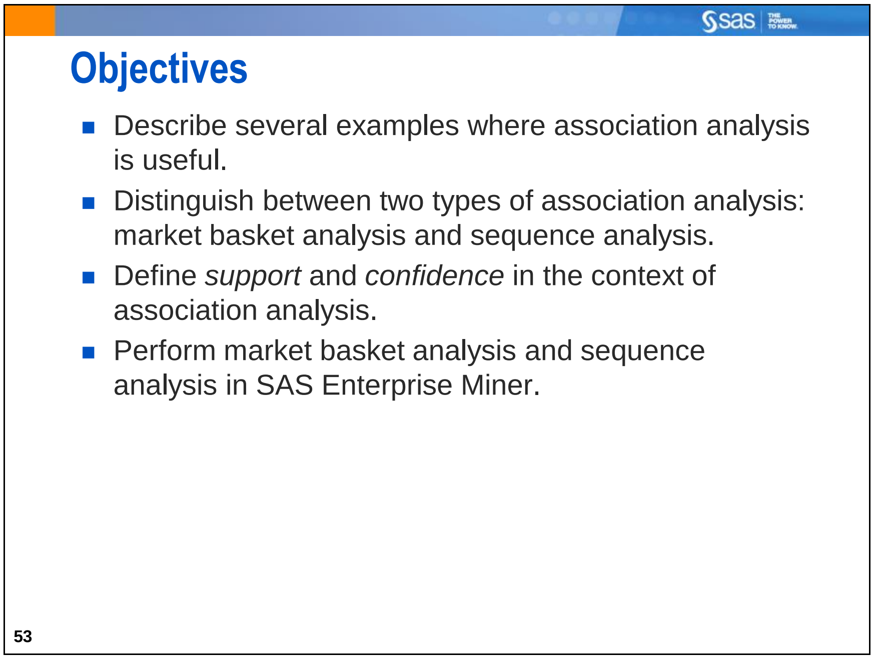
6.1 Market Basket Analysis



Objectives

- Describe several examples where association analysis is useful.
- Distinguish between two types of association analysis: market basket analysis and sequence analysis.
- Define *support* and *confidence* in the context of association analysis.
- Perform market basket analysis and sequence analysis in SAS Enterprise Miner.

53

Market Baskets for Grocery Groupings

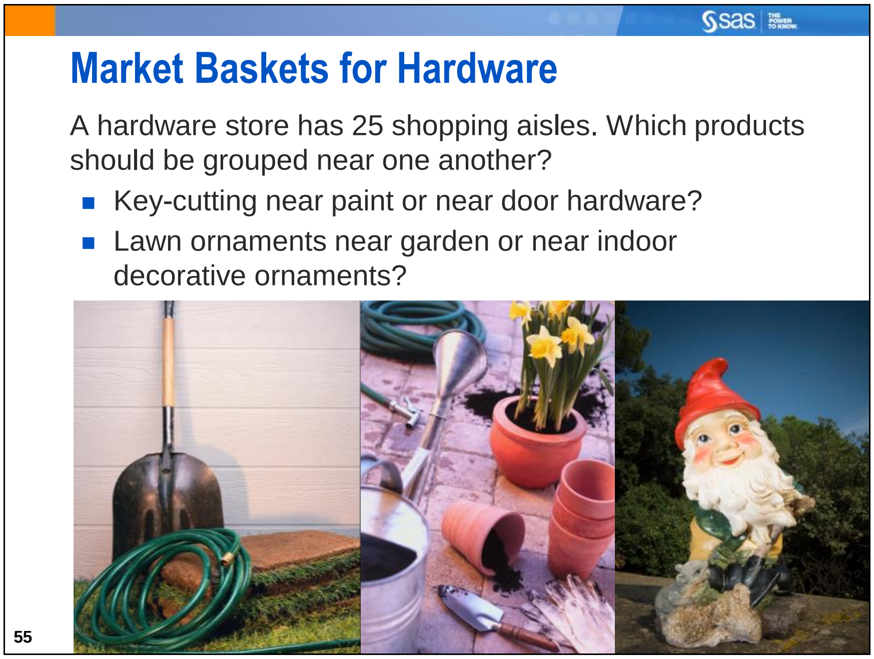
A classic application of market basket analysis addresses this question:

Which items are likely to be purchased together?

- If product A and product B often go together, then placing a more expensive alternative to B near the display for A can create an up-sell opportunity.
- If product A and B are often purchased together, putting them on sale at different times can drive purchases continually.

54

Up-sell refers to selling a new product to an existing customer. *Cross-sell* refers to attaching an additional product to an existing sale.

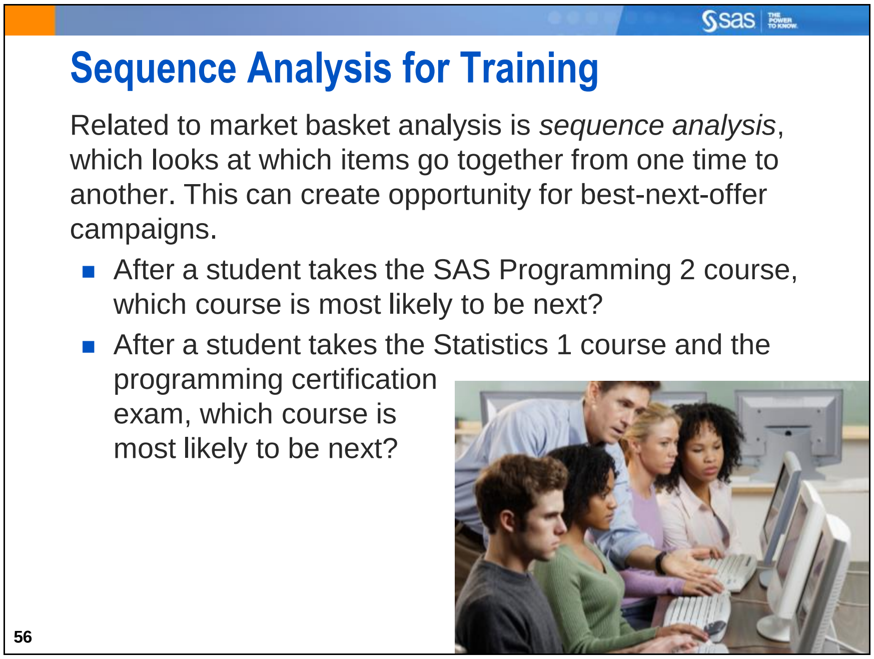


Market Baskets for Hardware

A hardware store has 25 shopping aisles. Which products should be grouped near one another?

- Key-cutting near paint or near door hardware?
- Lawn ornaments near garden or near indoor decorative ornaments?

