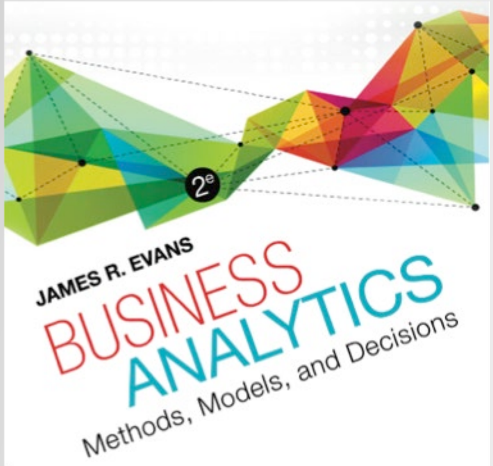


Chapter 10 Introduction to Data Mining

Data Mining

- **Data mining** is focused on better understanding of characteristics and patterns among variables in large databases using a variety of statistical and analytical tools.
  - It is used to identify relationships among variables in large data sets and understand hidden patterns that they may contain.
  - *XLMiner* software implement many basic data mining procedures in a spreadsheet environment.

The Scope of Data Mining

- *Data Exploration and Reduction*
  - identifying groups in which elements are in some way similar
- *Classification*
  - analyzing data to predict how to classify a new data element
- *Association*
  - analyzing databases to identify natural associations among variables and create rules for target marketing or buying recommendations
- *Cause-and-effect Modeling*
  - developing analytic models to describe relationships between metrics that drive business performance

Data Exploration in *XLMiner*

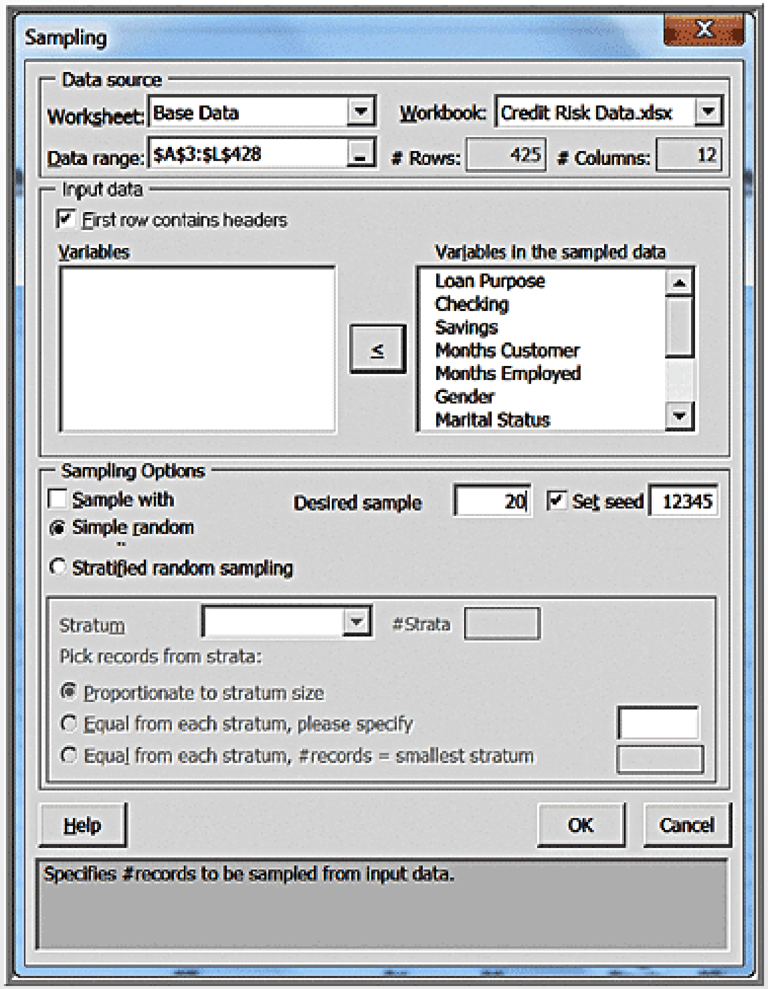
- *XLMiner* ribbon

  - *XLMiner* can sample from an Excel worksheet

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Credit Risk Data**											
2												
3	**Loan Purpose**	**Checking**	**Savings**	**Months Customer**	**Months Employed**	**Gender**	**Marital Status**	**Age**	**Housing**	**Years**	**Job**	**Credit Risk**
4	Small Appliance	$0	$739	13	12	M	Single	23	Own	3	Unskilled	Low
5	Furniture	$0	$1,230	25	0	M	Divorced	32	Own	1	Skilled	High
6	New Car	$0	$389	19	119	M	Single	38	Own	4	Management	High
7	Furniture	$638	$347	13	14	M	Single	36	Own	2	Unskilled	High
8	Education	$963	$4,754	40	45	M	Single	31	Rent	3	Skilled	Low
9	Furniture	$2,827	$0	11	13	M	Married	25	Own	1	Skilled	Low
10	New Car	$0	$229	13	16	M	Married	26	Own	3	Unskilled	Low
11	Business	$0	$533	14	2	M	Single	27	Own	1	Unskilled	Low

Example 10.1: Using *XLMiner* to Sample from a Worksheet

- Click inside the database
- *XLMiner > Data Analysis > Sample > Sample from Worksheet*
- Select variables and move to right pane
- Choose sampling options



Example 10.1 Continued

- Results

XLMiner : Sampling from Worksheet

Date : 27-Dec-2013 09:09:43 | (Ver: 12.5.3P)

Data	
Data Source	Base Data!$A$4:$L$428
Sampling Method	Simple Random Sampling
Sampling with replacement	FALSE
Random Seed	12345
#records in input data	425
Desired sample size	20

