Enterprise Miner: Cluster, SOM/Kohonen, and Segment Profile. The next section features a demonstration of the Cluster and Segment Profile tools.

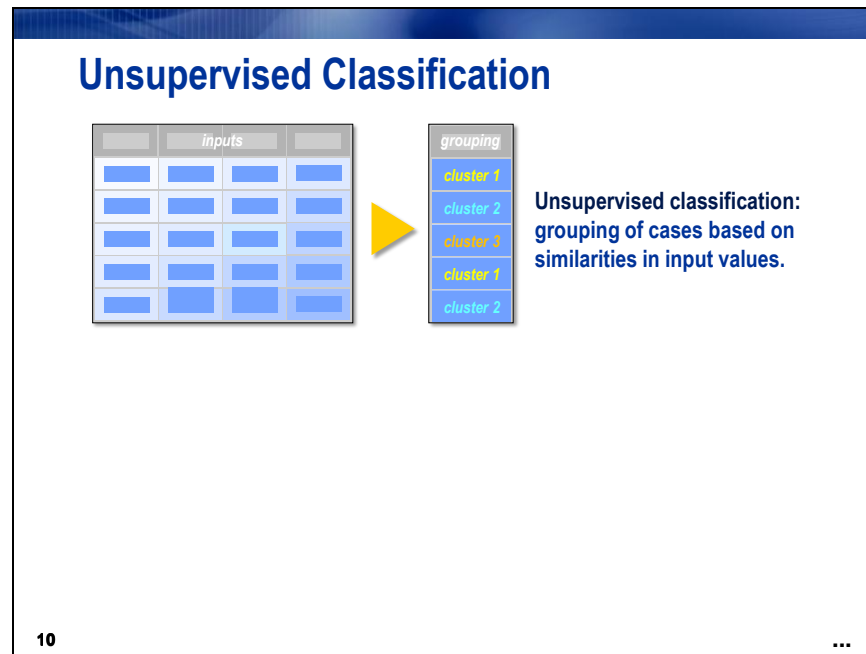
Pattern Discovery Tools

	Data reduction	 Cluster
	Novelty detection	 Segment Profile
	Profiling	 SOM/Kohonen
	Market basket analysis	 Association
	Sequence analysis	 Path Analysis

8

Market basket analysis and sequence analysis are performed by the Association tool. The Path Analysis tool can also be used to analyze sequence data. (An optional demonstration of the Association tool is presented at the end of this chapter.)

8.2 Cluster Analysis



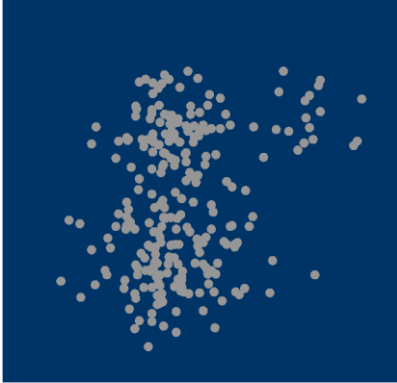
Unsupervised classification (also known as *clustering* and *segmenting*) attempts to group training data set cases based on similarities in **input** variables. It is a data reduction method because an entire training data set can be represented by a small number of clusters. The groupings are known as *clusters* or *segments*, and they can be applied to other data sets to classify new cases. It is distinguished from *supervised classification* (also known as *predictive modeling*), which is discussed in previous chapters.

The purpose of clustering is often description. For example, segmenting existing customers into groups and associating a distinct profile with each group might help future marketing strategies. However, there is no guarantee that the resulting clusters will be meaningful or useful.

Unsupervised classification is also useful as a step in predictive modeling. For example, customers can be clustered into homogenous groups based on sales of different items. Then a model can be built to predict the cluster membership based on more easily obtained input variables.

***k*-means Clustering Algorithm**

Training Data



1. **Select inputs.**
2. Select k cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. Re-assign cases.
6. Repeat steps 4 and 5 until convergence.

12

One of the most commonly used methods for clustering is the *k-means algorithm*. It is a straightforward algorithm that scales well to large data sets and is, therefore, the primary tool for clustering in SAS Enterprise Miner.

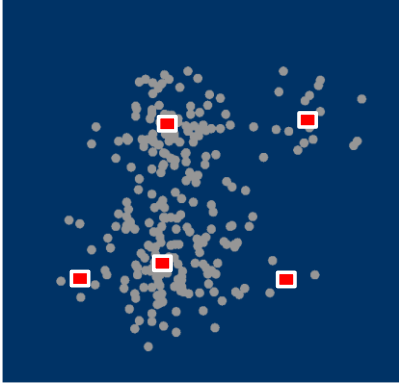
While often overlooked as an important part of a clustering process, the first step in using the *k-means* algorithm is to choose a set of inputs. In general, you should seek inputs that have the following attributes:

- are meaningful to the analysis objective
- are relatively independent
- are limited in number
- have a measurement level of Interval
- have low kurtosis and skewness statistics (at least in the training data)

Choosing meaningful inputs is clearly important for interpretation and explanation of the generated clusters. Independence and limited input count make the resulting clusters more stable. (Small perturbations of training data usually do not result in large changes to the generated clusters.) An interval measurement level is recommended for *k-means* to produce non-trivial clusters. Low kurtosis and skewness statistics on the inputs avoid creating single-case outlier clusters.

***k*-means Clustering Algorithm**

Training Data



1. Select inputs.
- 2. Select k cluster centers.**
3. Assign cases to closest center.
4. Update cluster centers.
5. Re-assign cases.
6. Repeat steps 4 and 5 until convergence.

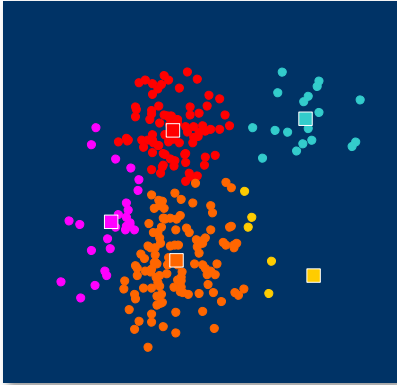
13

The next step in the k -means algorithm is to choose a value for k , the number of cluster centers. SAS Enterprise Miner features an automatic way to do this, assuming that the data has k distinct concentrations of cases. If this is not the case, you should choose k to be consistent with your analytic objectives.

With k selected, the k -means algorithm chooses cases to represent the initial *cluster centers* (also named *seeds*).

k-means Clustering Algorithm

Training Data



1. Select inputs.
2. Select k cluster centers.
3. **Assign cases to closest center.**
4. Update cluster centers.
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

14 ...

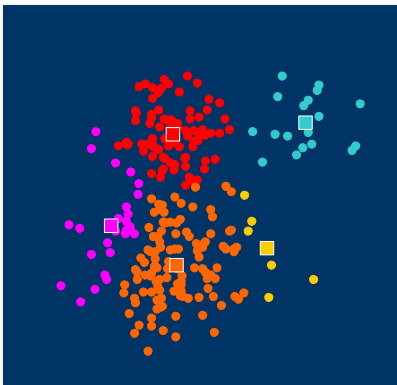
The Euclidean distance from each case in the training data to each cluster center is calculated. Cases are assigned to the closest cluster center.



Because the distance metric is Euclidean, it is important for the inputs to have compatible measurement scales. Unexpected results can occur if one input's measurement scale differs greatly from the others.

k-means Clustering Algorithm

Training Data



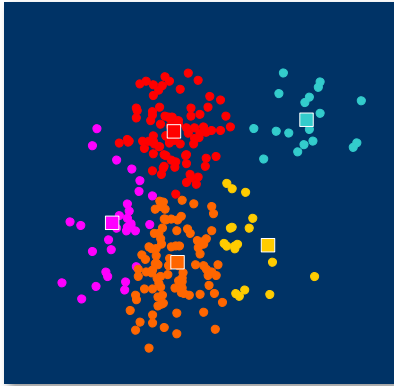
1. Select inputs.
2. Select k cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

15 ...

The cluster centers are updated to equal the average of the cases assigned to the cluster in the previous step.

***k*-means Clustering Algorithm**

Training Data



1. Select inputs.
2. Select k cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
- 5. Reassign cases.**
6. Repeat steps 4 and 5 until convergence.

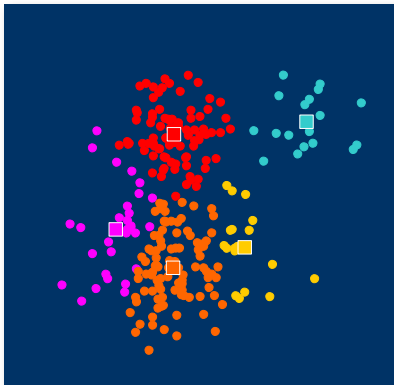
16

...

Cases are reassigned to the closest cluster center.

***k*-means Clustering Algorithm**

Training Data



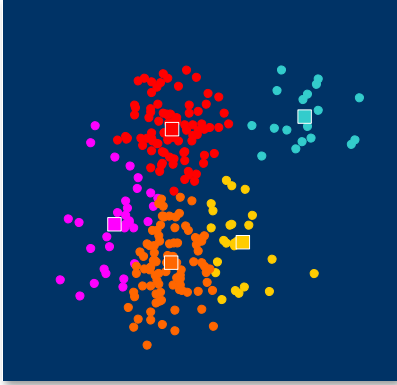
1. Select inputs.
2. Select k cluster centers.
3. Assign cases to closest center.
- 4. Update cluster centers.**
5. Reassign cases.
- 6. Repeat steps 4 and 5 until convergence.**

17

...

***k*-means Clustering Algorithm**

Training Data



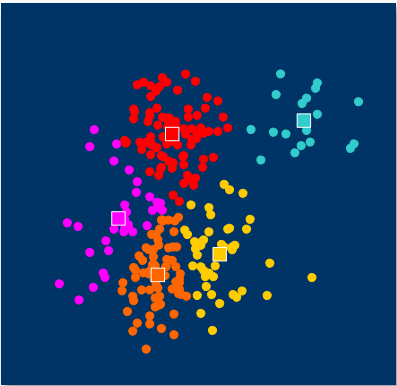
1. Select inputs.
2. Select k cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

18 ...

The update and reassign steps are repeated until the process converges.

***k*-means Clustering Algorithm**

Training Data



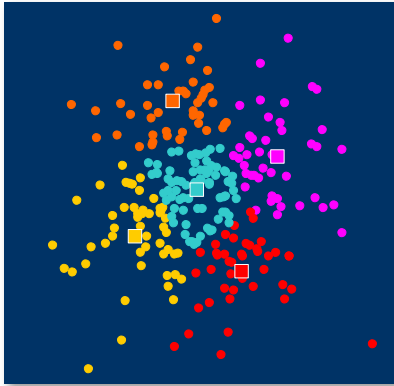
1. Select inputs.
2. Select k cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

25 ...

On convergence, final cluster assignments are made. Each case is assigned to a unique segment. The segment definitions can be stored and applied to new cases outside of the training data.

Segmentation Analysis

Training Data



When no clusters exist, use the *k*-means algorithm to partition cases into contiguous groups.

27

While they are often used synonymously, a segmentation analysis is distinct from a traditional cluster analysis. A cluster analysis is geared toward identifying distinct concentrations of cases in a data set. When no distinct concentrations exist, the best you can do is a segmentation analysis, that is, algorithmically partitioning the input space into contiguous groups of cases.

Demographic Segmentation Demonstration

Analysis goal:

Group geographic regions into segments based on income, household size, and population density.

Analysis plan:

- Select and transform segmentation inputs.
- Select the number of segments to create.
- Create segments with the Cluster tool.
- Interpret the segments.

28

The following demonstration illustrates the use of clustering tools. The goal of the analysis is to group people in the United States into distinct subsets based on urbanization, household size, and income factors. These factors are common to commercial lifestyle and life-stage segmentation products. (For examples, see www.claritas.com or www.spectramarketing.com.)



Segmenting Census Data

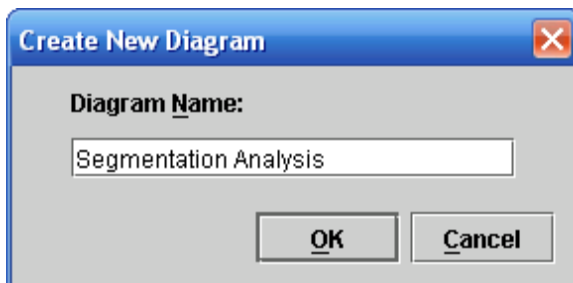
This demonstration introduces SAS Enterprise Miner tools and techniques for cluster and segmentation analysis. There are five parts:

- 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

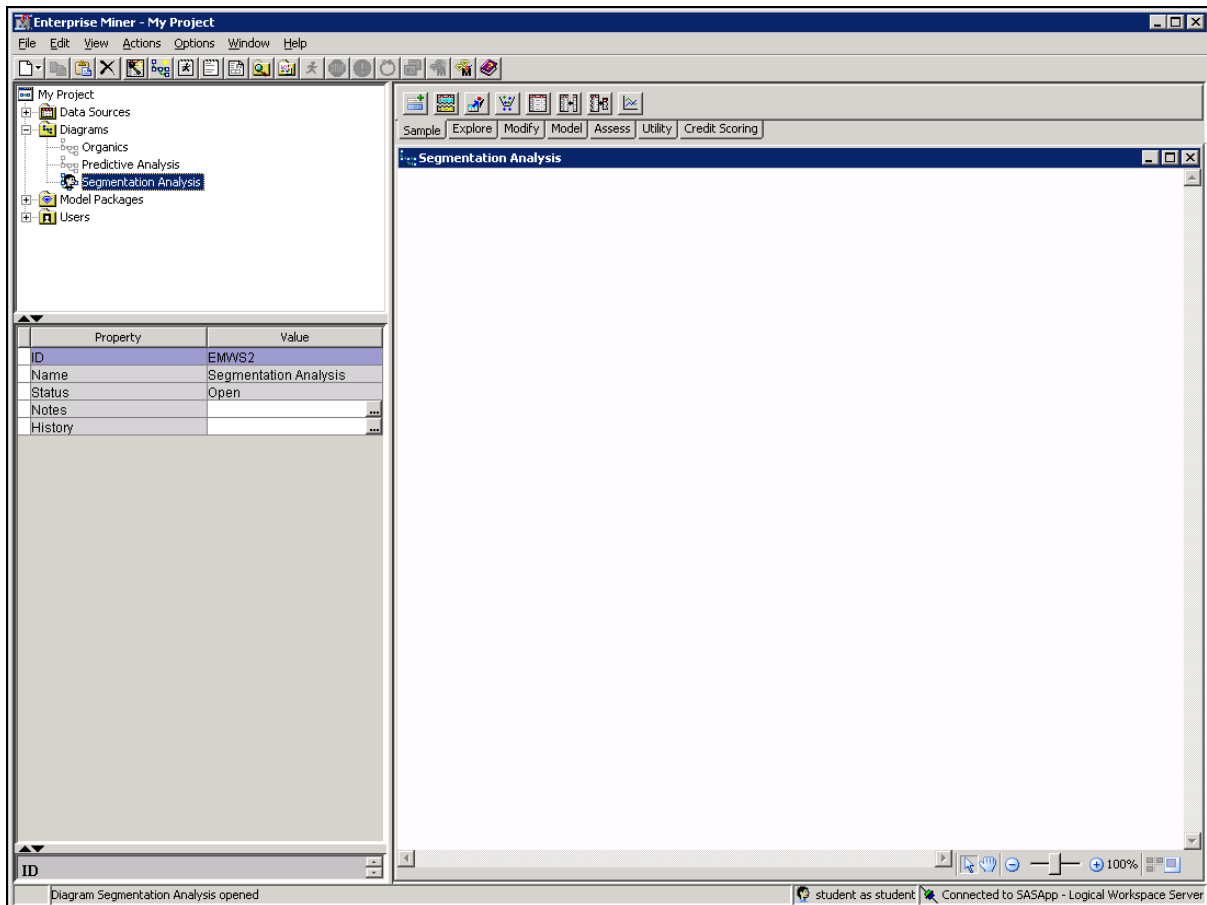
Diagram Definition

Use the following steps to define the diagram for the segmentation analysis.

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



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

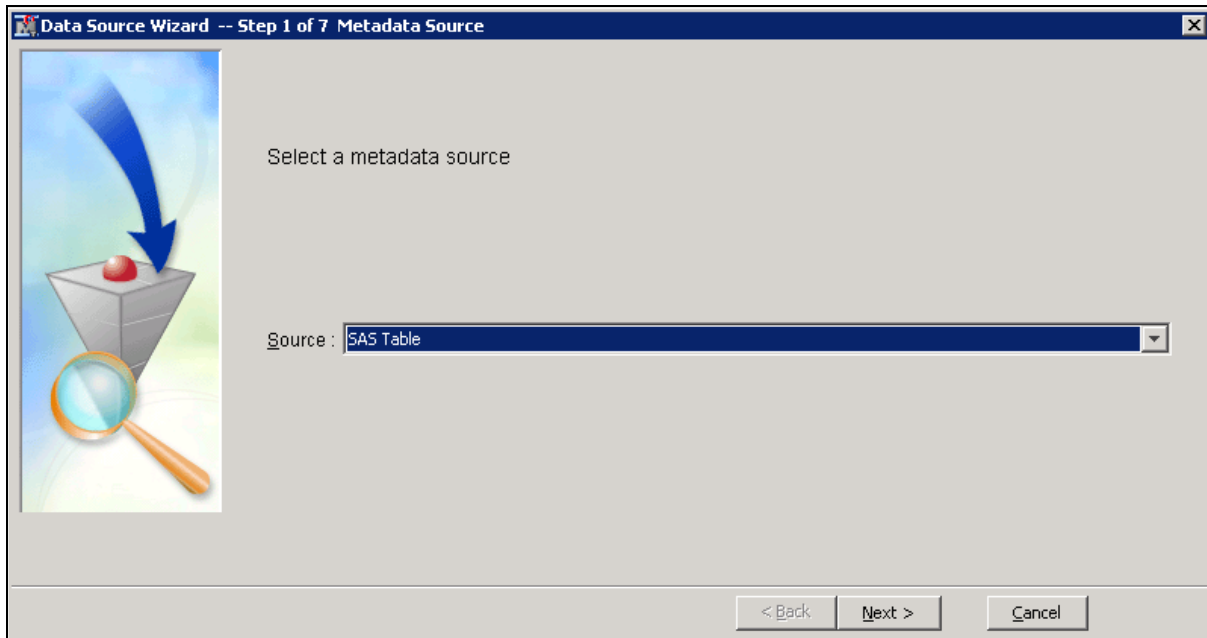


You use the Segmentation Analysis window to create process flow diagrams.

Data Source Definition

Follow these steps to create the segmentation analysis data source.

1. Right-click **Data Sources** in the Project panel and select **Create Data Source**. The Data Source Wizard opens.

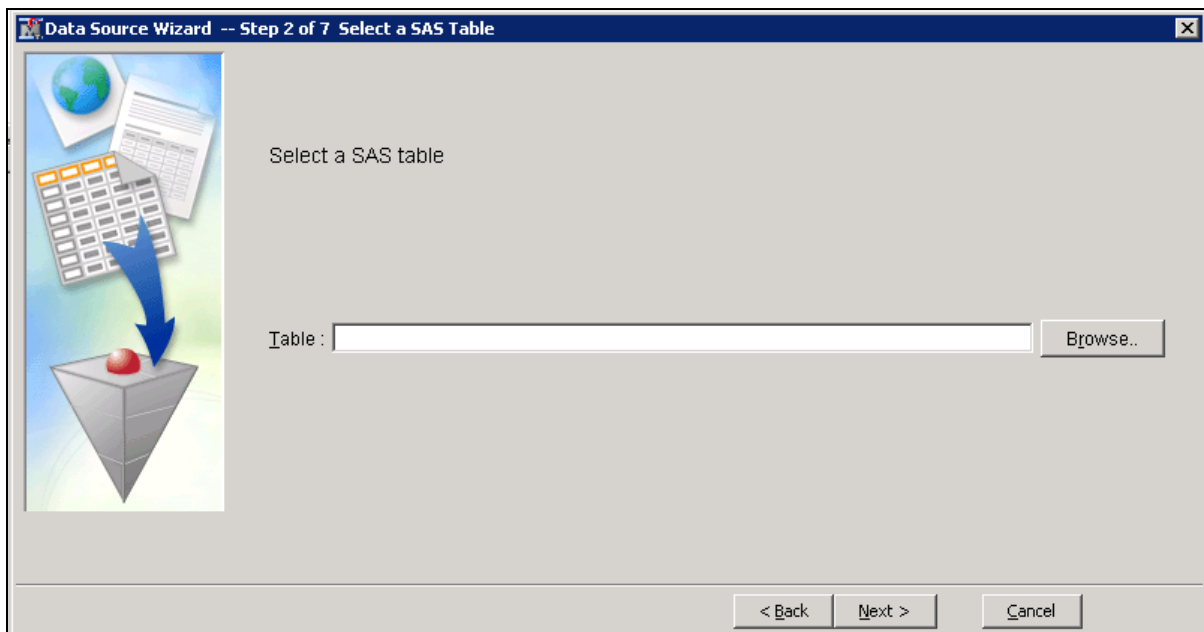


The Data Source Wizard guides you through a seven-step process to create a SAS Enterprise Miner data source.

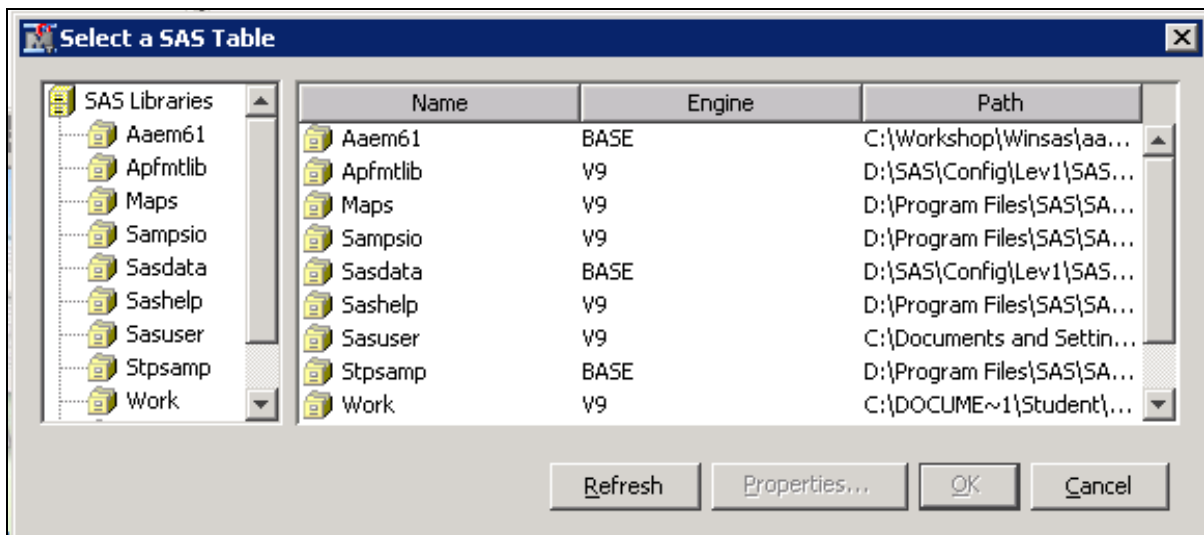
2. Select **Next >** to use a SAS table as the source for the metadata. (This is the usual choice.)

The Data Source Wizard proceeds to Step 2. In this step, select the SAS table that you want to make available to SAS Enterprise Miner. You can either type the library name and SAS table name as *libname.tablename* or select the SAS table from a list.

3. Select **Browse...** to choose a SAS table from the libraries visible to the SAS Foundation server.

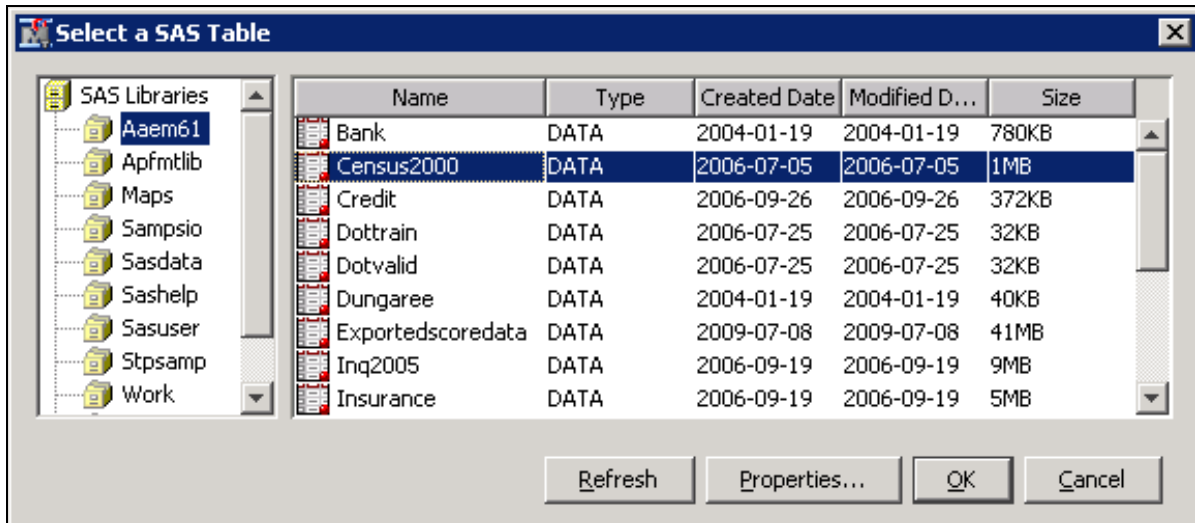


The Select a SAS Table dialog box opens.



One of the libraries listed is named AAEM61, which is the library name defined in the project start-up code.

4. Select the **Aaem61** library and the **Census2000** SAS table.



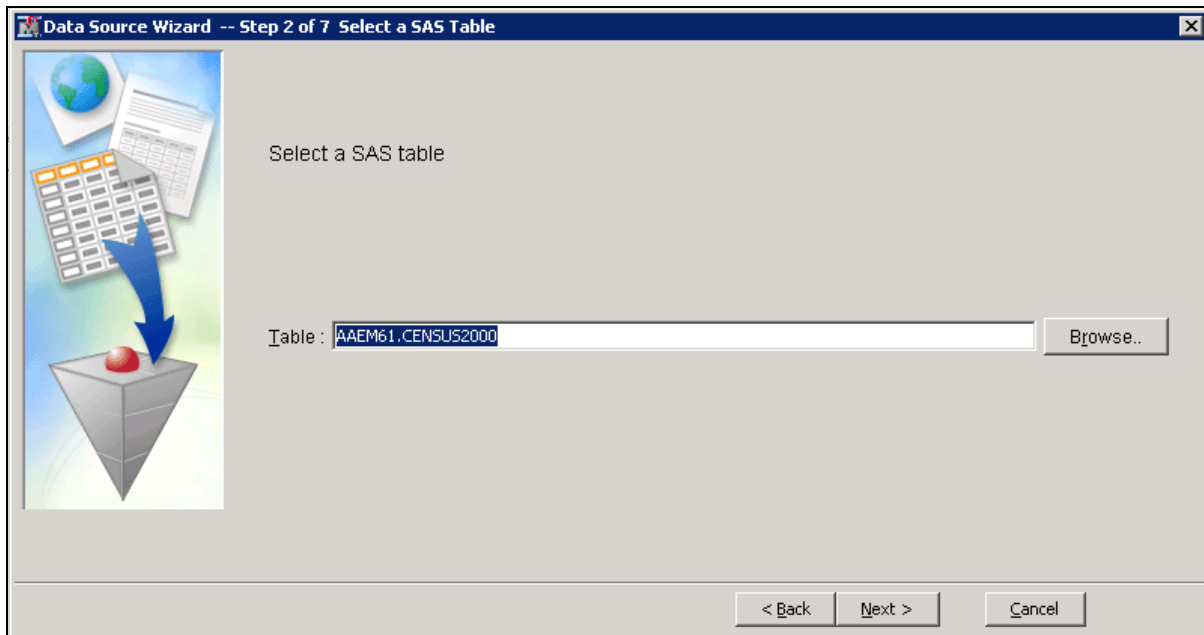
The **Census2000** data is a postal code-level summary of the entire 2000 United States Census. It features seven variables:

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

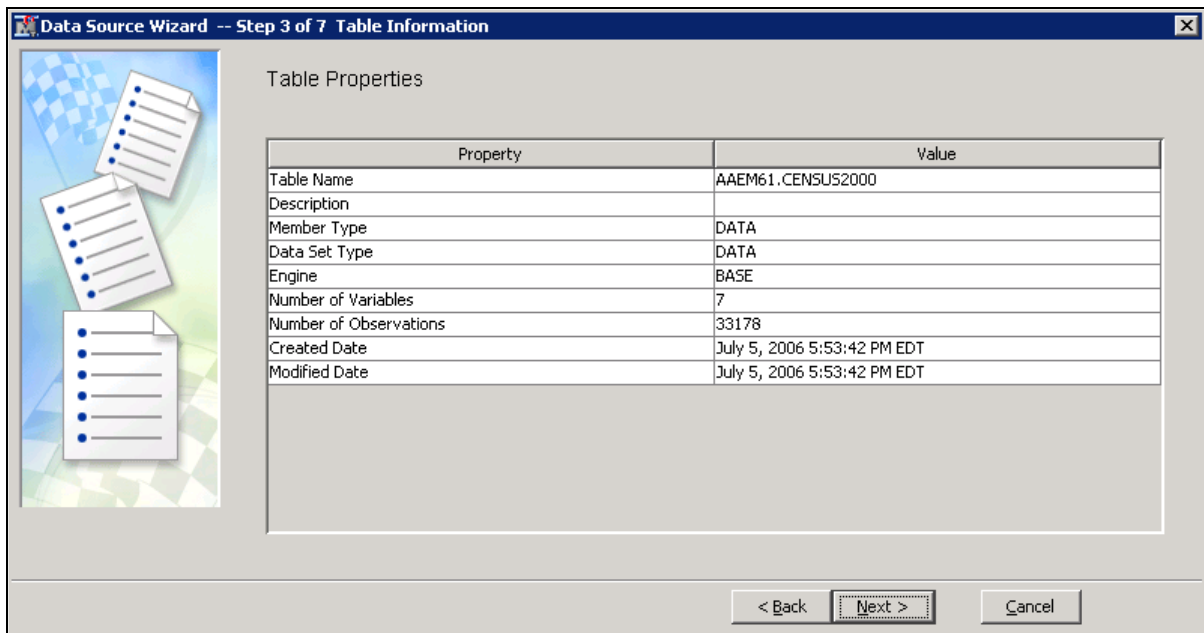
The data is suited for creation of life-stage, lifestyle segments using SAS Enterprise Miner's pattern discovery tools.