55

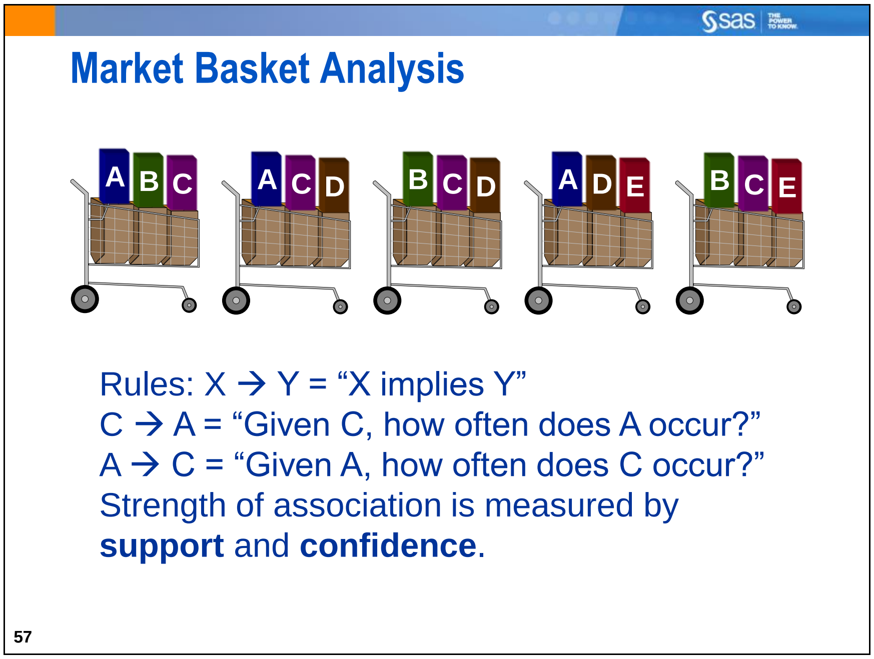


Sequence Analysis for Training

Related to market basket analysis is *sequence analysis*, which looks at which items go together from one time to another. This can create opportunity for best-next-offer campaigns.

- After a student takes the SAS Programming 2 course, which course is most likely to be next?
- After a student takes the Statistics 1 course and the programming certification exam, which course is most likely to be next?

56



Market Basket Analysis

A B C
A C D
B C D
A D E
B C E

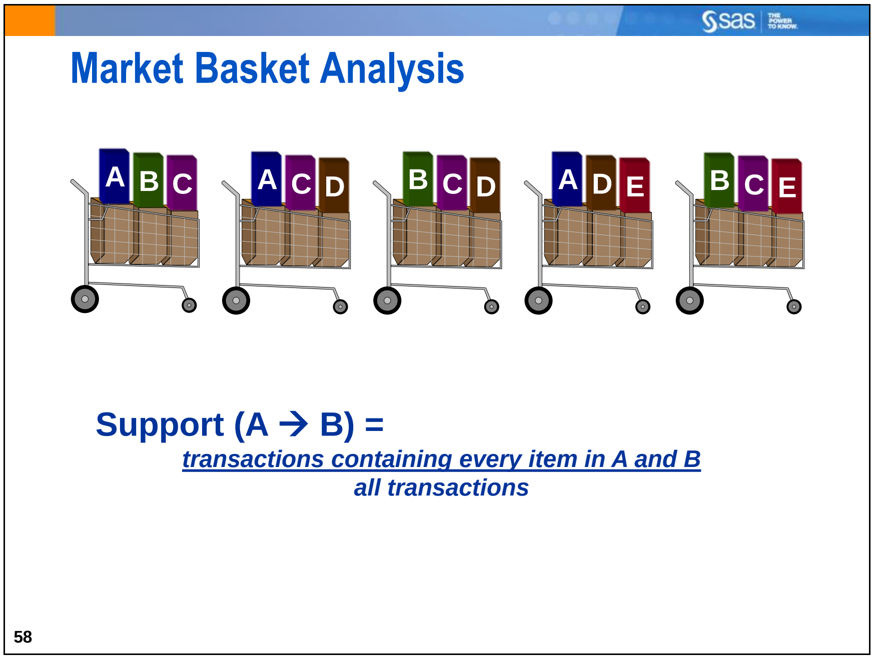
Rules: X → Y = “X implies Y”
C → A = “Given C, how often does A occur?”
A → C = “Given A, how often does C occur?”
Strength of association is measured by **support** and **confidence**.

57

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*) ⇒ (item set *B*).

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



Market Basket Analysis

Support (A → B) =

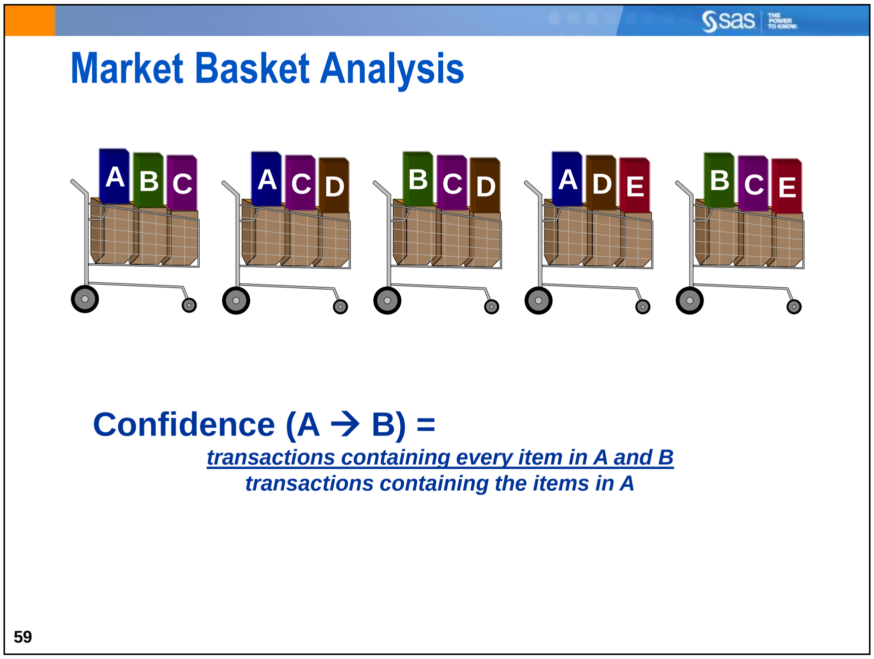
$$\frac{\textit{transactions containing every item in A and B}}{\textit{all transactions}}$$

58

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:

