

## INFS 5102 – Unsupervised Methods in Analytics

### Practical #8: Association analysis

#### Objective:

- Learn how to do association analysis using SAS Enterprise Miner.

#### Submission:

- What to submit:  
A PDF report generated by the Reporter node/tool of SAS EM, containing the information (diagram, results etc.) about the exercise done by you in this practical.
- Deadline of the submission: 11:59PM (Adelaide Time), Tuesday of **Week 12**.
- Submission link: “**Submission Link of Prac #9**” in **Week 11 section** on Learnonline course site.
- Marks: Prac#9 (part of the ongoing assessment of the course) is worth 2% of the total marks of the course.

#### Instructions:

**This practice will help you learn how to use the SAS EM Association tool by going through the steps given in the following steps.**

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. These requirements suggest a 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

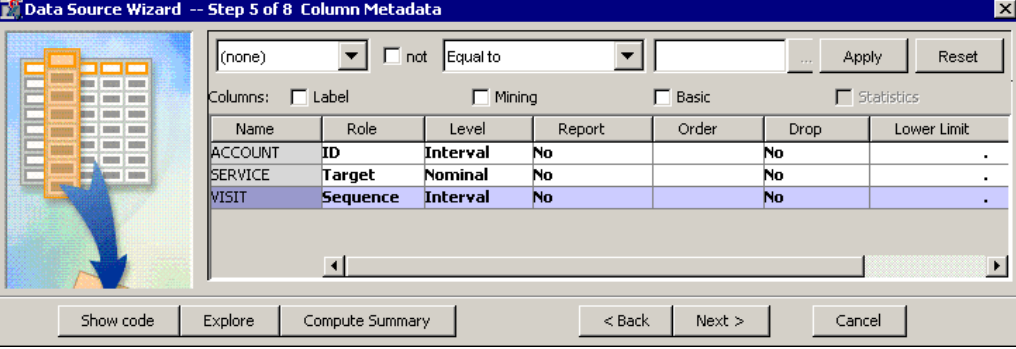
The **BANK** data set has over 32,000 rows. Each row of the data set represents a customer-service combination. Therefore, a single customer can have multiple rows in the data set, and each row represents one of the products he or she owns. The median number of products per customer is three.

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 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. (To go to the AAEM library, select Metadata Repository when the Data Source Wizard is opened, then Browse and select the Shared Data folder, then the Library folder)
4. Assign roles to the table variables as shown below.



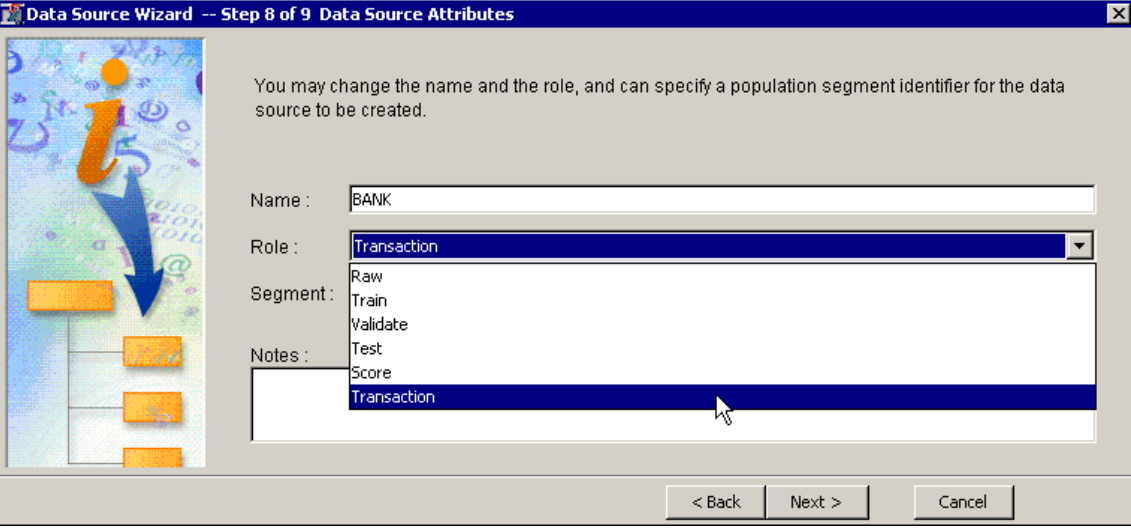
The screenshot shows the 'Data Source Wizard -- Step 5 of 8 Column Metadata' dialog. It features a table with columns: Name, Role, Level, Report, Order, Drop, and Lower Limit. The table contains three rows: ACCOUNT (ID, Interval, No, No, No, .), SERVICE (Target, Nominal, No, No, No, .), and VISIT (Sequence, Interval, No, No, No, .). The 'VISIT' row is highlighted. Above the table, there are checkboxes for 'Label', 'Mining', 'Basic', and 'Statistics'. Below the table, there are buttons for 'Show code', 'Explore', 'Compute Summary', '< Back', 'Next >', and 'Cancel'.