5. Select **OK**.

The Select a SAS Table dialog box closes and the selected table is entered in the Table field.

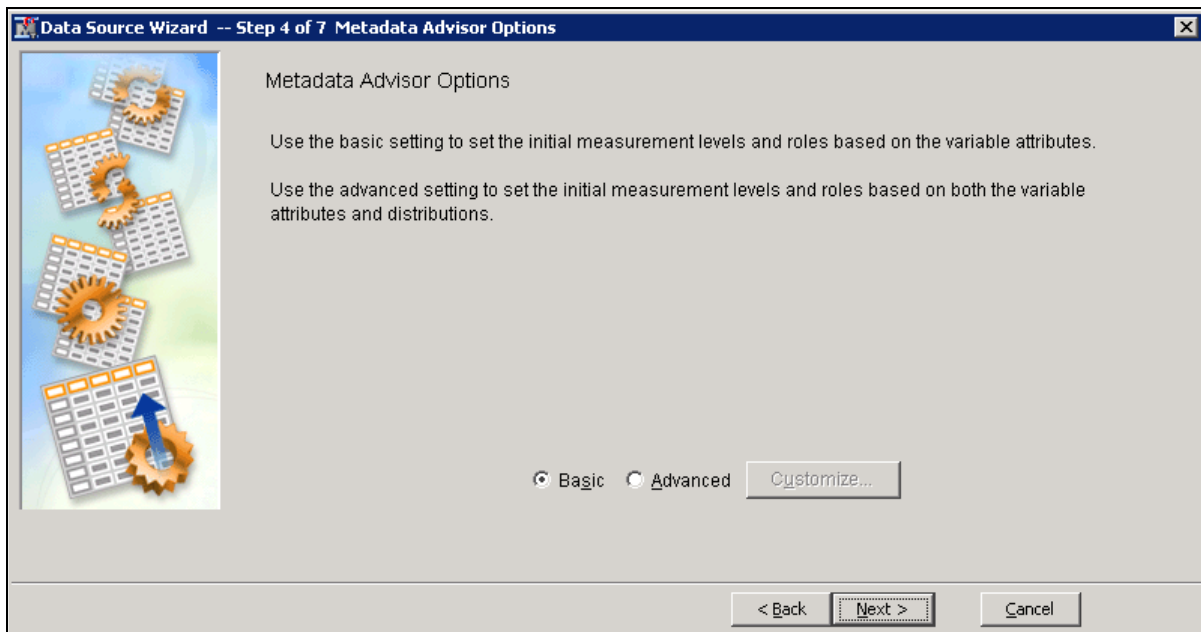


6. Select **Next >**. The Data Source Wizard proceeds to Step 3.



This step of the Data Source Wizard provides basic information about the selected table.

7. Select **Next >**. The Data Source Wizard proceeds to Step 4.



Step 4 of the Data Source Wizard starts the metadata definition process. SAS Enterprise Miner assigns initial values to the metadata based on characteristics of the selected SAS table. The Basic setting assigns initial values to the metadata based on variable attributes such as the variable name, data type, and assigned SAS format. The Advanced setting also includes information about the distribution of the variable to assign the initial metadata values.

8. Select **Next >** to use the Basic setting. The Data Source Wizard proceeds to Step 5.

Data Source Wizard -- Step 5 of 7 Column Metadata

(none) ☐ not Equal to ☐ Apply Reset

Columns: ☐ Label ☐ Mining ☒ Basic ☐ Statistics

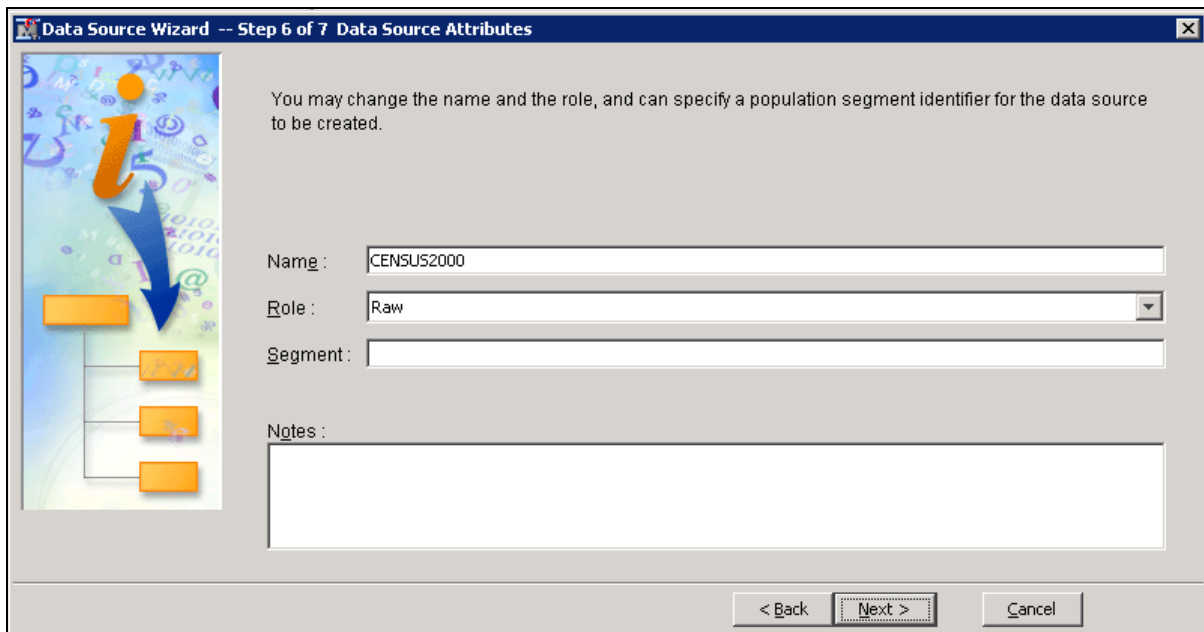
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ID	ID	Nominal	No		No	.	
LocX	Input	Interval	No		No	.	
LocY	Input	Interval	No		No	.	
MeanHHSz	Input	Interval	No		No	.	
MedHHInc	Input	Interval	No		No	.	
RegDens	Input	Interval	No		No	.	
RegPop	Input	Interval	No		No	.	

Show code Explore Compute Summary < Back Next > Cancel

Step 5 of the Data Source Wizard enables you to specify the role and level for each variable in the selected SAS table. A default role is assigned based on the name of a variable. 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.

The metadata settings are correct for the upcoming analysis.

9. Select **Next >**. The Data Source Wizard proceeds to Step 6.



Data Source Wizard -- Step 6 of 7 Data Source Attributes

You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : CENSUS2000

Role : Raw

Segment :

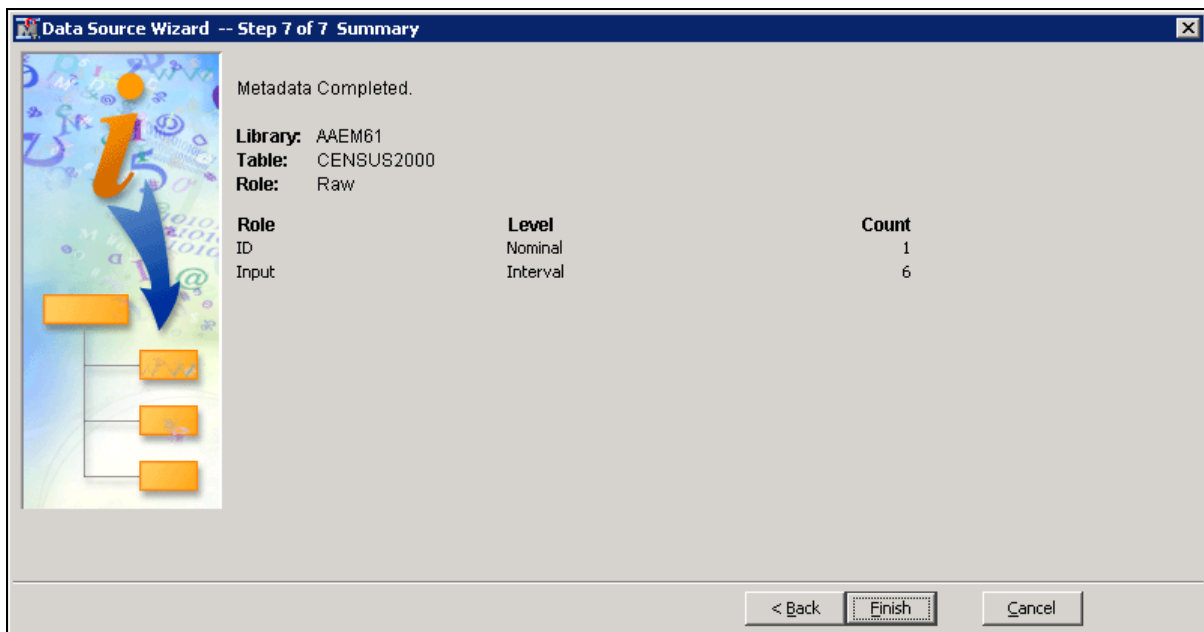
Notes :

< Back Next > Cancel

Step 6 in the Data Source Wizard enables you to set a role for the selected SAS table and provide descriptive comments about the data source definition.

For the impending analysis, a table role of Raw is acceptable.

Step 7 provides summary information about the created data set.



Data Source Wizard -- Step 7 of 7 Summary

Metadata Completed.

Library: AAEM61
Table: CENSUS2000
Role: Raw

Role	Level	Count
ID	Nominal	1
Input	Interval	6

< Back Finish Cancel

10. 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

A worthwhile next step in the process of defining a data source is to explore and validate its contents. 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.

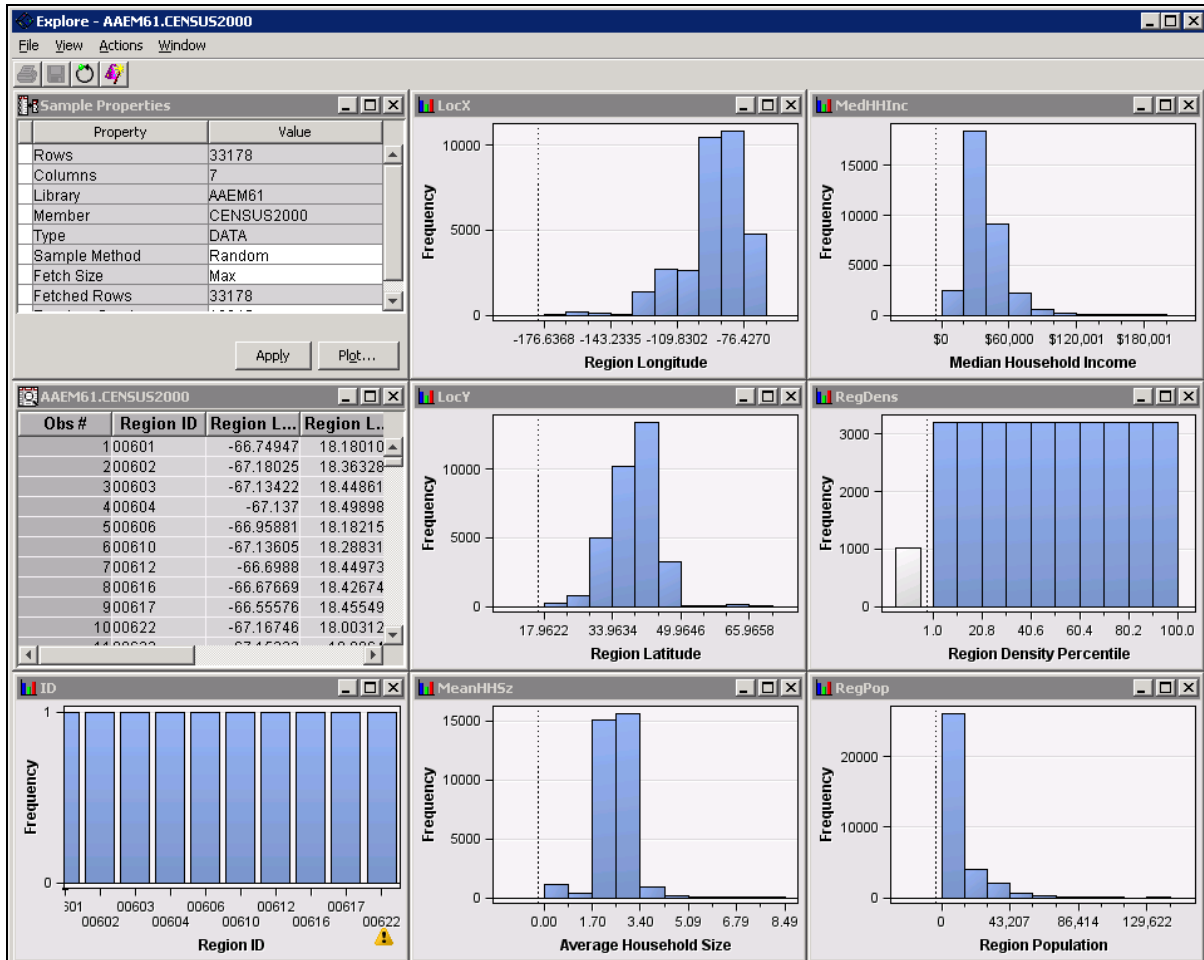
Data Source Exploration

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

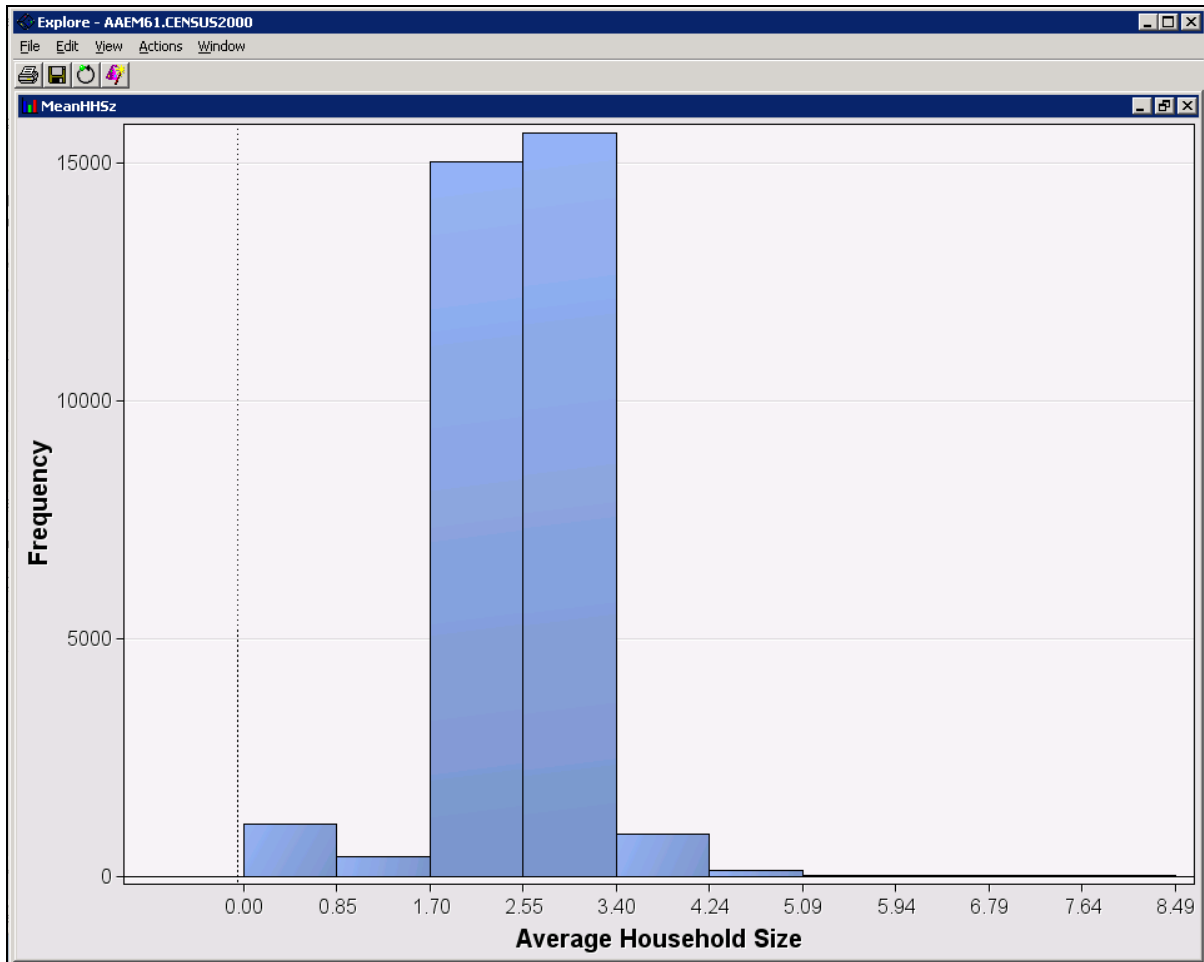
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ID	ID	Nominal	No		No	.	.
LocX	Input	Interval	No		No	.	.
LocY	Input	Interval	No		No	.	.
MeanHHSz	Input	Interval	No		No	.	.
MedHHInc	Input	Interval	No		No	.	.
RegDens	Input	Interval	No		No	.	.
RegPop	Input	Interval	No		No	.	.

2. Examine histograms for the available variables.
3. Select all listed inputs by dragging the cursor across all of the input names or by holding down the CTRL key and typing **A**.

4. Select **Explore...** The Explore window opens, and this time displays histograms for all of the variables in the **CENSUS2000** data source.

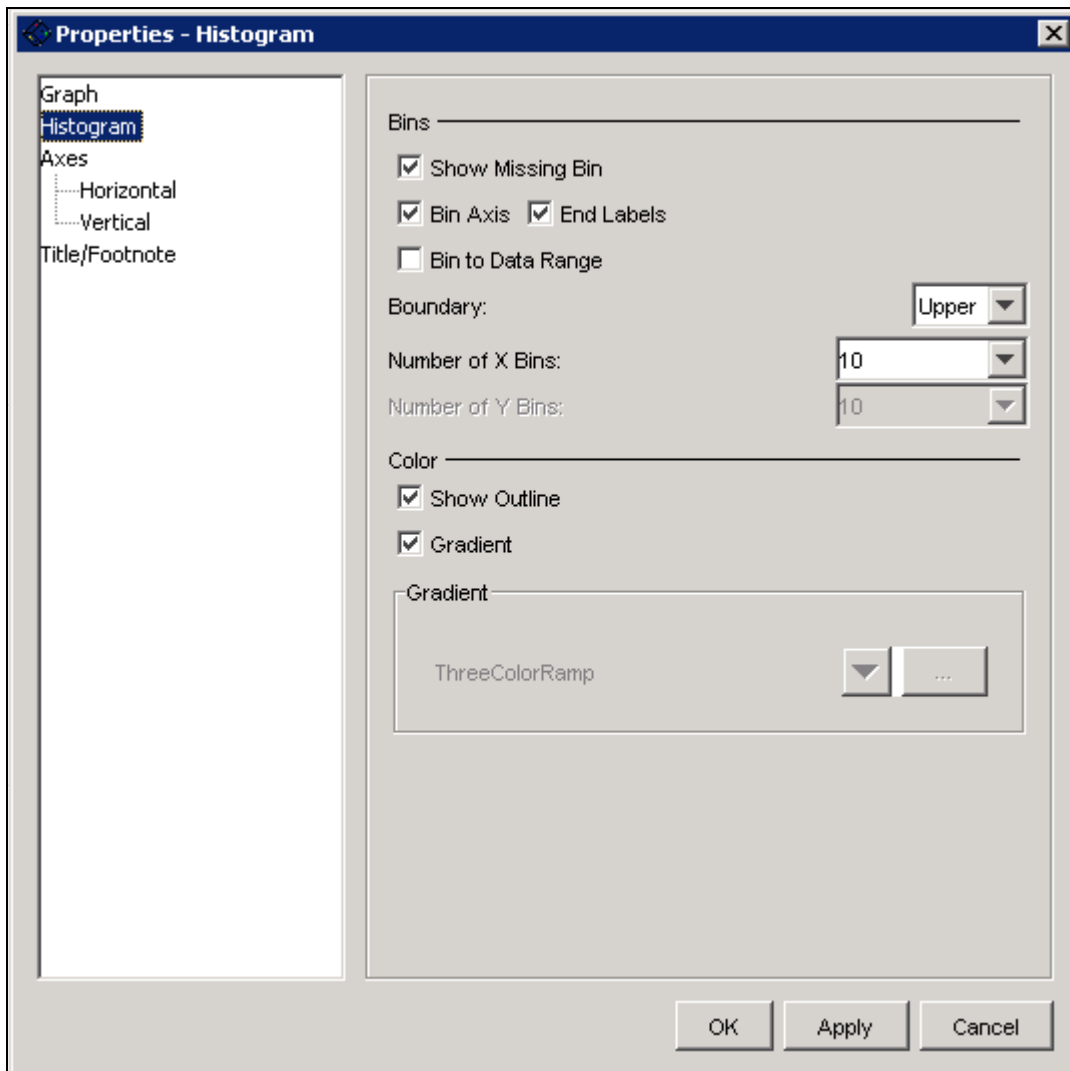


5. Maximize the MeanHHSz histogram by double-clicking its title bar. The histogram now fills the Explore window.



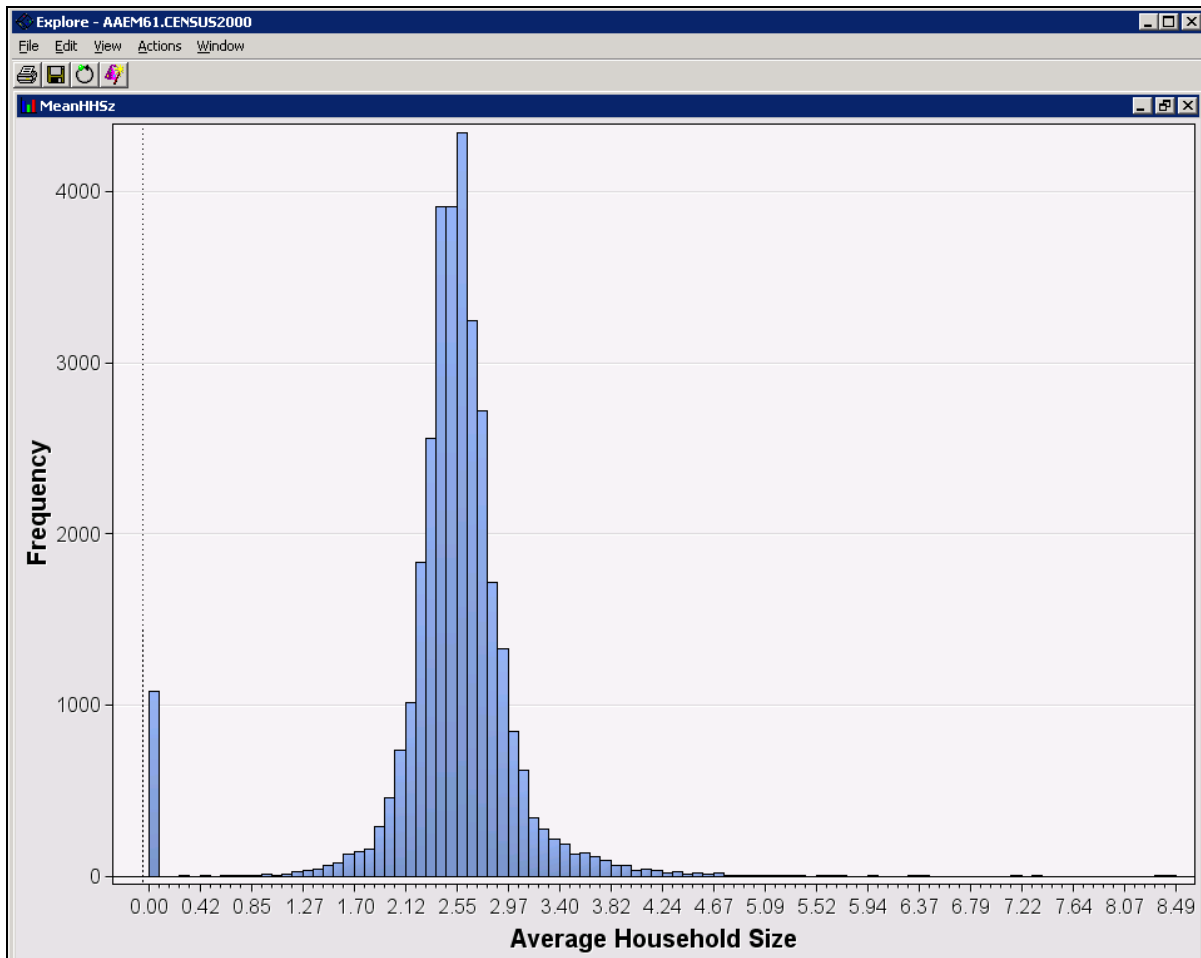
As before, increasing the number of histogram bins from the default of 10 increases your understanding of the data.

6. Right-click in the histogram window and select **Graph Properties...** from the shortcut menu. The Properties - Histogram dialog box opens.



You can use the Properties - Histogram dialog box to change the appearance of the corresponding histogram.

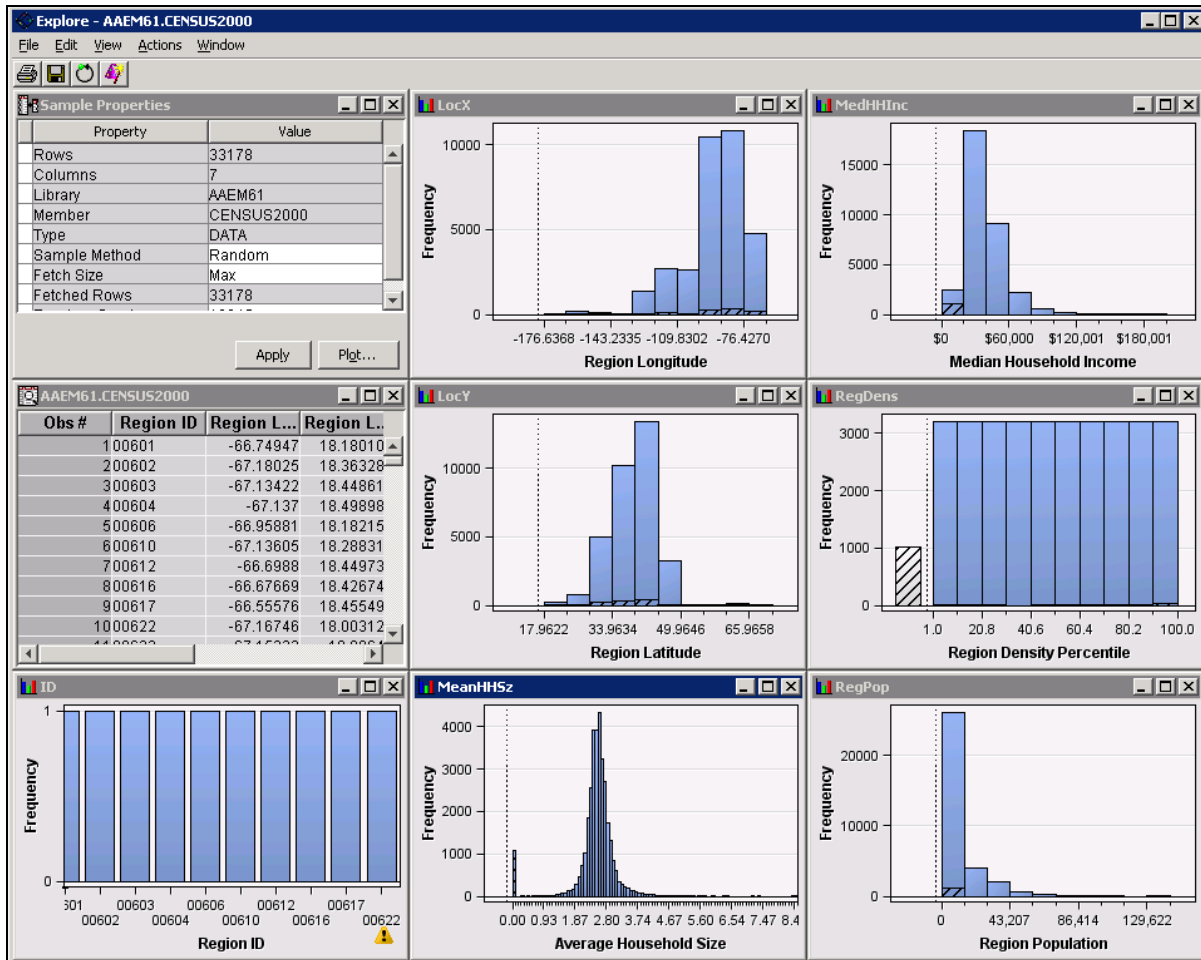
7. Type **100** in the Number of X Bins field and select **OK**. The histogram is updated to have 100 bins.