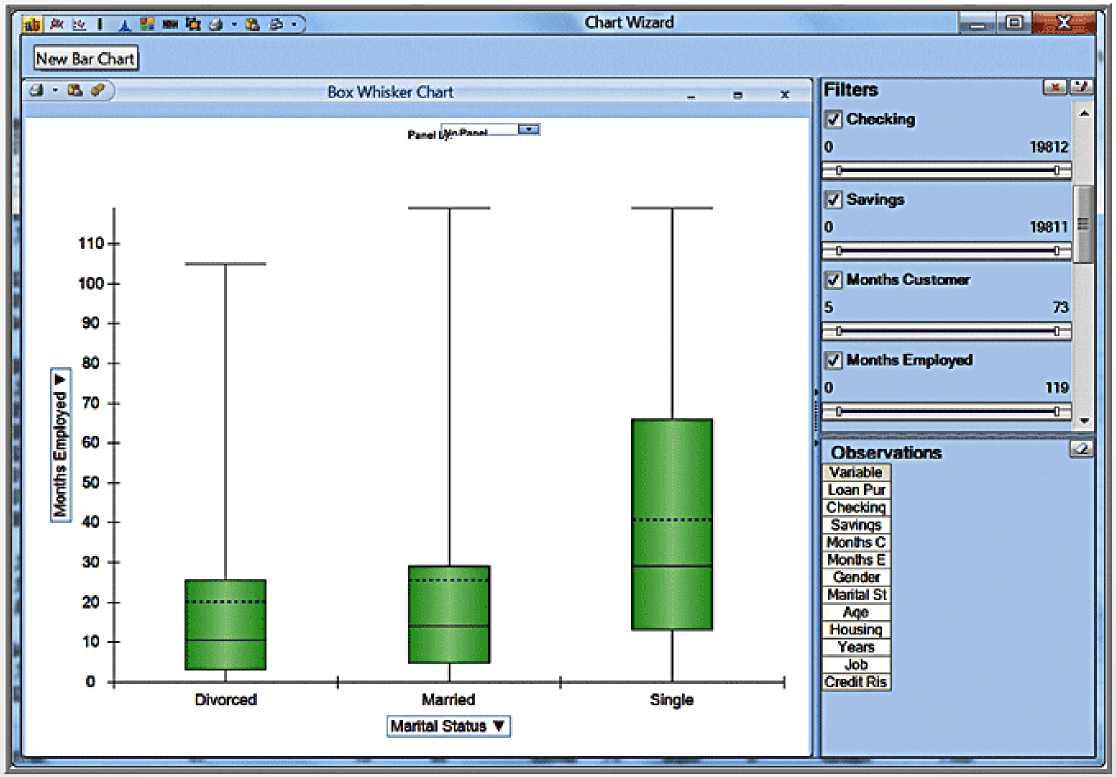
Row Id.	Loan Purpose	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
2	Furniture	$0	$1,230	25	0	M	Divorced	32	Own	1	Skilled	High
25	New Car	$0	$10,723	11	15	M	Single	39	Rent	2	Unskilled	Low
34	Business	$16,647	$895	16	34	M	Single	25	Rent	4	Skilled	Low
40	New Car	$0	$128	13	74	M	Single	34	Own	3	Skilled	High
56	Education	$0	$0	37	114	M	Single	39	Own	4	Management	High
69	Furniture	$513	$442	7	0	M	Single	34	Own	1	Management	Low
79	Repairs	$0	$626	43	0	M	Single	64	Own	4	Unemployed	Low
141	Small Appliance	$0	$325	19	13	F	Divorced	23	Own	2	Skilled	High
145	Business	$0	$265	13	10	F	Divorced	26	Own	2	Skilled	Low
170	Furniture	$113	$692	11	14	M	Divorced	30	Own	2	Unskilled	Low
195	Small Appliance	$586	$0	13	0	M	Single	51	Own	1	Management	High
203	New Car	$5,588	$0	22	10	F	Divorced	28	Own	4	Skilled	High
215	Used Car	$0	$10,099	16	108	M	Single	22	Rent	4	Skilled	Low
246	Furniture	$0	$736	13	6	F	Divorced	19	Rent	4	Skilled	High
261	Business	$0	$500	25	1	M	Single	26	Own	2	Skilled	High
278	New Car	$425	$0	19	7	F	Divorced	32	Own	2	Skilled	High
281	Education	$0	$164	13	65	F	Divorced	56	Other	4	Unskilled	Low
311	New Car	$19,766	$2,141	11	54	F	Divorced	47	Other	4	Unskilled	High
334	New Car	$0	$716	19	33	M	Single	30	Own	2	Skilled	High
367	New Car	$0	$862	49	62	M	Single	41	Other	4	Management	High

Data Visualization

- *XLMiner* has the capability to produce boxplots, parallel coordinate charts, scatterplot matrix charts, and variable charts.
  - These are found from the *Explore* button in the *Data Analysis* group.

Example 10.2: A Boxplot for *Credit Risk Data*

- *XLMiner >Data Analysis > Explore > Chart Wizard > Boxplot*
- In the second dialog, choose *Months Employed* as the variable to plot on the vertical axis.
- In the next dialog, choose *Marital Status* as the variable to plot on the horizontal axis.
- Click *Finish*