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

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


A B C
A C D
B C D
A D E
B C E

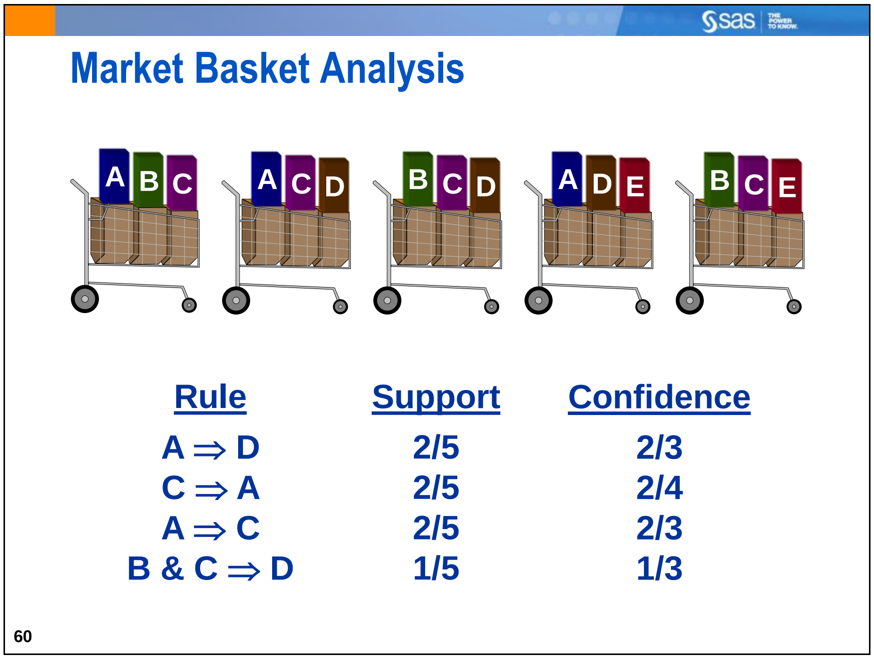
Market Basket Analysis

Confidence (A → B) =

$$\frac{\textit{\underline{transactions containing every item in A and B}}}{\textit{transactions containing the items in A}}$$

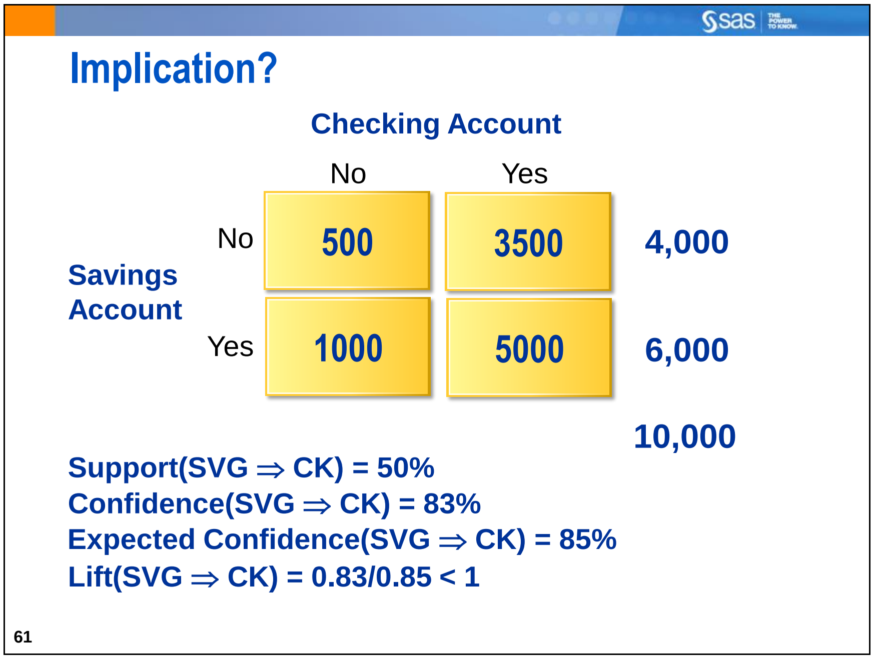
The confidence of an association rule $A \Rightarrow B$ is the conditional probability of a transaction containing item set B given that it contains item set A. The confidence is estimated by the following:

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


A B C
A C D
B C D
A D E
B C E

Market Basket Analysis

Rule	Support	Confidence
$A \Rightarrow D$	2/5	2/3
$C \Rightarrow A$	2/5	2/4
$A \Rightarrow C$	2/5	2/3
$B \,\&\, C \Rightarrow D$	1/5	1/3



Implication?

		Checking Account		
		No	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

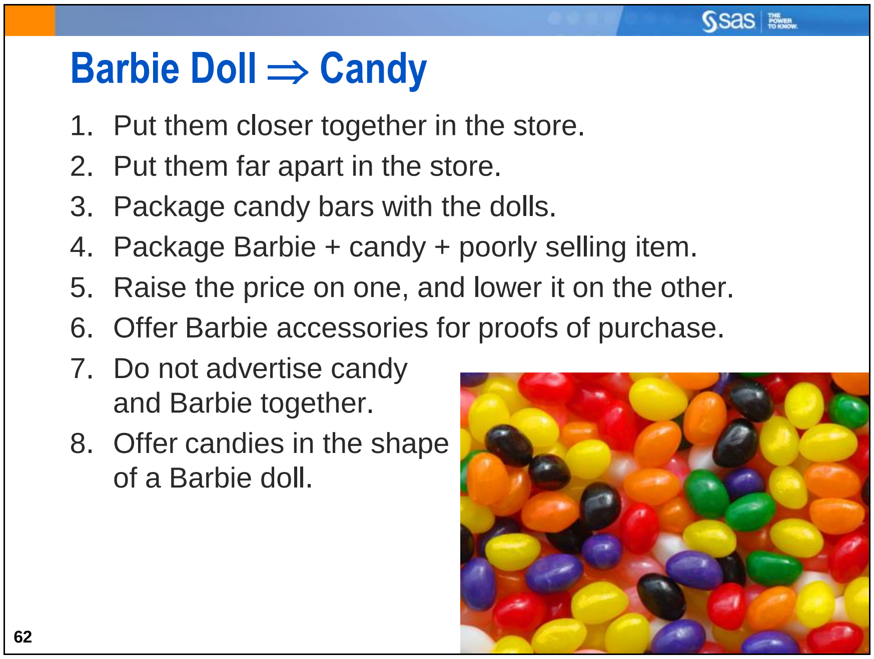
61

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 $B \Rightarrow A$.