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.

8. Select the bar near zero in the histogram.

9. 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 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.

10. Maximize the **CENSUS2000** data table.
11. Scroll in the data table until you see the first selected row.

Obs #	Region ID	Region L...	Region L...	Region ...	Region ...	Median ...	Average ...
1600638	-66.49835	18.308139		70	19,648	\$10,986	3.26
1700641	-66.70519	18.268896		71	35,095	\$9,901	3.14
1800646	-66.27689	18.442798		84	34,114	\$16,378	3.11
1900647	-66.93993	17.964529		73	5,352	\$9,607	2.87
2000650	-66.56773	18.363331		77	12,207	\$11,544	3.11
2100652	-66.61217	18.457453		75	2,903	\$12,097	2.84
2200653	-66.90098	17.992112		78	16,536	\$9,770	3.04
2300656	-66.79168	18.038866		76	23,072	\$11,361	3.19
2400659	-66.80039	18.432956		80	38,925	\$12,378	3.06
2500660	-67.12086	18.139108		84	16,614	\$16,745	2.82
2600662	-67.01974	18.478855		80	42,178	\$11,753	2.94
2700664	-66.59244	18.212565		73	17,318	\$11,220	3.35
2800667	-67.04227	18.017819		74	26,637	\$11,426	2.91
2900669	-66.87504	18.288418		77	32,246	\$9,758	3.12
3000670	-66.97604	18.241343		68	11,174	\$9,556	3.08
3100674	-66.48697	18.426137		81	45,874	\$12,820	2.97
3200676	-67.08425	18.37956		79	39,697	\$11,271	3.11
3300677	-67.23675	18.336121		81	14,767	\$11,460	2.86
3400678	-66.93276	18.442334		81	27,271	\$12,005	3.07
3500680	-67.12655	18.205232		84	75,090	\$12,042	2.77
3600682	-67.15429	18.208402		90	25,138	\$11,121	2.84
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
5300716	-66.59966	17.999066		89	38,859	\$15,610	3.06
5400717	-66.61375	18.004303		99	23,083	\$11,697	2.55
5500718	-65.74294	18.22048		74	23,753	\$11,461	2.97
5600719	-66.25000	18.204571		82	28,034	\$12,472	3.22

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

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

Explore - AAEM61.CENSUS2000

FileViewActionsWindow

AAEM61.CENSUS2000

Obs #	Region ID	Region L...	Region L...	Region ...	Region ...	Median ...	Average Household Size ▲
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	
649027HH	-71.0363	41.711052	.	0	\$0	0.00	
704028HH	-71.42673	41.651234	.	0	\$0	0.00	
721029HH	-71.40145	41.8141	.	0	\$0	0.00	
755030HH	-71.47319	42.81187	.	0	\$0	0.00	
763031HH	-71.47113	43.007183	.	0	\$0	0.00	
815032HH	-71.59953	43.612667	.	0	\$0	0.00	
816032XX	-71.52839	43.976513	.	0	\$0	0.00	
821033HH	-71.5597	43.246016	.	0	\$0	0.00	
847034HH	-71.93063	42.988155	.	0	\$0	0.00	
864035HH	-71.24389	44.636085	.	0	\$0	0.00	
865035XX	-71.29905	44.510468	.	0	\$0	0.00	
873036HH	-72.43502	43.255688	.	0	\$0	0.00	
897037HH	-72.1821	43.832093	.	0	\$0	0.00	
959038HH	-70.91394	43.311793	.	0	\$0	0.00	
969039HH	-70.69216	43.143285	.	0	\$0	0.00	
1028040HH	-70.27453	43.562653	.	0	\$0	0.00	
1037041HH	-70.26518	43.67038	.	0	\$0	0.00	
1082042HH	-70.39034	44.327523	.	0	\$0	0.00	
1103043HH	-69.85778	44.282949	.	0	\$0	0.00	
1163044HH	-68.76274	45.279499	.	0	\$0	0.00	
1191045HH	-69.6258	43.920971	.	0	\$0	0.00	
1253046HH	-67.89638	44.69515	.	0	\$0	0.00	
1292047HH	-68.3812	46.665748	.	0	\$0	0.00	
1312048HH	-69.0871	44.104367	.	0	\$0	0.00	
1373049HH	-69.71961	44.762422	.	0	\$0	0.00	
1424050HH	-72.44107	43.741081	.	0	\$0	0.00	
1472053HH	-72.89058	42.858851	.	0	\$0	0.00	
1473053XX	-72.99699	42.982934	.	0	\$0	0.00	
1516054HH	-73.0833	44.36125	.	0	\$0	0.00	
1550056HH	-72.71076	44.313574	.	0	\$0	0.00	
1590057HH	-72.20046	42.754454	.	0	\$0	0.00	

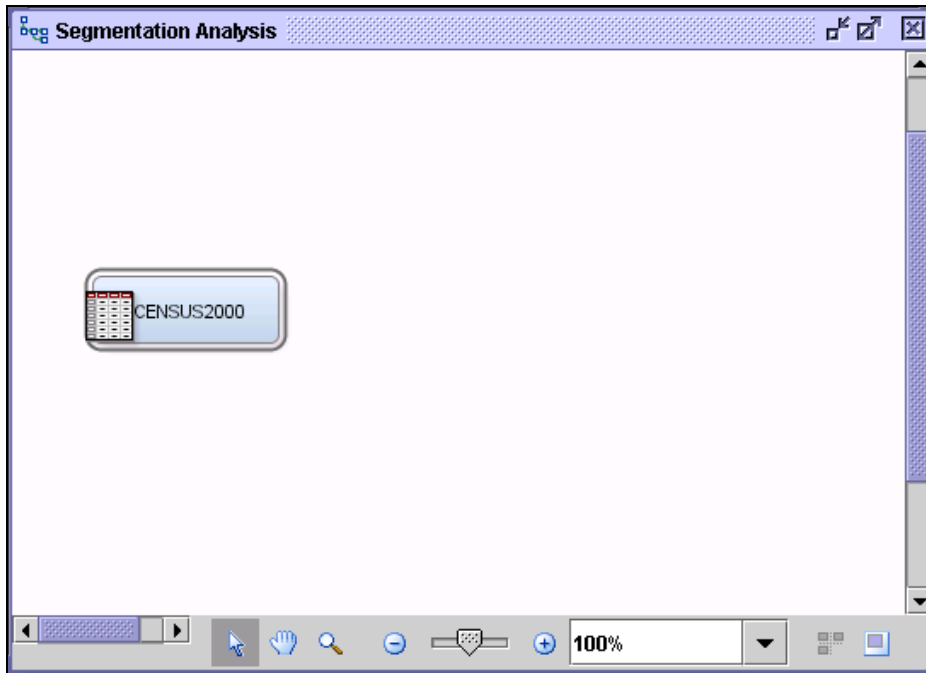
Most of the cases with zero Average Household Size have zero or missing on the remaining non-geographic attributes. There are some exceptions, but it could be argued that cases such as this are not of interest for analyzing household demographics. The next part of this demonstration shows how to remove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

13. Close the Explore and Variables windows.

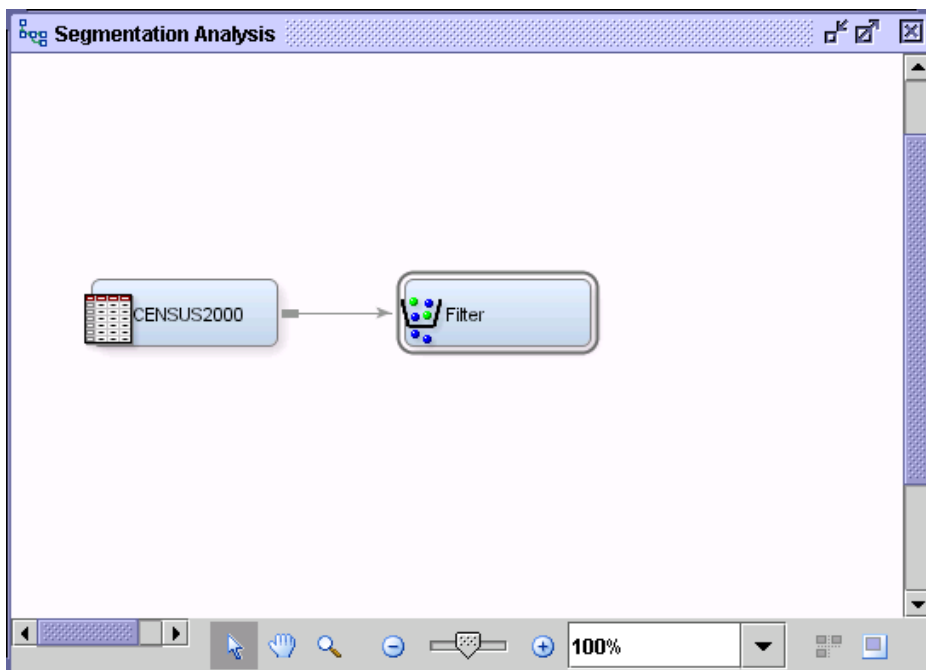
Case Filtering

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.



4. 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
<input type="checkbox"/> Class Variables	
Class Variables	...
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency C1	
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels	25
<input type="checkbox"/> Interval Variables	
Interval Variables	...
Default Filtering Method	Standard Deviations from Mean
Keep Missing Values	Yes
Tuning Parameters	...
Score	
Create score code	Yes

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.

5. Change the Default Filtering Method property to **User-Specified Limits**.

<input type="checkbox"/> Interval Variables	
Interval Variables	...
Default Filtering Method	User-Specified Limits
Keep Missing Values	Yes
Tuning Parameters	...

6. Select the Interval Variables ellipsis (...). The Interactive Interval Filter window opens.

Interactive Interval Filter

Train or raw data set does not exist.

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Report
LocX	Default	Default	.	.	No
LocY	Default	Default	.	.	No
MeanHHSz	Default	Default	.	.	No
MedHHInc	Default	Default	.	.	No
RegDens	Default	Default	.	.	No
RegPop	Default	Default	.	.	No

Generate Summary OK Cancel

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.

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

Interactive Interval Filter

Train or raw data set does not exist.

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

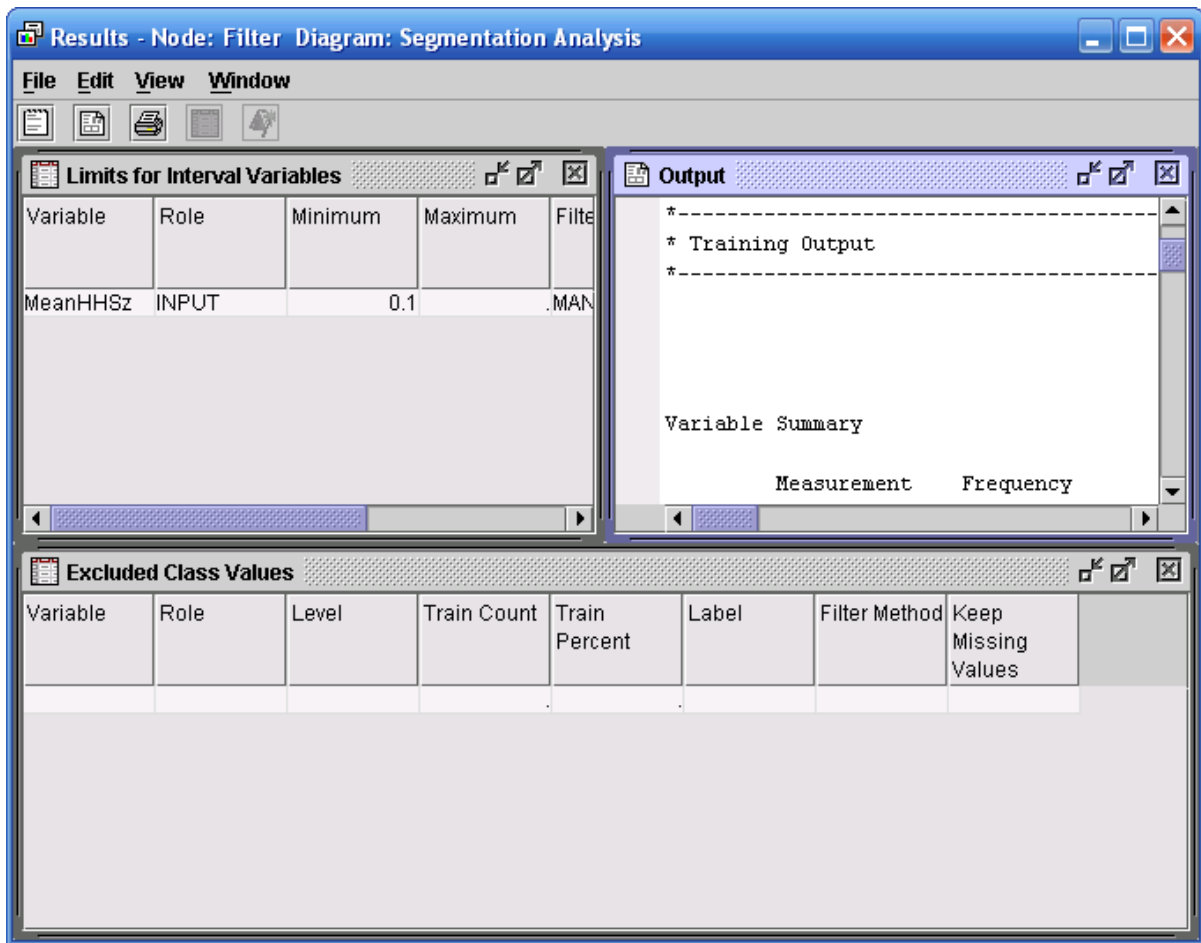
Name	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Report
LocX	Default	Default	.	.	No
LocY	Default	Default	.	.	No
MeanHHSz	Default	Default	0.1	.	No
MedHHInc	Default	Default	.	.	No
RegDens	Default	Default	.	.	No
RegPop	Default	Default	.	.	No

Generate Summary OK Cancel

8. 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.

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



10. Go to line 38 in the Output window.

Number Of Observations			
Data			
Role	Filtered	Excluded	DATA
TRAIN	32097	1081	33178

The Filter node removed 1081 cases with zero household size.

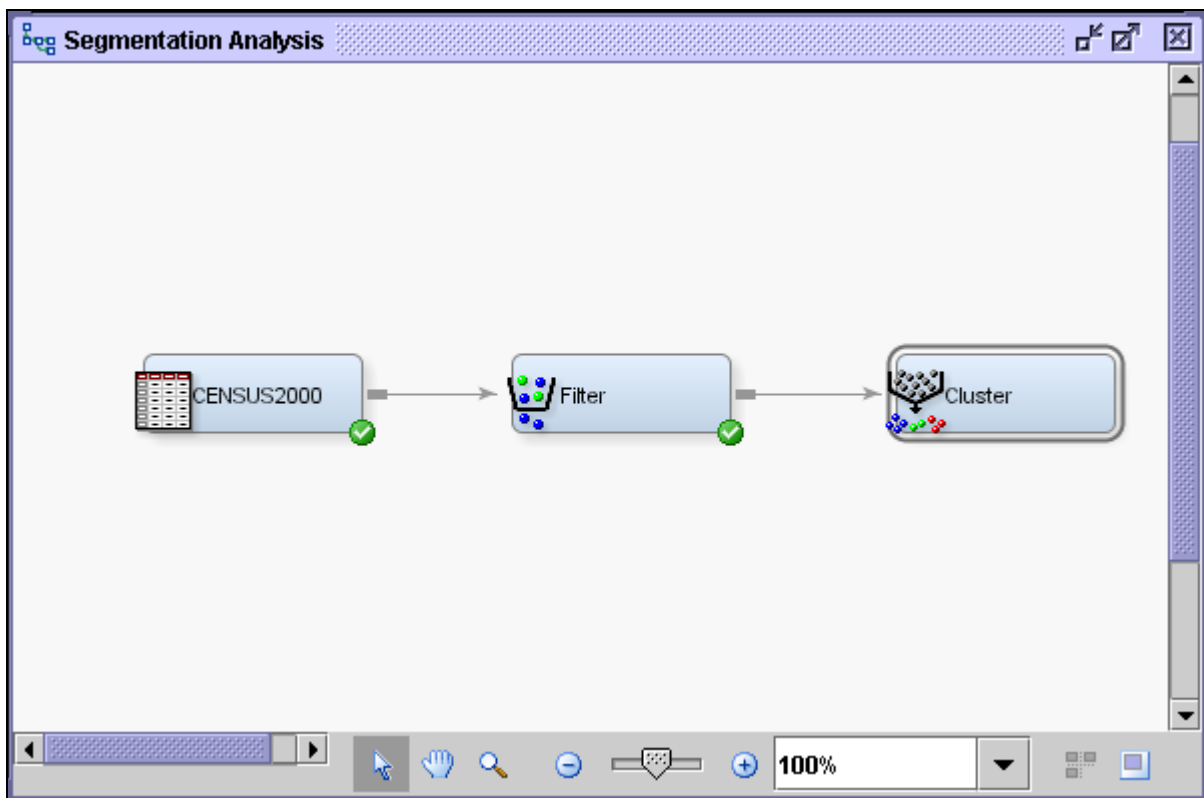
11. Close the Results window. The **CENSUS2000** data is ready for segmentation.



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
- standardize the inputs to have a similar range

4. Select the **Variables...** property for the Cluster node. The Variables window opens.

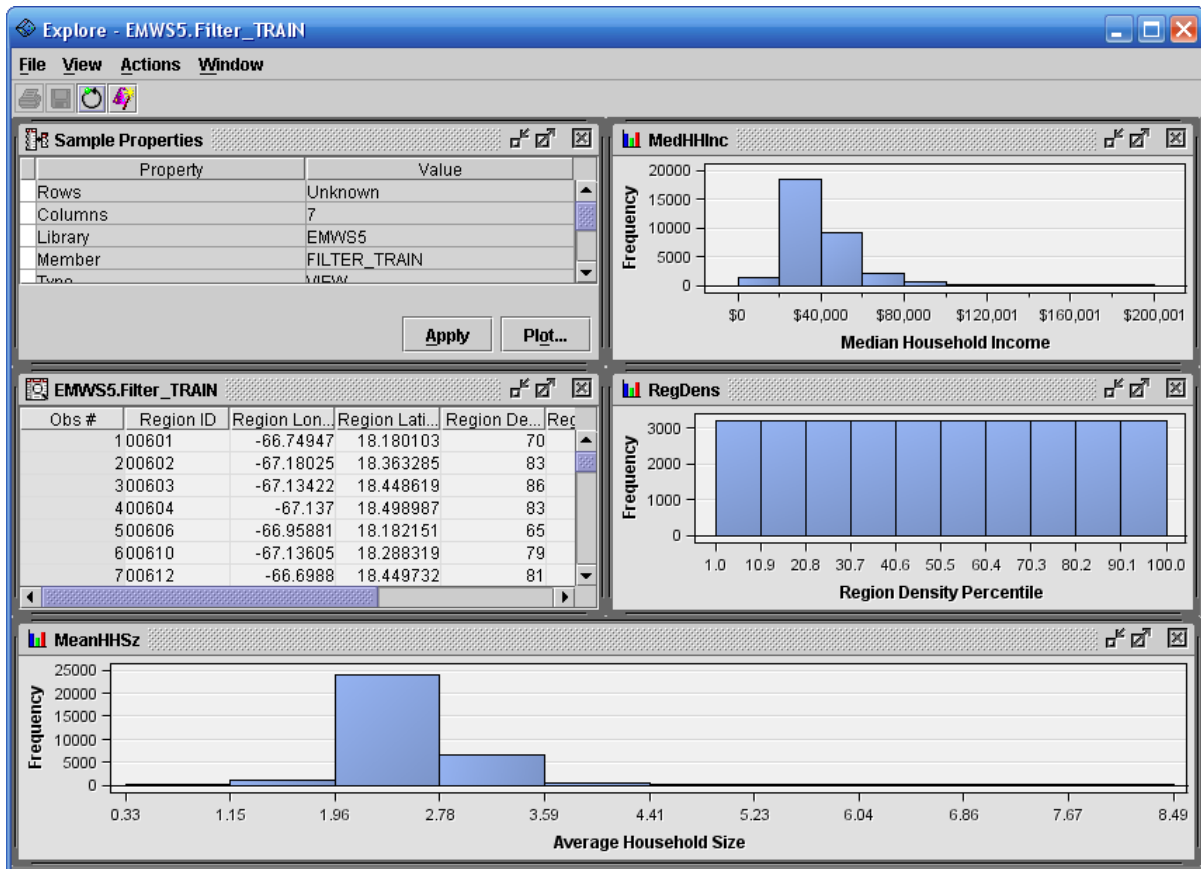
5. Select **Use** ⇒ **No** for **LocX**, **LocY**, and **RegPop**.

Name	Use	Report	Role	Level	Type	Order	Label	Format
ID	Yes	No	ID	Nominal	Character		Region ID	
LocX	No	No	Input	Interval	Numeric		Region Longit	BEST9.2
LocY	No	No	Input	Interval	Numeric		Region Latitud	BEST9.2
MeanHHSz	Default	No	Input	Interval	Numeric		Average Hous	COMMA9.2
MedHHInc	Default	No	Input	Interval	Numeric		Median House	DOLLAR9.0
RegDens	Default	No	Input	Interval	Numeric		Region Densi	BEST9.0
RegPop	No	No	Input	Interval	Numeric		Region Popul	COMMA9.0

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

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.

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



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

7. Close the Explore window.
8. Select **OK** to close the Variables window.

9. Select **Internal Standardization** ⇔ **Standardization**. 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
Maximum Number of C10	



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.

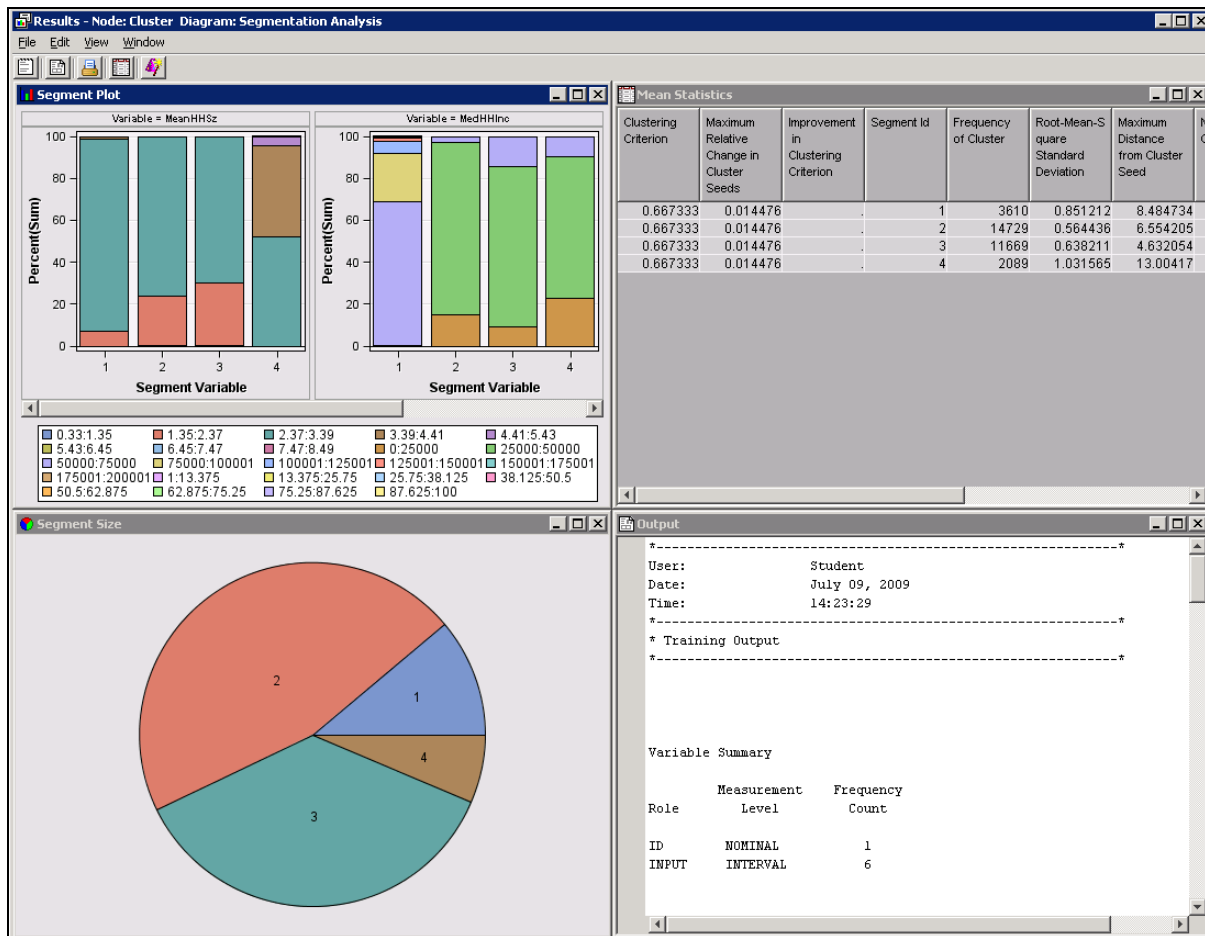


Creating Clusters with the Cluster Tool

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.

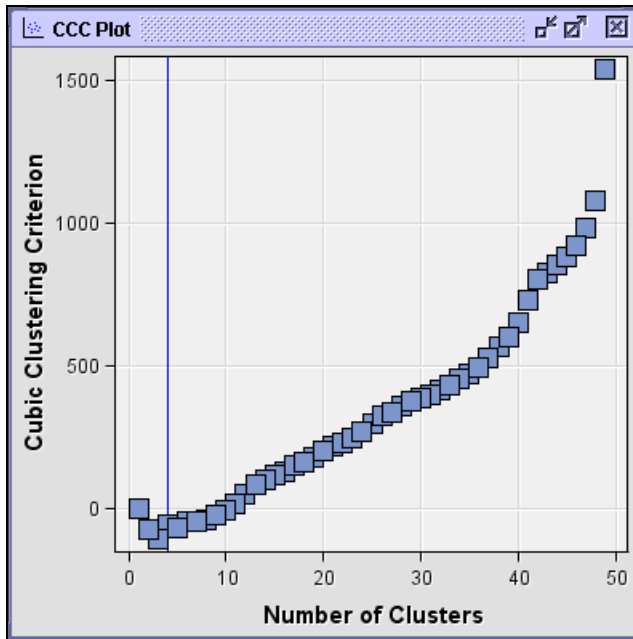


The Results - Cluster window contains four embedded windows.

- 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.

2. 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 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.



Specifying the Segment Count

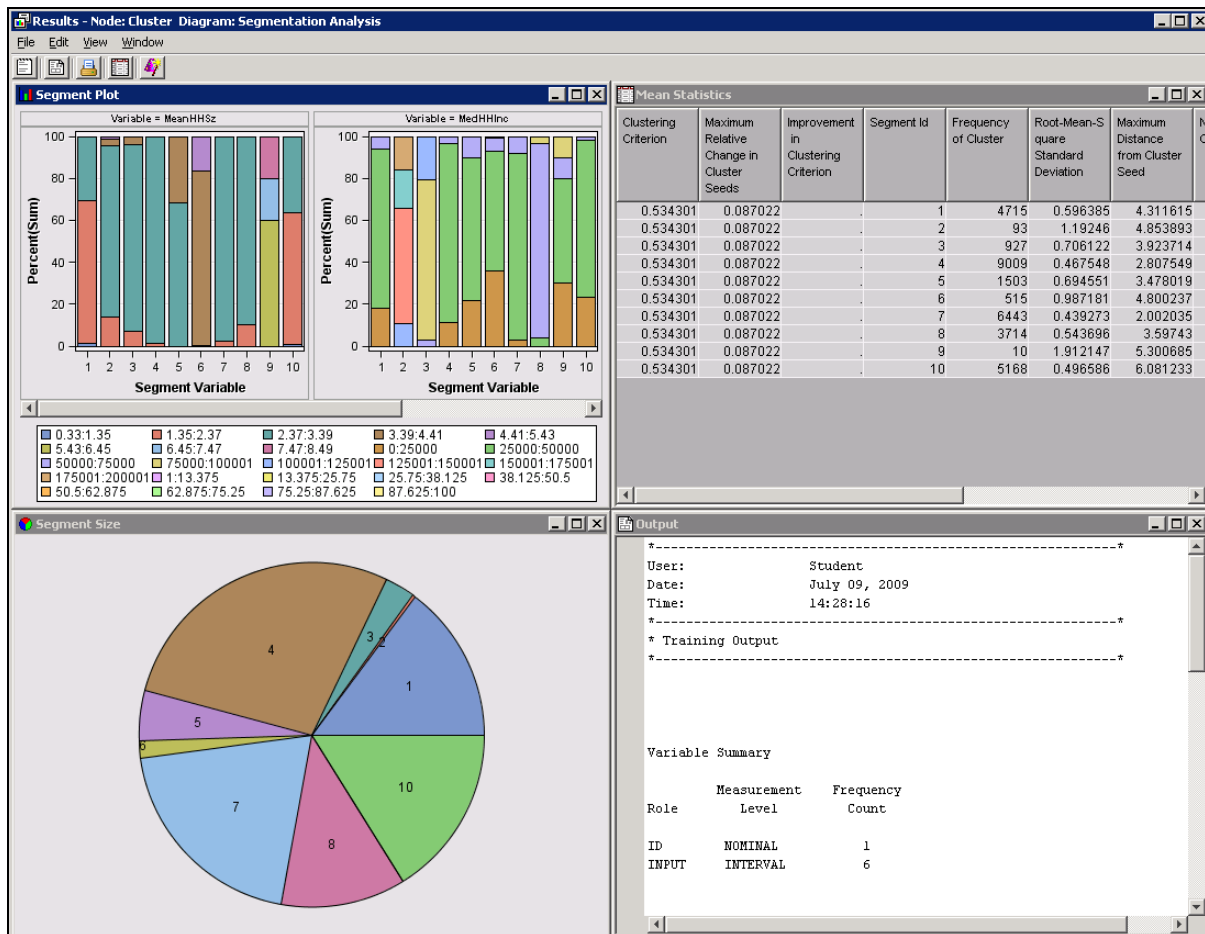
You might want to increase the number of clusters created by the Cluster node. You can do this by changing the CCC cutoff property or by specifying the desired number of clusters.