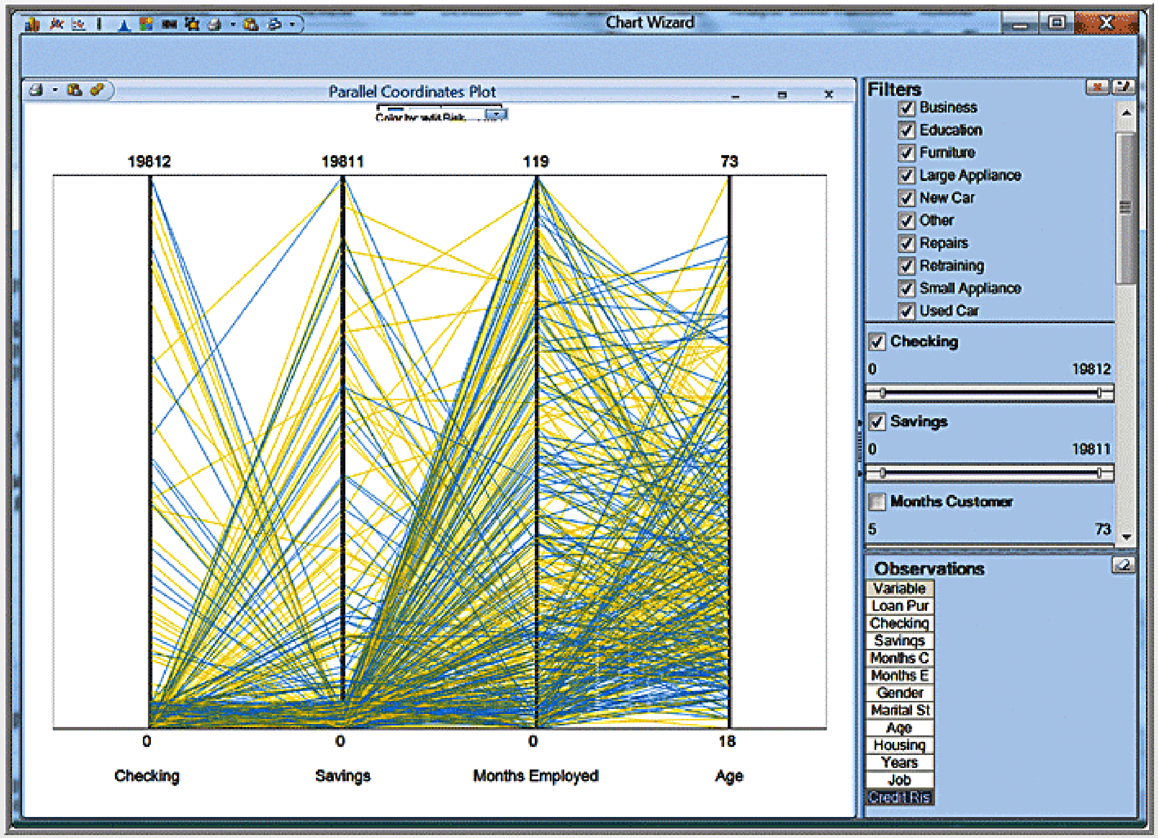


Parallel Coordinates Chart

- A **parallel coordinates chart** consists of a set of vertical axes, one for each variable selected. For each observation, a line is drawn connecting the vertical axes. The point at which the line crosses an axis represents the value for that variable.
- A parallel coordinates chart creates a "multivariate profile," and help an analyst to explore the data and draw basic conclusions.

Example 10.3: A Parallel Coordinates Chart for *Credit Risk Data*

- *XLMiner > Data Analysis > Explore > Chart Wizard > Parallel Coordinates*
- In the second dialog, choose *Checking, Savings, Months Employed*, and *Age* as the variables to include.



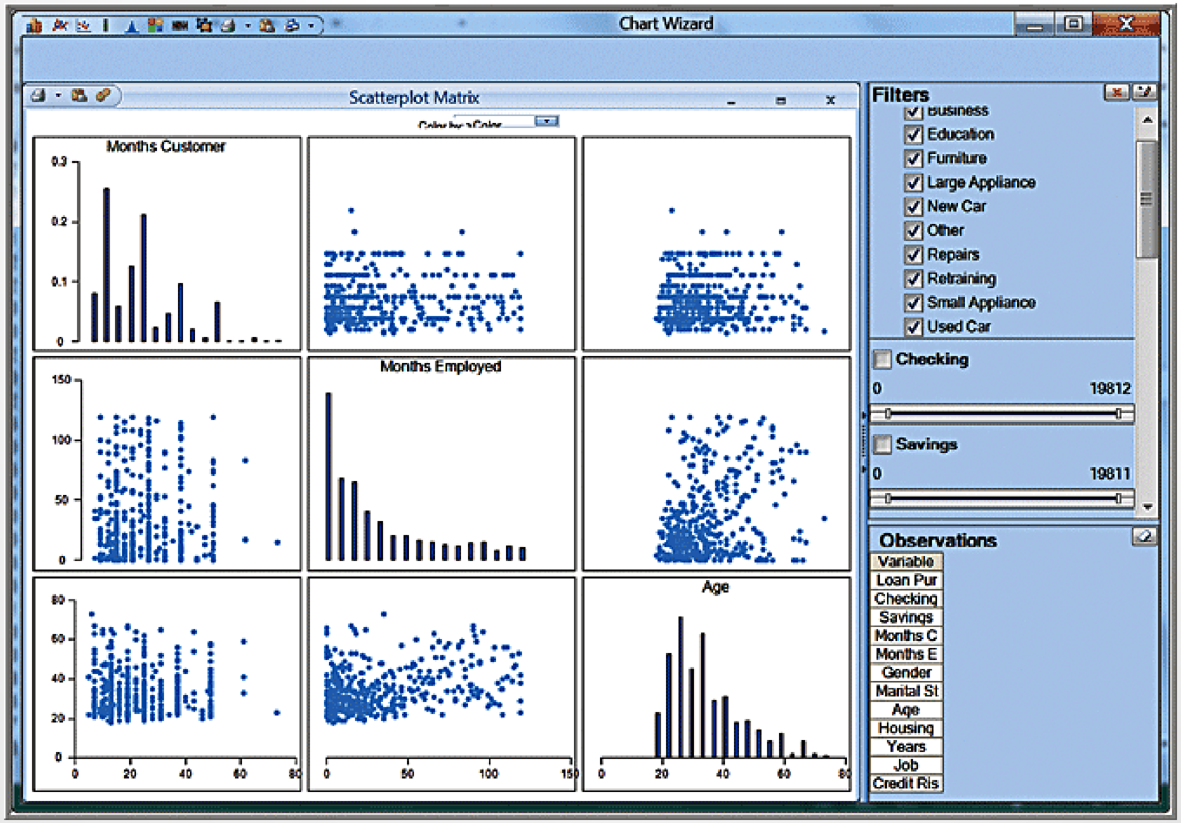
Yellow = low credit risk; blue = high

Scatterplot Matrix

- A **scatterplot matrix** combines several scatter charts into one panel, allowing the user to visualize pairwise relationships between variables.