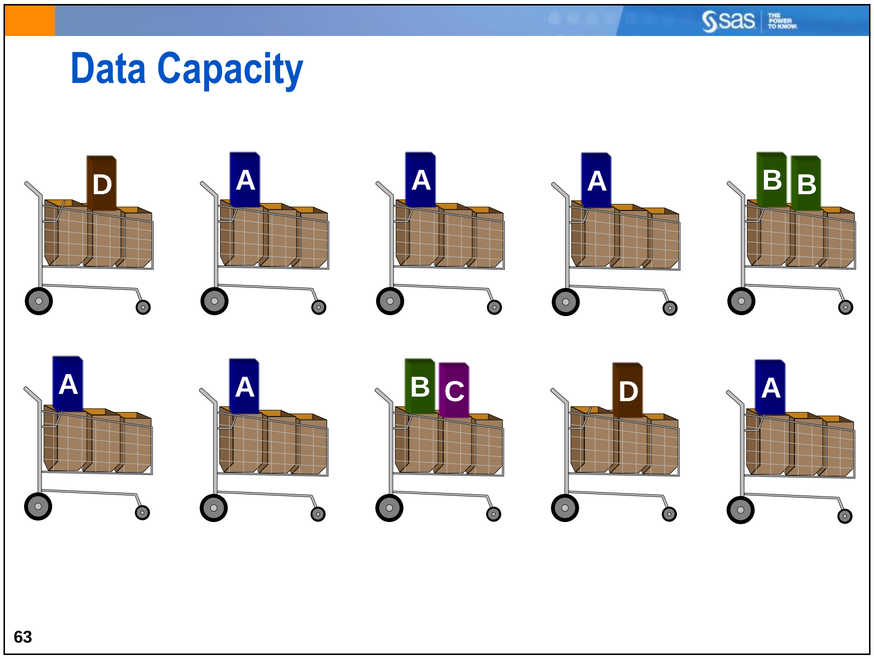


Barbie Doll ⇒ Candy

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.

62

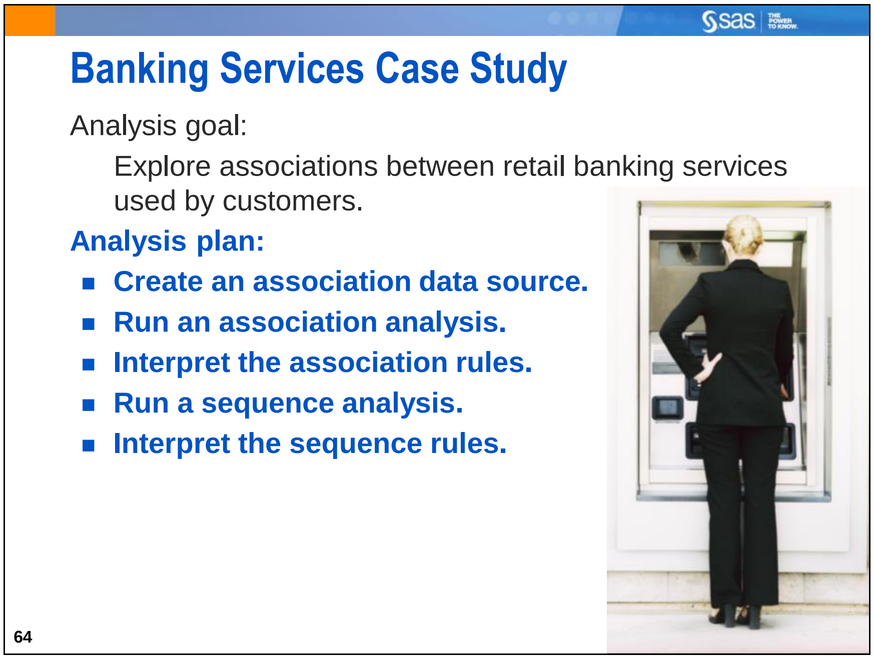
Forbes (Palmeri 1997) reported that a major retailer determined that customers who buy Barbie dolls have a 60% likelihood of buying one of three types of candy bars. The confidence of the rule Barbie ⇒ candy is 60%. The retailer was unsure what to do with this nugget. The online newsletter *Knowledge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



Data Capacity

63

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.



Banking Services Case Study

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.**

64

Extending from the knowledge gained in the banking segmentation, a bank’s Marketing Department is interested in examining associations between various retail banking services used by customers. This project hopes to improve cross-sell and up-sell opportunities by determining typical and atypical service combinations as well as the order in which the services were first used. This can be helpful in suggesting, for new and existing customers, services that match their personal banking habits.

These requirements suggest both a market basket analysis and a sequence 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 more than 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 3.

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	savings account
TRUST	personal trust account

Banking Services Case Study: Performing Association Analysis

Market Basket Analysis

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

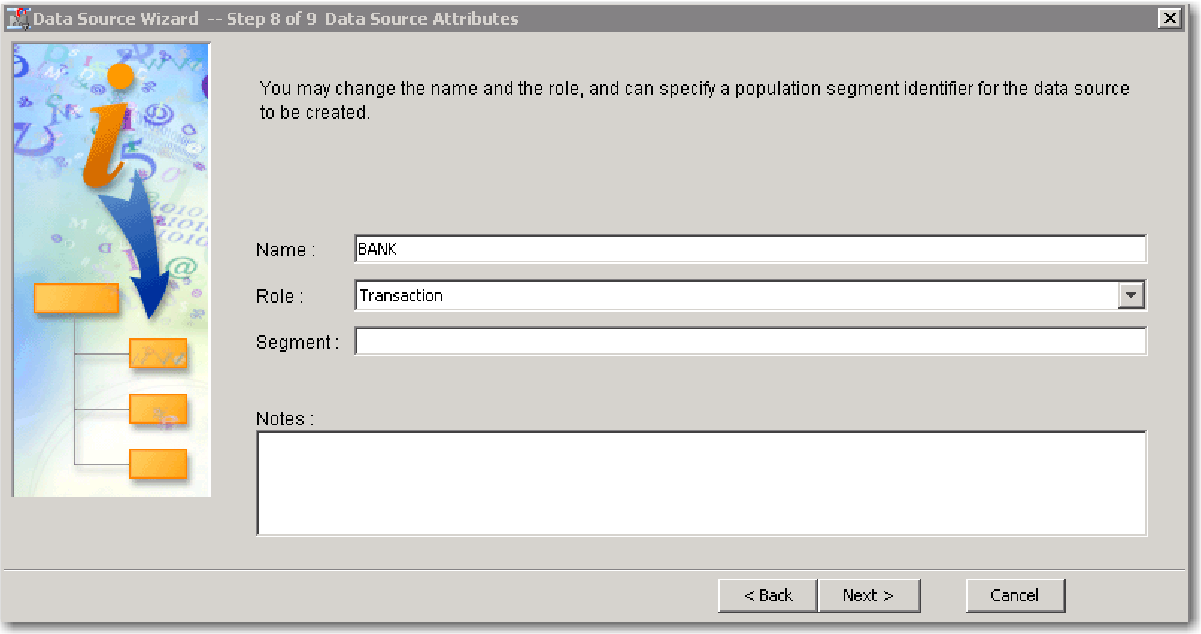
1. Create a new diagram named **Associations Analysis** to contain this analysis.
2. Right-click **Data Sources** and select **Create Data Source**.
3. Change the source to metadata repository and select the **BANK** table in the **ABA1** library.
4. Proceed to step 6 of the Data Source Wizard.
5. Assign metadata to the table variables as follows: change **ACCOUNT** to an **ID** role, change **SERVICE** to a **Target** role, and change **VISIT** to a **Sequence** role.