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

Property	Value
General	
Node ID	Clus
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Cluster Variable Role	Segment
Internal Standardization	Standardization
<input checked="" type="checkbox"/> Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	10

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

2. Run the Cluster node 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.

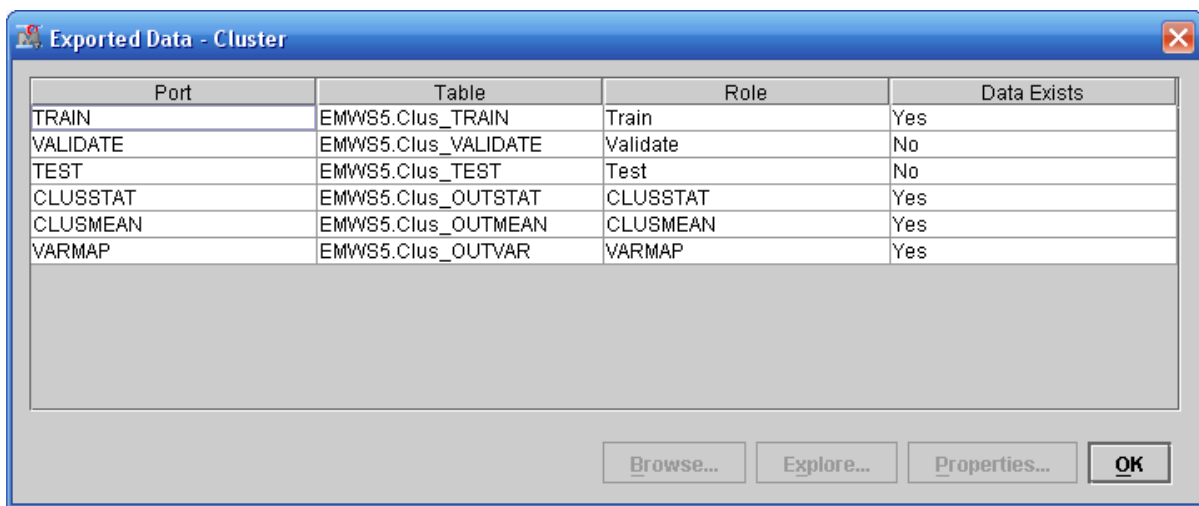
Mean Statistics						
Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	_SEGMENT_	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster
0.087022	.	1	4715	0.596385	4.311615	
0.087022	.	2	93	1.19246	4.853893	
0.087022	.	3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



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 - Cluster window.
2. Select **Exported Data** from the Properties panel for the Cluster node. The Exported Data - Cluster window opens.

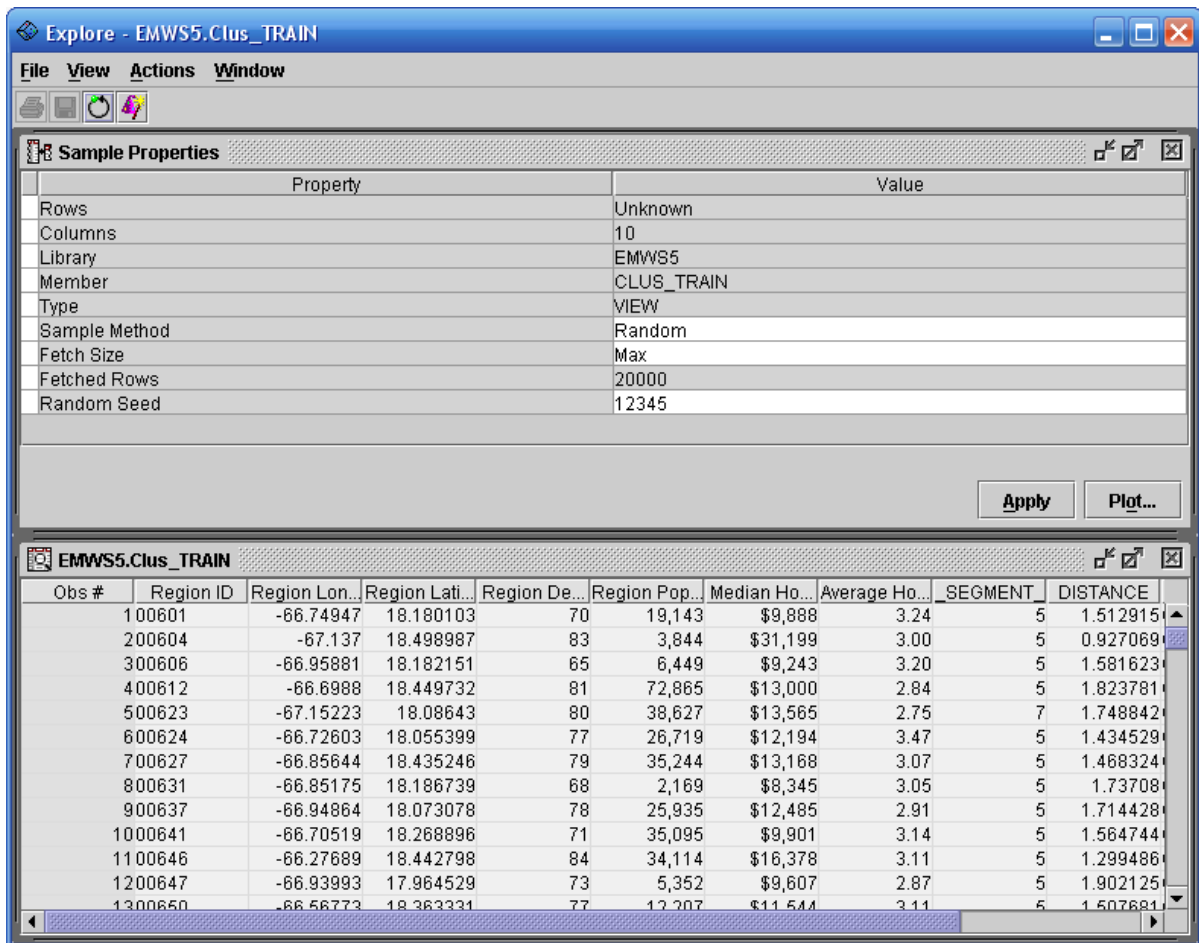


Port	Table	Role	Data Exists
TRAIN	EMWS5.Clus_TRAIN	Train	Yes
VALIDATE	EMWS5.Clus_VALIDATE	Validate	No
TEST	EMWS5.Clus_TEST	Test	No
CLUSSTAT	EMWS5.Clus_OUTSTAT	CLUSSTAT	Yes
CLUSMEAN	EMWS5.Clus_OUTMEAN	CLUSMEAN	Yes
VARMAP	EMWS5.Clus_OUTVAR	VARMAP	Yes

Buttons: Browse... Explore... Properties... OK

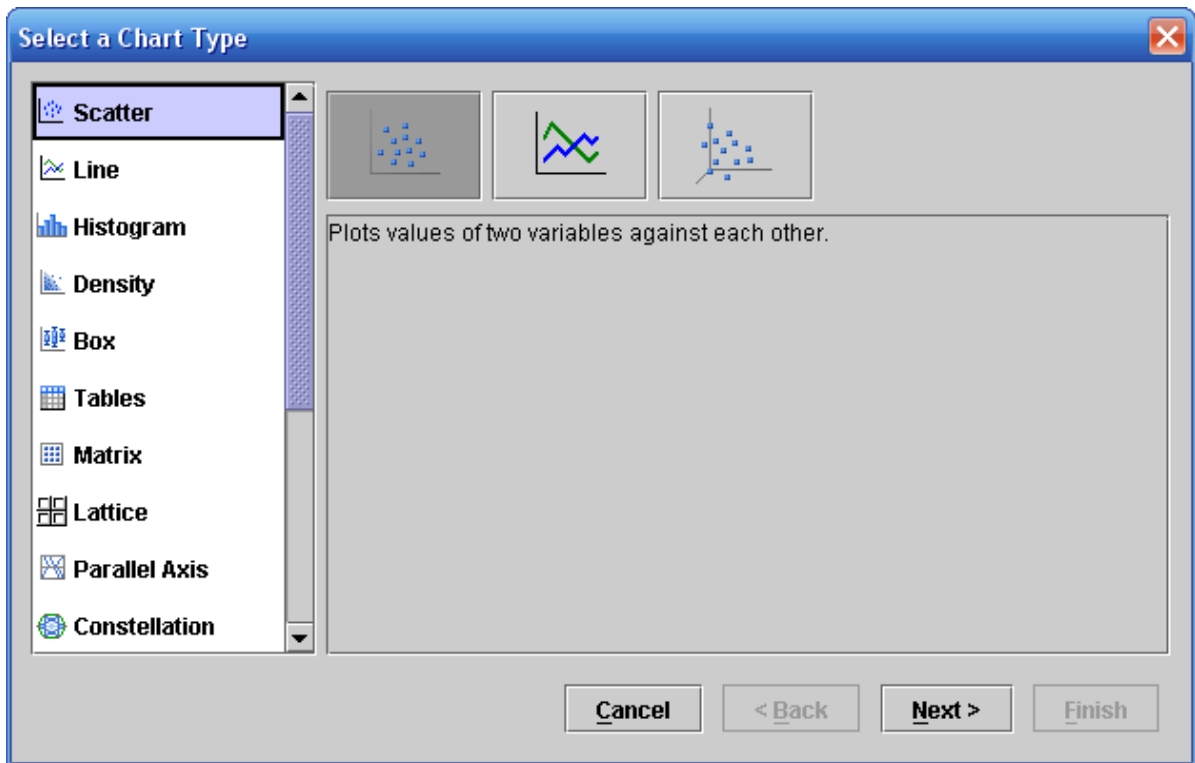
This window shows the data sets that are generated and exported by the Cluster node.

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

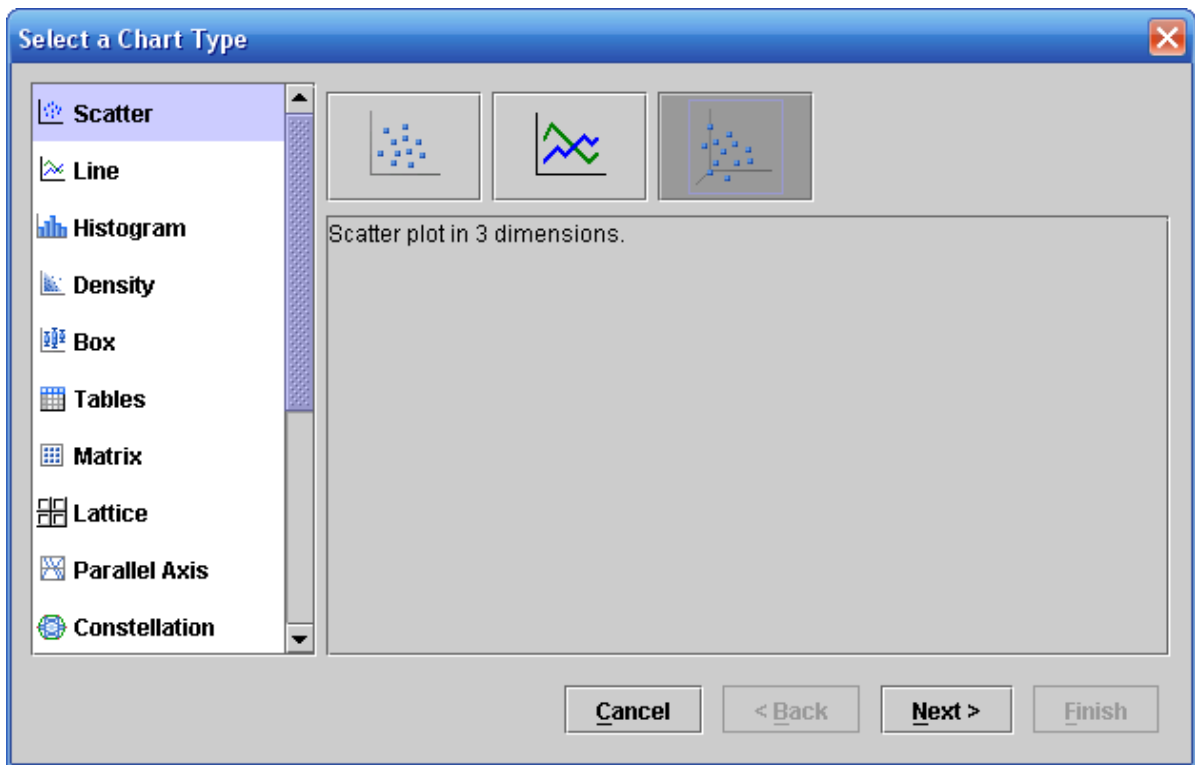


You can use the Graph Wizard to generate a three-dimensional plot of the **CENSUS2000** data.

4. Select **Actions** ⇒ **Plot**. The Select a Chart Type window opens.



5. Select the icon for a three-dimensional scatter plot.



6. Select **Next >**. The Graph Wizard proceeds to the next step. Select **Chart Roles**.

7. Select roles of **X**, **Y**, and **Z** for **MeanHHSz**, **MedHHInc**, and **RegDens**, respectively.
8. Select **Role** ⇒ **Color** for **_SEGMENT_**.

Select Chart Roles X

Use default assignments

▲ Variable	Role	Type	Description	Format
SEGMENT	Color	Numeric	Segment Id	
_SEGMENT_LABEL_		Character	Segment Description	
Distance		Numeric	Distance	
ID		Character	Region ID	
LocX		Numeric	Region Longitude	BEST9
LocY		Numeric	Region Latitude	BEST9
MeanHHSz	X	Numeric	Average Household S...	COMMA9.2
MedHHInc	Y	Numeric	Median Household Inc...	DOLLAR9
RegDens	Z	Numeric	Region Density Perce...	BEST9
RegPop		Numeric	Region Population	COMMA9

☐ Allow multiple role assignments

9. Select **Finish**.

The Explore window opens with a three-dimensional plot of the **CENSUS2000** data.

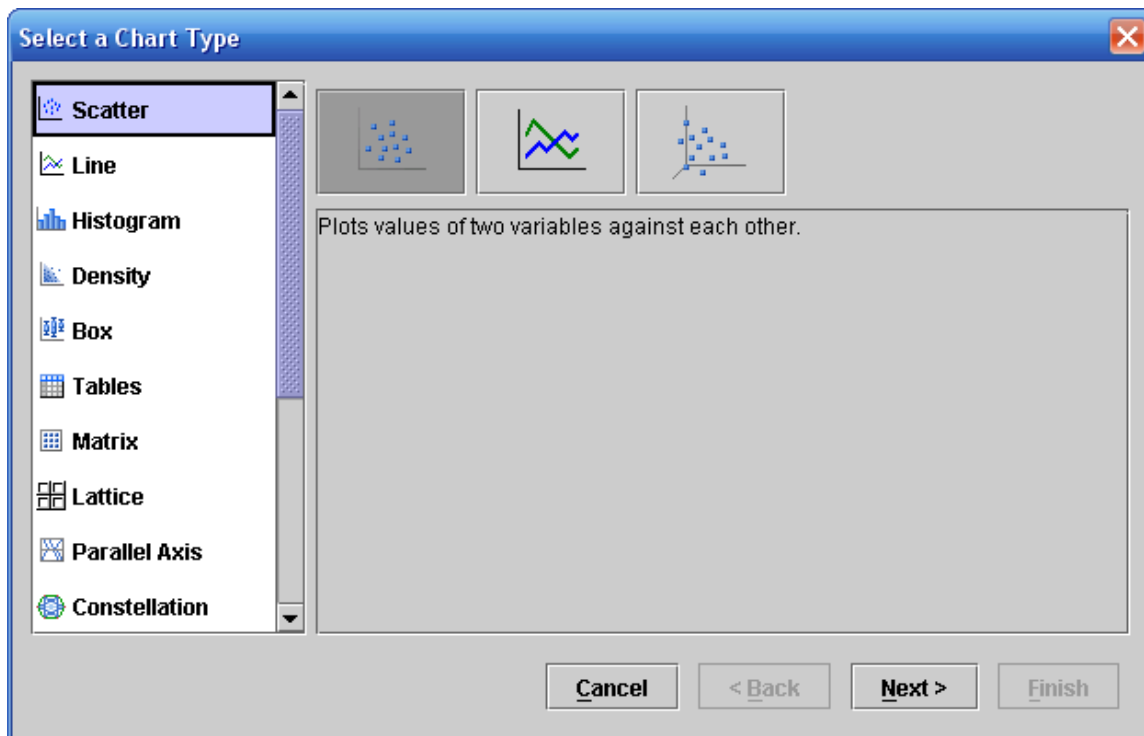


10. Rotate the plot by holding down the CTRL key and dragging the mouse.

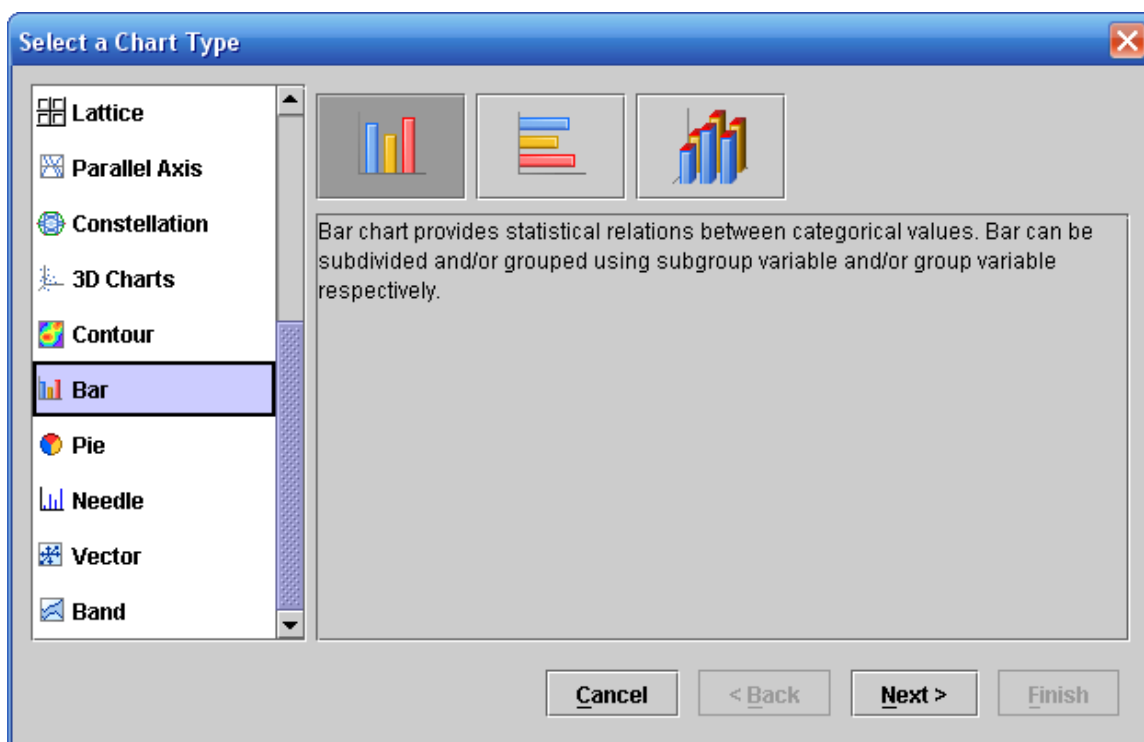
Each square in the plot represents a unique postal code. The squares are color-coded by cluster segment.

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.



3. Select **Next >**.

4. Select **Role** ⇒ **Category** for the variable **_SEGMENT_**.

Select Chart Roles

Use default assignments

Variable	Role	Type	Description	Format
SEGMENT	Category	Numeric	Segment Id	
_SEGMENT_LABEL_		Character	Segment Description	
Distance		Numeric	Distance	
ID		Character	Region ID	
LocX		Numeric	Region Longitude	BEST9
LocY		Numeric	Region Latitude	BEST9
MeanHHSz		Numeric	Average Household S...	COMMA9.2
MedHHInc		Numeric	Median Household Inc...	DOLLAR9
RegDens		Numeric	Region Density Perce...	BEST9
RegPop		Numeric	Region Population	COMMA9

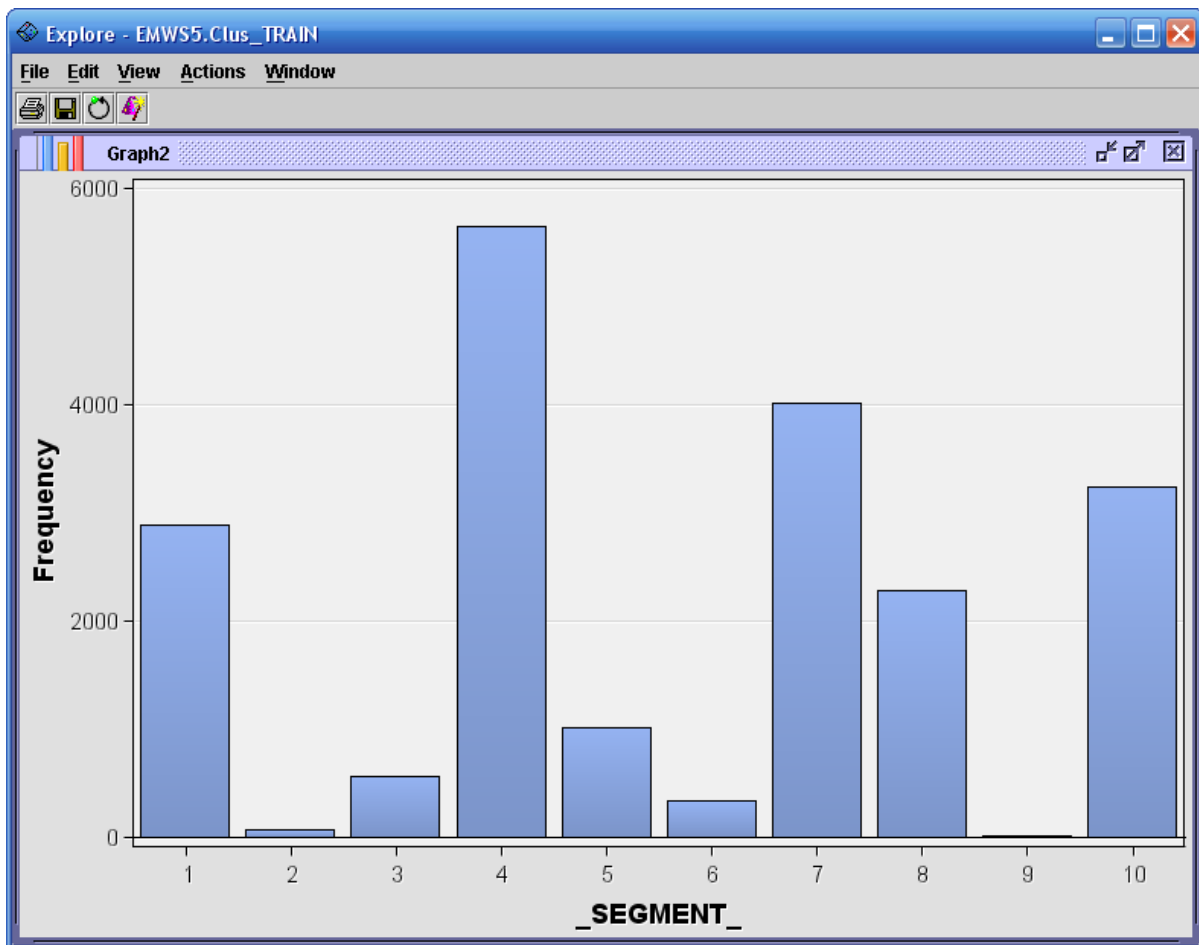
Response statistic: Frequency

☐ Allow multiple role assignments

Cancel < Back Next > Finish

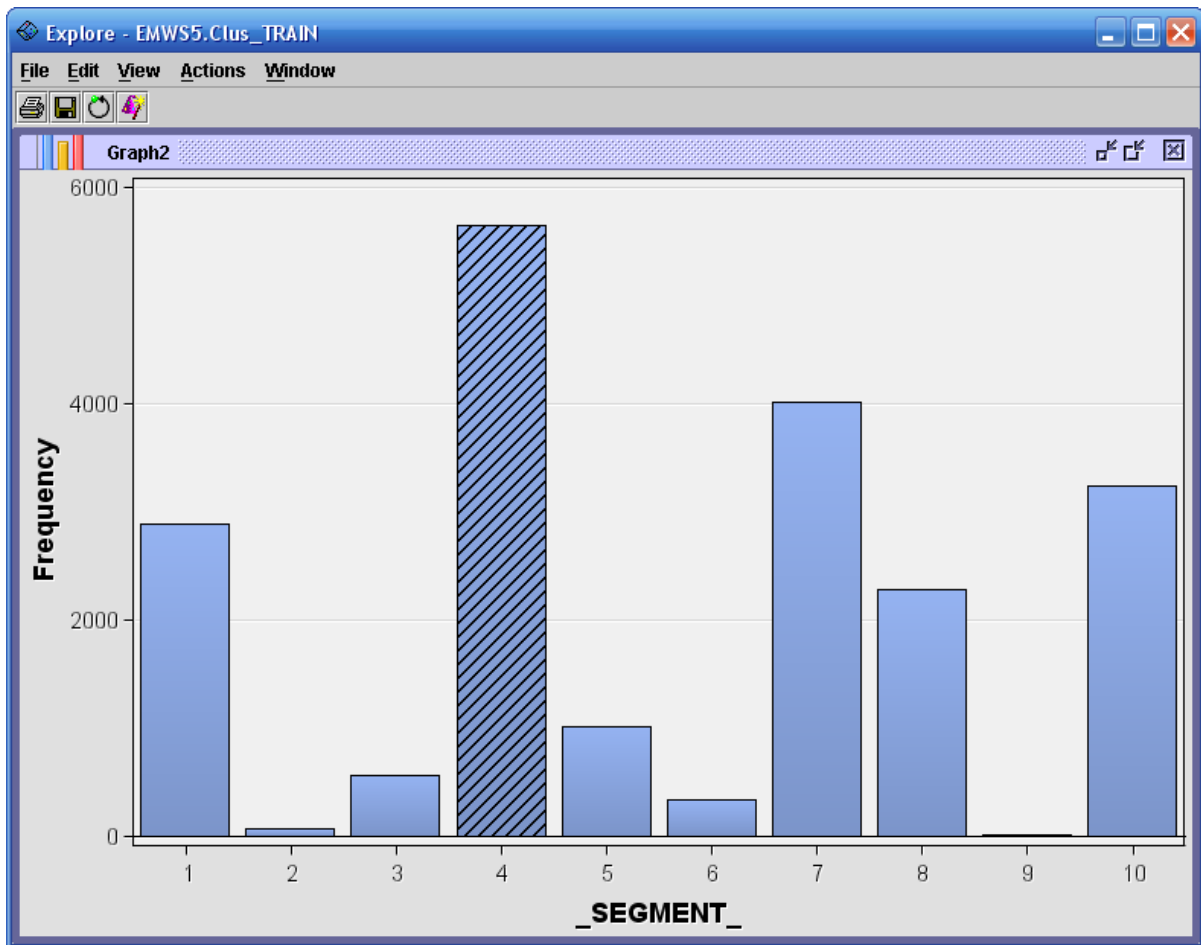
5. Select **Finish**.

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.

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



7. Select the three-dimensional plot. Cases corresponding to segment 4 are highlighted.
8. 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.

9. Close the Explore, Exported Data, and Results windows.

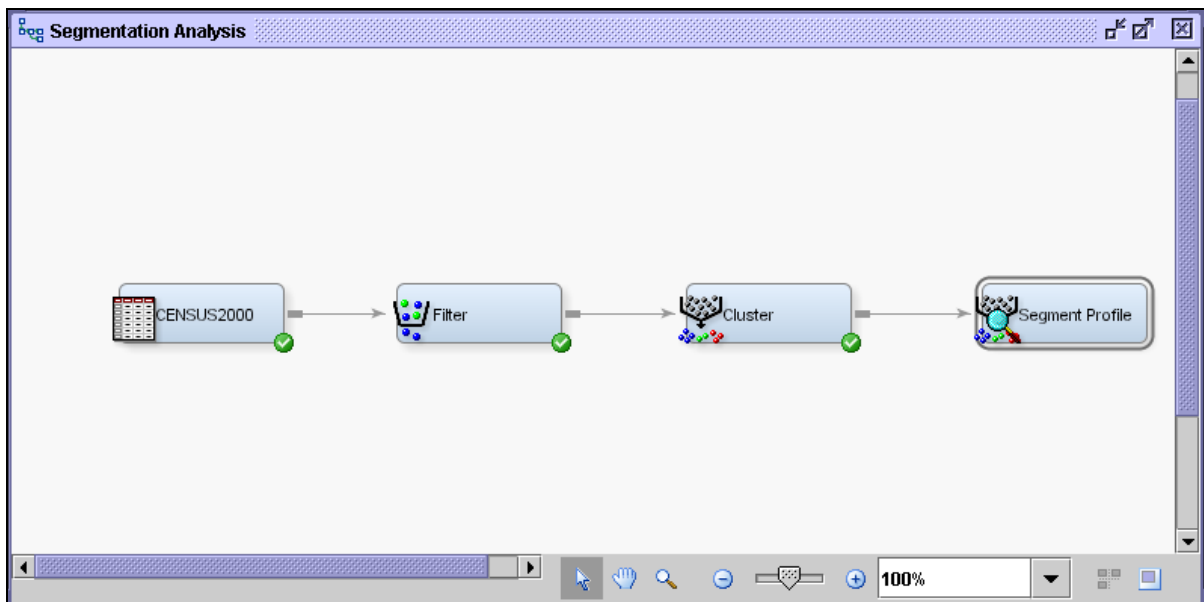


Profiling Segments

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.