Name	Role	Level	Report	Order	Drop	Lower Limit
ACCOUNT	ID	Interval	No		No	.
SERVICE	Target	Nominal	No		No	.
VISIT	Sequence	Interval	No		No	.

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**.



The screenshot shows the 'Data Source Wizard -- Step 8 of 9 Data Source Attributes' dialog. It contains fields for 'Name' (BANK), 'Role' (Transaction), 'Segment' (Raw, Train, Validate, Test, Score, Transaction), and 'Notes'. The 'Transaction' role is selected in the dropdown menu. Below the 'Notes' field, there are buttons for '< Back', 'Next >', and 'Cancel'.

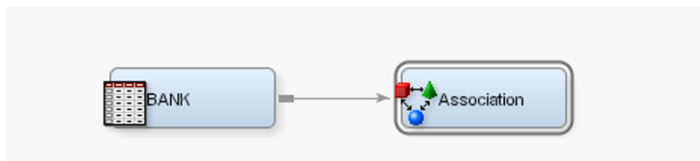
Name : BANK

Role : Transaction

Segment : Raw, Train, Validate, Test, Score, Transaction

Notes :

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
<b>General</b>	
Node ID	Assoc
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
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

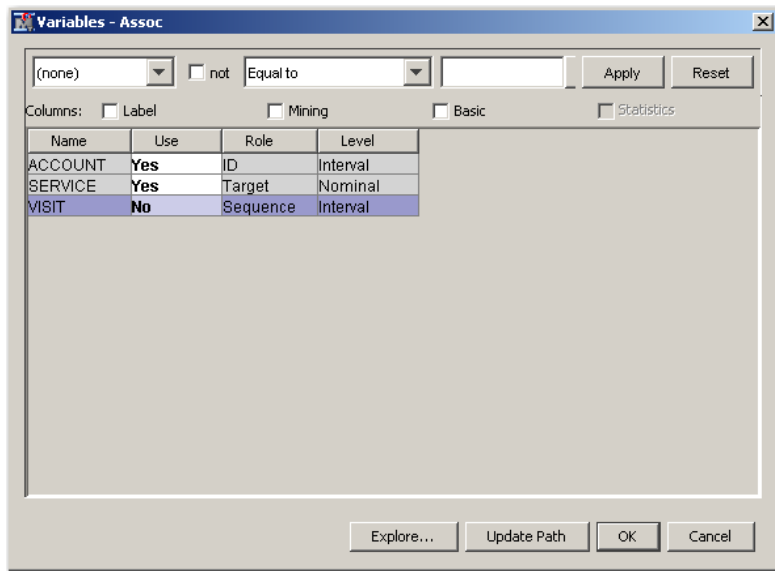
Other options in the Properties panel include the following:

- **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.