   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.

6. Proceed through the Data Source Wizard to the Data Source Attributes step.

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

7. Select **Transaction** for the value of the **Role** field.



Data Source Wizard -- Step 8 of 9 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 : BANK

Role : Transaction

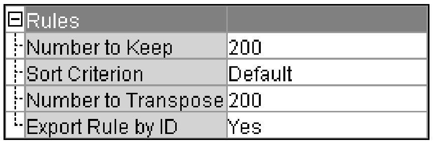
Segment :

Notes :

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

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

13. The Export Rule by ID property determines whether the **Rule-by-ID** data is exported from the node and whether the Rule Description table is 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). The default count is 1.
- ***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 is 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 a possible solution.

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

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

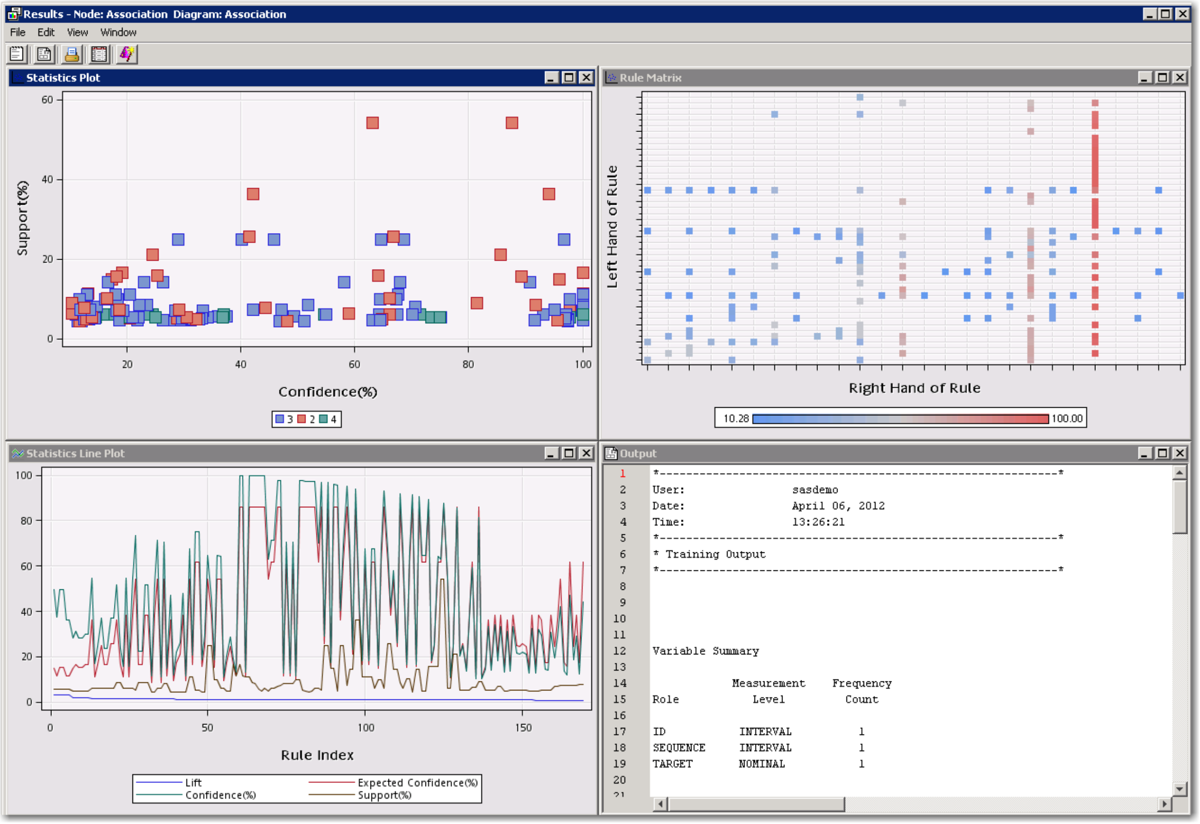
Select **Update Path** if no variables appear.

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

16. Click **OK** to close the Variables dialog box.

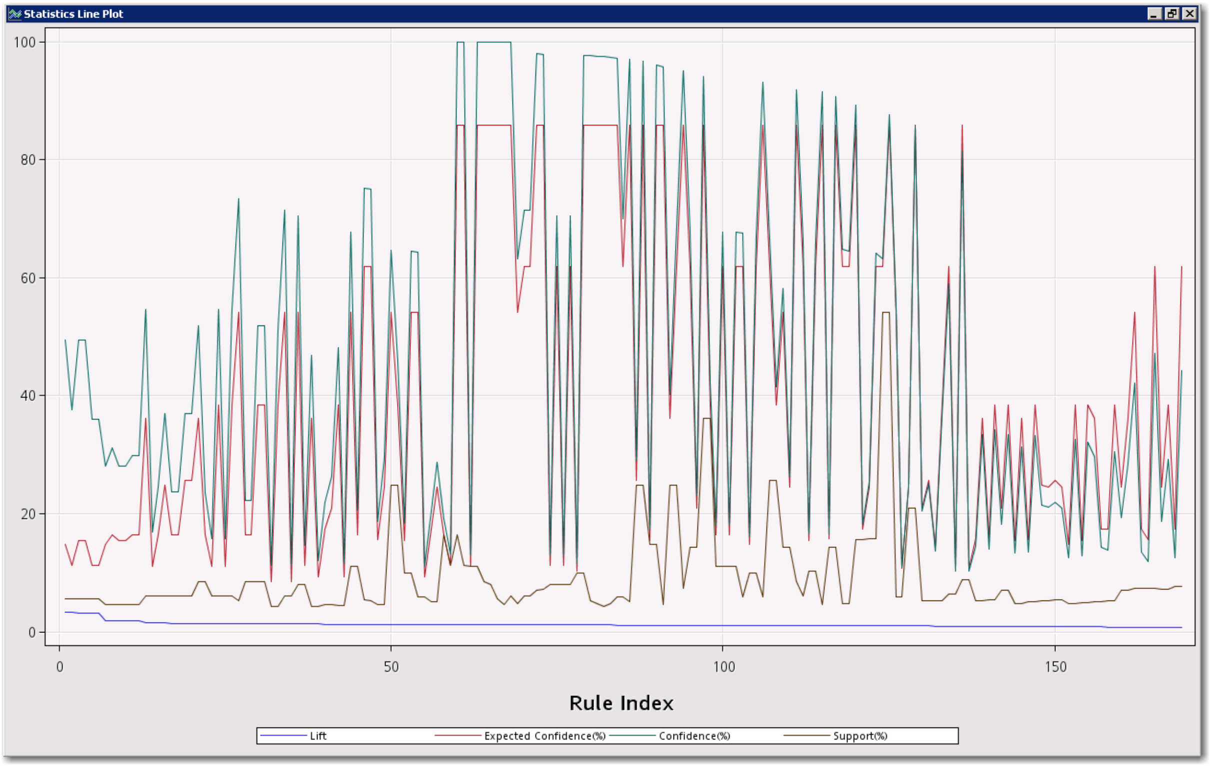
17. Run the diagram from the Association node and view the results.

