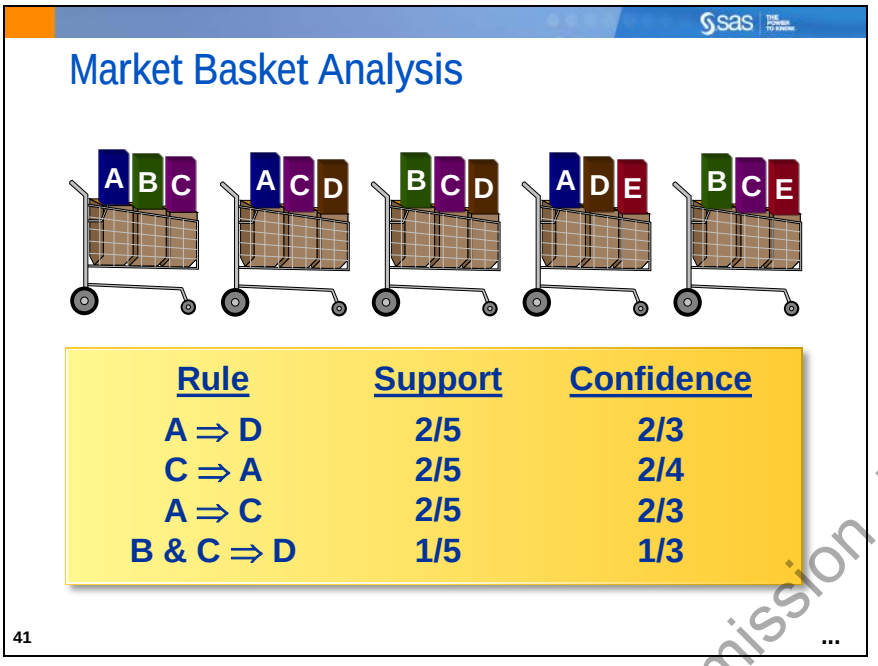
Note on Market basket Analysis and Pattern Discovery

Market Basket Analysis



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:

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

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:

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

Implication?

		Checking Account: No	Checking Account: Yes	
Savings Account	No	500	3500	4,000
	Yes	1000	5000	6,000
				10,000

Support(SVG ⇒ CK) = 50%
Confidence(SVG ⇒ CK) = 83%
Expected Confidence(SVG ⇒ CK) = 85%
Lift(SVG ⇒ CK) = 0.83/0.85 < 1

43

The interpretation of the implication (⇒) 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) ⇒ (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 ⇒ 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 $B \Rightarrow A$.

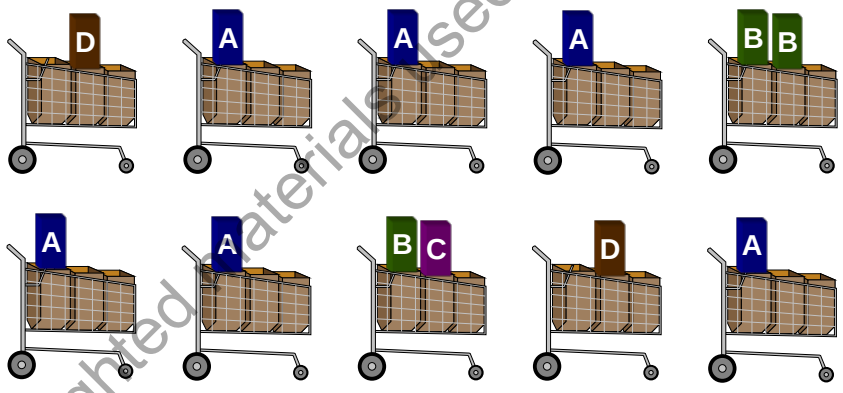
Barbie Doll ⇒ Candy

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.

44

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 ⇒ 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



45

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.**

46

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.

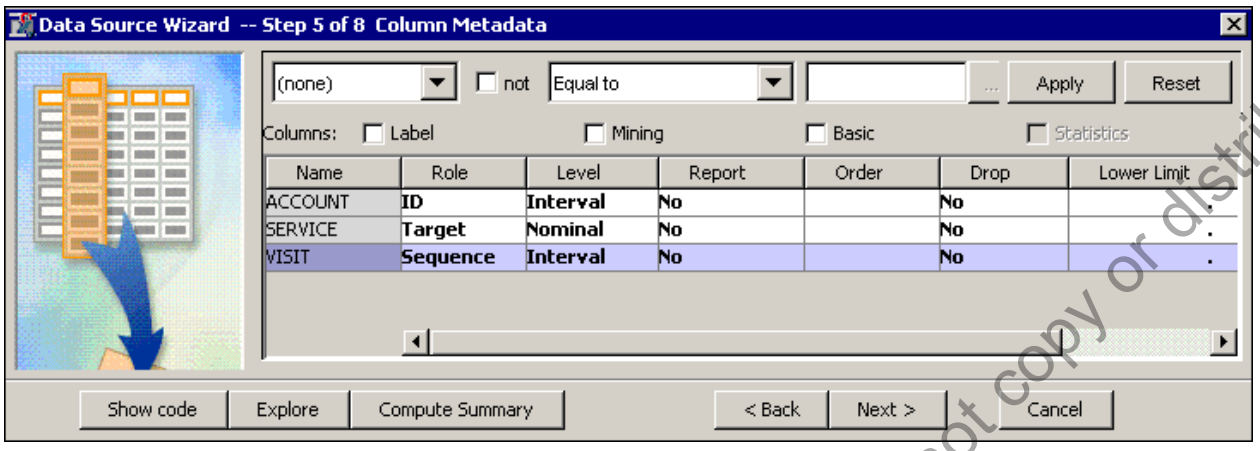
The 13 products are represented in the data set using the following abbreviations:

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. Select the **BANK** table from the **AAEM** library.

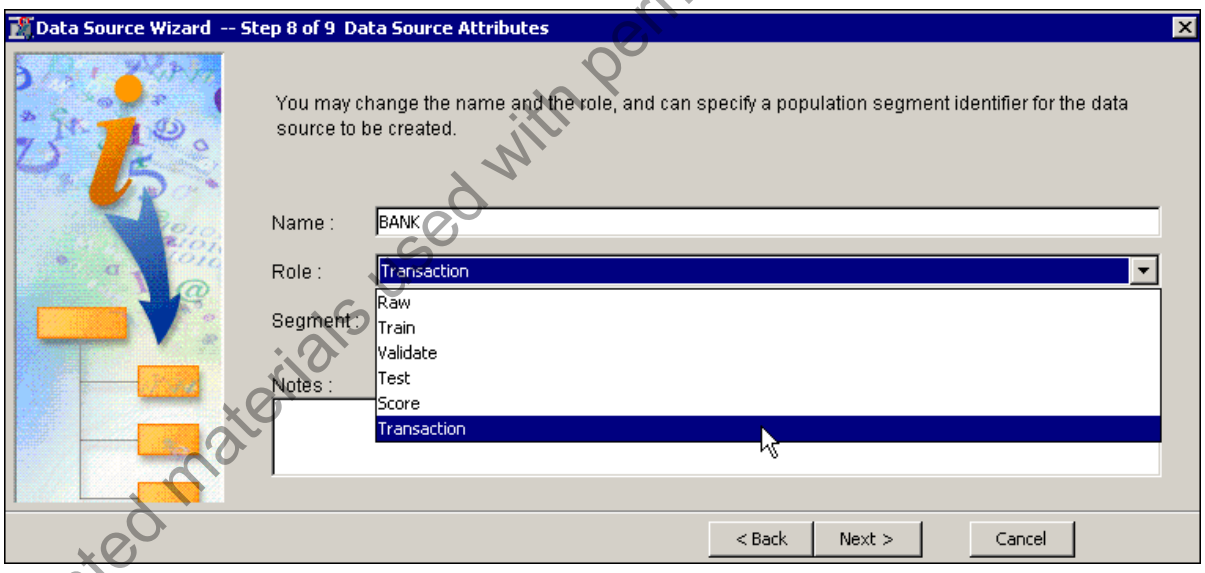
4. In Step 5, assign roles to the table variables as shown below.



