Market Basket Analysis

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

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 | | Interval | Account Number |
| **SERVICE** | | Target | | Nominal | Type of Service |
| **VISIT** | | Sequence | | Interval | 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

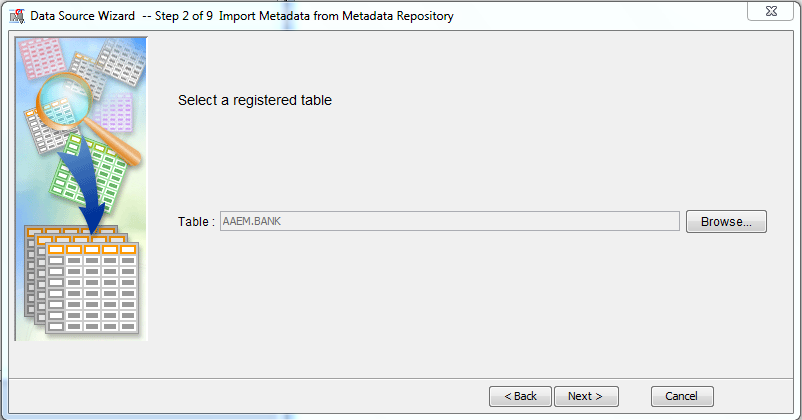
PLOAN personal/consumer installment loan

SVG saving account

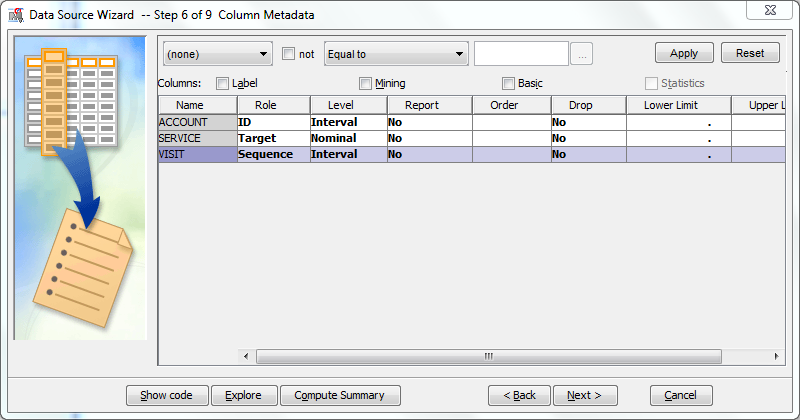
TRUST personal trust account

Your first task is to create a new analysis diagram and data source for the **BANK** data set. You can use the project you’ve been working with so far.

1. Create a new diagram named **Associations Analysis** to contain this analysis.
2. Select **Create Data Source** by right-clicking on the Data Sources project property.
3. At Step 1, make sure “Metadata Repository” is selected and click “Next.” Proceed to Step 2 of the Data Source Wizard.
4. At Step 2, click “Browse” and select the folder “Shared Data.” Then select the folder “Libraries.” Then select the folder “AAEM.” Finally, select the table called “BANK.” Click “OK.”Select the **BANK** table in the AAEM library.