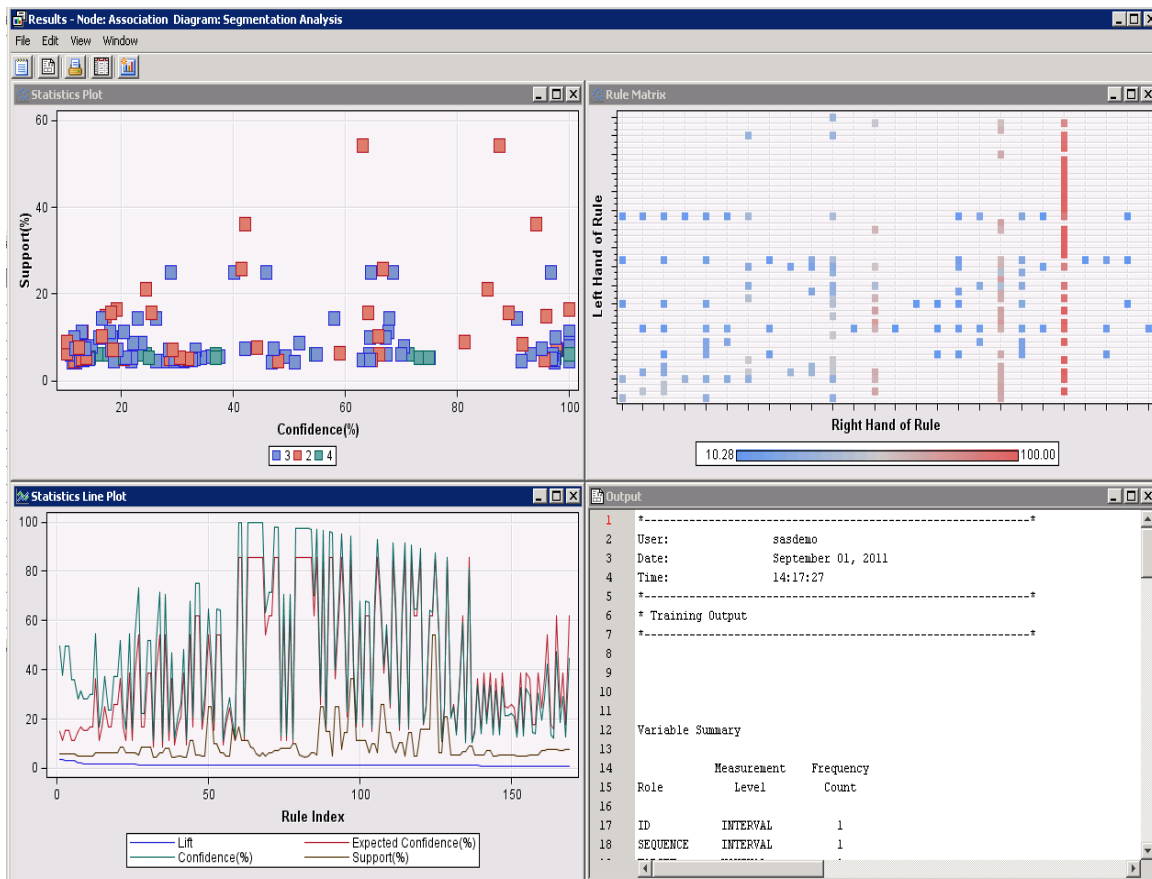
If you are interested in associations that involve fairly rare products, you should consider reducing the support count or percentage when you run the Association node. If you obtain too many rules to be practically useful, you should consider raising the minimum support count or percentage as one possible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

12. Access the Variables dialog box for the Association node.
13. Select **Use**  $\Rightarrow$  **No** for the **VISIT** variable. Because you want to perform a market basket analysis, you do not need the sequence variable.