An association analysis requires exactly one target variable and at least one ID variable. Both should have a nominal measurement level; however, a level of Interval for the ID variable is sufficient. A sequence analysis also requires a sequence variable. It usually has an ordinal measurement scale; however, in SAS Enterprise Miner the sequence variable must be assigned the level Interval.

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

   Select **Role** ⇨ **Transaction**.

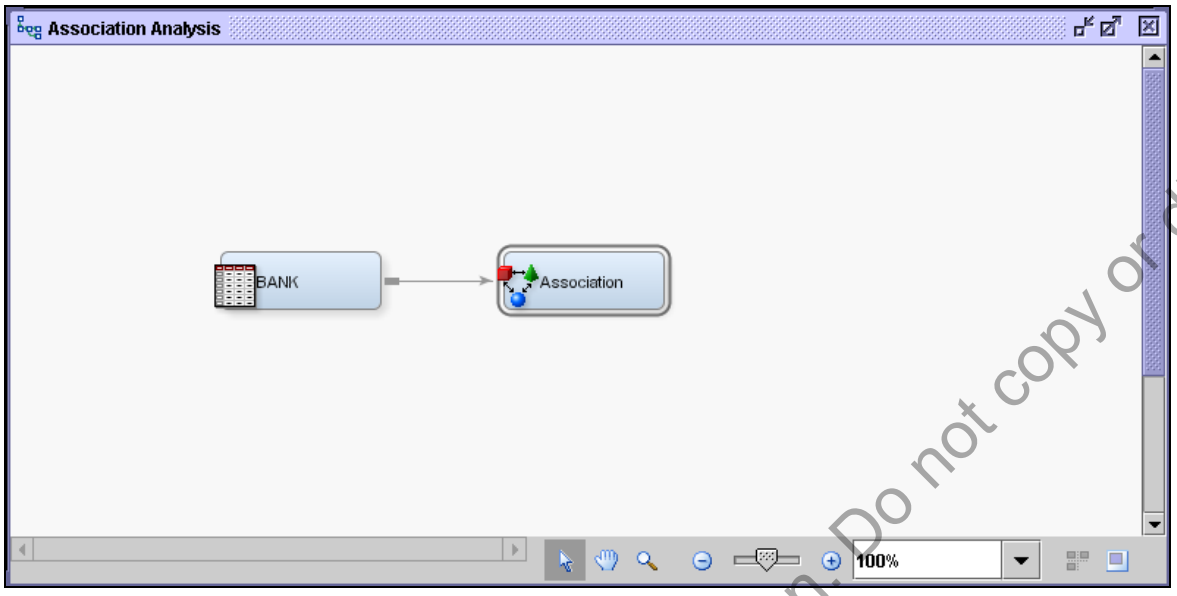


6. Select **Finish** to close the Data Source Wizard.

7. Drag a **BANK** data source into the diagram workspace.

8. Select the **Explore** tab and drag an **Association** tool into the diagram workspace.

9. Connect the **BANK** node to the **Association** node.



10. 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 Items to Process	100000
Rules	...
Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	.
Support Type	Percent
Support Count	.
Support Percentage	2.0
Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	No
Recommendation	No

11. 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**.

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:

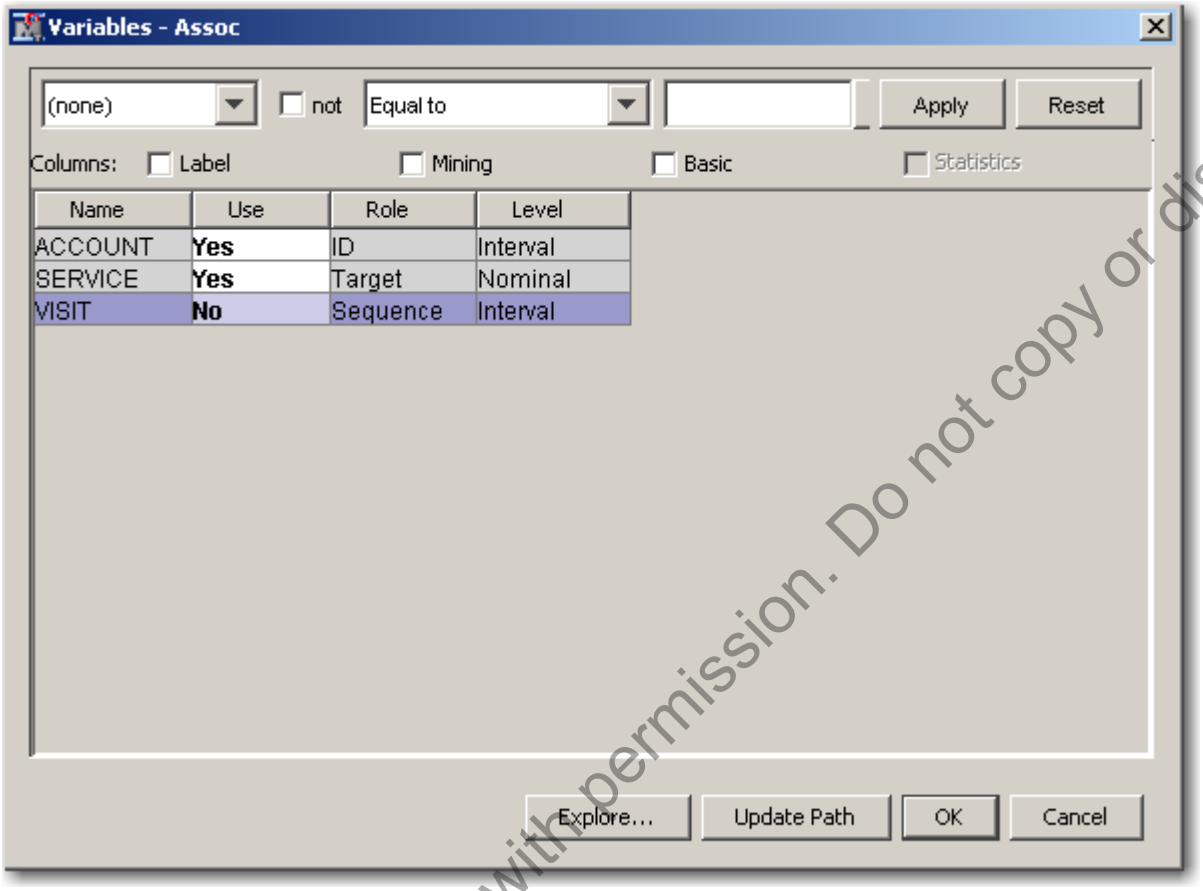
- **Minimum Confidence Level** specifies the minimum confidence level to generate a rule. The default level is **10%**.
- **Support Type** specifies whether the analysis should use the support count or support percentage property. The default setting is **Percent**.
- **Support Count** specifies a minimum level of support to claim that items are associated (that is, they occur together in the database).
- **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.
- **Maximum Items** determines 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.