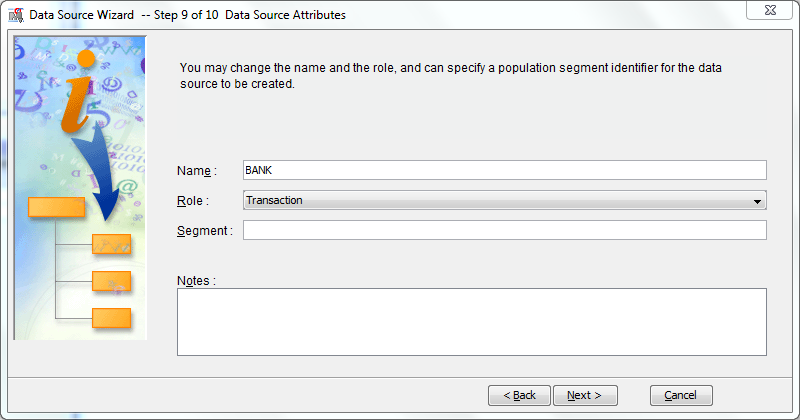


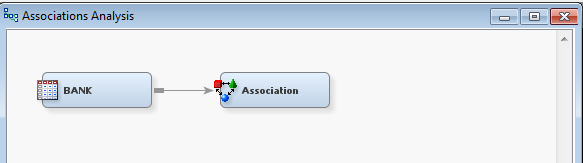
1. Proceed to Step 6 of the Data Source Wizard.
2. In Step 6, assign metadata to the table variables as shown below. **BE CAREFUL! You’ll have to make several changes in your metadata window to make it look like this!**

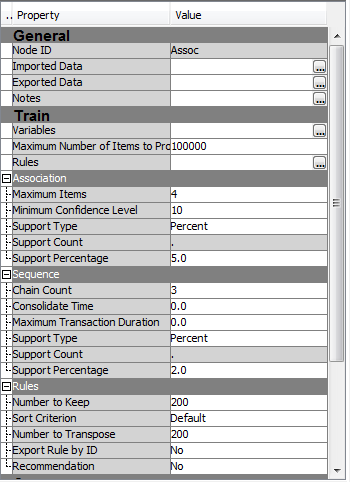


An association analysis requires exactly one target variable and at least one ID variable. ACCOUNT should have an interval measurement level and SERVICE should have a nominal measurement level (unordered values). A sequence analysis also requires a sequence variable. It usually has an ordinal measurement scale (ordered values).

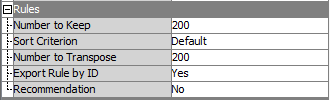
1. Proceed to Step 9 of the Data Source Wizard. For an association analysis, the data source should have a role of Transaction.
2. Select **Role** ⇨ **Transaction**. **DON’T SKIP THIS STEP! IT’S IMPORTANT!**

S

1. Select **Next>** and then **Finish** to close the Data Source Wizard.
2. Drag the **BANK** data source into the diagram workspace.
3. Select the **Explore** tab and drag an **Association** tool (the first icon from the left) into the diagram workspace.
4. Connect the **BANK** node to the **Association** node.  
     
   
5. Select the **Association** node and examine its Properties panel.



1. 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**.



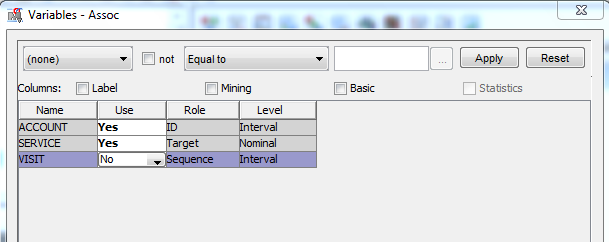
Other options in the Properties panel include the following:

* 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.