3. Select the **Variables** property for the Segment Profile node.

Variables - Prof

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

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

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

Variables - Prof

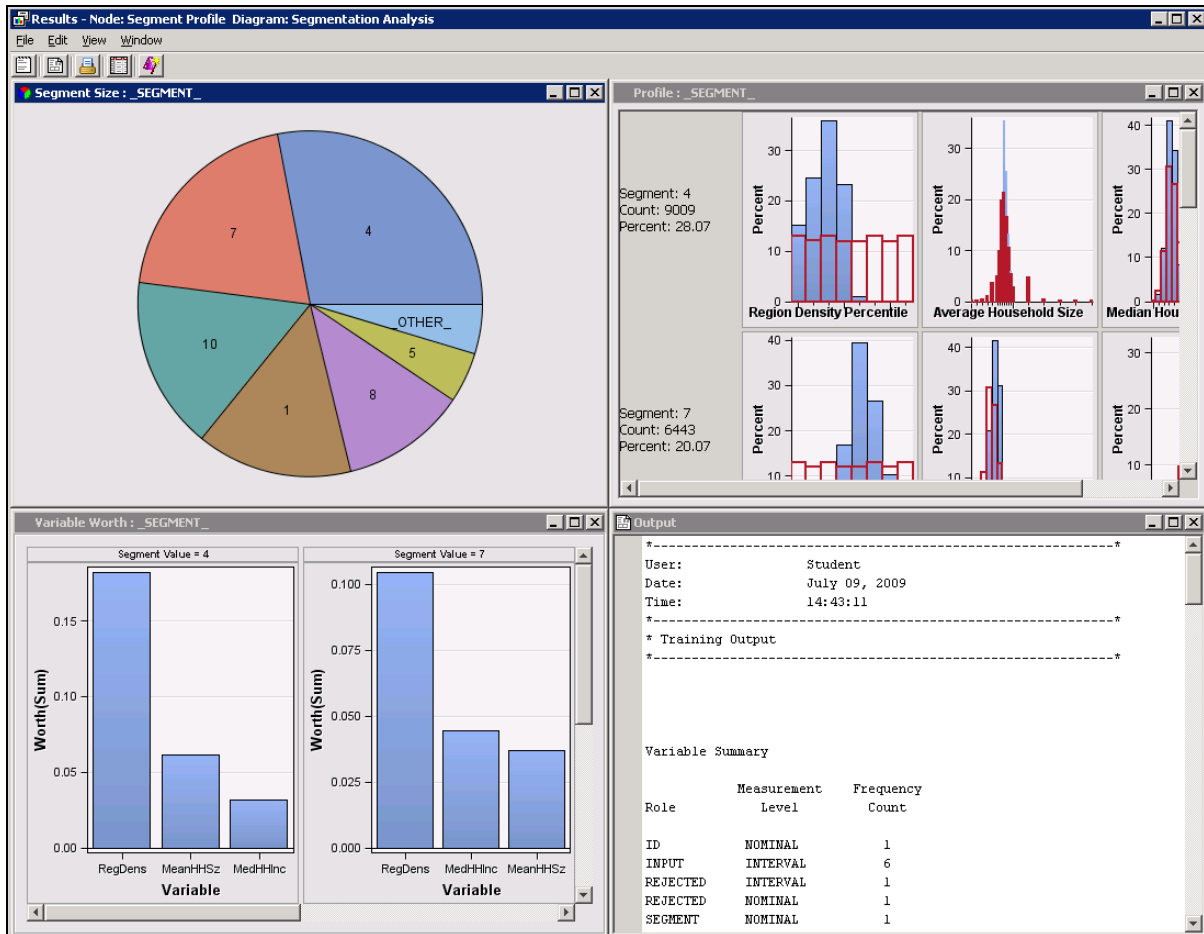
(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

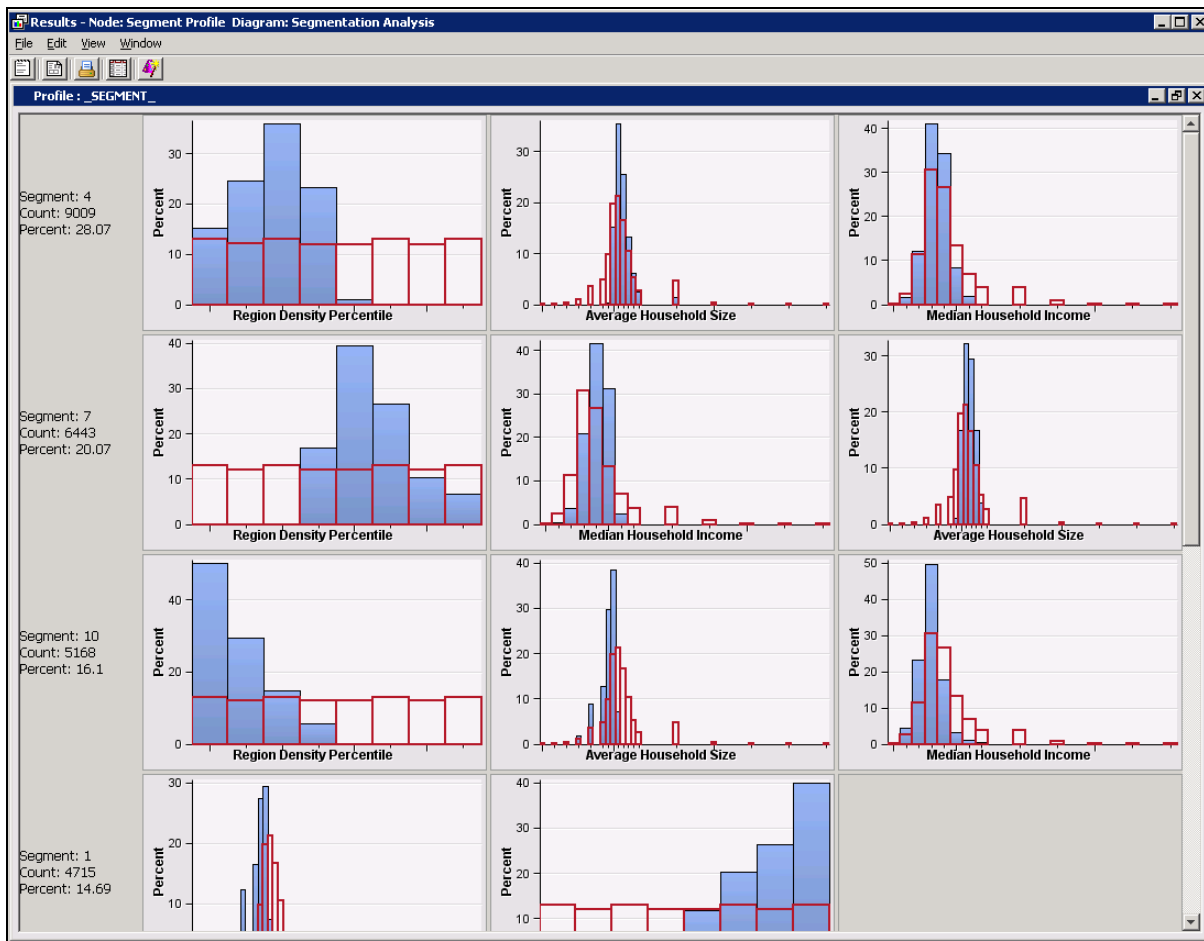
Name /	Use	Report	Role	Level
Distance	Default	No	Rejected	Interval
ID	No	No	ID	Nominal
LocX	No	No	Input	Interval
LocY	No	No	Input	Interval
MeanHHSz	Default	No	Input	Interval
MedHHInc	Default	No	Input	Interval
RegDens	Default	No	Input	Interval
RegPop	No	No	Input	Interval
SEGMENT	Default	No	Segment	Nominal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

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

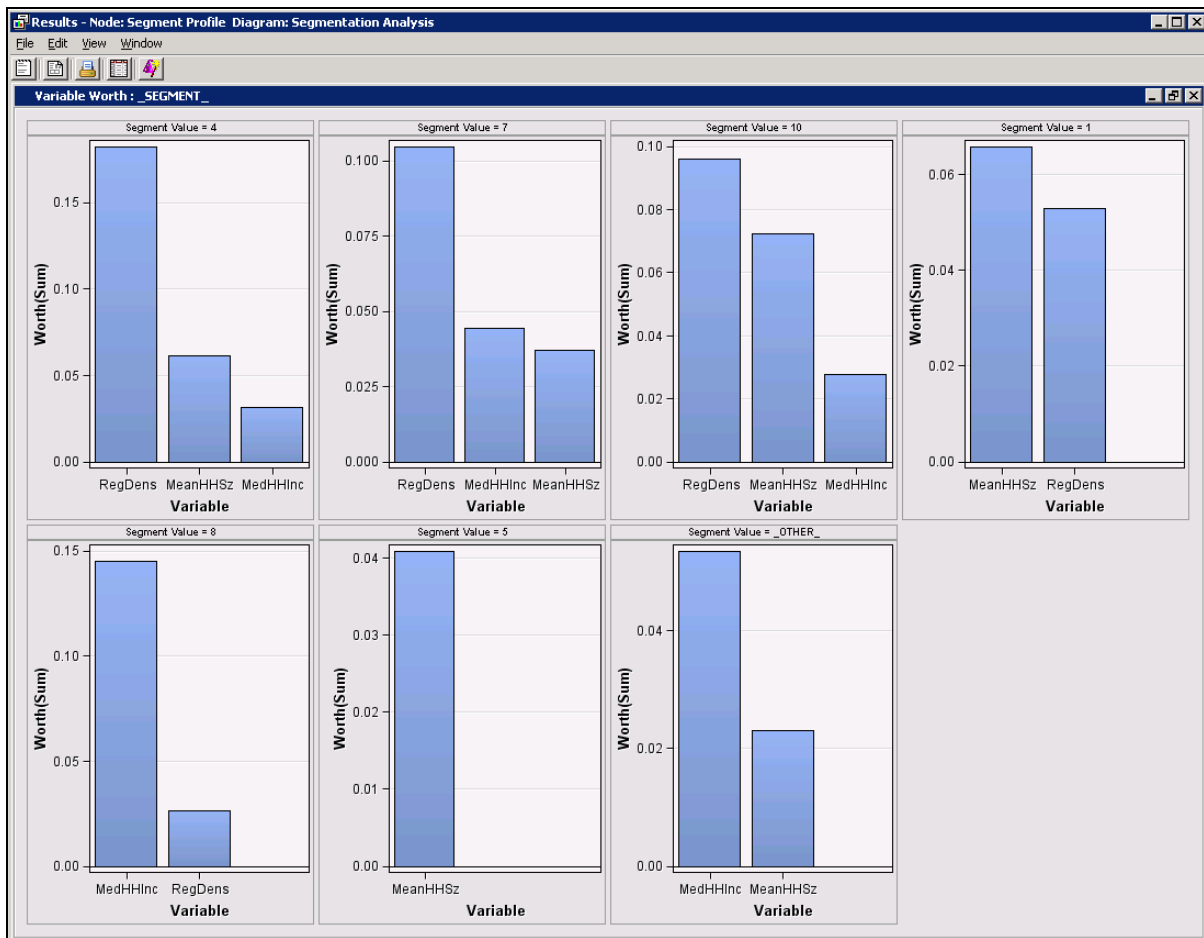
6. Run the Segment Profile node and select **Results...**. The Results - Node: Segment Profile Diagram window opens.



7. Maximize the Profile 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.

11. 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.



Exercises

1. Conducting Cluster Analysis

The **DUNGAREE** data set gives the number of pairs of four different types of dungarees sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans sold.

Name	Model Role	Measurement Level	Description
STOREID	ID	Nominal	Identification number of the store
FASHION	Input	Interval	Number of pairs of fashion jeans sold at the store
LEISURE	Input	Interval	Number of pairs of leisure jeans sold at the store
STRETCH	Input	Interval	Number of pairs of stretch jeans sold at the store
ORIGINAL	Input	Interval	Number of pairs of original jeans sold at the store
SALESTOT	Rejected	Interval	Total number of pairs of jeans sold (the sum of FASHION, LEISURE, STRETCH, and ORIGINAL)

- Create a new diagram in your project. Name the diagram **Jeans**.
- Define the data set **DUNGAREE** as a data source.
- Determine whether the model roles and measurement levels assigned to the variables are appropriate.

Examine the distribution of the variables.

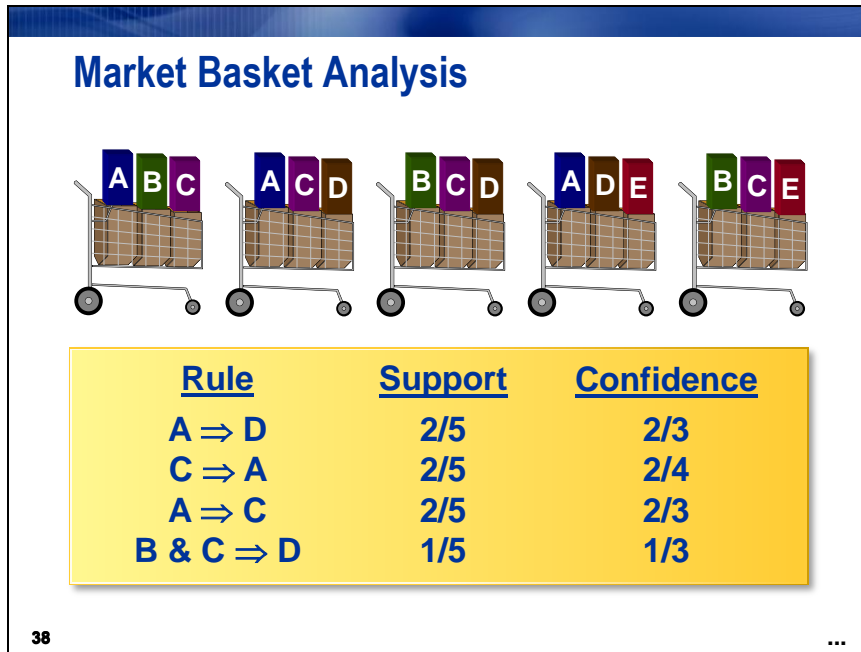
- Are there any unusual data values? _____
- Are there missing values that should be replaced? _____

- Assign the variable **STOREID** the model role **ID** and the variable **SALESTOT** the model role **Rejected**. Make sure that the remaining variables have the Input model role and the Interval measurement level. Why should the variable **SALESTOT** be rejected? _____

- Add an Input Data Source node to the diagram workspace and select the **DUNGAREE** data table as the data source.

- f. Add a Cluster node to the diagram workspace and connect it to the Input Data node.
- g. Select the **Cluster** node and select **Internal Standardization** ⇔ **Standardization**. What would happen if you did not standardize your inputs? _____
- _____
- h. Run the diagram from the Cluster node and examine the results.
Does the number of clusters created seem reasonable? _____
- i. Specify a maximum of six clusters and rerun the Cluster node.
How does the number and quality of clusters compare to that obtained in part h? _____
- _____
- j. Use the Segment Profile node to summarize the nature of the clusters.

8.3 Market Basket Analysis (Self-Study)



Market basket analysis (also known as *association rule discovery* or *affinity analysis*) is a popular data mining method. In the simplest situation, the data consists of two variables: a *transaction* and an *item*.

For each transaction, there is a list of items. Typically, a transaction is a single customer purchase, and the items are the things that were bought. An *association rule* is a statement of the form (item set A) \Rightarrow (item set B).

The aim of the analysis is to determine the strength of all the association rules among a set of items.

The strength of the association is measured by the *support* and *confidence* of the rule. The support for the rule $A \Rightarrow B$ is the probability that the two item sets occur together. The support of the rule $A \Rightarrow B$ is estimated by the following:

$$\frac{\text{transactions that contain every item in } A \text{ and } B}{\text{all transactions}}$$

Notice that support is symmetric. That is, the support of the rule $A \Rightarrow B$ is the same as the support of the rule $B \Rightarrow A$.

The confidence of an association rule $A \Rightarrow B$ is the conditional probability of a transaction containing item set B given that it contains item set A . The confidence is estimated by the following:

$$\frac{\text{transactions that contain every item in } A \text{ and } B}{\text{transactions that contain the items in } A}$$

		Checking Account		
		No	Yes	
Savings Account	No	500	3500	4,000
	Yes	1000	5000	6,000
				10,000
Support(SVG \Rightarrow CK) = 50% Confidence(SVG \Rightarrow CK) = 83% Expected Confidence(SVG \Rightarrow CK) = 85% Lift(SVG \Rightarrow CK) = $0.83/0.85 < 1$				

The interpretation of the implication (\Rightarrow) in association rules is precarious. High confidence and support does not imply cause and effect. The rule is not necessarily interesting. The two items might not even be correlated. The term *confidence* is not related to the statistical usage; therefore, there is no repeated sampling interpretation.

Consider the association rule (saving account) \Rightarrow (checking account). This rule has 50% support (5,000/10,000) and 83% confidence (5,000/6,000). Based on these two measures, this might be considered a strong rule. On the contrary, those **without** a savings account are even more likely to have a checking account (87.5%). Saving and checking are, in fact, negatively correlated.

If the two accounts were independent, then knowing that a person has a saving account does not help in knowing whether that person has a checking account. The expected confidence if the two accounts were independent is 85% (8,500/10,000). This is higher than the confidence of SVG \Rightarrow CK.

The *lift* of the rule $A \Rightarrow B$ is the confidence of the rule divided by the expected confidence, assuming that the item sets are independent. The lift can be interpreted as a general measure of association between the two item sets. Values greater than 1 indicate positive correlation, values equal to 1 indicate zero correlation, and values less than 1 indicate negative correlation. Notice that lift is symmetric. That is, the lift of the rule $A \Rightarrow B$ is the same as the lift of the rule $B \Rightarrow A$.

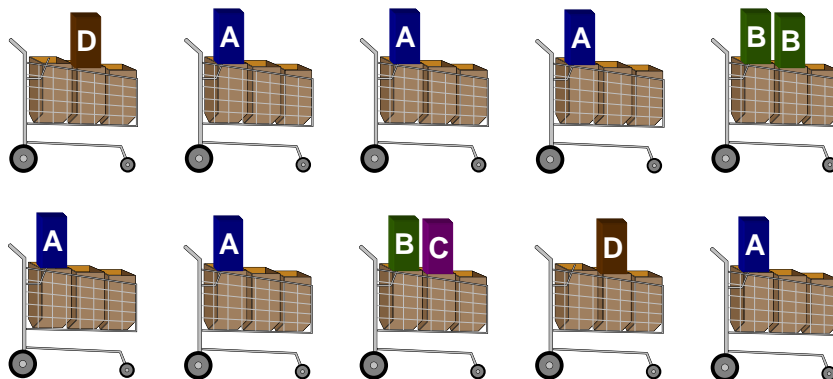
Barbie Doll \Rightarrow Candy

1. Put them closer together in the store.
2. Put them far apart in the store.
3. Package candy bars with the dolls.
4. Package Barbie + candy + poorly selling item.
5. Raise the price on one, and lower it on the other.
6. Offer Barbie accessories for proofs of purchase.
7. Do not advertise candy and Barbie together.
8. Offer candies in the shape of a Barbie doll.

41

Forbes (Palmeri 1997) reported that a major retailer determined that customers who buy Barbie dolls have a 60% likelihood of buying one of three types of candy bars. The confidence of the rule Barbie \Rightarrow candy is 60%. The retailer was unsure what to do with this nugget. The online newsletter *Knowledge Discovery Nuggets* invited suggestions (Piatesky-Shapiro 1998).

Data Capacity



42

In data mining, the data is not generated to meet the objectives of the analysis. It must be determined whether the data, as it exists, has the capacity to meet the objectives. For example, quantifying affinities among related items would be pointless if very few transactions involved multiple items. Therefore, it is important to do some initial examination of the data before attempting to do association analysis.

Association Tool Demonstration

Analysis goal:

Explore associations between retail banking services used by customers.

Analysis plan:

- Create an association data source.
- Run an association analysis.
- Interpret the association rules.
- Run a sequence analysis.
- Interpret the sequence rules.

43

A bank's Marketing Department is interested in examining associations between various retail banking services used by customers. Marketing would like to determine both typical and atypical service combinations as well as the order in which the services were first used.

These requirements suggest both a market basket analysis and a sequence analysis.



Market Basket Analysis

The **BANK** data set contains service information for nearly 8,000 customers. There are three variables in the data set, as shown in the table below.

Name	Model Role	Measurement Level	Description
ACCOUNT	ID	Nominal	Account Number
SERVICE	Target	Nominal	Type of Service
VISIT	Sequence	Ordinal	Order of Product Purchase

The **BANK** data set has over 32,000 rows. Each row of the data set represents a customer-service combination. Therefore, a single customer can have multiple rows in the data set, and each row represents one of the products he or she owns. The median number of products per customer is three.

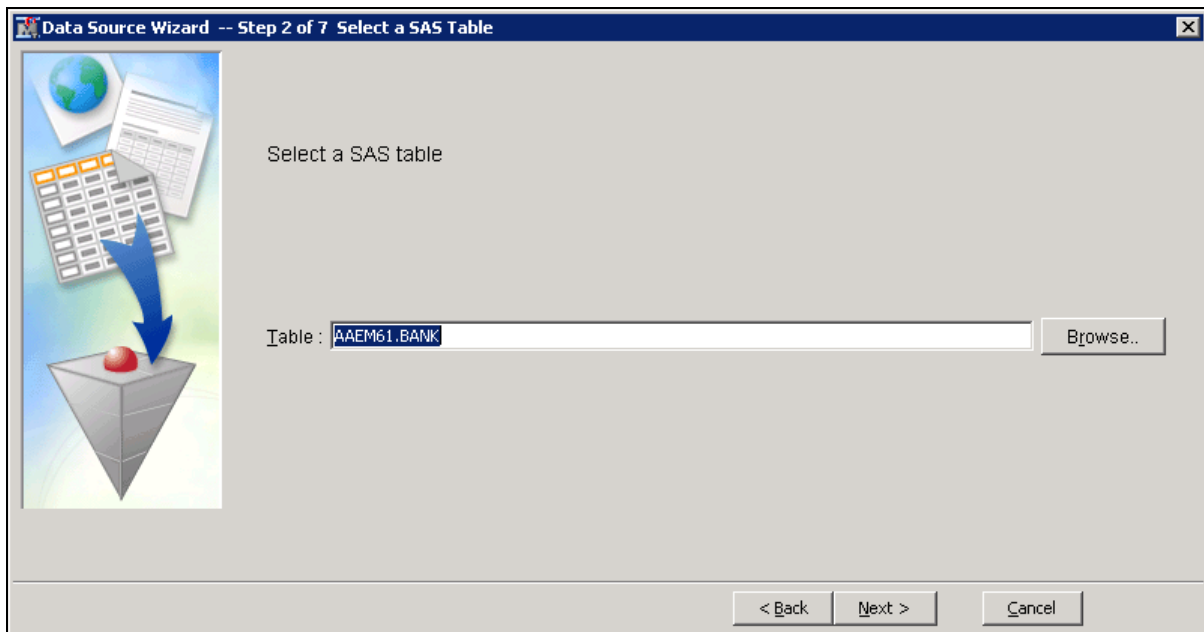
The 13 products are represented in the data set using the following abbreviations:

ATM	automated teller machine debit card
AUTO	automobile installment loan
CCRD	credit card
CD	certificate of deposit
CKCRD	check/debit card
CKING	checking account
HMEQLC	home equity line of credit
IRA	individual retirement account
MMDA	money market deposit account
MTG	mortgage
PLOAN	personal/consumer installment loan
SVG	saving account
TRUST	personal trust account

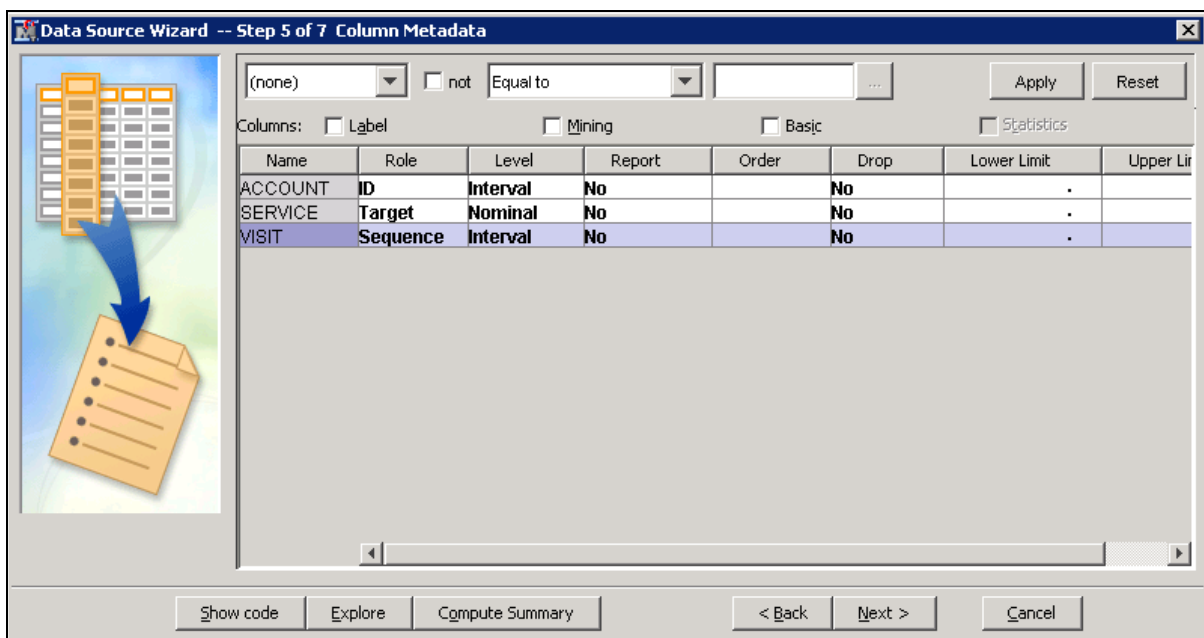
Your first task is to create a new analysis diagram and data source for the **BANK** data set.

1. Create a new diagram named **Associations Analysis** to contain this analysis.
2. Select **Create Data Source** from the Data Sources project property.
3. Proceed to Step 2 of the Data Source Wizard.

4. Select the **BANK** table in the AAEM61 library.



5. Proceed to Step 5 of the Data Source Wizard.
6. In Step 5, assign metadata to the table variables as shown below.

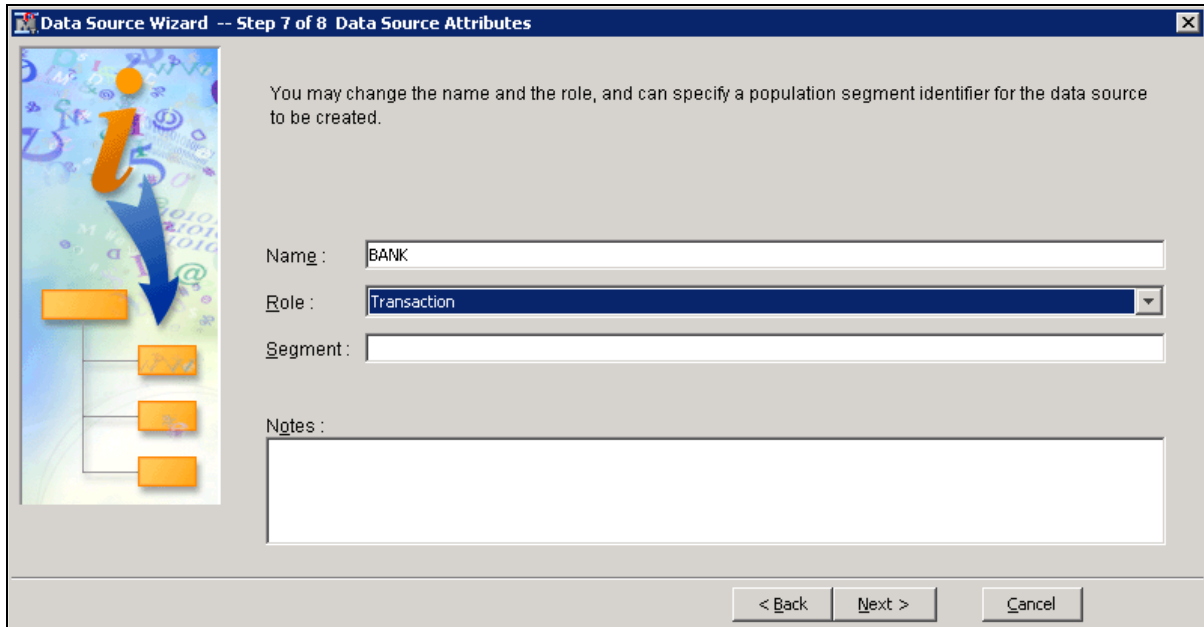


An association analysis requires exactly one target variable and at least one ID variable. Both should have a nominal measurement level. A sequence analysis also requires a sequence variable. It usually has an ordinal measurement scale.