12. Access the Variables dialog box for the Association node.

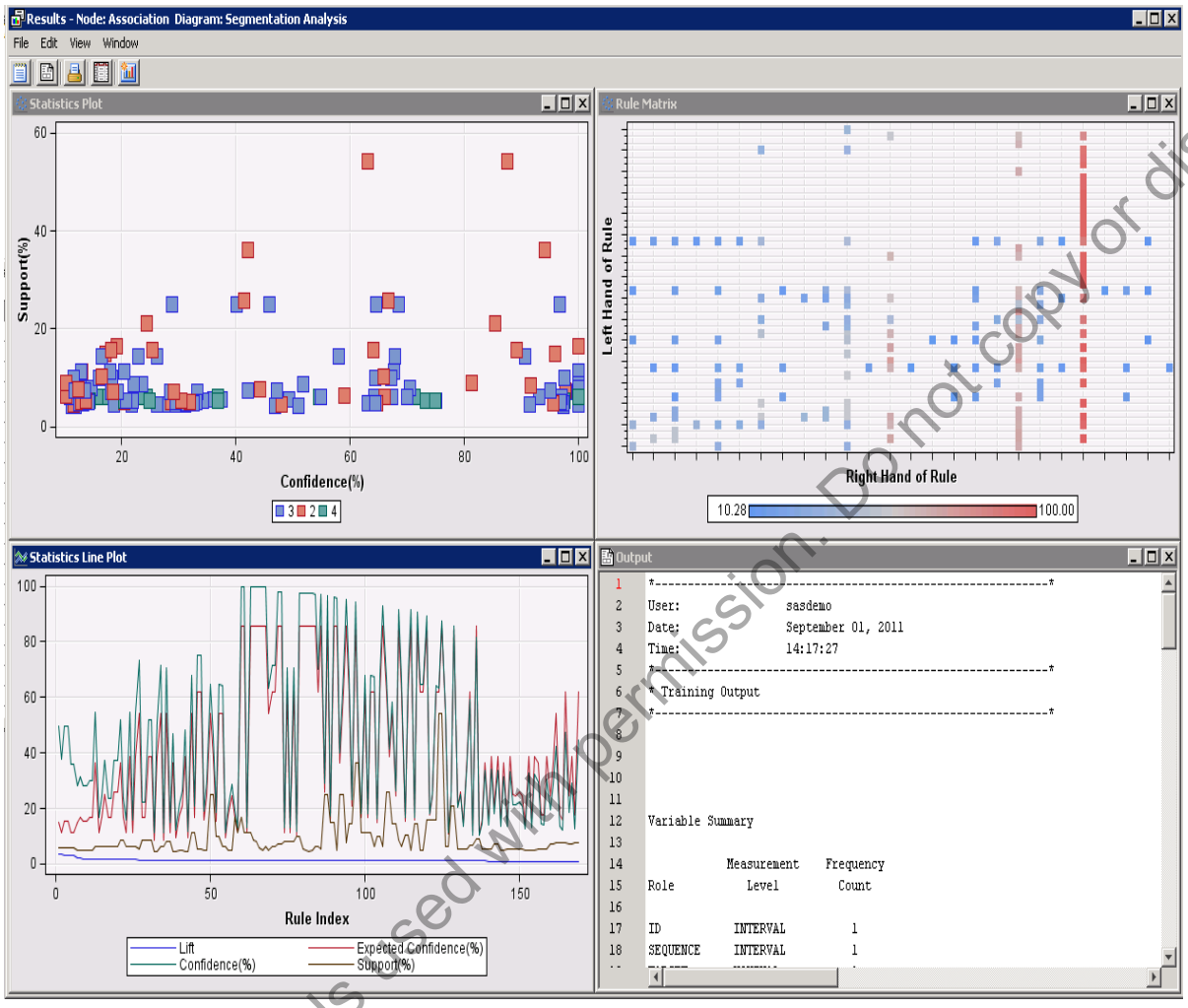
13. Select **Use** ⇨ **No** for the **VISIT** variable.



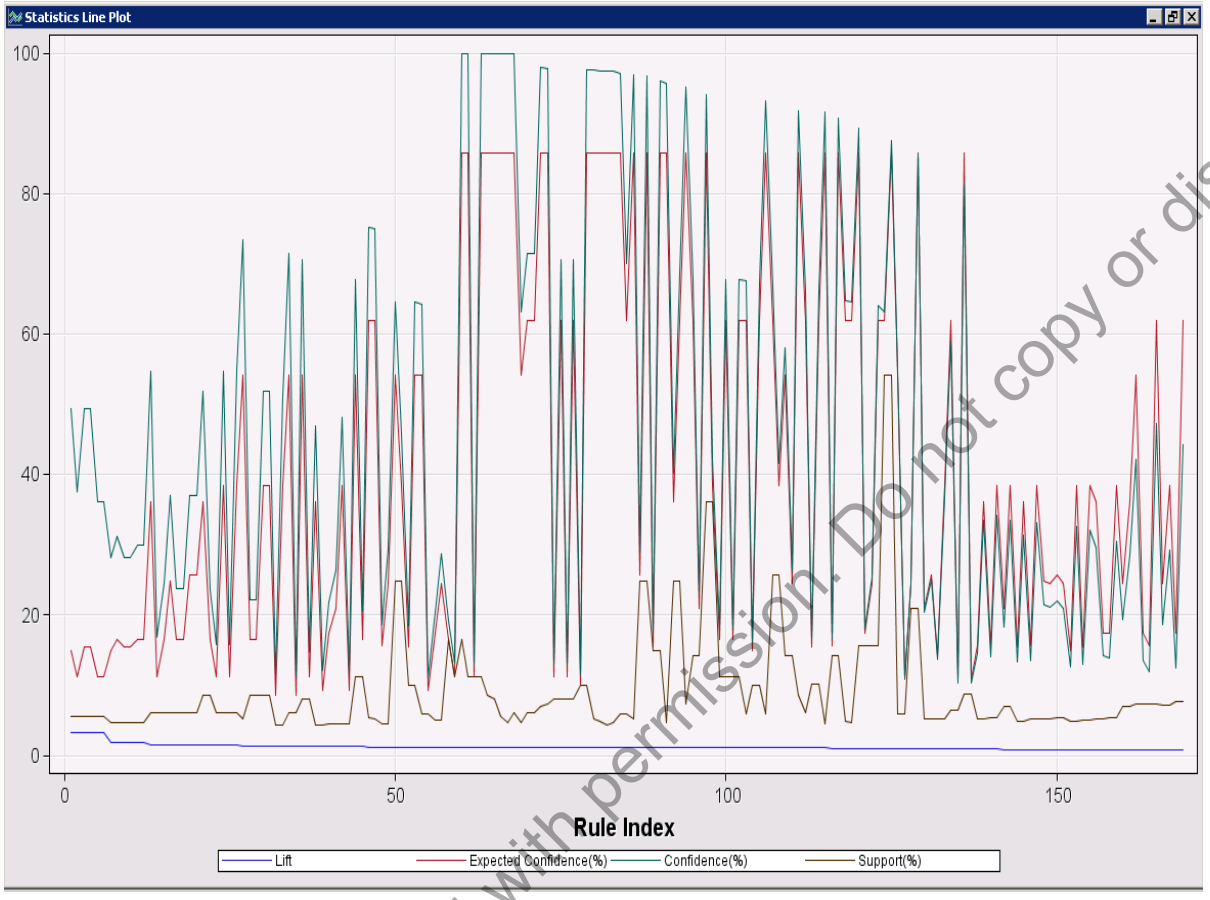
14. Select **OK** to close the Variables dialog box.