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



```
*------------------------------------------------------------*
User:                 sasdemo
Date:                 April 06, 2012
Time:                 13:26:21
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

ID           INTERVAL          1
SEQUENCE     INTERVAL          1
TARGET       NOMINAL           1

```

18. Maximize the Statistics Line Plot window.



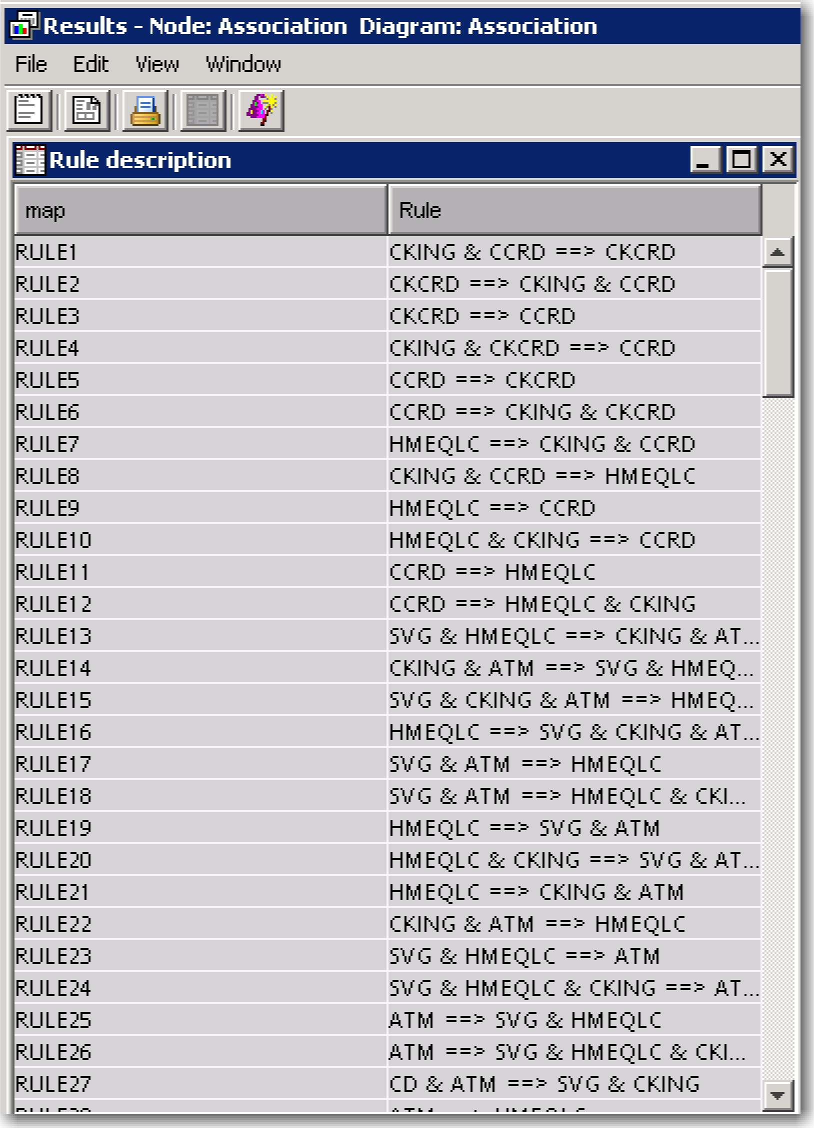
The statistics line plot graphs the lift, expected confidence, confidence, and support for each of the rules by rule index number.

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 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.