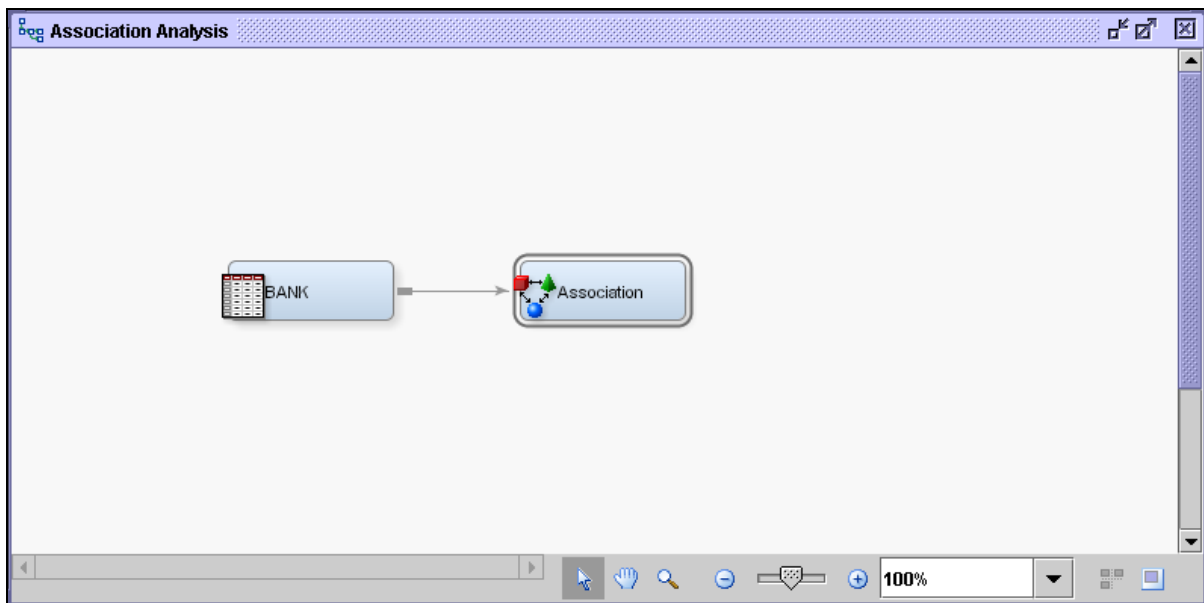
7. Proceed to Step 7 of the Data Source Wizard.

For an association analysis, the data source should have a role of Transaction.

8. Select **Role** ⇒ **Transaction**.



9. Select **Finish** to close the Data Source Wizard.
10. Drag a **BANK** data source into the diagram workspace.
11. Select the **Explore** tab and drag an **Association** tool into the diagram workspace.
12. Connect the **BANK** node to the **Association** node.



13. Select the **Association** node and examine its Properties panel.

Property	Value
General	
Node ID	Assoc
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Maximum Number of Iterations	100000
Rules	...
<input checked="" type="checkbox"/> Association	
Maximum Items	4
Minimum Confidence	10
Support Type	Percent
Support Count	1
Support Percentage	5.0
<input checked="" type="checkbox"/> Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction	0.0
Support Type	Percent
Support Count	1
Support Percentage	2.0
<input checked="" type="checkbox"/> Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	No

14. The Export Rule by ID property determines whether the **Rule-by-ID** data is exported from the node and if the **Rule Description** table will be available for display in the Results window. Set the value for Export Rule by ID to **Yes**.

<input checked="" type="checkbox"/> Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes

Other options in the Properties panel include the following:

- The **Minimum Confidence Level** specifies the minimum confidence level to generate a rule. The default level is 10%.
- The **Support Type** specifies whether the analysis should use the support count or support percentage property. The default setting is `Percent`.
- The **Support Count** specifies a minimum level of support to claim that items are associated (that is, they occur together in the database). The default count is 2.
- The **Support Percentage** specifies a minimum level of support to claim that items are associated (that is, they occur together in the database). The default frequency is 5%. The support percentage figure that you specify refers to the proportion of the largest single item frequency, and not the end support.
- The **Maximum Items** determine the maximum size of the item set to be considered. For example, the default of four items indicates that a maximum of four items will be included in a single association rule.

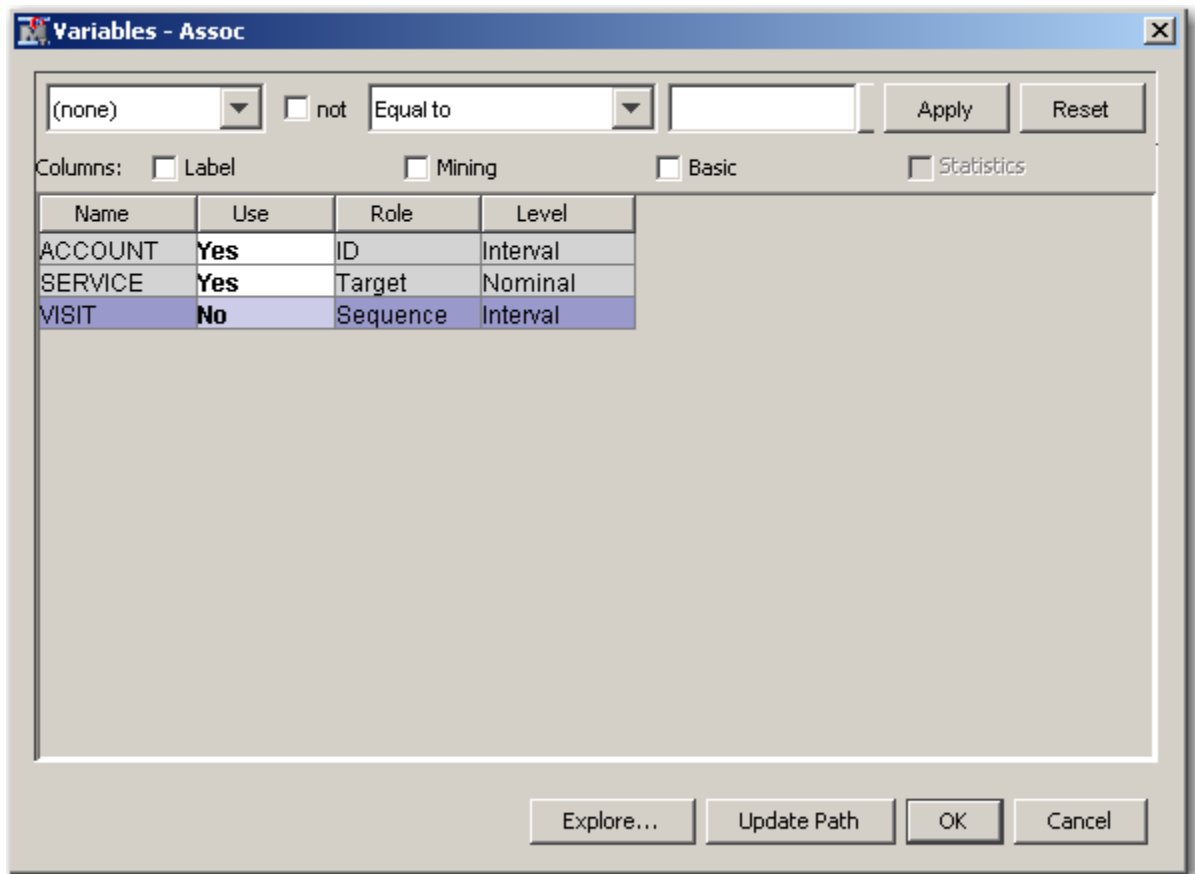


If you are interested in associations that involve fairly rare products, you should consider reducing the support count or percentage when you run the Association node. If you obtain too many rules to be practically useful, you should consider raising the minimum support count or percentage as one possible solution.

Because you first want to perform a market basket analysis, you do not need the sequence variable.

15. Open the Variables dialog box for the Association node.

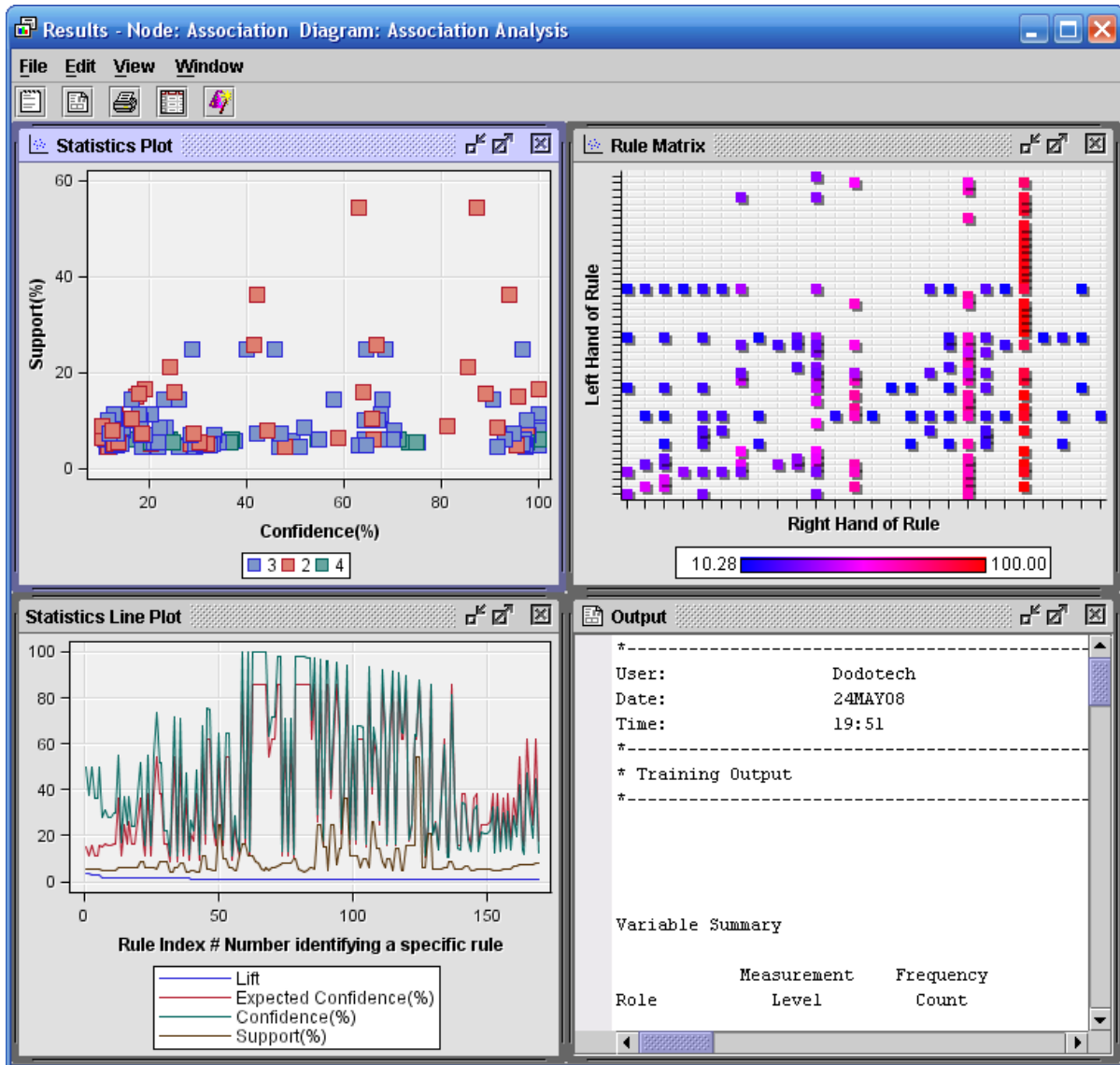
16. Select **Use** ⇒ **No** for the **visit** variable.



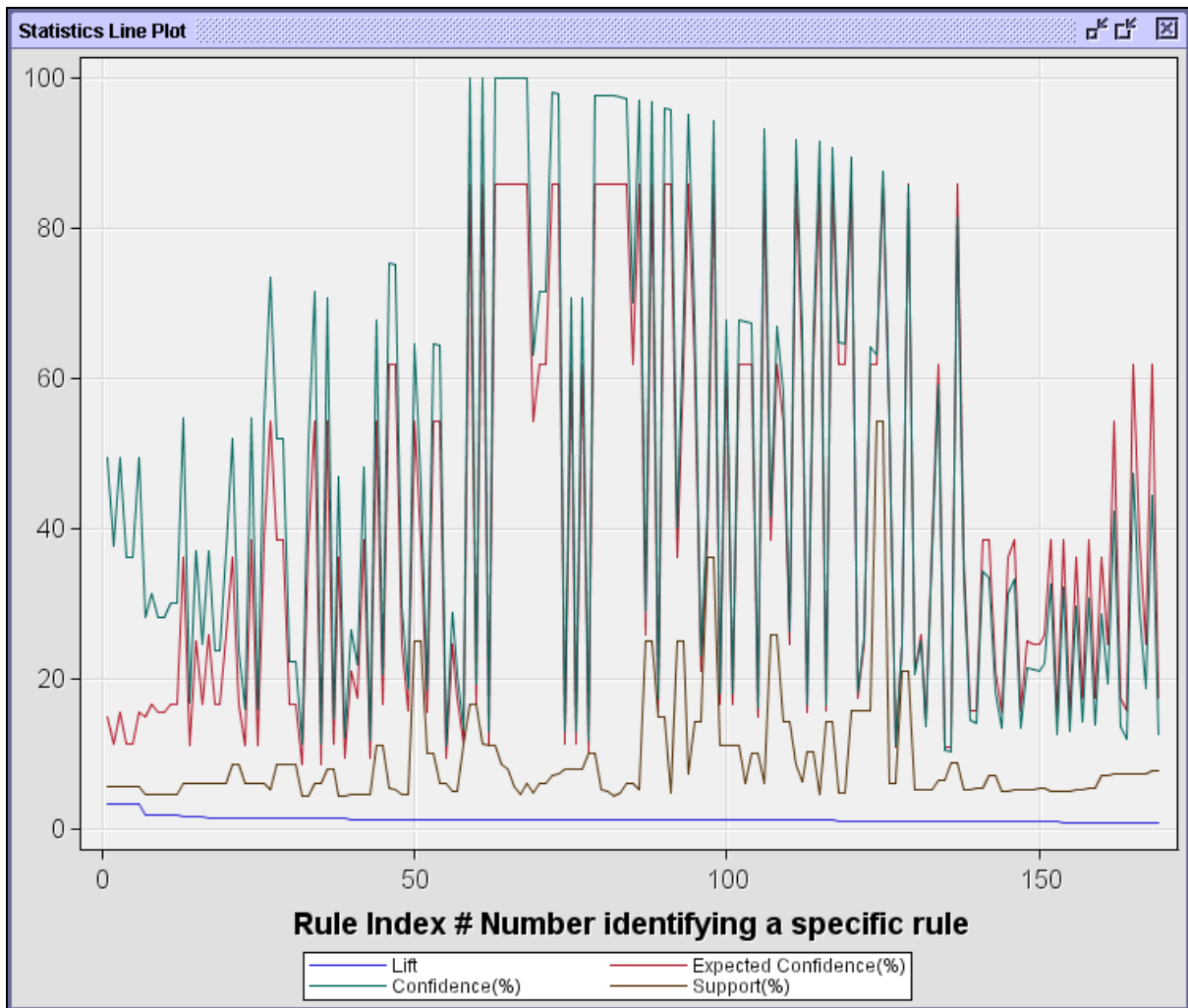
17. Select **OK** to close the Variables dialog box.

18. Run the diagram from the Association node and view the results.

The Results - Node: Association Diagram window opens with the Statistics Plot, Statistics Line Plot, Rule Matrix, and Output windows visible.



19. Maximize the Statistics Line Plot window.



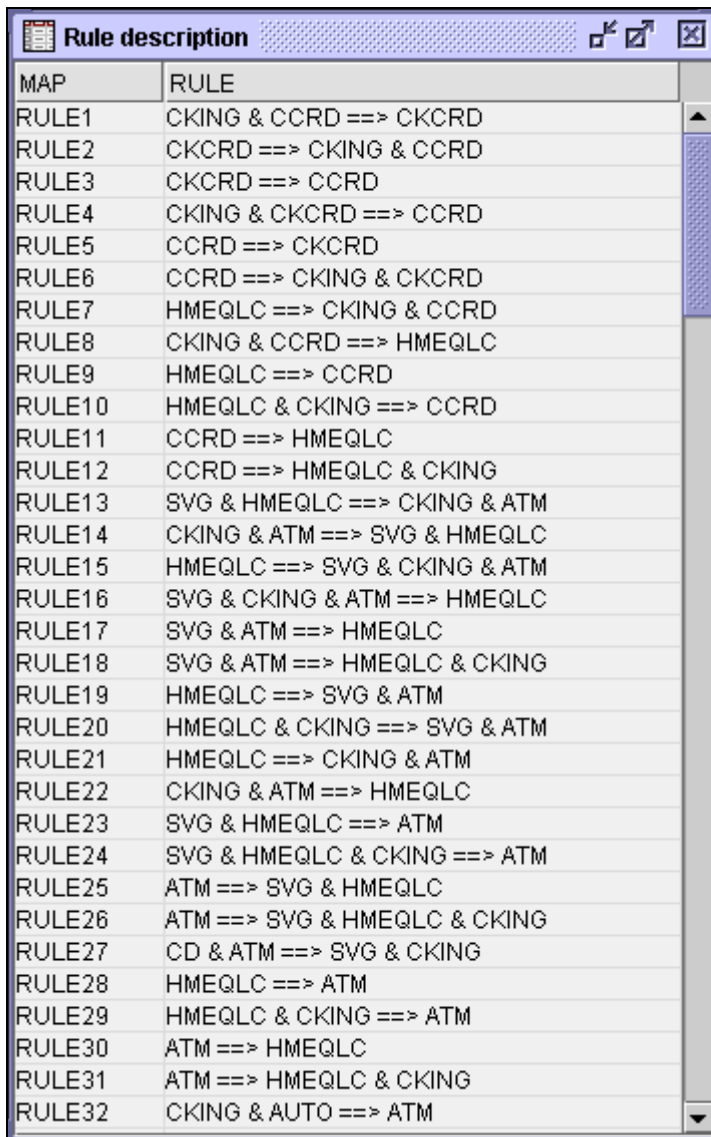
The statistics line plot graphs the lift, expected confidence, confidence, and support for each of the rules by rule index number.

Consider the rule $A \Rightarrow B$. Recall the following:

- **Support** of $A \Rightarrow B$ is the probability that a customer has both A and B.
- **Confidence** of $A \Rightarrow B$ is the probability that a customer has B given that the customer has A.
- **Expected Confidence** of $A \Rightarrow B$ is the probability that a customer has B.
- **Lift** of $A \Rightarrow B$ is a measure of the strength of the association. If $\text{Lift}=2$ for the rule $A \Rightarrow B$, then a customer having A is twice as likely to have B than a customer chosen at random. *Lift* is the confidence divided by the expected confidence.

Notice that the rules are ordered in descending order of lift.

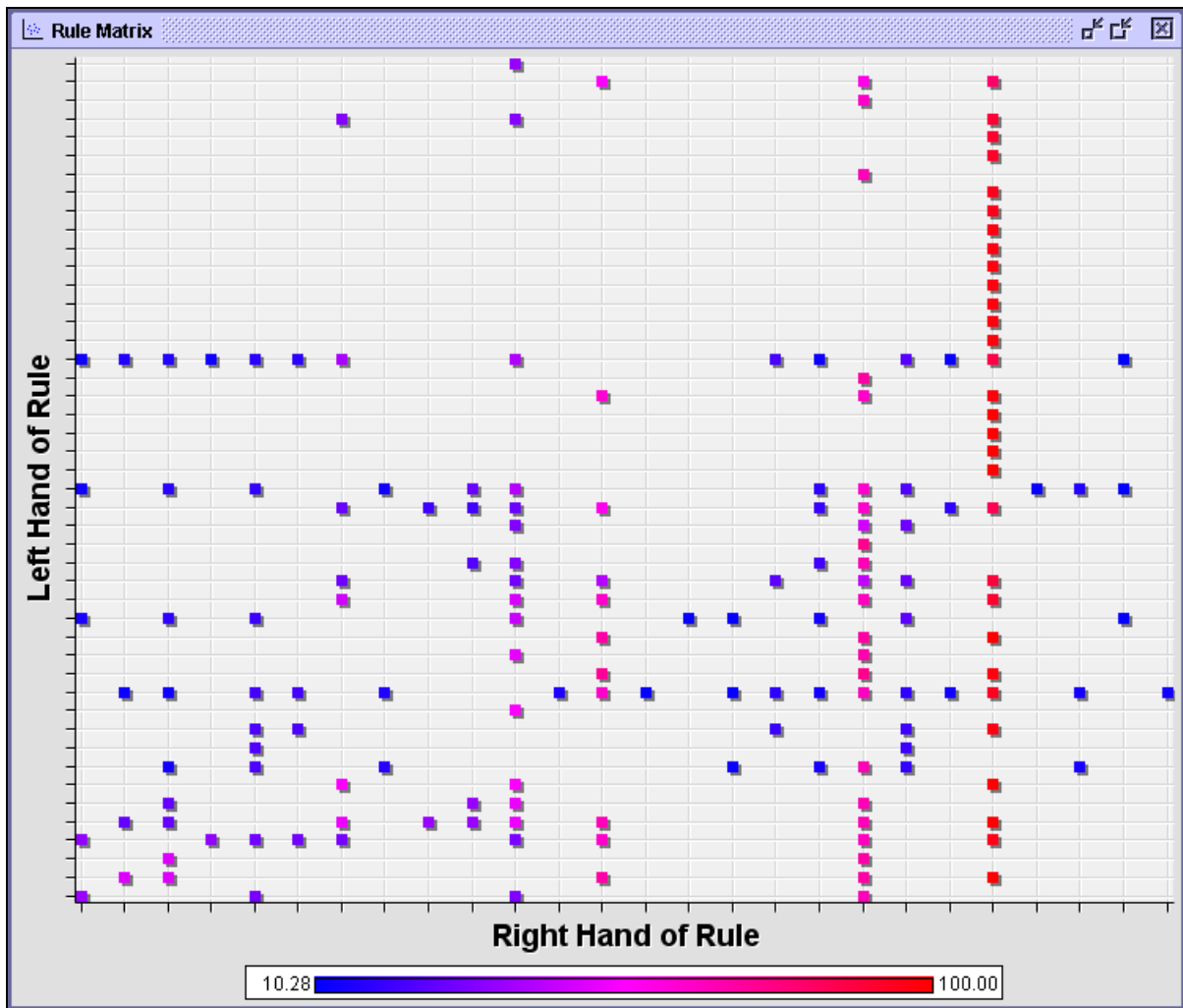
20. To view the descriptions of the rules, select **View** ⇒ **Rules** ⇒ **Rule description**.



MAP	RULE
RULE1	CKING & CCRD ==> CKCRD
RULE2	CKCRD ==> CKING & CCRD
RULE3	CKCRD ==> CCRD
RULE4	CKING & CKCRD ==> CCRD
RULE5	CCRD ==> CKCRD
RULE6	CCRD ==> CKING & CKCRD
RULE7	HMEQLC ==> CKING & CCRD
RULE8	CKING & CCRD ==> HMEQLC
RULE9	HMEQLC ==> CCRD
RULE10	HMEQLC & CKING ==> CCRD
RULE11	CCRD ==> HMEQLC
RULE12	CCRD ==> HMEQLC & CKING
RULE13	SVG & HMEQLC ==> CKING & ATM
RULE14	CKING & ATM ==> SVG & HMEQLC
RULE15	HMEQLC ==> SVG & CKING & ATM
RULE16	SVG & CKING & ATM ==> HMEQLC
RULE17	SVG & ATM ==> HMEQLC
RULE18	SVG & ATM ==> HMEQLC & CKING
RULE19	HMEQLC ==> SVG & ATM
RULE20	HMEQLC & CKING ==> SVG & ATM
RULE21	HMEQLC ==> CKING & ATM
RULE22	CKING & ATM ==> HMEQLC
RULE23	SVG & HMEQLC ==> ATM
RULE24	SVG & HMEQLC & CKING ==> ATM
RULE25	ATM ==> SVG & HMEQLC
RULE26	ATM ==> SVG & HMEQLC & CKING
RULE27	CD & ATM ==> SVG & CKING
RULE28	HMEQLC ==> ATM
RULE29	HMEQLC & CKING ==> ATM
RULE30	ATM ==> HMEQLC
RULE31	ATM ==> HMEQLC & CKING
RULE32	CKING & AUTO ==> ATM

The highest lift rule is checking, and credit card implies check card. This is not surprising given that many check cards include credit card logos. Notice the symmetry in rules 1 and 2. This is not accidental because, as noted earlier, lift is symmetric.

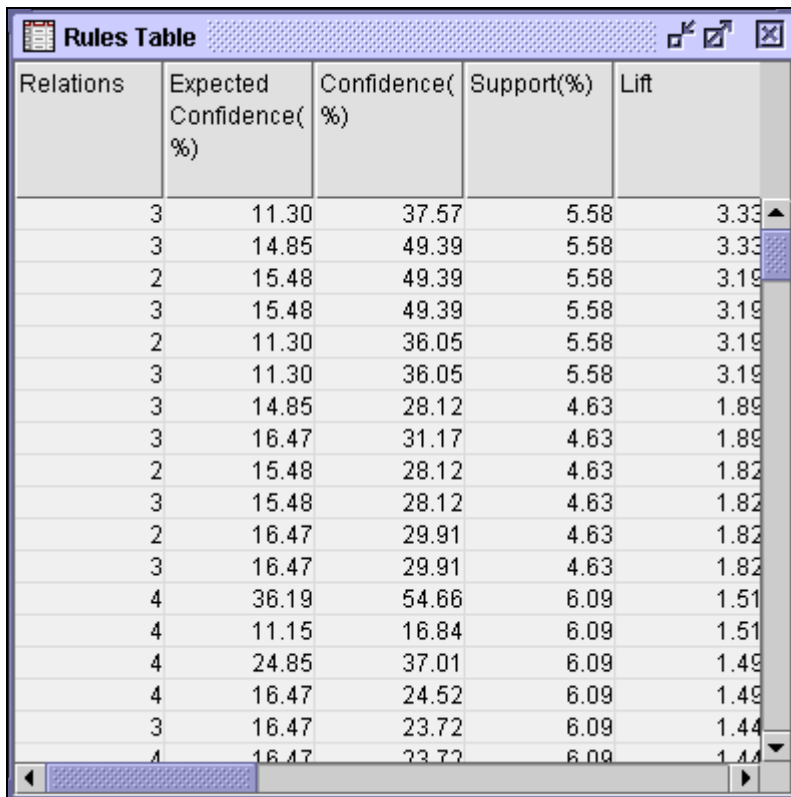
21. Examine the rule matrix.



The rule matrix plots the rules based on the items on the left side of the rule and the items on the right side of the rule. The points are colored, based on the confidence of the rules. For example, the rules with the highest confidence are in the column in the picture above. Using the interactive feature of the graph, you discover that these rules all have checking on the right side of the rule.

Another way to explore the rules found in the analysis is by plotting the Rules table.