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

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



16. 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 ⇒ B. Recall the following:

- ***Support*** of A ⇒ B is the probability that a customer has both A and B.
- ***Confidence*** of A ⇒ B is the probability that a customer has B given that the customer has A.
- ***Expected Confidence*** of A ⇒ B is the probability that a customer has B.
- ***Lift*** of A ⇒ B is a measure of the strength of the association. If Lift=2 for the rule A=>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.

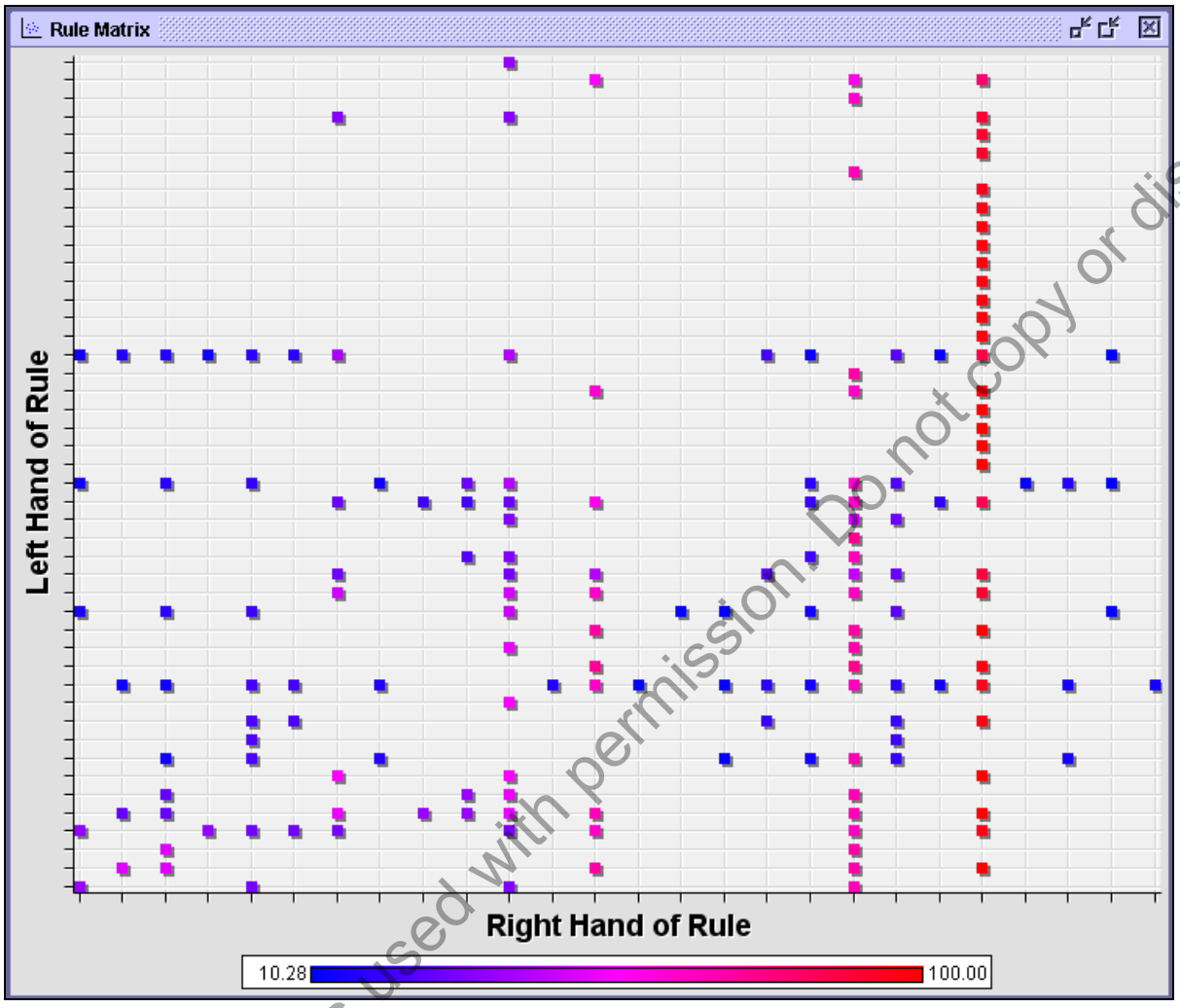
17. To view the descriptions of the rules, select **View** ⇨ **Rules** ⇨ **Rule description**.

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.

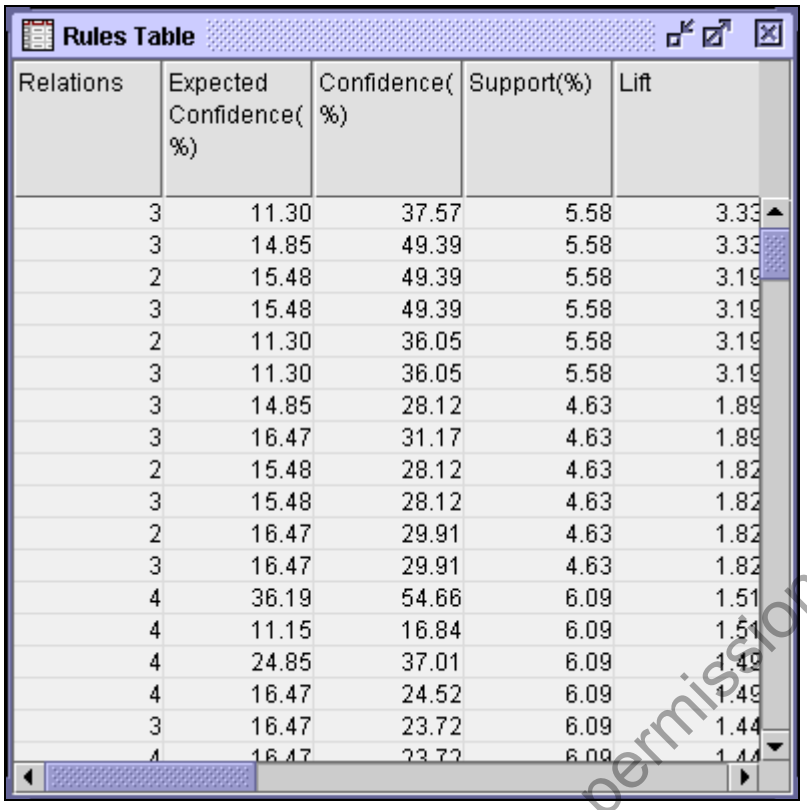
18. (Optional) Examine the rule matrix.



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.