Example 10.4: A Scatterplot Matrix for *Credit Risk Data*

- *XLMiner > Data Analysis > Explore > Chart Wizard > Scatterplot Matrix*
- In the next dialog, check the boxes for *Months Customer, Months Employed*, and *Age* and click *Finish*.

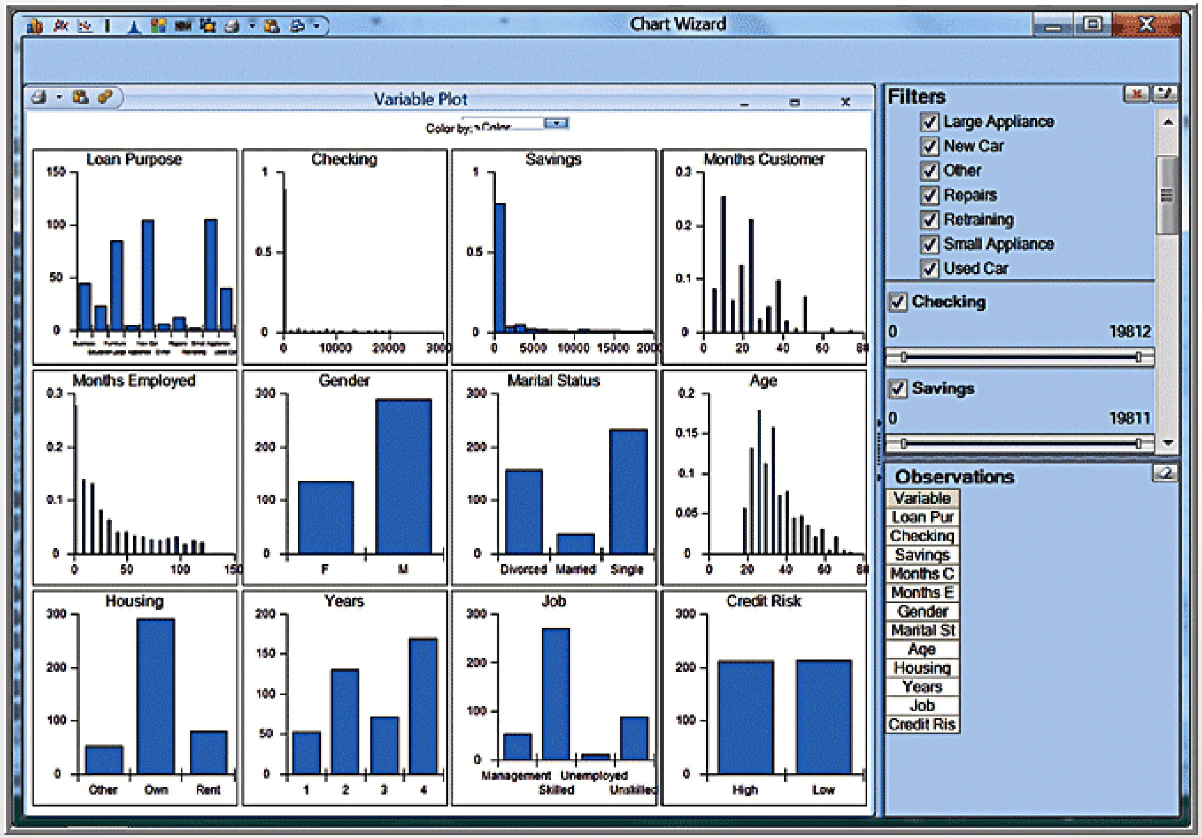


Variable Plot

- A **variable plot** plots a matrix of histograms for the variables selected.

Example 10.5: A Variable Plot of *Credit Risk Data*

- *XLMiner > Data Analysis > Chart Explore > Chart Wizard > Variable Plot*
- In the next dialog, check the boxes for the variables you wish to include and click *Finish*.



Dirty Data

- Real data sets that have missing values or errors. Such data sets are called "dirty" and need to be "cleaned" prior to analyzing them.
- Approaches for handling missing data.
  - Eliminate the records that contain missing data
  - Estimate reasonable values for missing observations, such as the mean or median value
  - Use a data mining procedure to deal with them. *XLMiner* has the capability to deal with missing data in the *Transform* menu in the *Data Analysis* group.
- Try to understand whether missing data are simply random events or if there is a logical reason. Eliminating sample data indiscriminately could result in misleading information and conclusions about the data.