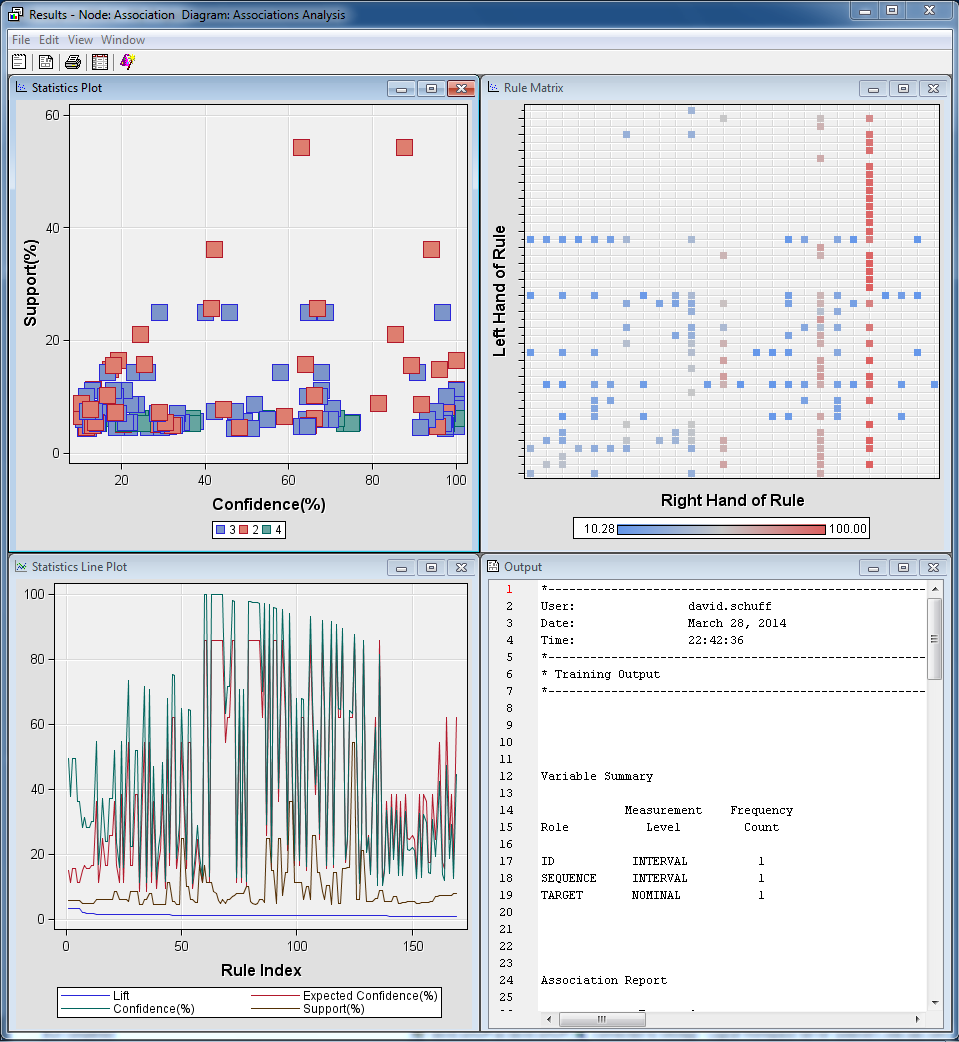
Because you first want to perform a market basket analysis, you do not need the sequence variable.

1. Open the Variables dialog box for the Association node. Right-click on the node and select **Edit Variables.**
2. Select **Use** ⇨ **No** for the **Visit** variable.