19. Select **View** ⇨ **Rules** ⇨ **Rules Table**. The Rules Table window appears.



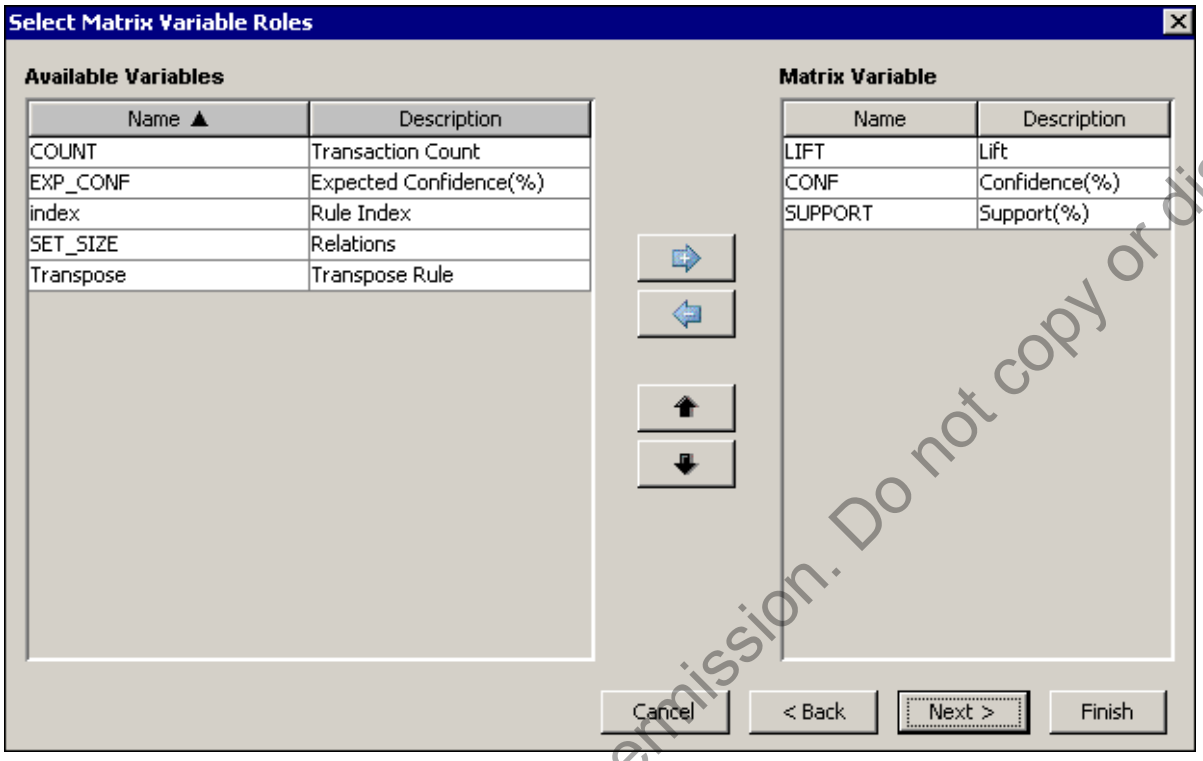
Rules Table

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

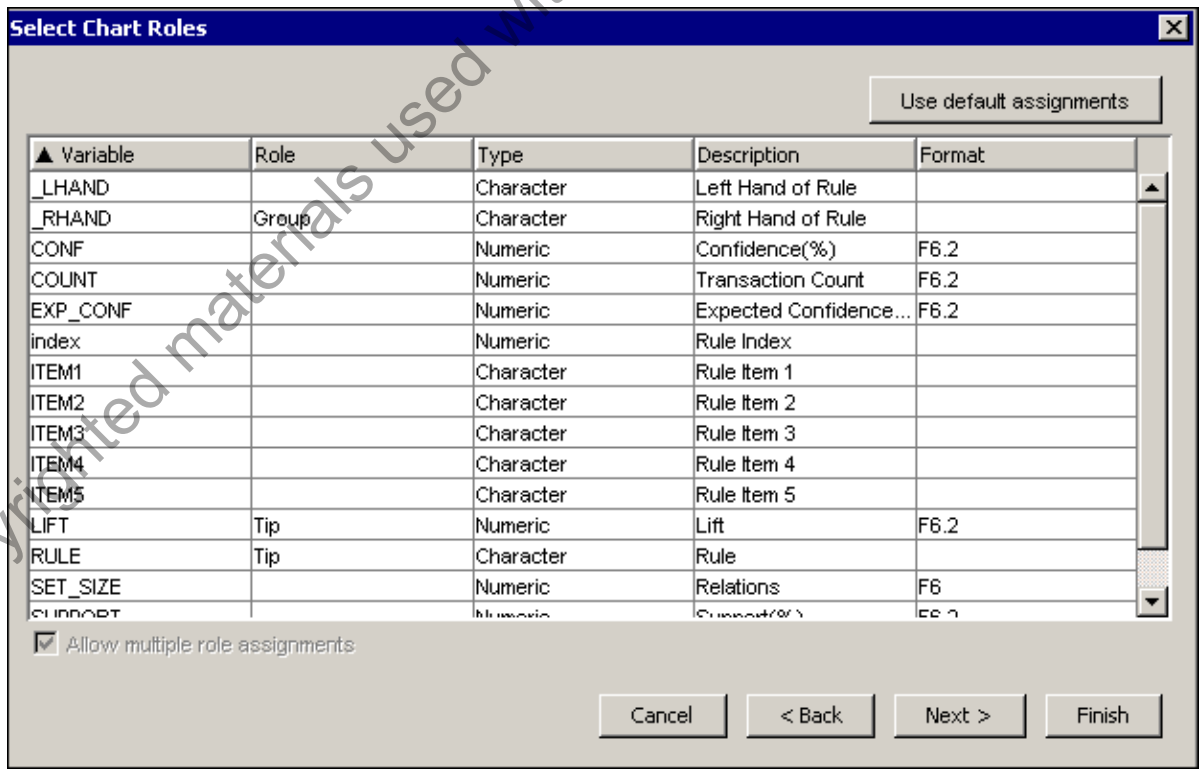
20. Select  (the Plot Wizard icon).