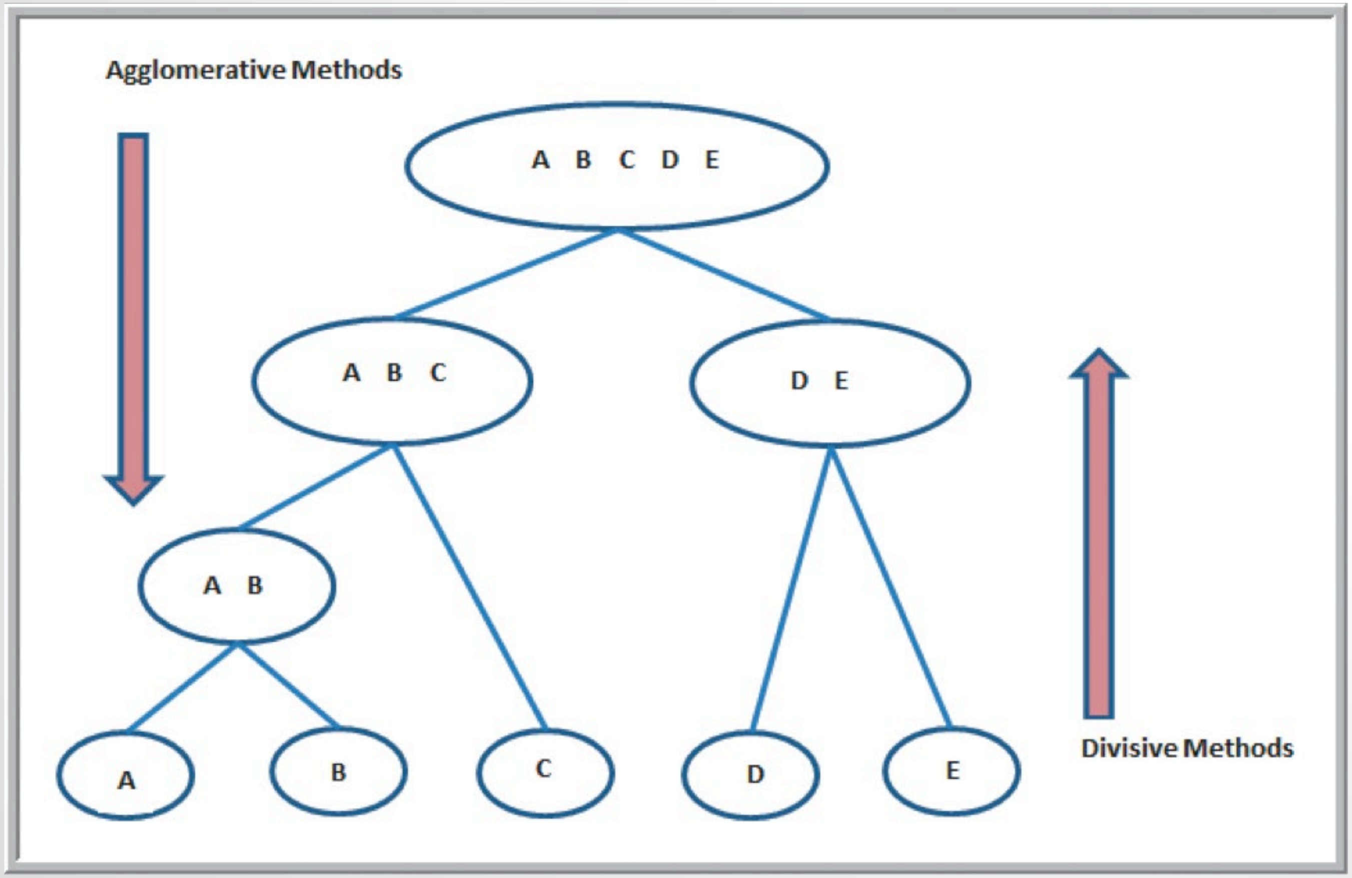
Cluster Analysis

- **Cluster analysis**, also called *data segmentation*, is a collection of techniques that seek to group or segment a collection of objects (observations or records) into subsets or clusters, such that those within each cluster are more closely related to one another than objects assigned to different clusters.
  - The objects within clusters should exhibit a high amount of similarity, whereas those in different clusters will be dissimilar.

Cluster Analysis Methods

- In **hierarchical clustering**, the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to *n* clusters, each containing a single object.
  - **Agglomerative clustering** methods proceed by series of fusions of the *n* objects into groups.
  - **Divisive clustering** methods separate *n* objects successively into finer groupings.
- Hierarchical clustering may be represented by a two-dimensional diagram known as a **dendrogram**, which illustrates the fusions or divisions made at each successive stage of analysis.

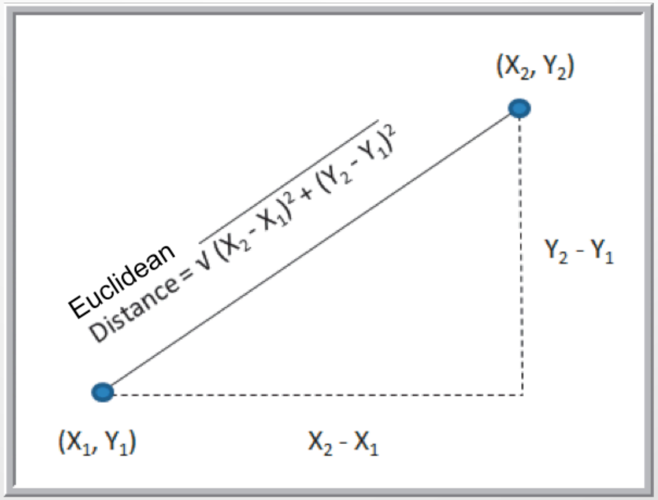
Agglomerative vs. Divisive Clustering



Distance Measures

- **Euclidean distance** is the straight-line distance between two points
- The Euclidean distance measure between two points $(x_1, x_2, . . . , x_n)$ and $(y_1, y_2, . . . , y_n)$ is