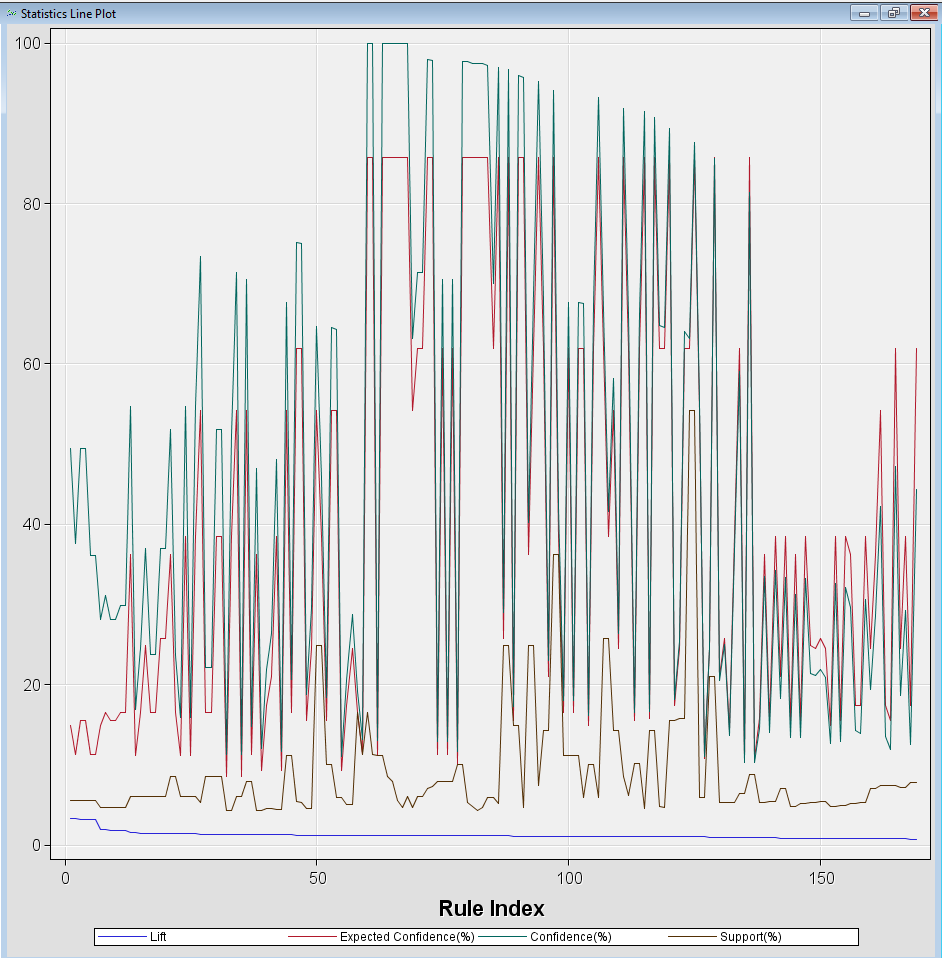


1. Select **OK** to close the Variables dialog box.
2. Run the diagram from the Association node and view the results.

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



1. Maximize the Statistics Line Plot window.



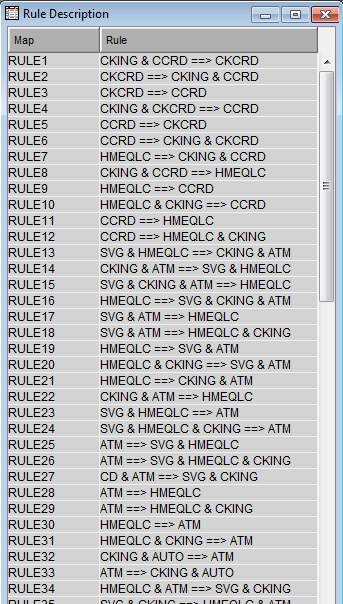
The statistics line plot graphs the lift, expected confidence, confidence, and support for each of the rules by rule index number.

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.