22. Select **View** ⇒ **Rules** ⇒ **Rules Table**. The Rules Table window opens.

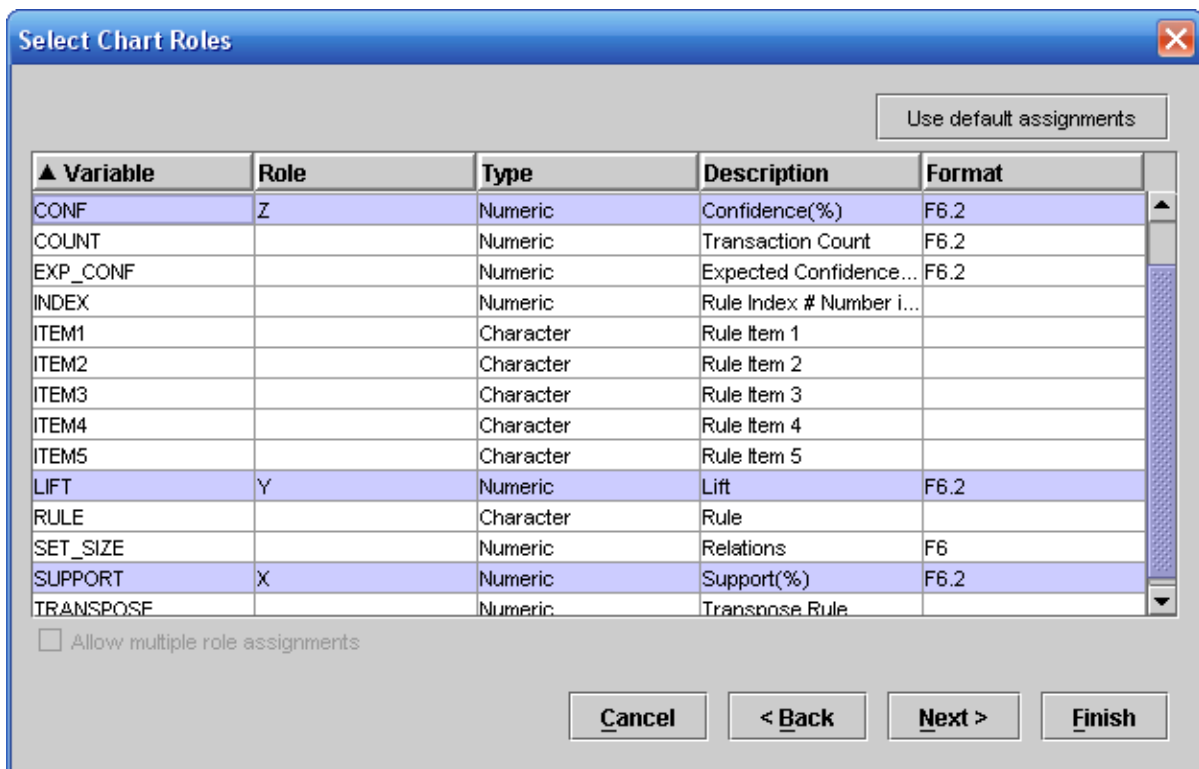


Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift
3	11.30	37.57	5.58	3.33
3	14.85	49.39	5.58	3.33
2	15.48	49.39	5.58	3.19
3	15.48	49.39	5.58	3.19
2	11.30	36.05	5.58	3.19
3	11.30	36.05	5.58	3.19
3	14.85	28.12	4.63	1.89
3	16.47	31.17	4.63	1.89
2	15.48	28.12	4.63	1.82
3	15.48	28.12	4.63	1.82
2	16.47	29.91	4.63	1.82
3	16.47	29.91	4.63	1.82
4	36.19	54.66	6.09	1.51
4	11.15	16.84	6.09	1.51
4	24.85	37.01	6.09	1.49
4	16.47	24.52	6.09	1.49
3	16.47	23.72	6.09	1.44
4	16.47	23.72	6.09	1.44

23. Select the Plot Wizard icon, .

24. Choose a three-dimensional scatter plot for the type of chart, and select **Next >**.

25. Select the roles **X**, **Y**, and **Z** for the variables **SUPPORT**, **LIFT**, and **CONF**, respectively.



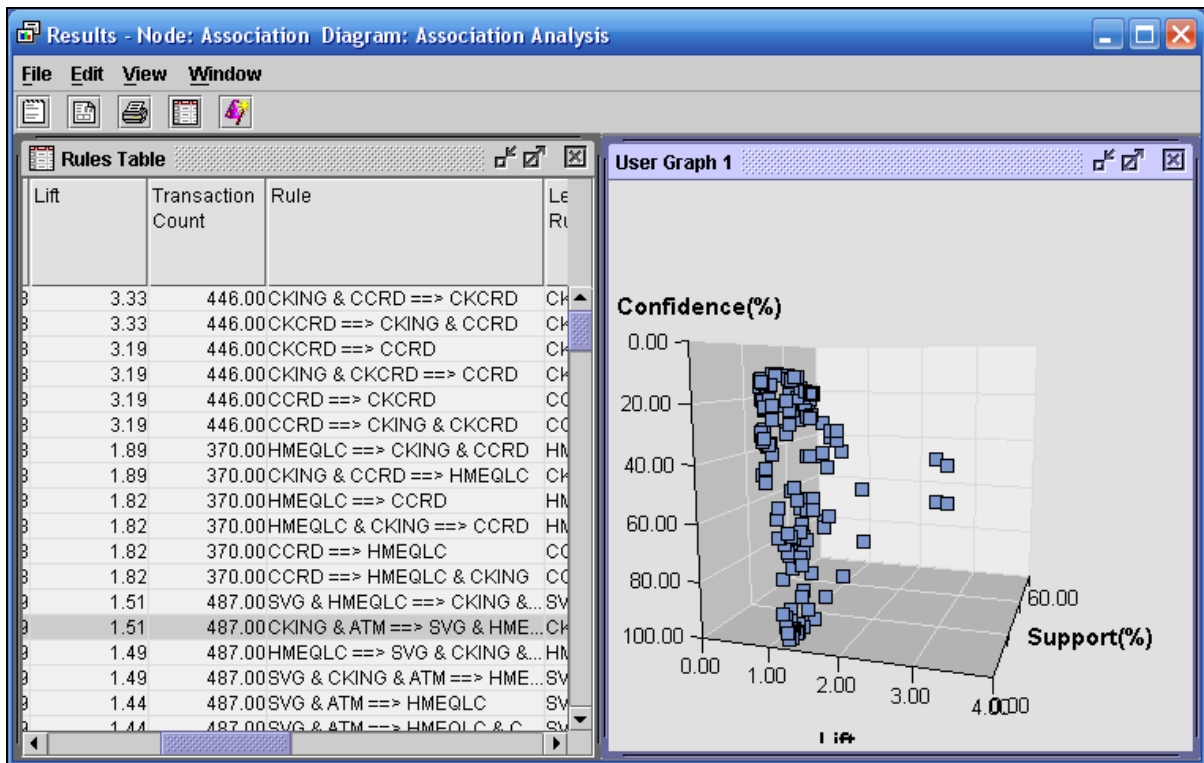
The dialog box titled "Select Chart Roles" contains a table with the following data:

Variable	Role	Type	Description	Format
CONF	Z	Numeric	Confidence(%)	F6.2
COUNT		Numeric	Transaction Count	F6.2
EXP_CONF		Numeric	Expected Confidence...	F6.2
INDEX		Numeric	Rule Index # Number i...	
ITEM1		Character	Rule Item 1	
ITEM2		Character	Rule Item 2	
ITEM3		Character	Rule Item 3	
ITEM4		Character	Rule Item 4	
ITEM5		Character	Rule Item 5	
LIFT	Y	Numeric	Lift	F6.2
RULE		Character	Rule	
SET_SIZE		Numeric	Relations	F6
SUPPORT	X	Numeric	Support(%)	F6.2
TRANSPOSE		Numeric	Transpose Rule	

Below the table, there is a checkbox labeled "Allow multiple role assignments" which is currently unchecked. At the bottom right, there are four buttons: "Cancel", "< Back", "Next >", and "Finish". A "Use default assignments" button is located at the top right of the dialog.

26. Select **Finish** to generate the plot.

27. Rearrange the windows to view the data and the plot simultaneously.



Expanding the Rule column in the data table and selecting points in the three-dimensional plot enable you to quickly uncover high lift rules from the market basket analysis while judging their confidence and support. You can use WHERE clauses in the Data Options dialog box to subset cases in which you are interested.

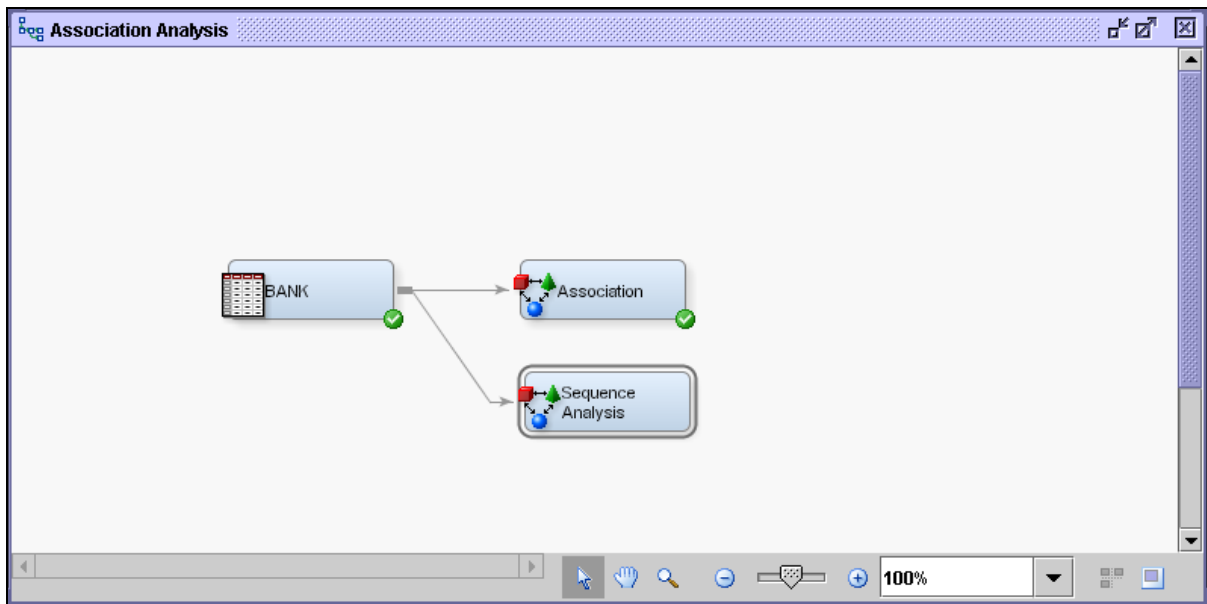
28. Close the Results window.



Sequence Analysis

In addition to the products owned by its customers, the bank is interested in examining the order in which the products are purchased. The sequence variable in the data set enables you to conduct a sequence analysis.

1. Add an Association node to the diagram workspace and connect it to the **BANK** node.
2. Rename the new node **Sequence Analysis**.



3. Set Export Rule by ID to **Yes**.

Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes

4. Examine the Sequence panel in the Properties panel.

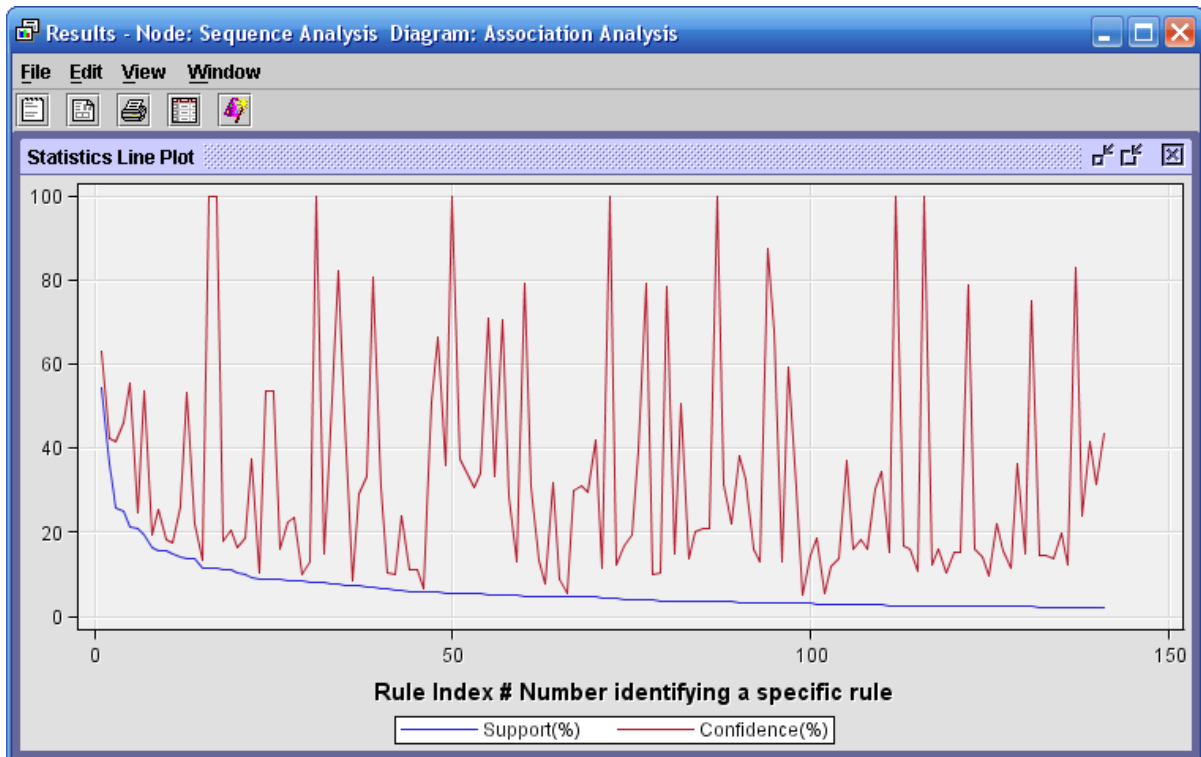
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction	0.0
Support Type	Percent
Support Count	1
Support Percentage	2.0

The options in the Sequence panel enable you to specify the following properties:

- **Chain Count** is the maximum number of items that can be included in a sequence. The default value is 3 and the maximum value is 10.
- **Consolidate Time** enables you to specify whether consecutive visits to a location or consecutive purchases over a given interval can be consolidated into a single visit for analysis purposes. For example, two products purchased less than a day apart might be considered to be a single transaction.
- **Maximum Transaction Duration** enables you to specify the maximum length of time for a series of transactions to be considered a sequence. For example, you might want to specify that the purchase of two products more than three months apart does not constitute a sequence.
- **Support Type** specifies whether the sequence analysis should use the Support Count or Support Percentage property. The default setting is Percent.
- **Support Count** specifies the minimum frequency required to include a sequence in the sequence analysis when the Sequence Support Type is set to Count. If a sequence has a count less than the specified value, that sequence is excluded from the output. The default setting is 2.
- **Support Percentage** specifies the minimum level of support to include the sequence in the analysis when the Support Type is set to Percent. If a sequence has a frequency that is less than the specified percentage of the total number of transactions, then that sequence is excluded from the output. The default percentage is 2%. Permissible values are real numbers between 0 and 100.

5. Run the diagram from the Sequence Analysis node and view the results.

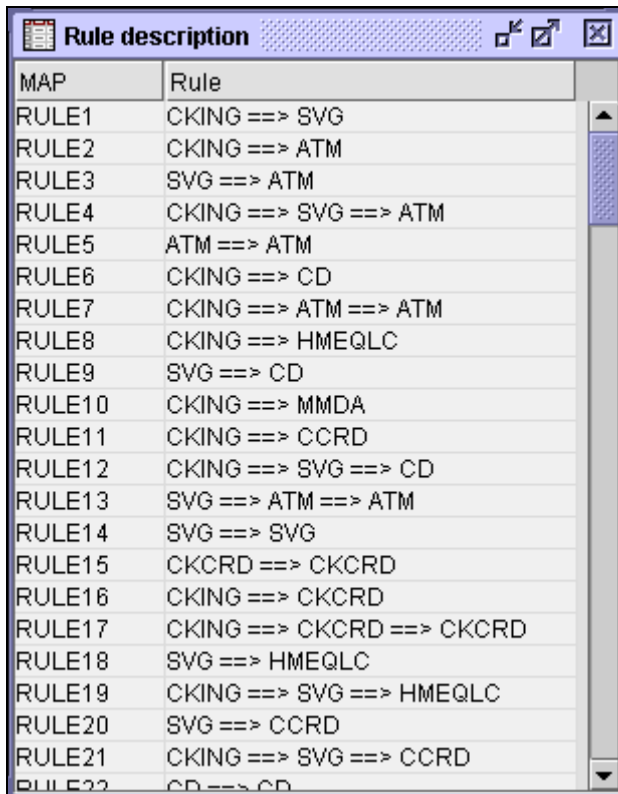
6. Maximize the Statistics Line Plot window.



The statistics line plot graphs the confidence and support for each of the rules by rule index number.

The *percent support* is the transaction count divided by the total number of customers, which would be the maximum transaction count. The *percent confidence* is the transaction count divided by the transaction count for the left side of the sequence.

7. Select **View** ⇒ **Rules** ⇒ **Rule description** to view the descriptions of the rules.



MAP	Rule
RULE1	CKING ==> SVG
RULE2	CKING ==> ATM
RULE3	SVG ==> ATM
RULE4	CKING ==> SVG ==> ATM
RULE5	ATM ==> ATM
RULE6	CKING ==> CD
RULE7	CKING ==> ATM ==> ATM
RULE8	CKING ==> HMEQLC
RULE9	SVG ==> CD
RULE10	CKING ==> MMDA
RULE11	CKING ==> CCRD
RULE12	CKING ==> SVG ==> CD
RULE13	SVG ==> ATM ==> ATM
RULE14	SVG ==> SVG
RULE15	CKCRD ==> CKCRD
RULE16	CKING ==> CKCRD
RULE17	CKING ==> CKCRD ==> CKCRD
RULE18	SVG ==> HMEQLC
RULE19	CKING ==> SVG ==> HMEQLC
RULE20	SVG ==> CCRD
RULE21	CKING ==> SVG ==> CCRD
RULE22	CD ==> CD

The confidence for many of the rules changes after the order of service acquisition is considered.



Exercises

2. Conducting an Association Analysis

A store is interested in determining the associations between items purchased from the Health and Beauty Aids department and the Stationery Department. The store chose to conduct a market basket analysis of specific items purchased from these two departments. The **ASSOCIATIONS** data set contains information about over 400,000 transactions made over the past three months. The following products are represented in the data set:

1. bar soap
2. bows
3. candy bars
4. deodorant
5. greeting cards
6. magazines
7. markers
8. pain relievers
9. pencils
10. pens
11. perfume
12. photo processing
13. prescription medications
14. shampoo
15. toothbrushes
16. toothpaste
17. wrapping paper

There are four variables in the data set:

Name	Model Role	Measurement Level	Description
STORE	Rejected	Nominal	Identification number of the store
TRANSACTION	ID	Nominal	Transaction identification number
PRODUCT	Target	Nominal	Product purchased
QUANTITY	Rejected	Interval	Quantity of this product purchased

- Create a new diagram. Name the diagram **Transactions**.
- Create a new data source using the data set **AAEM61 . TRANSACTIONS**.
- Assign the variables **STORE** and **QUANTITY** the model role Rejected. These variables will not be used in this analysis. Assign the ID model role to the variable **TRANSACTION** and the Target model role to the variable **PRODUCT**.
- Add the node for the **TRANSACTIONS** data set and an Association node to the diagram.
- Change the setting for Export Rule by ID to Yes.
- Leave the remaining default settings for the Association node and run the analysis.
- Examine the results of the association analysis.

What is the highest lift value for the resulting rules? _____

Which rule has this value? _____

8.4 Chapter Summary

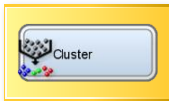
Pattern discovery seems to embody the promise of data mining, but there are many ways for an analysis to fail. SAS Enterprise Miner provides tools to help with data reduction, novelty detection, profiling, market basket analysis, and sequence analysis.

Cluster and segmentation analyses are similar in intent but differ in execution. In cluster analysis, the goal is to identify distinct groupings of cases across a set of inputs. In segmentation analysis, the goal is to partition cases from a single cluster into contiguous groups.

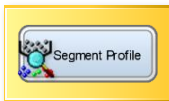
SAS Enterprise Miner offers several tools for exploring the results of a cluster and segmentation analysis. For low dimension data, you can use capabilities provided by the Graph Wizard and the Explore window. For higher dimensional data, you can choose the Segment Profile tool to understand the generated partitions.

Market basket and sequence analyses are handled by the Association tool. This tool transforms transaction data sets into rules. The value of the generated rules is gauged by confidence, support, and lift. The Association tool features a variety of plots and tables to help you explore the analysis results.

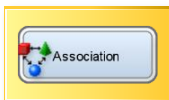
Pattern Discovery Tools: Review



Generate cluster models using automatic settings and segmentation models with user-defined settings.



Compare within-segment distributions of selected inputs to overall distributions. This helps you understand segment definition.



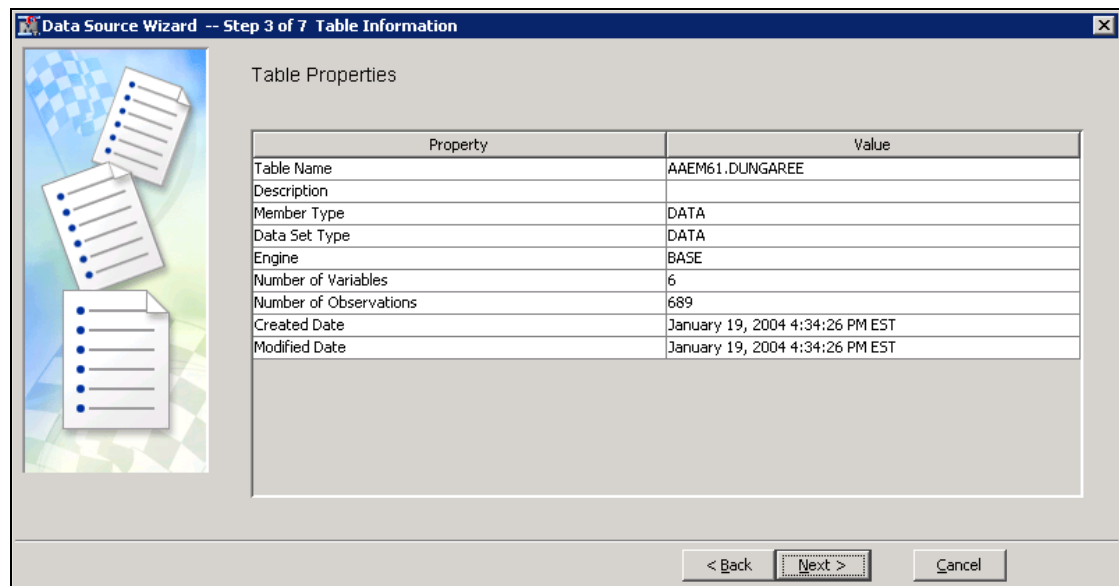
Conduct market basket and sequence analysis on transactions data. A data source must have one target, one ID, and (if desired) one sequence variable in the data source.

8.5 Solutions

Solutions to Exercises

1. Conducting Cluster Analysis

- a. Create a new diagram in your project.
 - 1) To open a diagram, select **File** ⇒ **New** ⇒ **Diagram**.
 - 2) Type the name of the new diagram, **Jeans**, and select **OK**.
- b. Define the data set **DUNGAREE** as a data source.
 - 1) Select **File** ⇒ **New** ⇒ **Data Source...**.
 - 2) In the Data Source Wizard - Metadata Source window, make sure that **SAS Table** is selected as the source and select **Next >**.
 - 3) To choose the desired data table, select **Browse...**.
 - 4) Double-click on the **AAEM61** library to see the data tables in the library.
 - 5) Select the **DUNGAREE** data set, and then select **OK**.
 - 6) Select **Next >**.

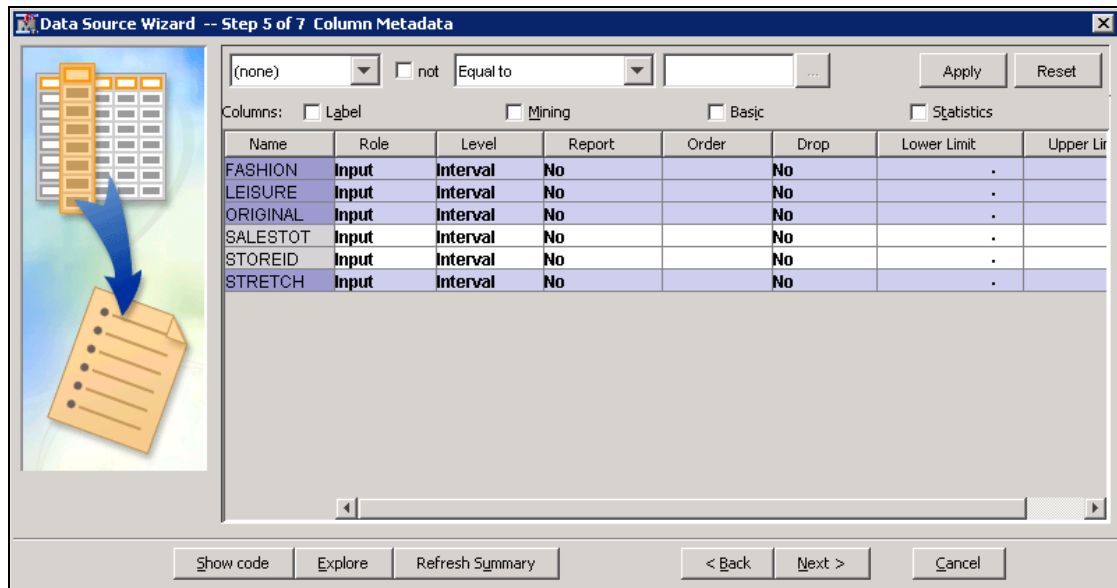


- 7) Select **Next >**.
- 8) Select **Advanced** to use the Advanced Advisor, and then select **Next >**.

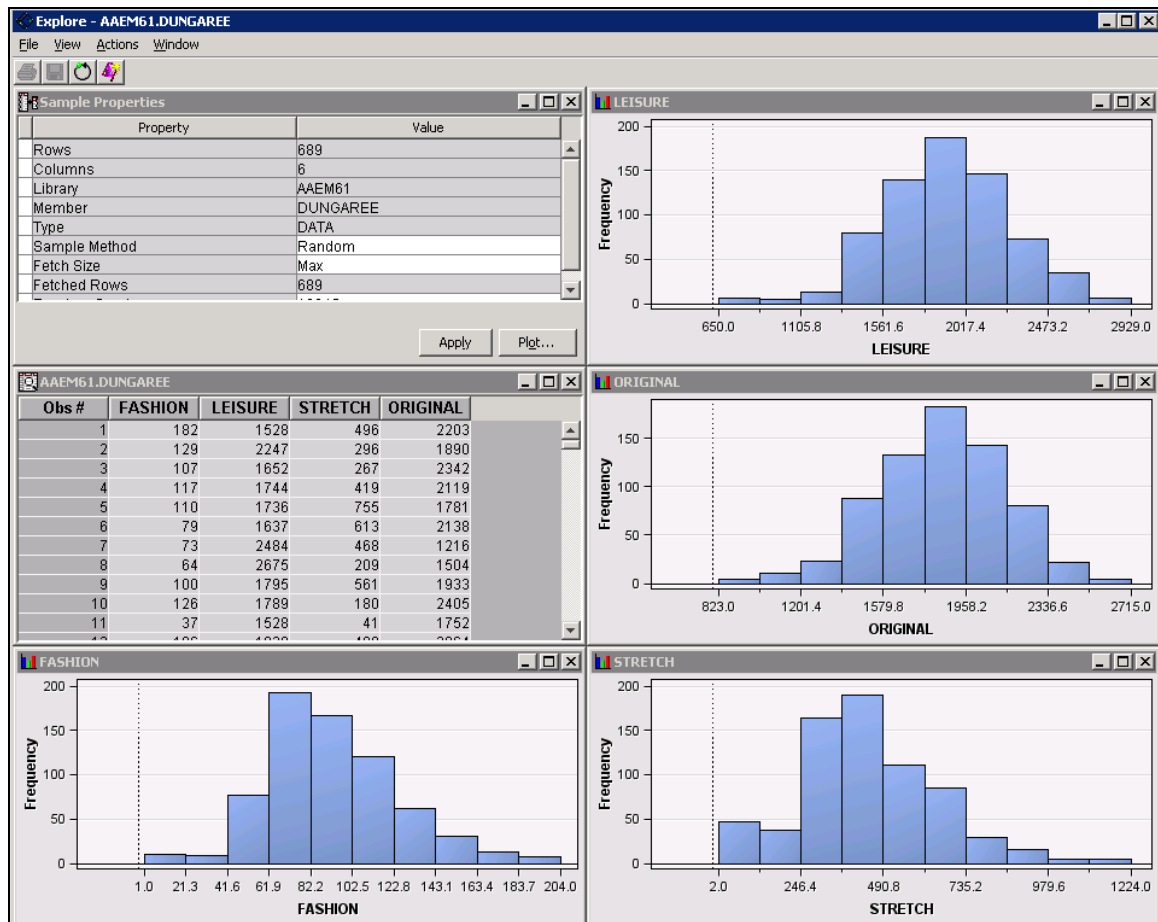
- c. Determine whether the model roles and measurement levels assigned to the variables are appropriate.

Examine the distribution of the variables.

- 1) Hold down the CTRL key and click to select the variables of interest.



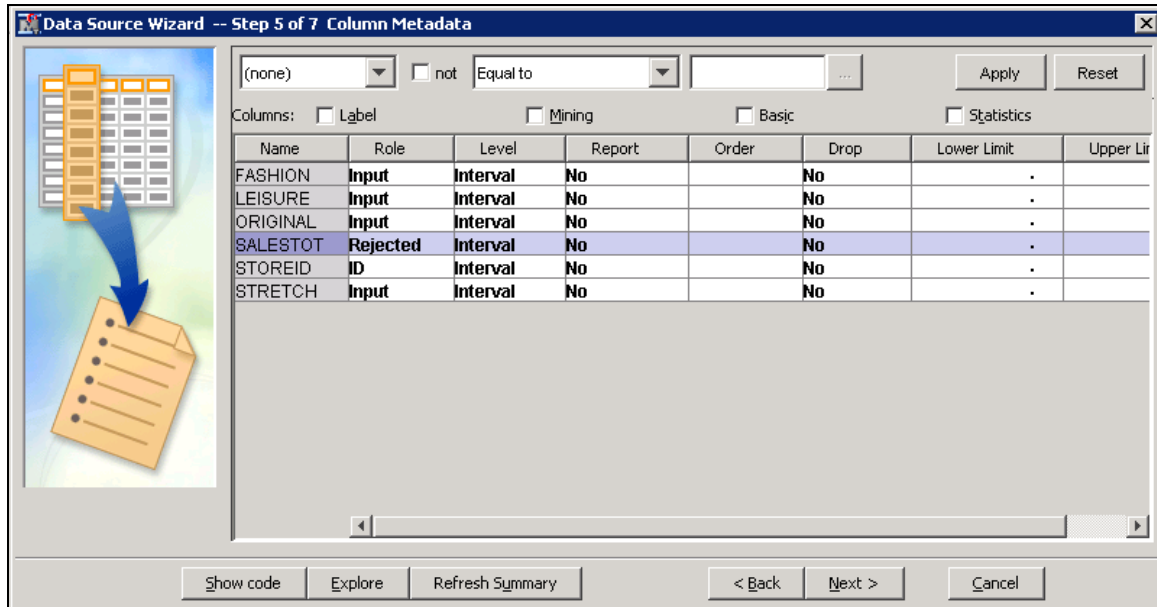
- 2) Select **Explore**.
- 3) Select **Next >** and then **Finish** to complete the data source creation.