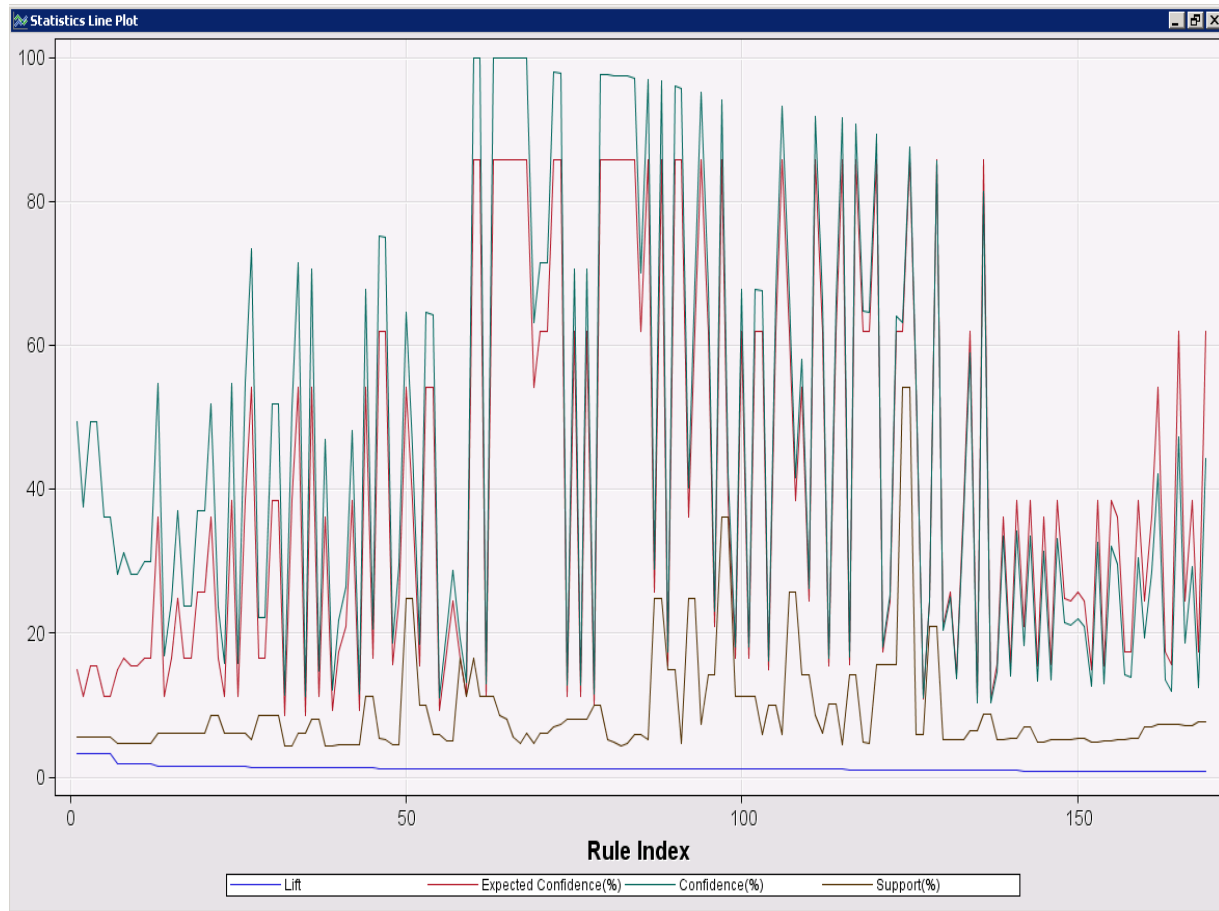


14. Select **OK** to close the Variables dialog box.
15. Run the diagram from the Association node and view the results.

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



16. Maximize the Statistics Line Plot window.



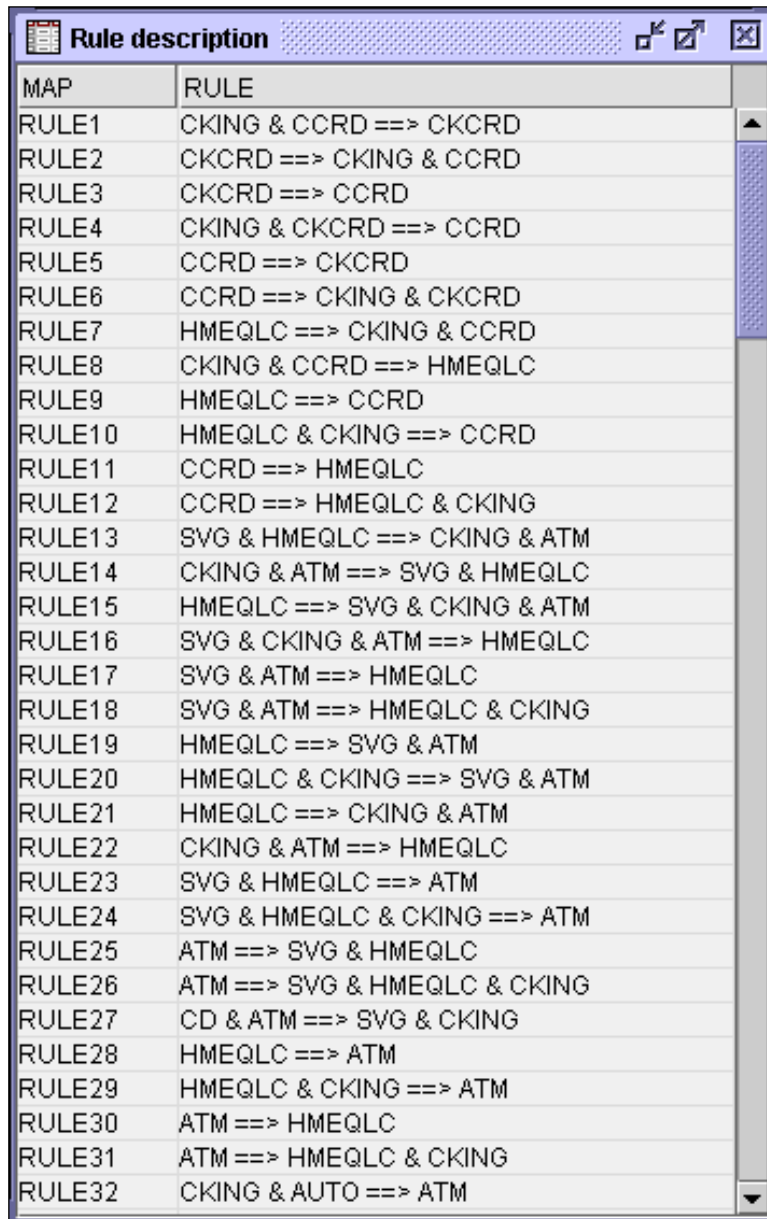
The statistics line plot graphs the lift, expected confidence, confidence, and support for each of the rules by rule index number.