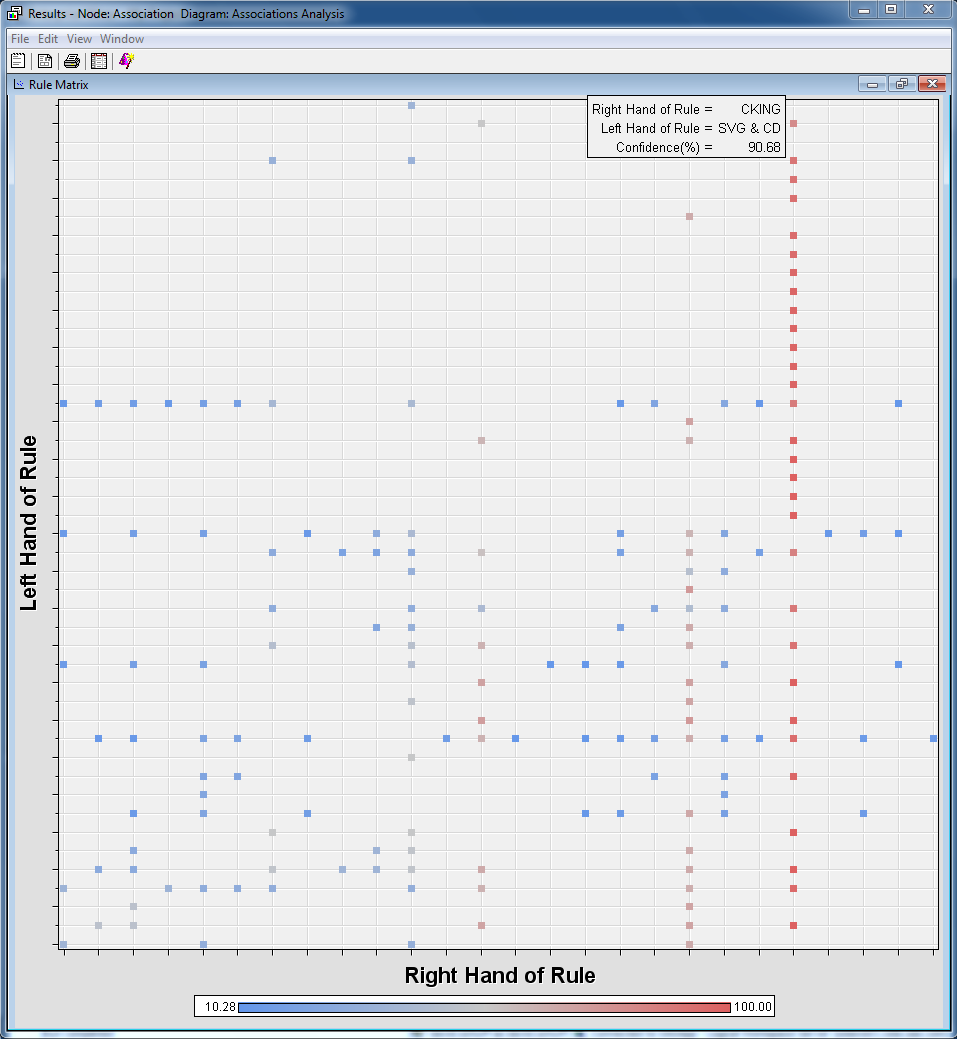
1. To view the descriptions of the rules, select **View** ⇨ **Rules** ⇨ **Rule Description**.



The highest lift rule is checking and credit card implies check card. In other words, those customers that have a checking account and a credit card also have a 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 lift is symmetric (in other words, if B is associated with A, then A is associated with B).

1. Examine the rule matrix.

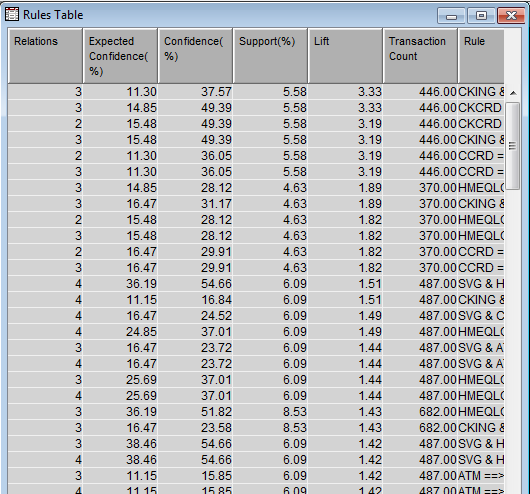
s

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.