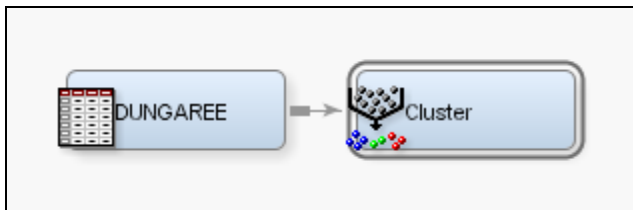
There do not appear to be any unusual or missing data values.

- d. The variable **STOREID** should have the ID model role and the variable **SALESTOT** should have the Rejected model role.

The variable **SALESTOT** should be rejected because it is the sum of the other input variables in the data set. Therefore, it should not be considered as an independent input value.



- e. To add an Input Data node to the diagram workspace and select the **DUNGAREE** data table as the data source, drag the **DUNGAREE** data source onto the diagram workspace.
- f. Add a Cluster node to the diagram workspace. The workspace should appear as shown.

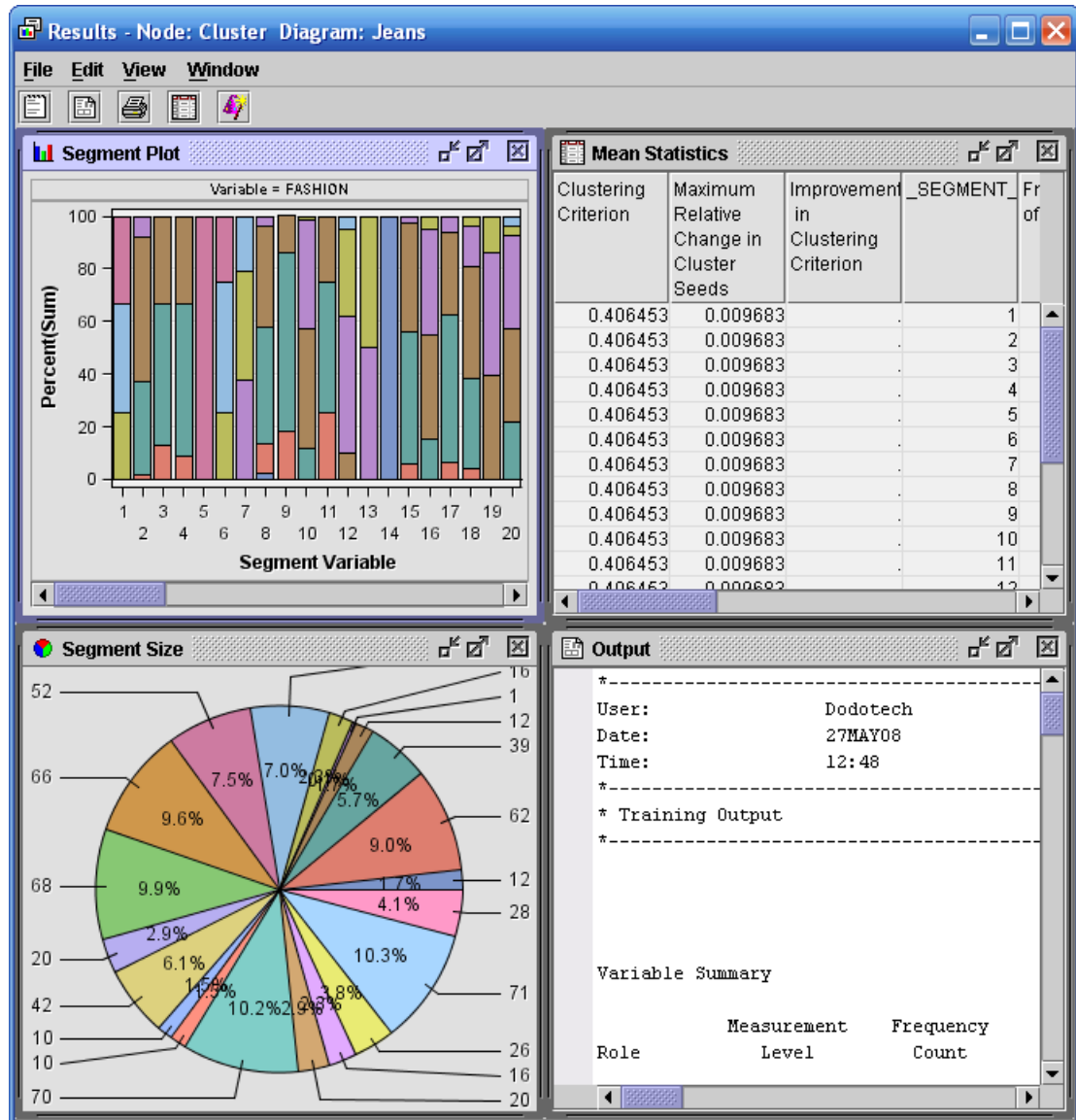


- g. Select the **Cluster** node.

In the property sheet, select **Internal Standardization** ⇒ **Standardization**.

If you do not standardize, the clustering will occur strictly on the inputs with the largest range (Original and Leisure).

- h. Run the diagram from the Cluster node and examine the results.
- 1) Run the Cluster node and view the results.
 - 2) To view the results, right-click the **Cluster** node and select **Results...**



The Cluster node's Automatic number of cluster specification method seems to generate an excessive number of clusters.

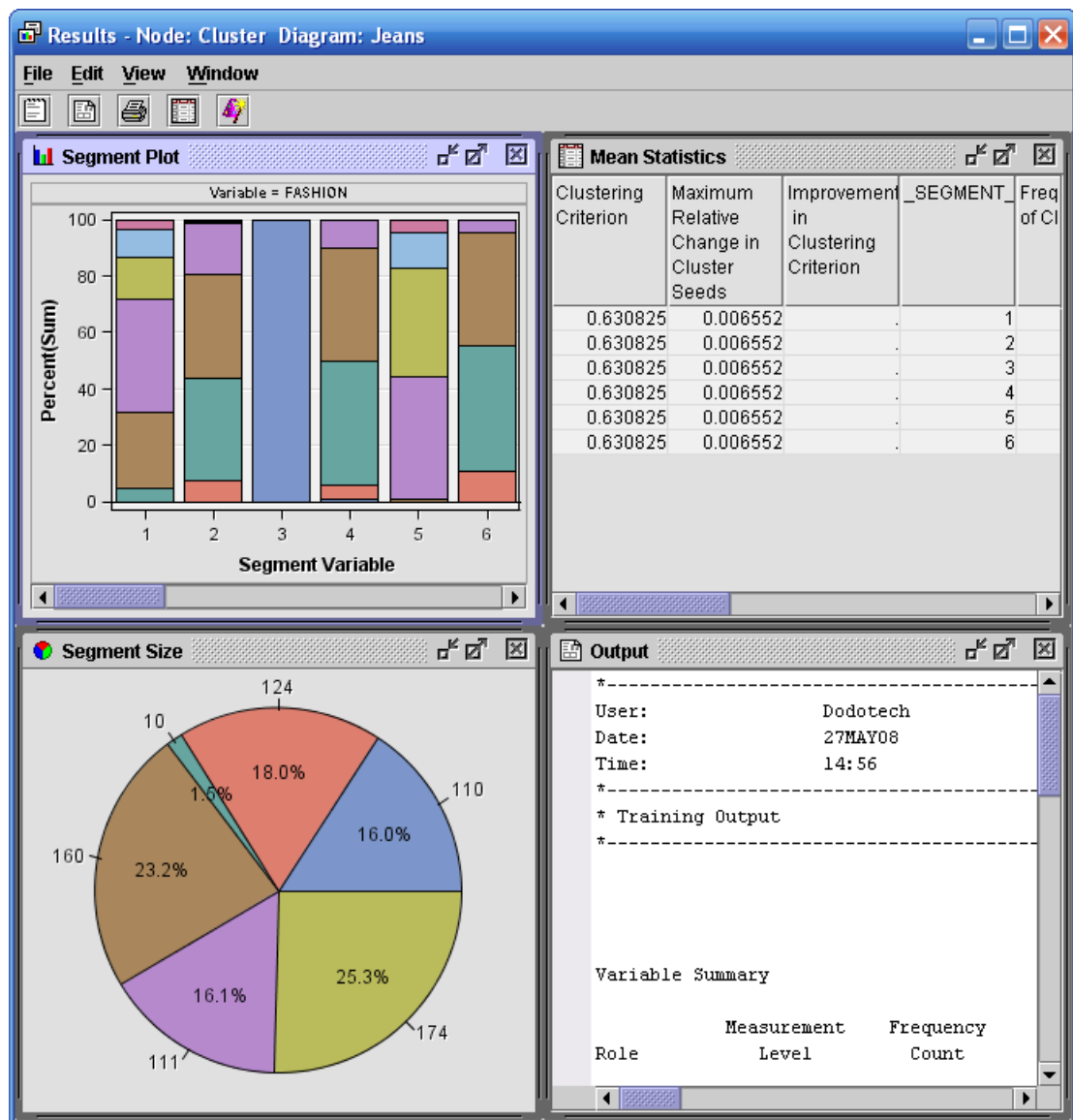
i. Specify a maximum of six clusters.

1) Select **Specification Method** ⇒ **User Specify**.

2) Select **Maximum Number of Clusters** ⇒ **6**.

Train	
Variables	...
Cluster Variable Role	Segment
Internal Standardization	Standardization
<input checked="" type="checkbox"/> Number of Clusters	
<input checked="" type="checkbox"/> Specification Method	User Specify
<input checked="" type="checkbox"/> Maximum Number of Clusters	6

3) Run the Cluster node and view the results.

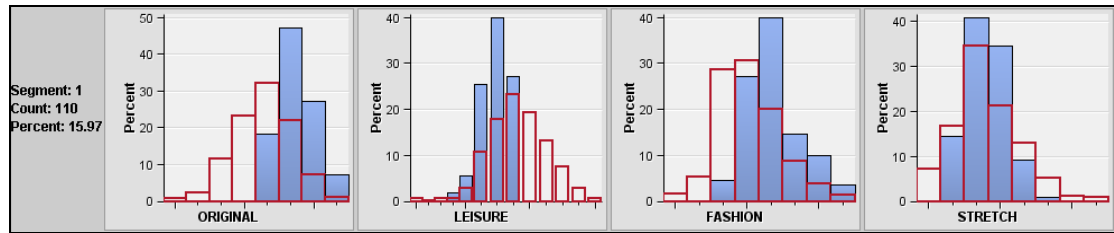


Apparently, all but one of the segments is well populated. There are more details about the segment composition in the next step.

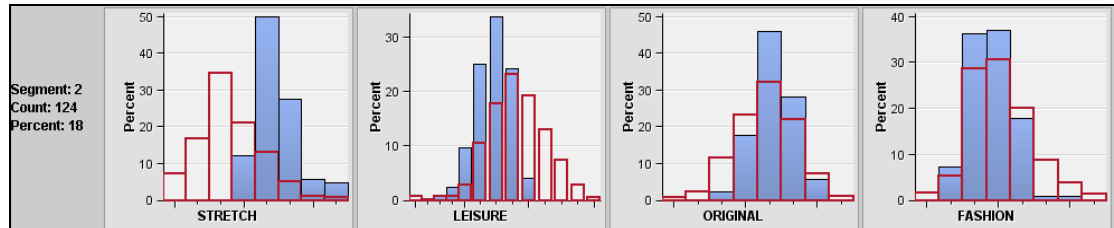
j. Connect a Segment Profile node to the Cluster node.



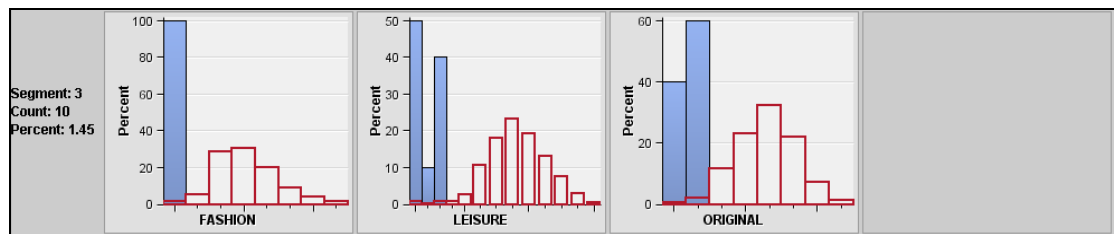
1) Run the Segment Profile node and view the results.



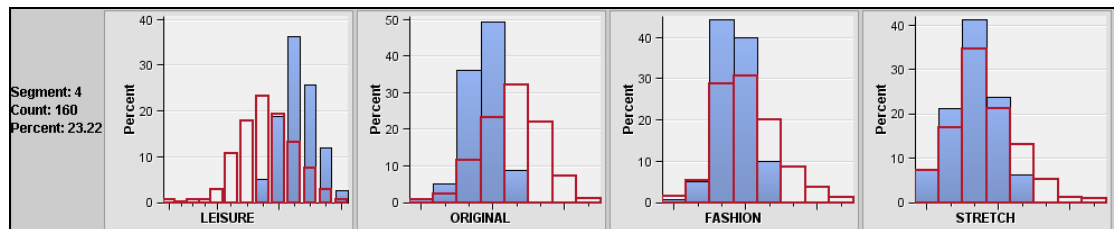
Segment 1 contains stores selling a higher-than-average number of original jeans.



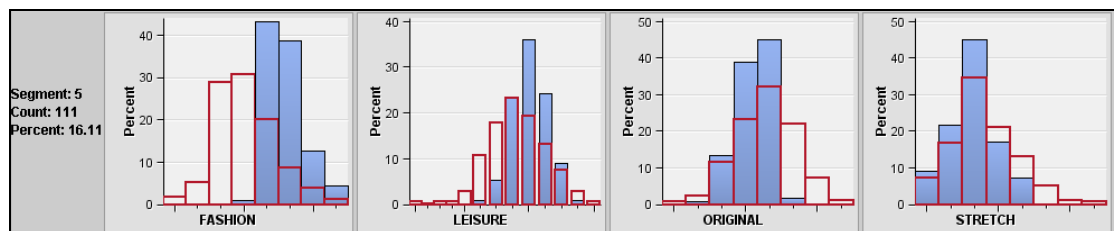
Segment 2 contains stores selling a higher-than-average number of stretch jeans.



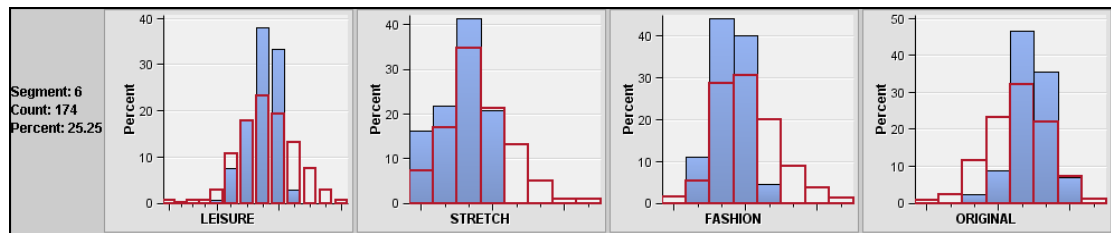
Segment 3 contains stores selling small numbers of all jean styles.



Segment 4 contains stores selling a higher-than-average number of leisure jeans.



Segment 5 contains stores selling a higher-than-average number of fashion jeans.



Segment 6 contains stores selling a higher-than-average number of original jeans, but lower-than-average number of stretch and fashion.

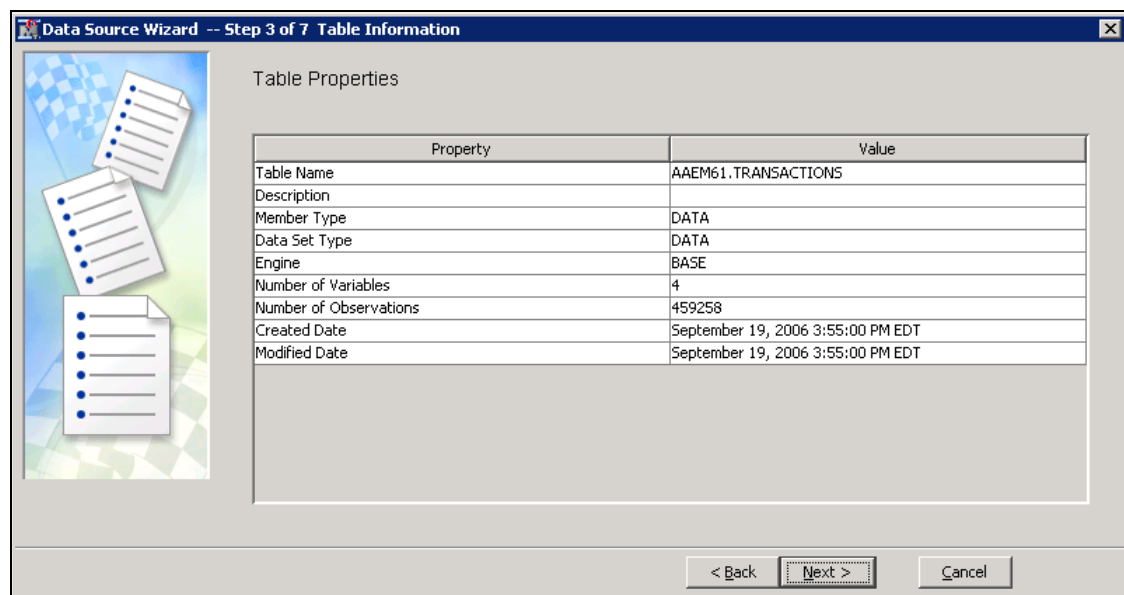
2. Conducting an Association Analysis

a. Create the Transactions diagram.

- 1) To open a new diagram in the project, select **File** ⇒ **New** ⇒ **Diagram...**.
- 2) Name the new diagram **Transactions** and select **OK**.

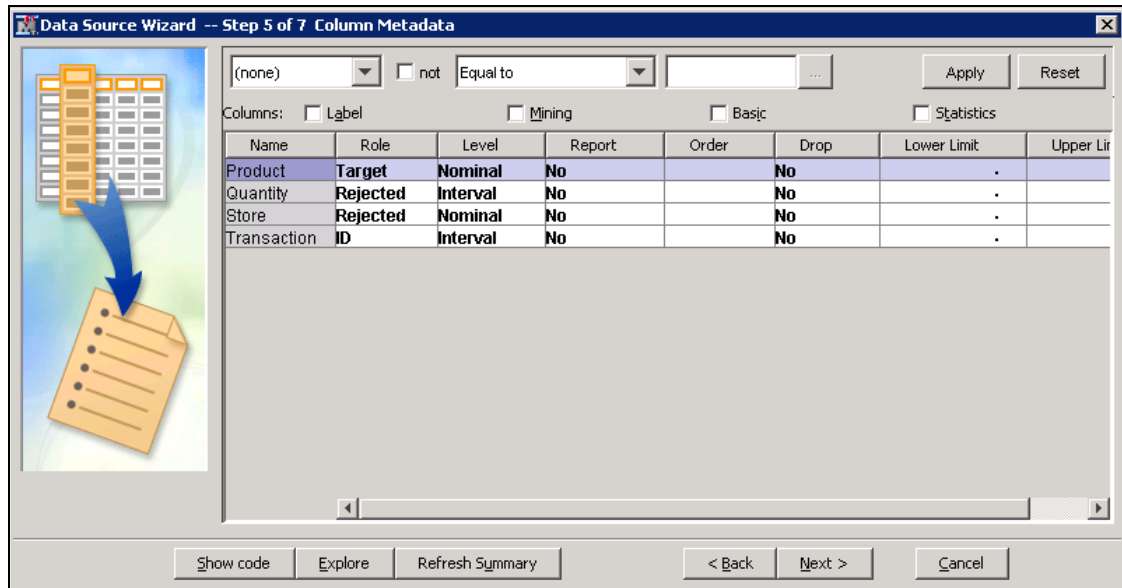
b. Create a new data source using the data set **AAEM61 . TRANSACTIONS**.

- 1) Right-click **Data Sources** in the project tree and select **Create Data Source**.
- 2) In the Data Source Wizard - Metadata Source window, make sure that **SAS Table** is selected as the source and select **Next >**.
- 3) Select **Browse...** to choose a data set.
- 4) Double-click on the **AAEM61** library and select the **TRANSACTIONS** data set.
- 5) Select **OK**.
- 6) Select **Next >**.



- 7) Examine the data table properties, and then select **Next >**.
- 8) Select **Advanced** to use the Advanced Advisor, and then select **Next >**.

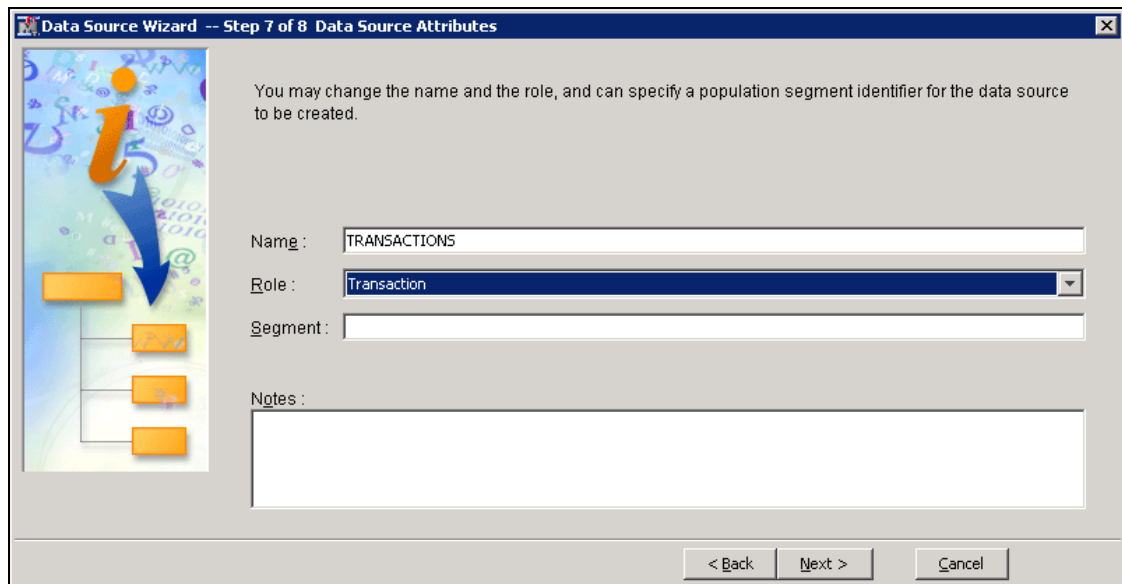
- c. Assign appropriate model roles to the variables.
- 1) Hold down the CTRL key and select the rows for the variables **STORE** and **QUANTITY**.
In the Role column of one of these rows, select **Rejected**.
 - 2) Select the **TRANSACTION** row and select **ID** as the role.
 - 3) Select the **PRODUCT** row and select **Target** as the role.



The screenshot shows the 'Data Source Wizard' window at Step 5 of 7, titled 'Column Metadata'. On the left is a graphic of a folder with a blue arrow pointing to a document. The main area contains a table with columns: Name, Role, Level, Report, Order, Drop, Lower Limit, and Upper Limit. The table has four rows: Product, Quantity, Store, and Transaction. The roles are assigned as follows: Product (Target), Quantity (Rejected), Store (Rejected), and Transaction (ID). The levels are: Product (Nominal), Quantity (Interval), Store (Nominal), and Transaction (Interval). The Report, Order, Drop, Lower Limit, and Upper Limit columns are all set to 'No' or empty. At the bottom are buttons for 'Show code', 'Explore', 'Refresh Summary', '< Back', 'Next >', and 'Cancel'.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Product	Target	Nominal	No		No	.	
Quantity	Rejected	Interval	No		No	.	
Store	Rejected	Nominal	No		No	.	
Transaction	ID	Interval	No		No	.	

- 4) Select **Next >**.
- 5) To skip decision processing, select **Next >**.
- 6) Change the role to **Transaction**.



The screenshot shows the 'Data Source Wizard' window at Step 7 of 8, titled 'Data Source Attributes'. On the left is a graphic of a folder with a blue arrow pointing to a document. The main area contains a text box with the instruction: 'You may change the name and the role, and can specify a population segment identifier for the data source to be created.' Below this are three fields: 'Name' (TRANSACTIONS), 'Role' (Transaction), and 'Segment' (empty). At the bottom is a 'Notes' text area. At the bottom right are buttons for '< Back', 'Next >', and 'Cancel'.

Name : TRANSACTIONS

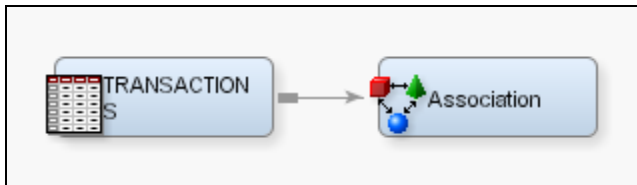
Role : Transaction

Segment :

Notes :

- 7) Select **Finish**.

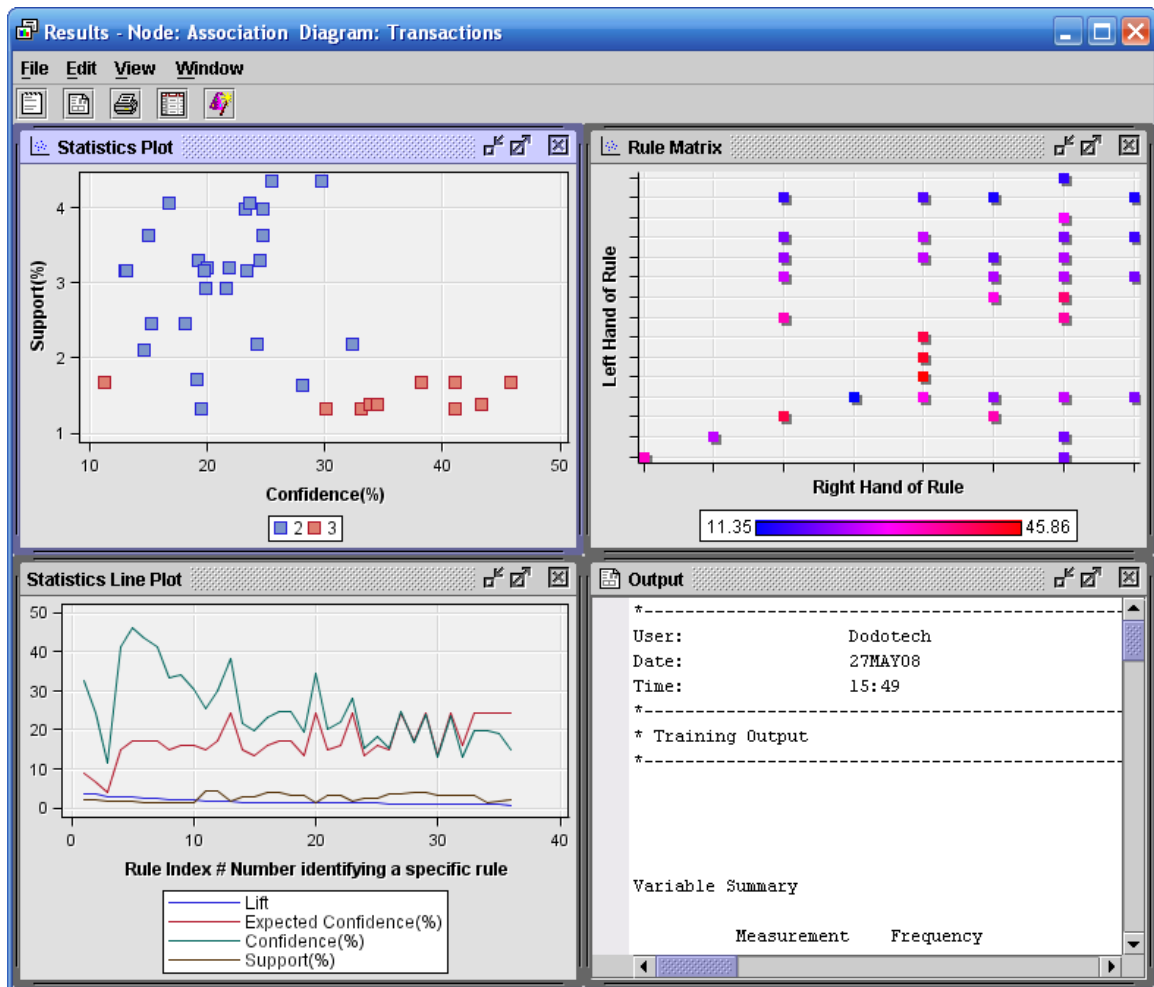
- d. Add the node for the **TRANSACTIONS** data set and an Association node to the diagram. The workspace should appear as shown.



- e. Change the setting for Export Rule by ID to Yes.

Property	Value
Train	
Variables	...
Maximum Number of Items	100000
Rules	...
<input checked="" type="checkbox"/> Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	1
Support Percentage	5.0
<input checked="" type="checkbox"/> Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	0.0
Support Type	Percent
Support Count	1
Support Percentage	2.0
<input checked="" type="checkbox"/> Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes

- f. Run the Association node and view the results.



- g. Examine the results of the association analysis.

Examine the Statistics Line plot.



Rule 1 has the highest lift value, 3.60.

Looking at the output reveals that Rule 1 is the rule Toothbrush \Rightarrow Perfume.