21. Choose a Matrix graph for the type of chart, and select **Next >**.

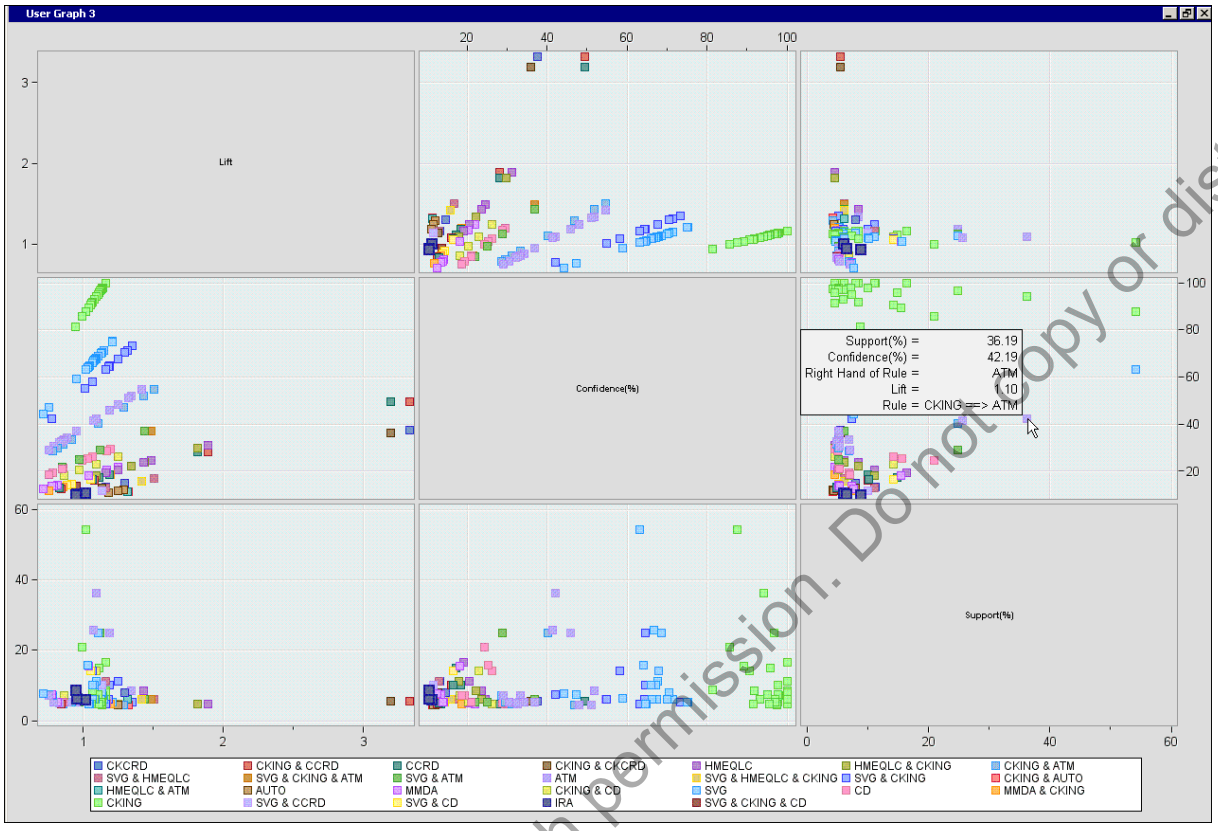
22. Select the matrix variables: **Lift, Conf** and **Support** as shown below right. Select **Next** .



23. Select the **Group** role for **_RHAND** and the **Tip** role for **LIFT** and **RULE** to add these details to the tooltip action.



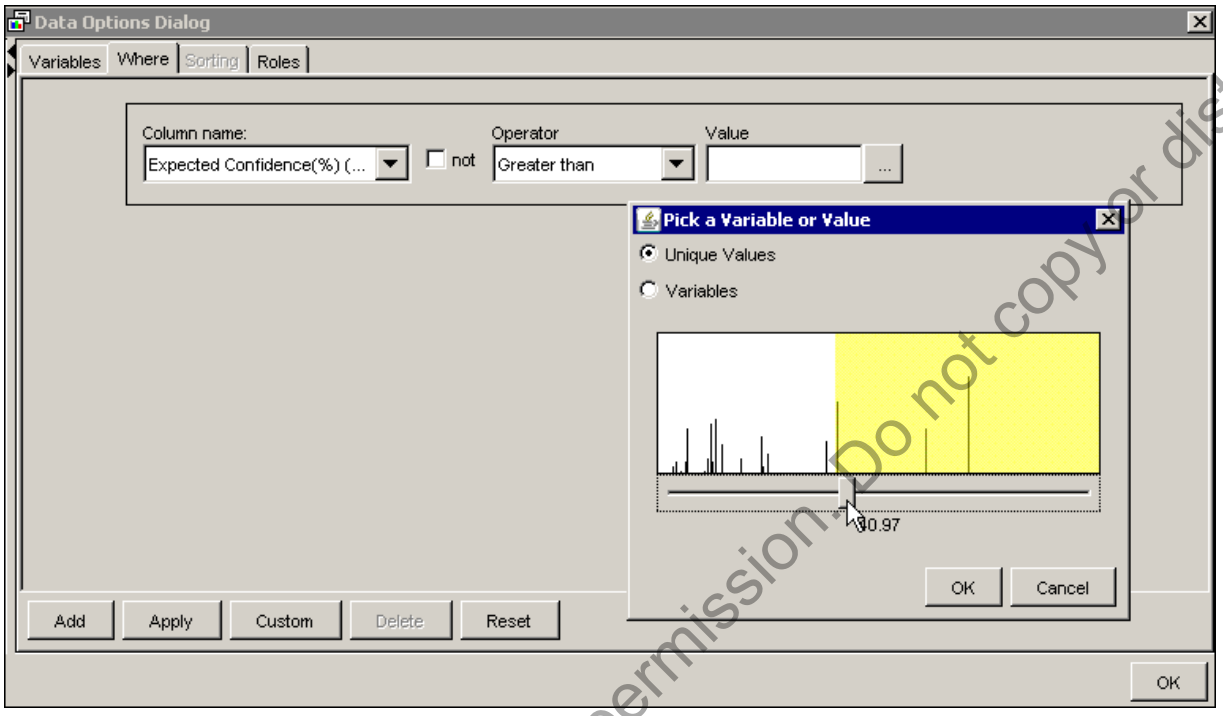
24. Select **Finish** to generate the plot.



The legend shows the right hand of the rule. When you click a service or group of services in the legend, the points in the matrix graphs are highlighted. This plot enables you to explore the relationships among the various metrics in association analysis.

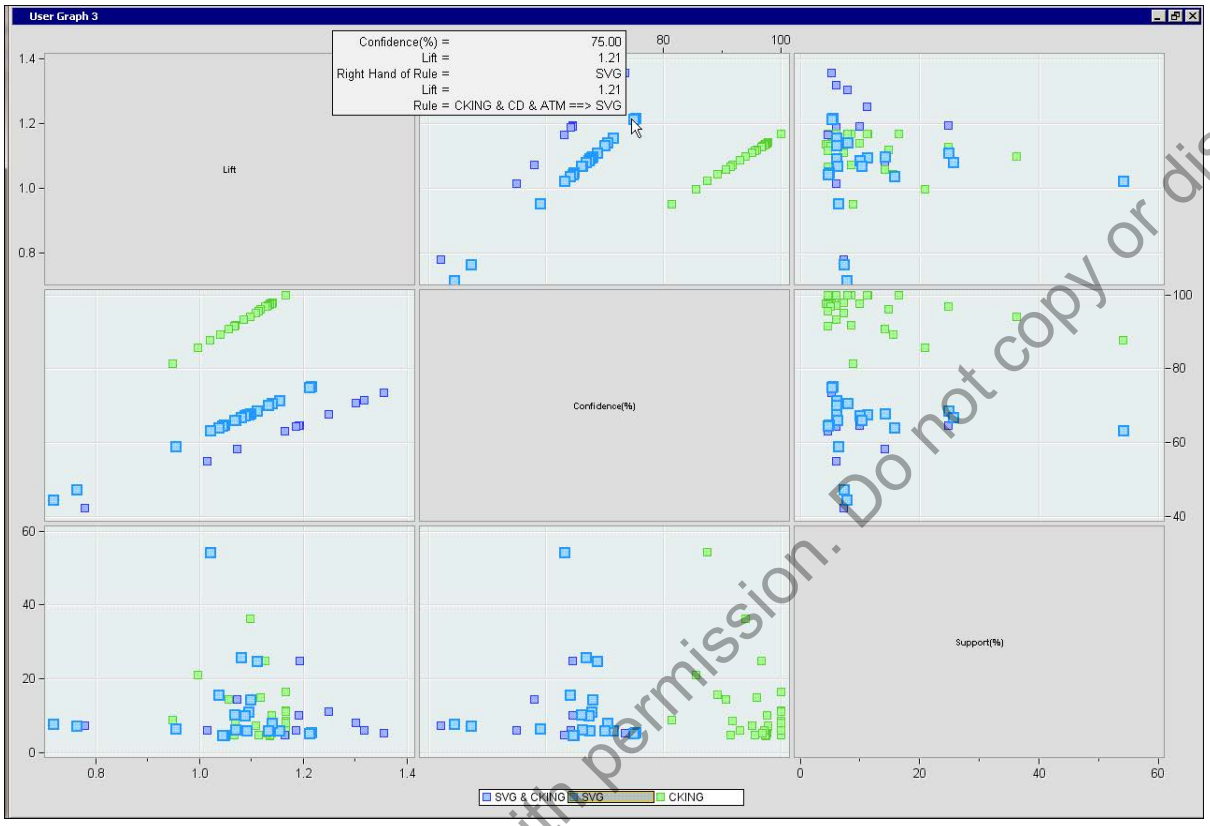
When you hover the cursor over a selected point in the plot, the tooltips show the details of the point, including the full rule.