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad \textbf{(10.1)}$$



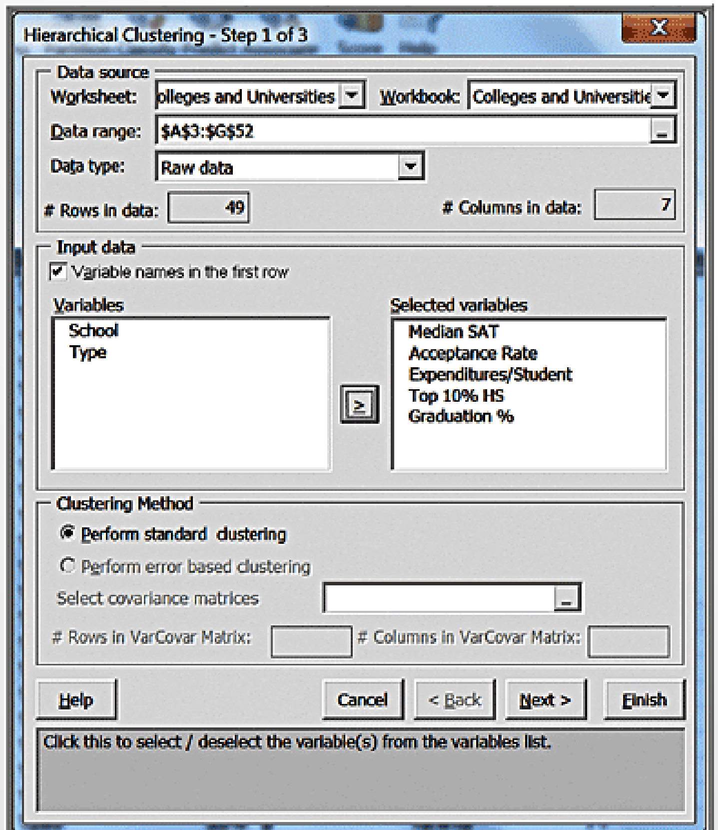
Agglomerative Clustering Methods

- Single linkage clustering (*nearest-neighbor*)
  - The distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered.
  - At each stage, the closest 2 clusters are merged
- Complete linkage clustering
  - The distance between groups is the distance between the most distant pair of objects, one from each group
- Average linkage clustering
  - Uses the mean values for each variable to compute distance between clusters
- Ward's hierarchical clustering
  - Uses a sum of squares criterion

Example 10.6: Clustering *Colleges and Universities* Data

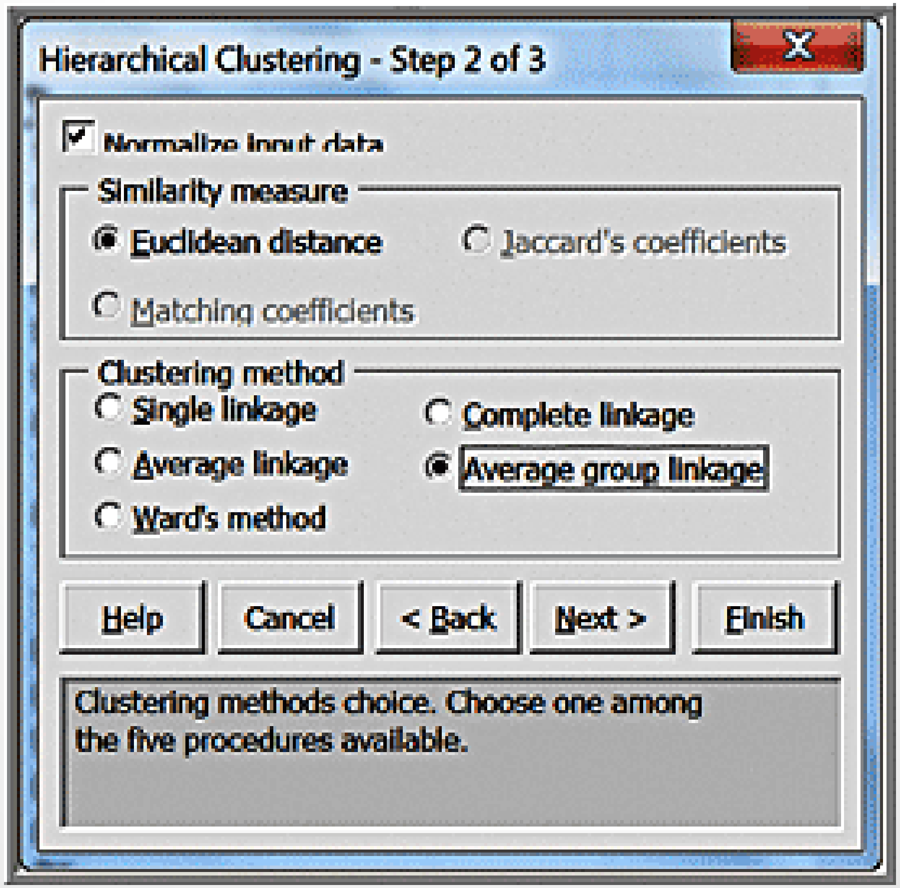
- Cluster the institutions using the five numeric columns in the data set.
- *XLMiner > Data Analysis > Cluster < Hierarchical Clustering*

	A	B	C	D	E	F	G
1	**Colleges and Universities**						
2							
3	**School**	**Type**	**Median SAT**	**Acceptance Rate**	**Expenditures/Student**	**Top 10% HS**	**Graduation %**
4	Amherst	Lib Arts	1315	22%	$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	$ 17,554	58	88
7	Berkeley	University	1176	37%	$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
9	Brown	University	1281	24%	$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	$ 18,847	70	84