Consider the rule  $A \Rightarrow B$ . 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.

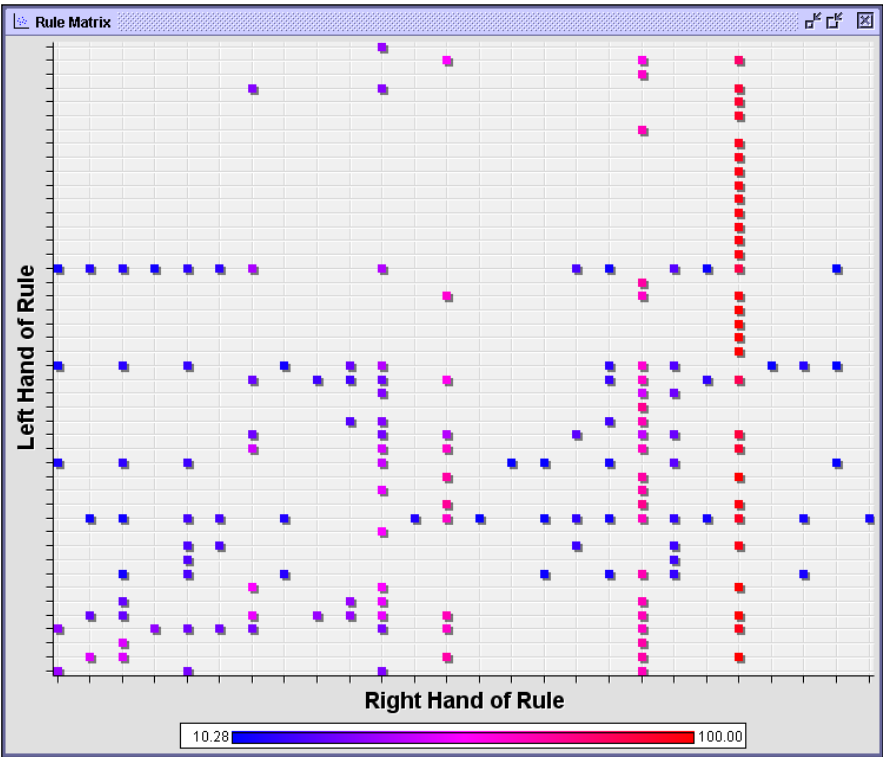
17. To view the descriptions of the rules, select **View** ⇒ **Rules** ⇒ **Rule description**.



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.

18. 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.

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**.

**Select Matrix Variable Roles**

Available Variables		Matrix Variable	
Name ▲	Description	Name	Description
COUNT	Transaction Count	LIFT	Lift
EXP_CONF	Expected Confidence(%)	CONF	Confidence(%)
index	Rule Index	SUPPORT	Support(%)
SET_SIZE	Relations		
Transpose	Transpose Rule		

Navigation buttons:    