25. Right-click in the graph and select **Data Options**. Select the **Where** tab. Specify **Expected Confidence(%)** as the column name and **Greater than** as the operator. Click the elipses next to **Value**. Set the slider to include values greater than 40, or type **40** for the value.

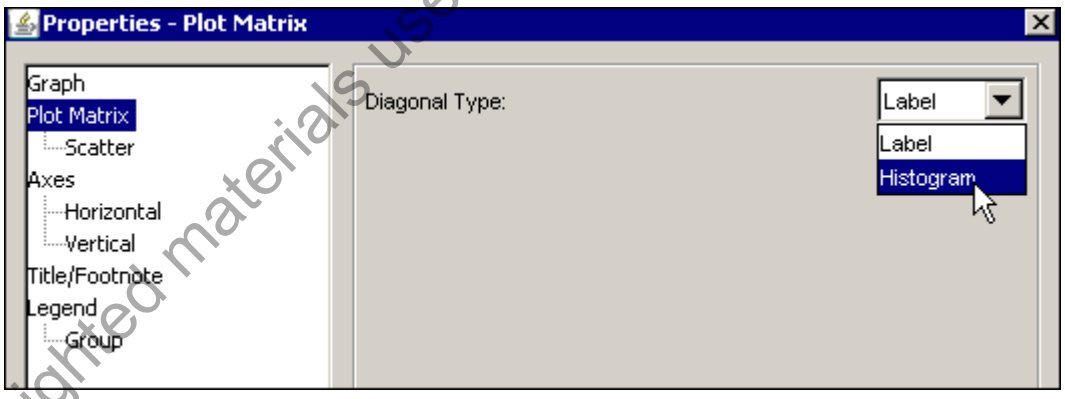


26. Select **OK** and **Apply**.

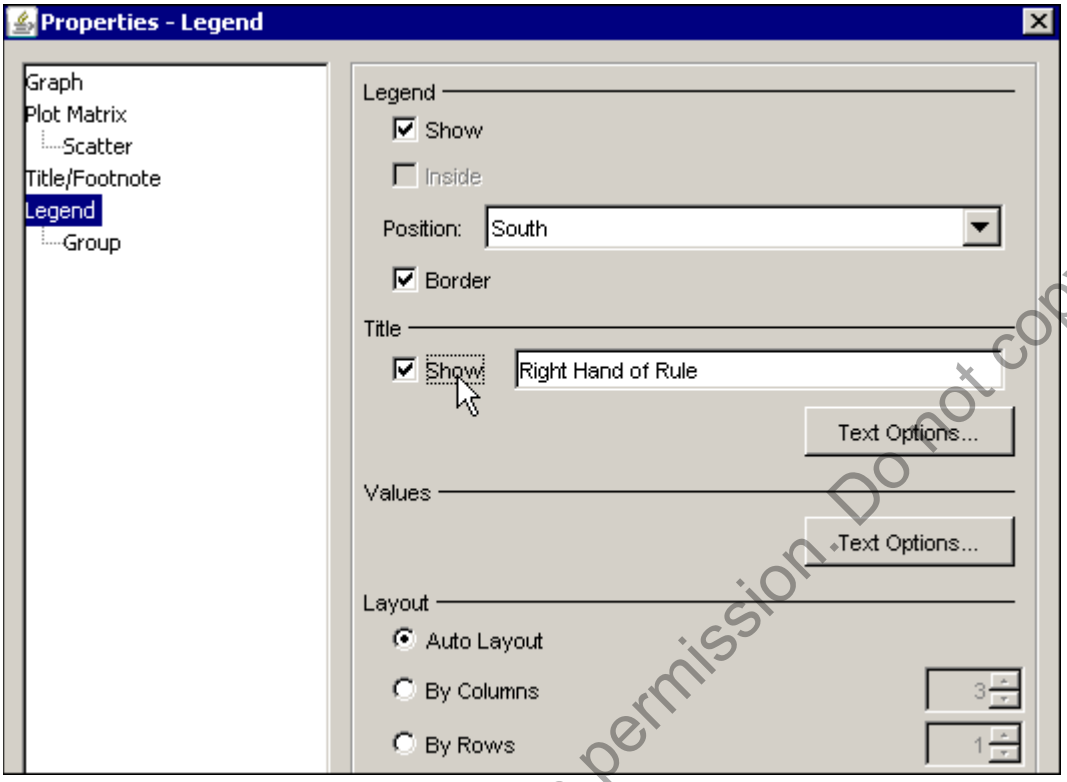
27. Select **OK** . The subset selected cases represent three different sets of services in the legend for the right hand of the rules.