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

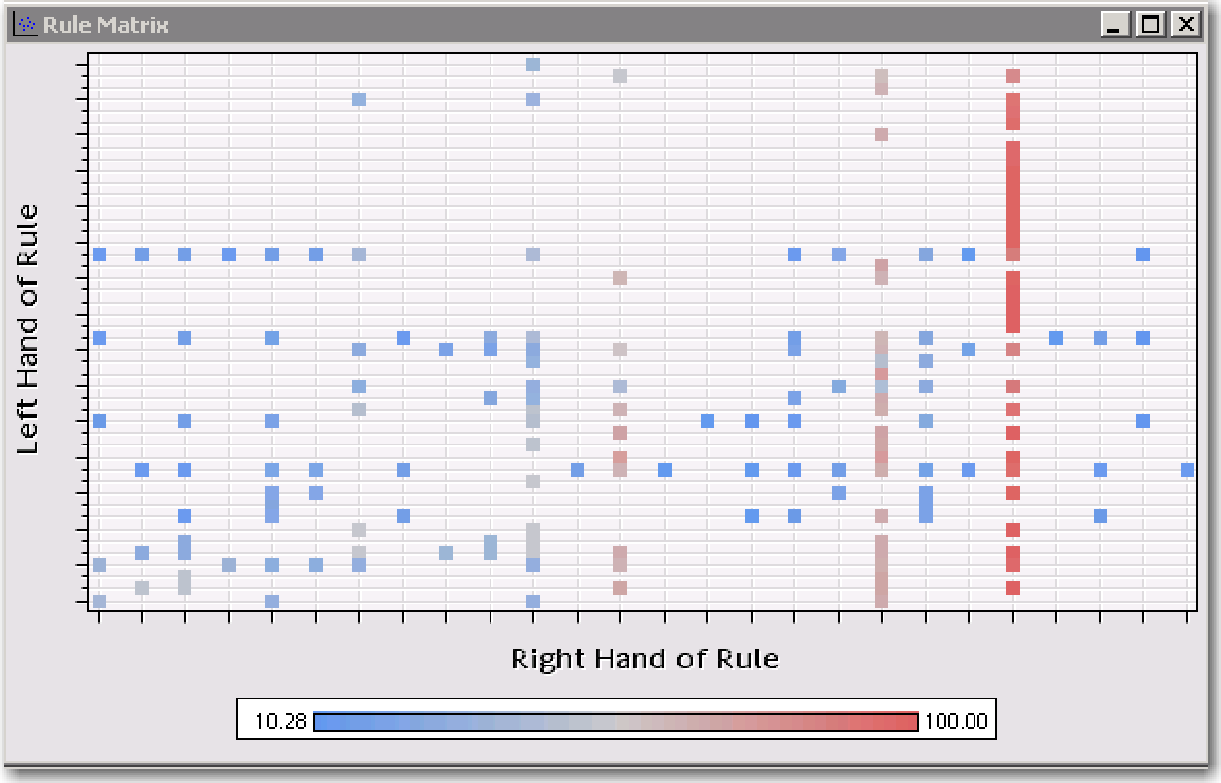


map	Rule
RULE1	CKING & CCRD ==> CKCRD
RULE2	CKCRD ==> CKING & CCRD
RULE3	CKCRD ==> CCRD
RULE4	CKING & CKCRD ==> CCRD
RULE5	CCRD ==> CKCRD
RULE6	CCRD ==> CKING & CKCRD
RULE7	HMEQLC ==> CKING & CCRD
RULE8	CKING & CCRD ==> HMEQLC
RULE9	HMEQLC ==> CCRD
RULE10	HMEQLC & CKING ==> CCRD
RULE11	CCRD ==> HMEQLC
RULE12	CCRD ==> HMEQLC & CKING
RULE13	SVG & HMEQLC ==> CKING & AT...
RULE14	CKING & ATM ==> SVG & HMEQ...
RULE15	SVG & CKING & ATM ==> HMEQ...
RULE16	HMEQLC ==> SVG & CKING & AT...
RULE17	SVG & ATM ==> HMEQLC
RULE18	SVG & ATM ==> HMEQLC & CKI...
RULE19	HMEQLC ==> SVG & ATM
RULE20	HMEQLC & CKING ==> SVG & AT...
RULE21	HMEQLC ==> CKING & ATM
RULE22	CKING & ATM ==> HMEQLC
RULE23	SVG & HMEQLC ==> ATM
RULE24	SVG & HMEQLC & CKING ==> AT...
RULE25	ATM ==> SVG & HMEQLC
RULE26	ATM ==> SVG & HMEQLC & CKI...
RULE27	CD & ATM ==> SVG & CKING

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.

One of the higher lift rules is that a home equity line of credit (LOC) implies checking and check card (and vice versa). Perhaps customers with a home equity LOC, who do not already have a checking account, should be offered a checking account and check card with a special promotion.

20. Examine the rule matrix.



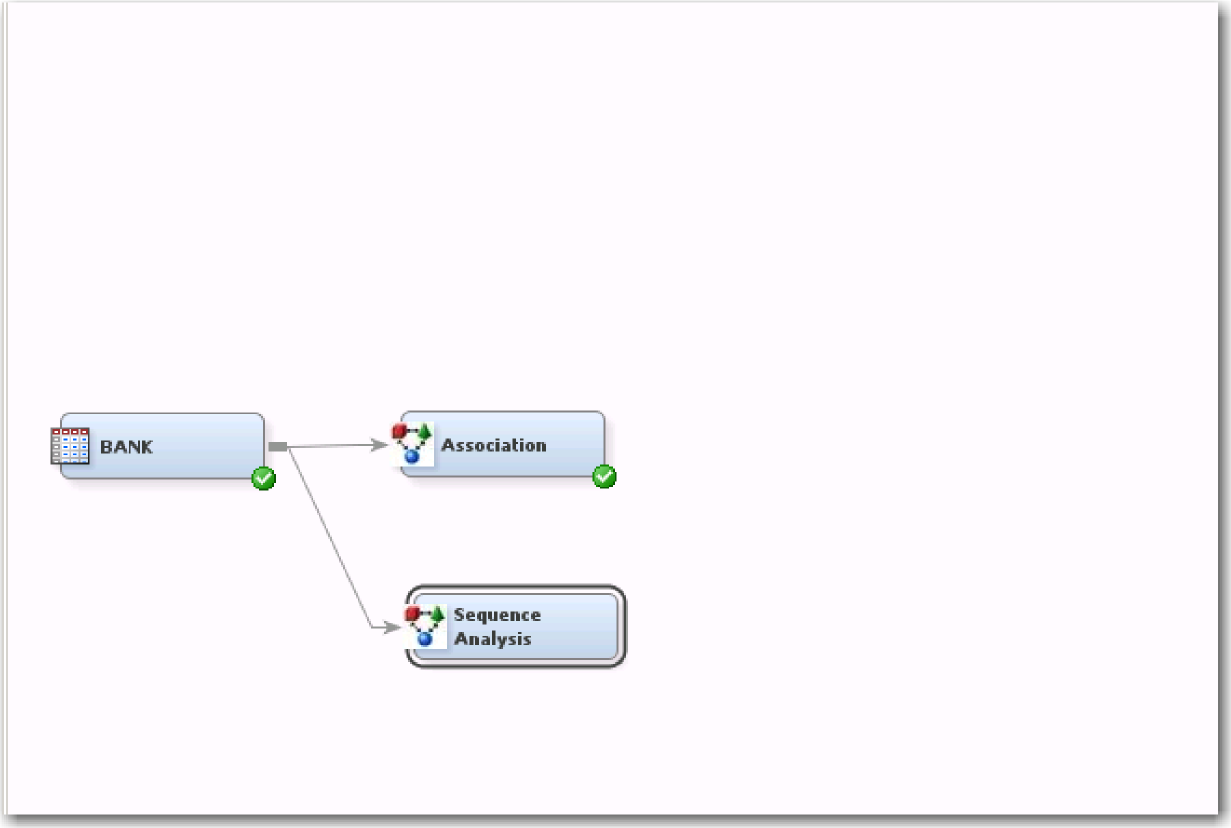
The rule matrix plots the rules based on the items on the left side of the rule and the items on the right side of the rule. The points are colored, based on the confidence of the rules. For example, the rules with the highest confidence are in the column in the picture above. Using the interactive feature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

21. Close the Results window.

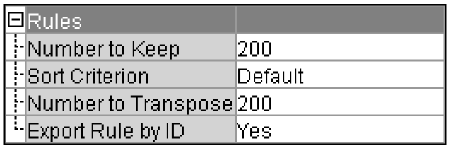
Sequence Analysis

In addition to the products owned by its customers, the bank is interested in examining the order in which the products are purchased to help with a best-next-offer (up-sell) campaign. 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 the **BANK** data source node to it.
2. Rename the new node Sequence Analysis. All variables now have a **Use** value of **Yes**.