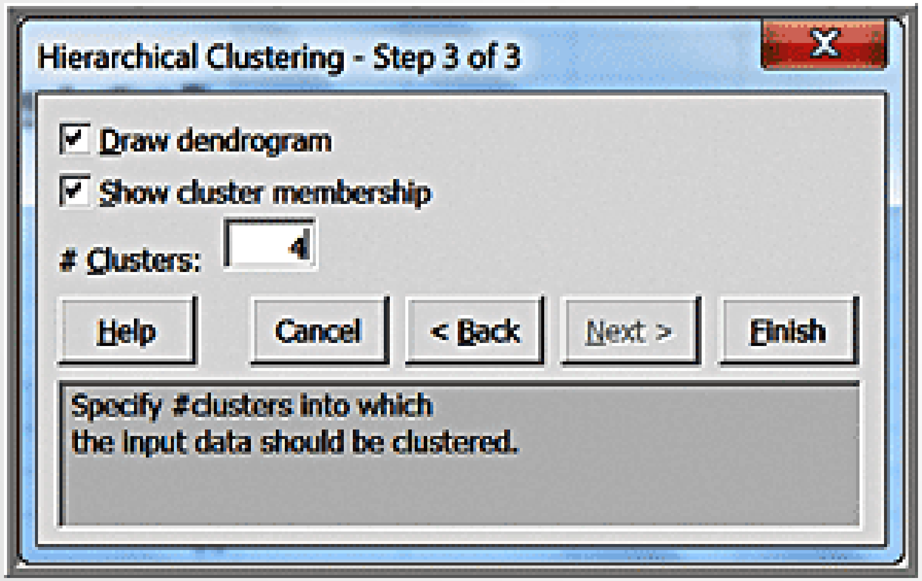
Example 10.6 Continued

- Second dialog
- Check the box *Normalize input data* to ensure that the distance measure accords equal weight to each variable



Example 10.6 Continued

- Step 3
- Select the number of clusters



Example 10.6 Continued

- Results

XLMiner : Hierarchical Clustering

Output Navigator		
Inputs	Clustering Stages	Dendrogram
Elapsed Time	Predicted Clusters	

Inputs

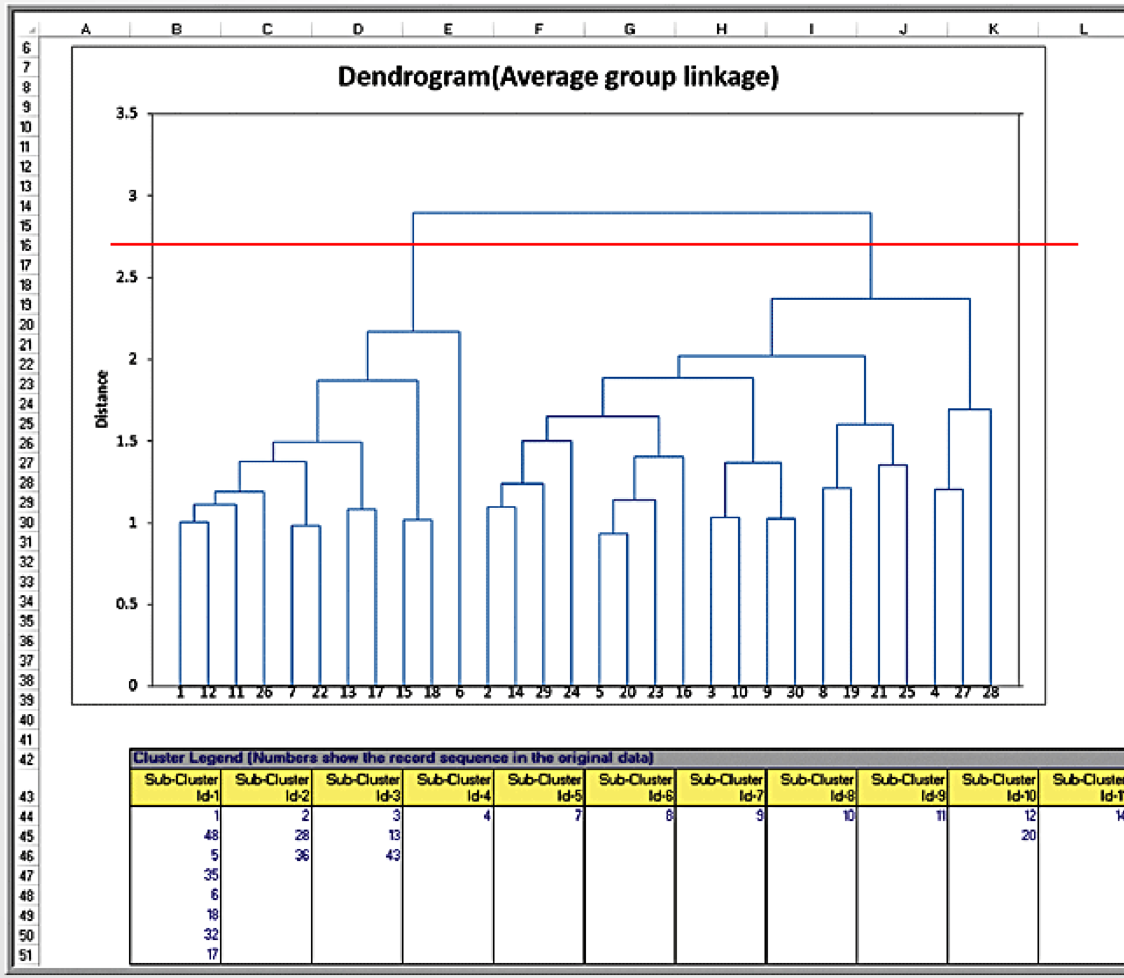
Data	
Input data	[Colleges and Universities.xlsx]Colleges and Universities'!$A$4:$G$52
# Records in the input data	49
Input variables normalized	Yes
Data Type	Raw data

Variables					
# Selected Variables	5				
Selected variables	Median SAT	Acceptance Rate	Expenditures/ Student	Top 10% HS	Graduation %