Buttons: **Cancel** **< Back** **Next >** **Finish**

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

**Select Chart Roles**

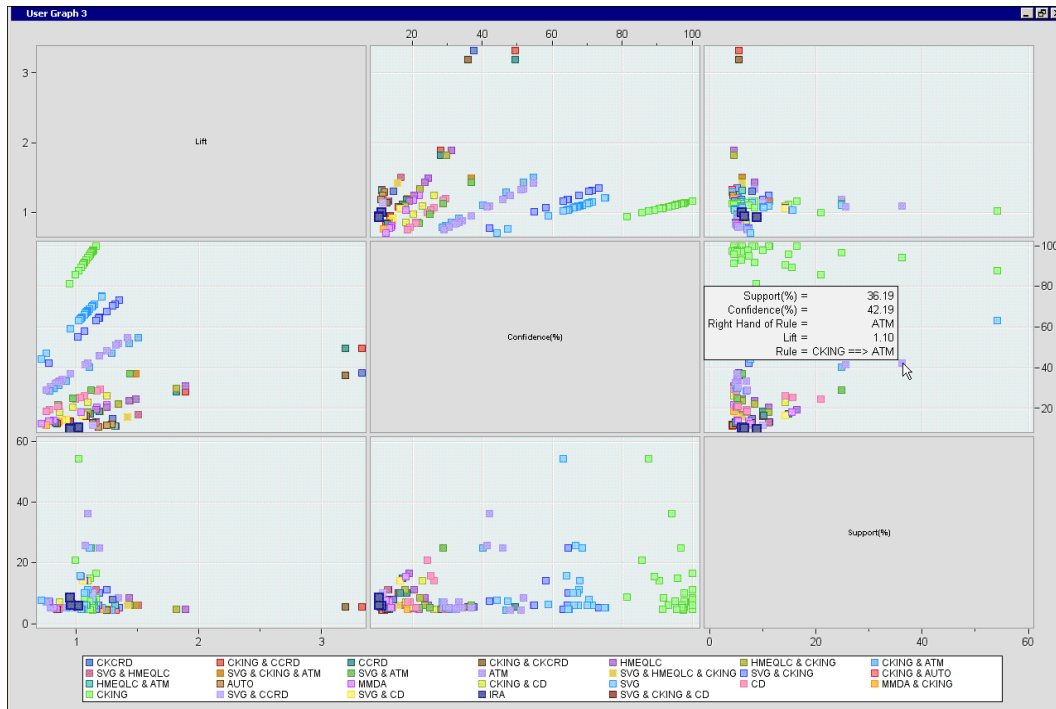
▲ Variable	Role	Type	Description	Format
_LHAND		Character	Left Hand of Rule	
_RHAND	Group	Character	Right Hand of Rule	
CONF		Numeric	Confidence(%)	F6.2
COUNT		Numeric	Transaction Count	F6.2
EXP_CONF		Numeric	Expected Confidence...	F6.2
index		Numeric	Rule Index	
ITEM1		Character	Rule Item 1	
ITEM2		Character	Rule Item 2	
ITEM3		Character	Rule Item 3	
ITEM4		Character	Rule Item 4	
ITEM5		Character	Rule Item 5	
LIFT	Tip	Numeric	Lift	F6.2
RULE	Tip	Character	Rule	
SET_SIZE		Numeric	Relations	F6
SUPPORT		Numeric	Support(%)	F6.2

☒ Allow multiple role assignments

Buttons: **Cancel** **< Back** **Next >** **Finish**



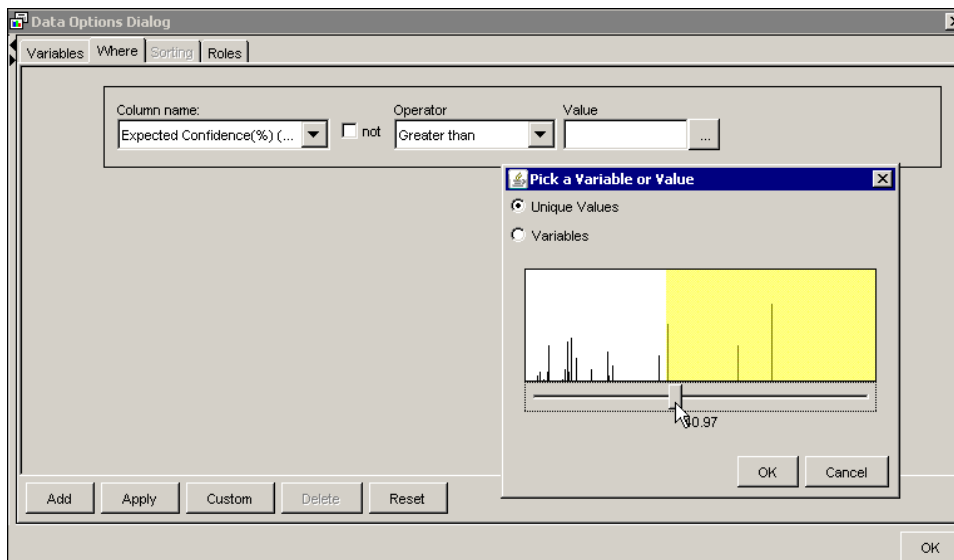
24. Select **Finish** to generate the plot.



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.