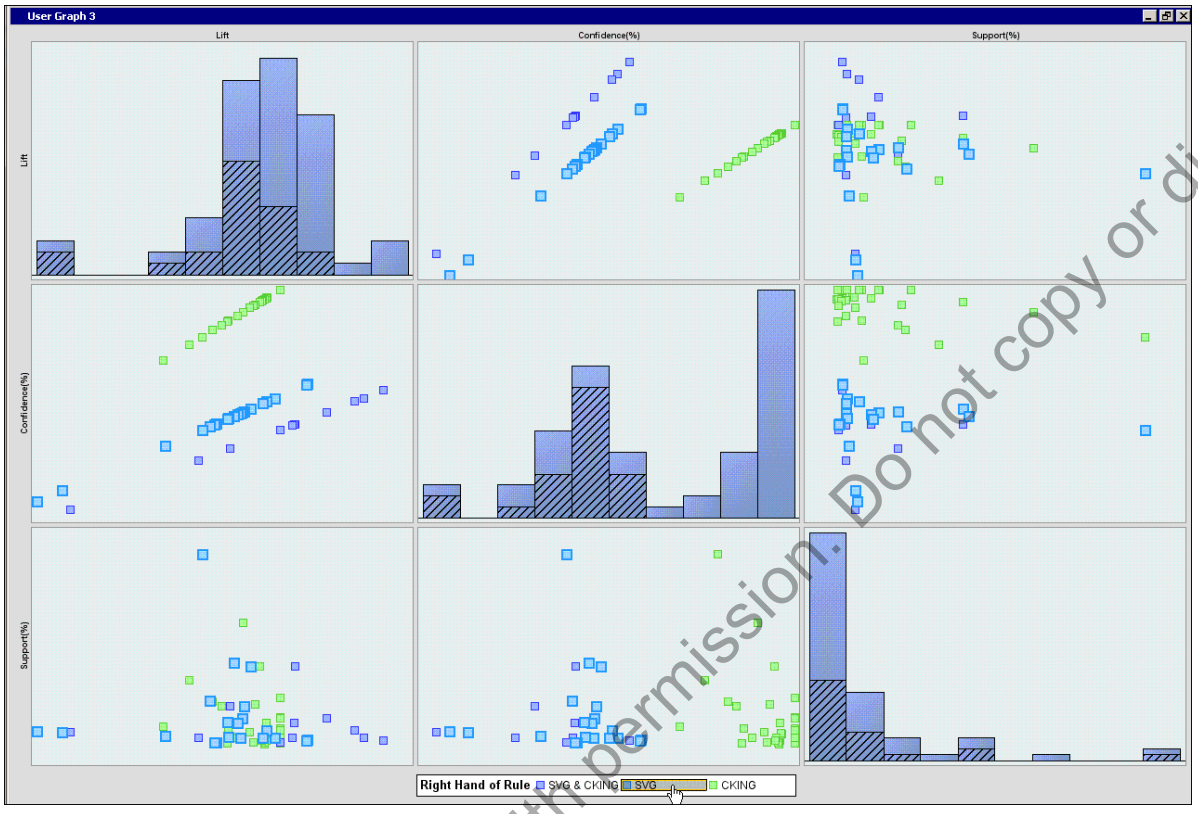
28. You can modify the look of the graph. Right-click the graph and select **Graph Properties**. Change the plot matrix diagonal type to **Histogram**.



29. Label the legend by selecting **Legend** and selecting the check box in the Title are next to **Show (Right hand of Rule)**. Select **OK**.



30. Click the **SVG** (Savings Account) category in the legend and notice that the histograms show the distribution of the selected rules in the diagonal.

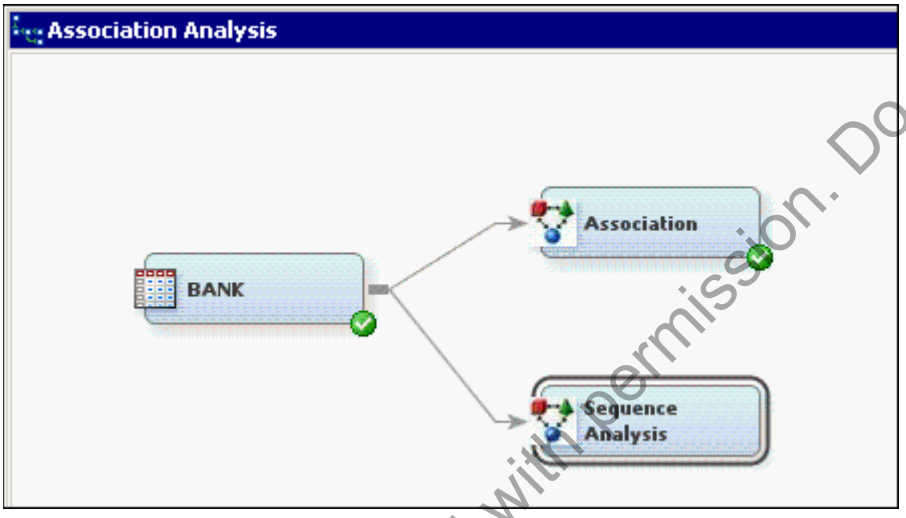


31. Close the Results window.

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.

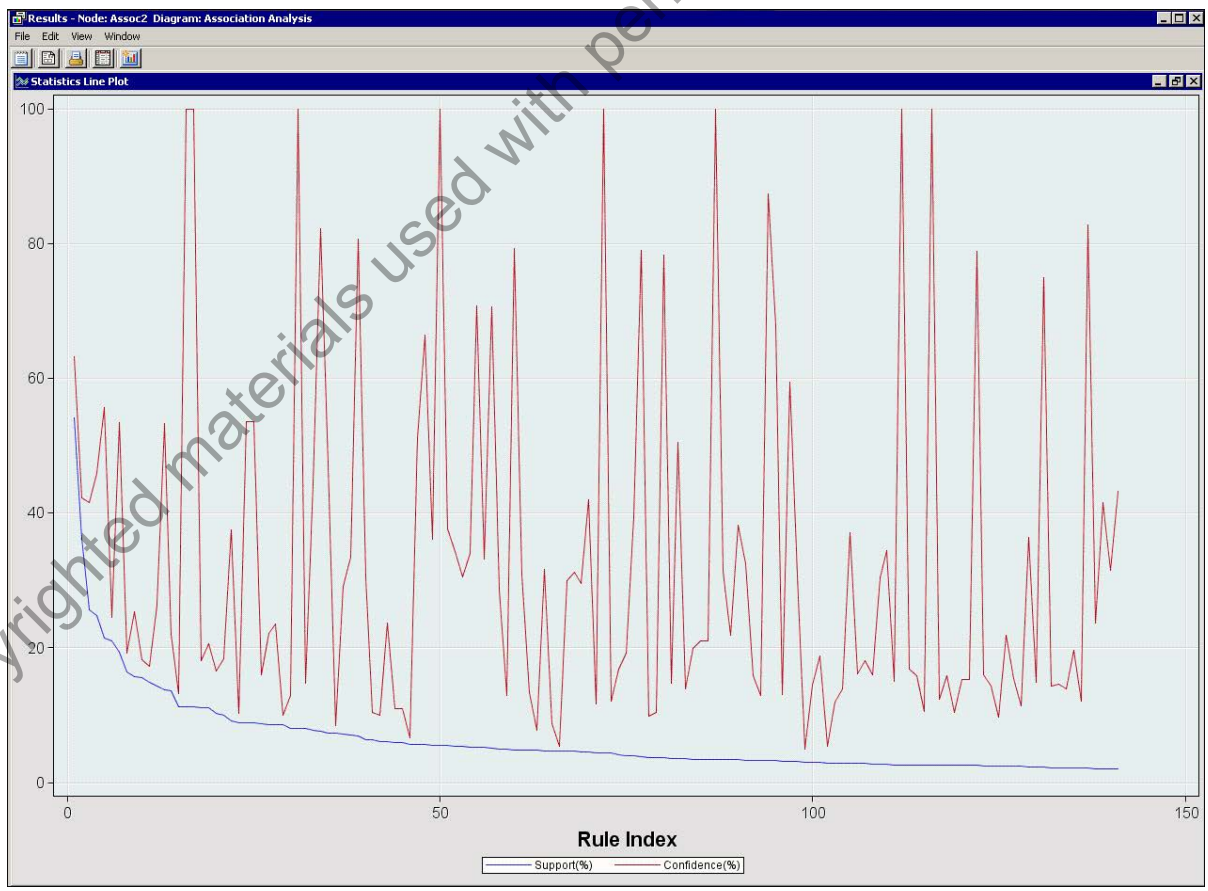
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	.
Support Type	Percent
Support Count	.
Support Percentage	2.0

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 property is set to **Count**. If a sequence has a count less than the specified value, that sequence is excluded from the output.
- **Support Percentage** specifies the minimum level of support to include the sequence in the analysis when the Support Type property 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.

Rule description

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

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