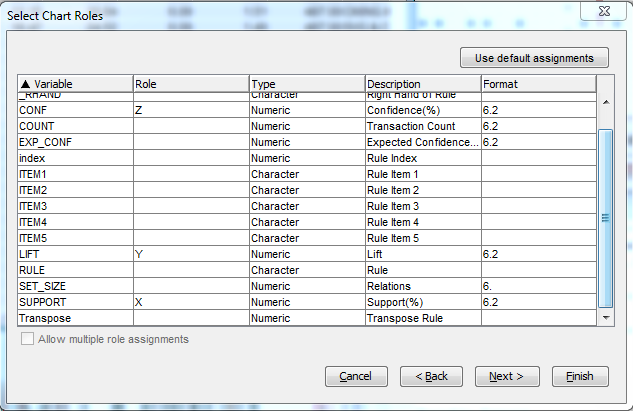
Using the interactive feature of the graph, you can see that these rules all have checking on the right side of the rule. Click on a rule (one of the dots) and then hover your mouse over it. You’ll see the left and right hand side of that rule.

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

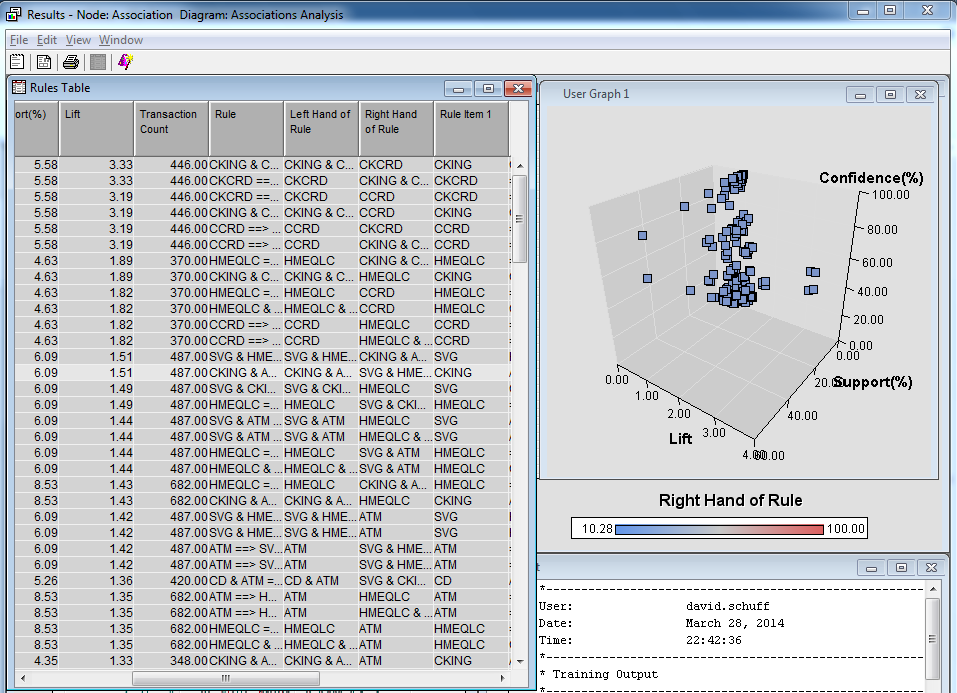
1. Select **View** ⇨ **Rules** ⇨ **Rules Table**. The Rules Table window opens.



1. Select the Plot Wizard icon, .
2. Choose a three-dimensional scatter plot for the type of chart, and select **Next >**.
3. Select the roles **X**, **Y**, and **Z** for the variables **SUPPORT**, **LIFT**, and **CONF**, respectively.



1. Select **Finish** to generate the plot.
2. 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 also drag your mouse to draw a box around those high-lift rules on the right side of the graph. You’ll see the corresponding rules become highlighted in the rules table.

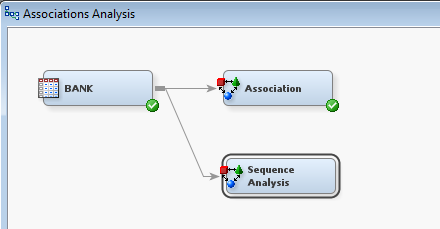
You can use WHERE clauses in the Data Options dialog box to subset cases in which you are interested.

1. Close the Results window.

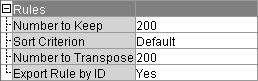
Sequence Analysis

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.