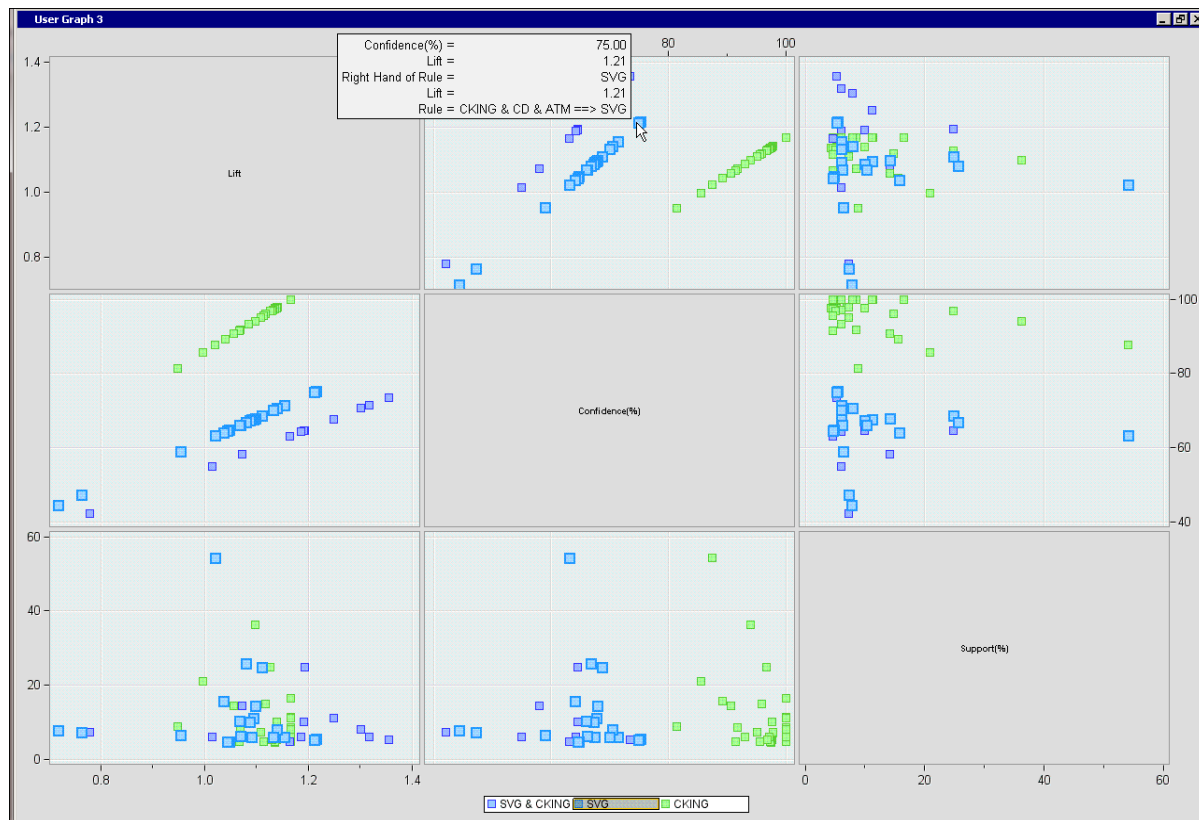
When you hover the cursor over a selected point in the plot, the tooltips show the details of the point, including the full rule.

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 ellipses next to **Value**. Set the slider to include values greater than 40, or type **40** for the value.