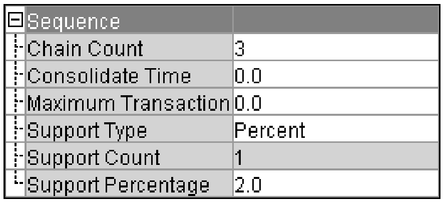


3. Set the Export Rule by ID property to **Yes**.



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

4. Examine the Sequence section in the Properties panel.



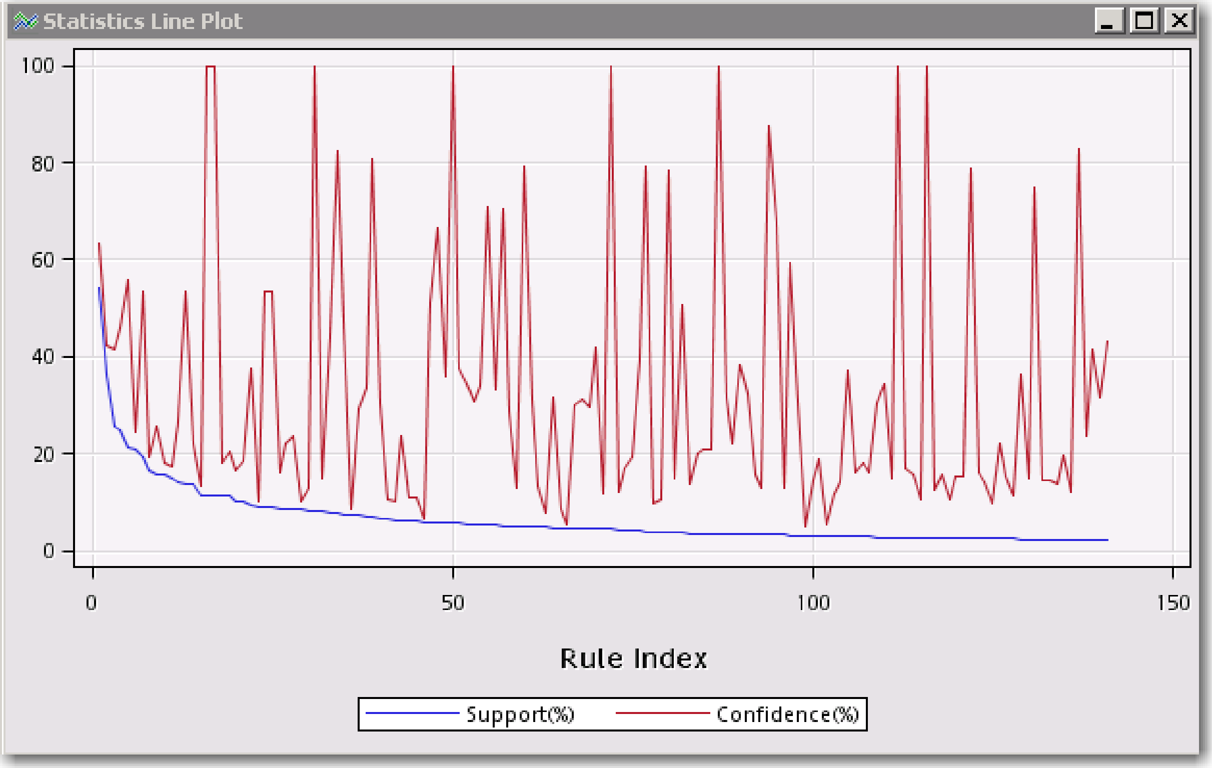
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction	0.0
Support Type	Percent
Support Count	1
Support Percentage	2.0

The options in the Sequence section 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. The default setting is 1.
- ***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.

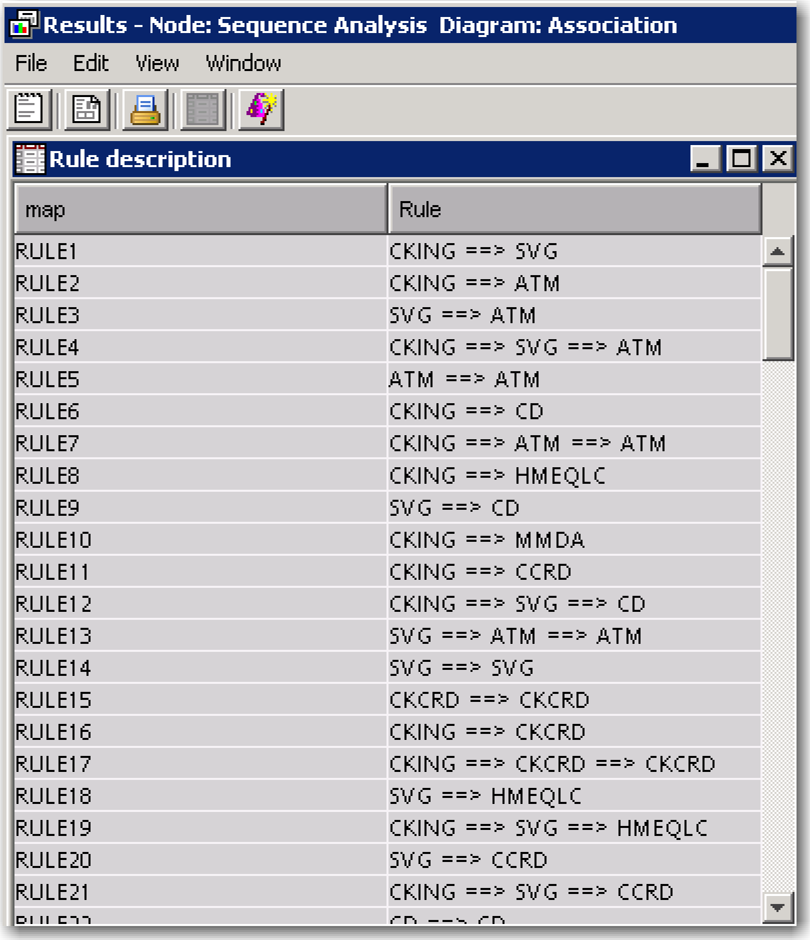
6. Maximize the Statistics Line Plot window.



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.



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. For example, from the rule description above, if a customer already has checking and savings, they are likely to get an ATM card next. Perhaps tying the ATM card to an additional offer for cross-sell would be beneficial.