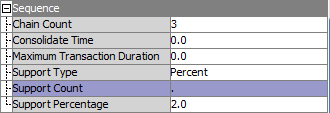
1. Add an Association node to the diagram workspace and connect it to the **BANK** node.
2. Rename the new node **Sequence Analysis**. Note that this time you are NOT excluding the Visit variable. You’re now including the sequence information – the order in which services are purchased – in the analysis



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



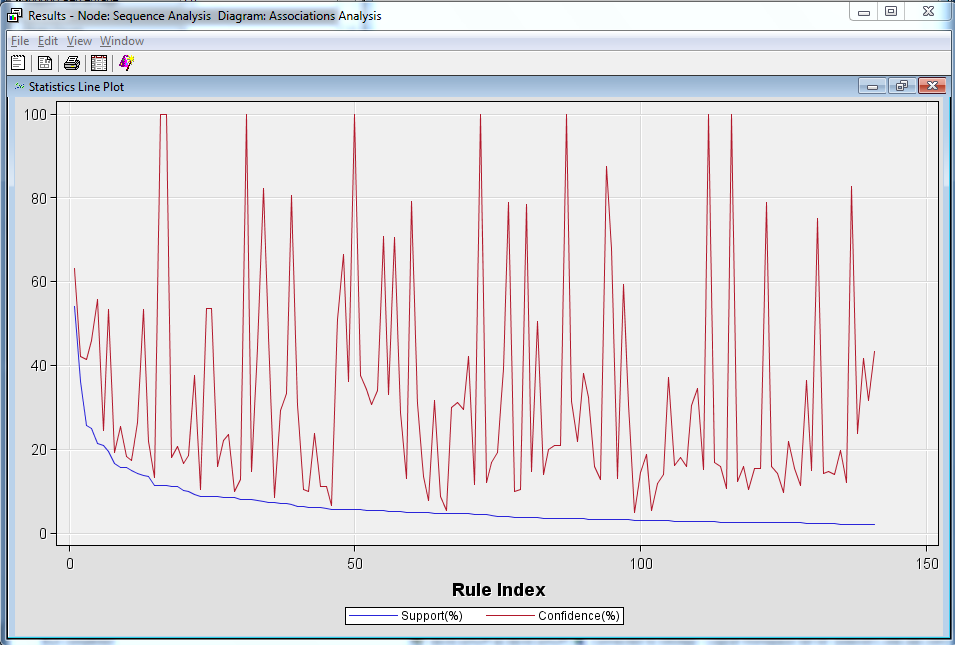
1. Examine the Sequence panel in the Properties panel.



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.
* **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.

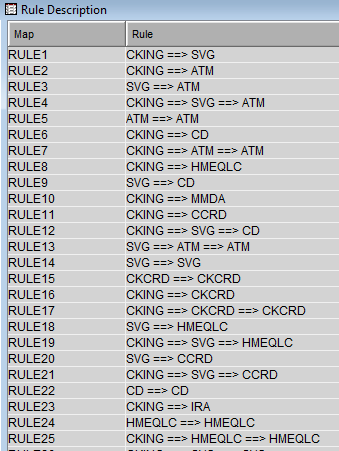
1. Run the diagram from the Sequence Analysis node and view the results.
2. Maximize the Statistics Line Plot window.



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.

1. Select **View** ⇨ **Rules** ⇨ **Rule description** to view the descriptions of the rules. You may have to expand the Rule column see the full rule. You do this by clicking on the right side of the column heading and dragging it a little to the right.



The confidence for many of the rules changes after the order of service acquisition is considered. For example, notice that this now says that customers who have a checking account first are likely to open a savings account (RULE1). However, customers who already have a savings account are likely to use an ATM card (RULE3).

You can see that doing a sequence analysis is more powerful than a regular association rule analysis. We’re not just showing what things happen together, but also *why* they happen.