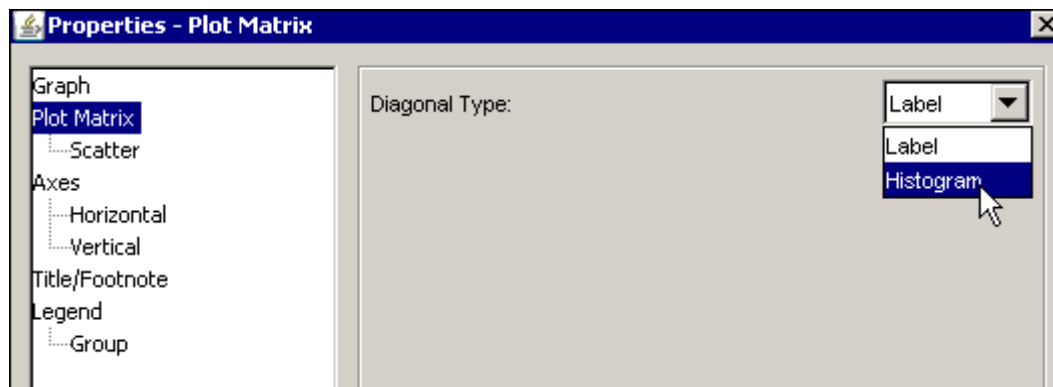


26. Select **OK** and **Apply**.

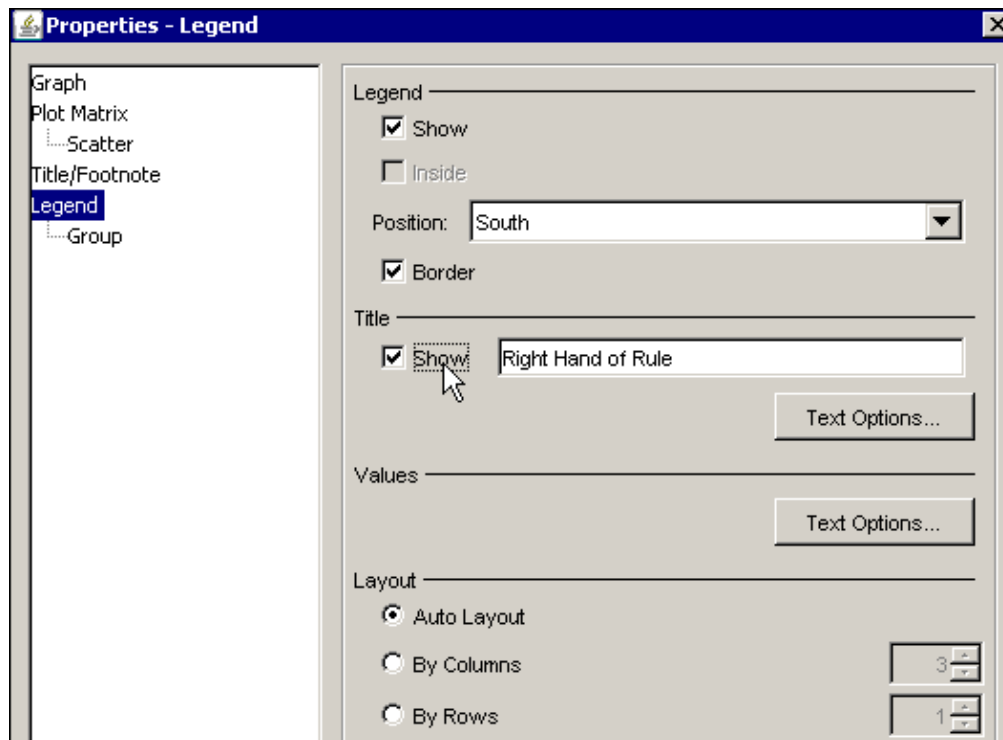
27. Select **OK** . The subset selected cases represent three different sets of services in the legend for the right hand of the rules.



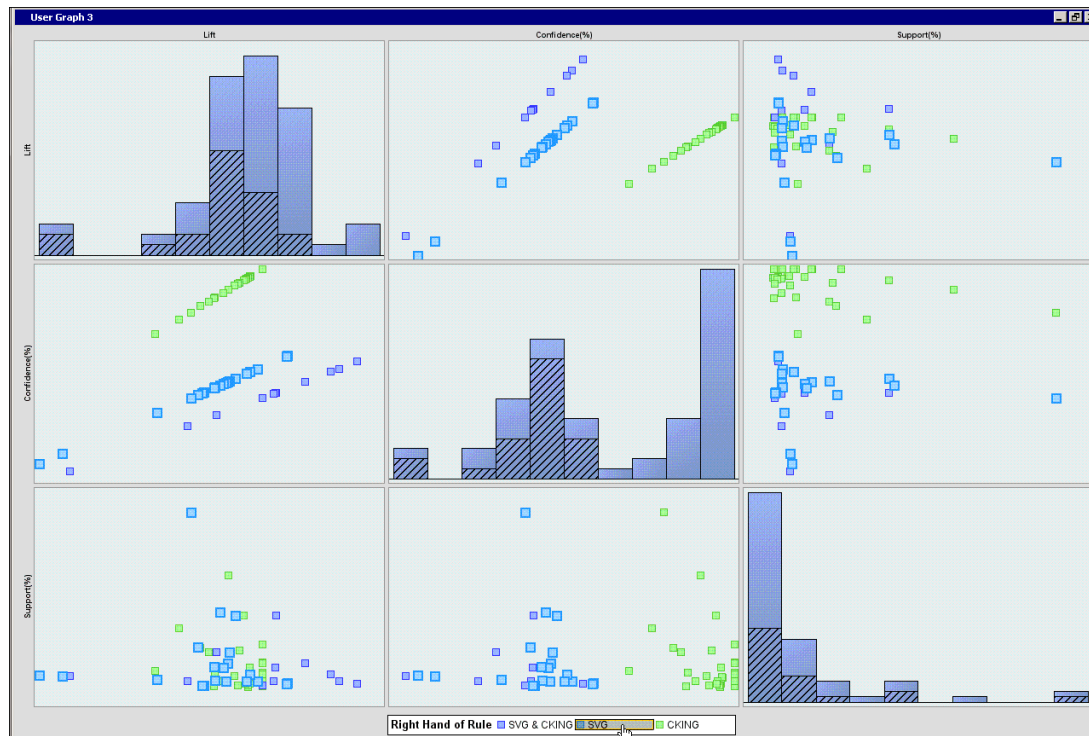
28. You can modify the look of the graph. Right-click the graph and select **Graph Properties**. Change the plot matrix diagonal type to **Histogram**.



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.

**Save the PDF report, and include the report as part of your submission by following the instruction given on page 1 of the practical document.**