Parameters/Options	
Draw dendrogram	Yes
Show cluster membership	Yes
# Clusters	4
Selected Similarity measure	Euclidean distance
Selected clustering method	Average group linkage

Example 10.6 Continued

- Dendogram
  - A horizontal line shows the cluster partitions



Example 10.6 Continued

- Predicted clusters
  - shows the assignment of observations to the number of clusters we specified in the input dialog, (in this case four)

Cluster	# Colleges
1	23
2	22
3	3
4	1

XLMiner : Hierarchical Clustering - Predicted clusters

Back to Navigator

Row Id.	Cluster Id	Sub Cluster	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
1	1	1	1315	0.22	$26,636.00	85	93
2	2	2	1220	0.53	$17,653.00	69	80
3	2	3	1240	0.36	$17,554.00	58	88
4	3	4	1176	0.37	$23,665.00	95	68
5	1	1	1300	0.24	$25,703.00	78	90
6	1	1	1281	0.24	$24,201.00	80	90
7	2	5	1255	0.56	$18,847.00	70	84
8	4	6	1400	0.31	$102,262.00	98	75
9	1	7	1300	0.4	$15,904.00	75	80
10	2	8	1225	0.64	$33,607.00	52	77
11	2	9	1260	0.36	$20,377.00	68	74
12	2	10	1200	0.46	$18,872.00	52	84
13	2	3	1258	0.38	$17,520.00	61	85
14	1	11	1268	0.29	$45,879.00	78	90
15	1	12	1290	0.3	$37,137.00	85	83
16	1	13	1230	0.36	$17,721.00	77	89
17	1	1	1310	0.25	$39,504.00	91	91
18	1	1	1278	0.24	$23,115.00	79	89
19	2	14	1244	0.67	$22,301.00	65	73
20	2	10	1215	0.38	$20,722.00	51	85
21	1	15	1370	0.18	$46,918.00	90	90
22	1	13	1285	0.35	$19,418.00	71	87
23	2	16	1290	0.48	$45,460.00	69	86
24	1	17	1255	0.25	$24,718.00	65	92
25	1	18	1357	0.3	$56,766.00	95	86
26	2	19	1200	0.61	$23,358.00	47	83
27	2	20	1230	0.47	$28,851.00	77	82
28	2	2	1247	0.54	$23,591.00	64	77
29	2	21	1170	0.49	$20,192.00	54	72
30	1	22	1320	0.33	$26,668.00	79	80
31	1	15	1340	0.17	$48,123.00	89	93
32	1	1	1327	0.24	$26,730.00	85	88
33	2	23	1195	0.57	$25,271.00	65	87
34	1	15	1370	0.18	$61,921.00	92	88
35	1	1	1310	0.24	$27,487.00	78	88
36	2	2	1195	0.6	$21,853.00	71	77
37	2	24	1300	0.45	$38,937.00	74	73
38	2	25	1155	0.56	$38,597.00	52	73
39	1	26	1280	0.41	$30,882.00	87	86
40	1	13	1218	0.37	$19,365.00	77	88
41	3	27	1142	0.43	$26,859.00	96	61
42	3	28	1109	0.32	$19,684.00	82	73
43	2	3	1287	0.43	$20,179.00	53	84
44	2	29	1225	0.54	$39,883.00	71	76
45	2	30	1234	0.29	$17,998.00	61	78
46	2	20	1250	0.49	$27,879.00	76	86
47	1	13	1290	0.35	$19,948.00	73	91
48	1	1	1336	0.28	$23,772.00	86	93
49	1	15	1350	0.19	$52,468.00	90	93

Classification

- **Classification methods** seek to classify a categorical outcome into one of two or more categories based on various data attributes.
- For each record in a database, we have a categorical variable of interest and a number of additional predictor variables.
- For a given set of predictor variables, we would like to assign the best value of the categorical variable.

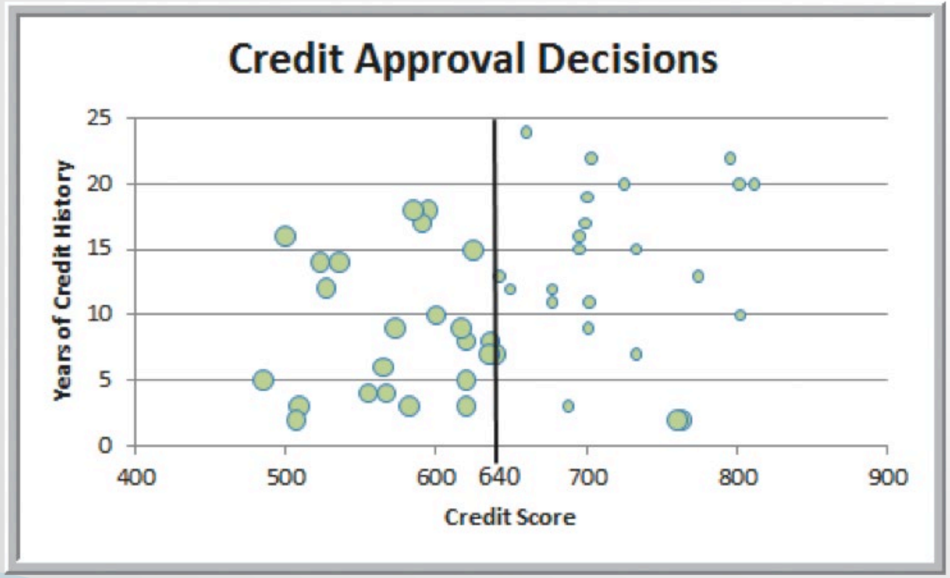
Credit Approval Decisions Data

- Categorical variable of interest: *Decision* (whether to approve – coded as 1 – or reject – coded as 0 – a credit application)
- Predictor variables: shown in columns A-E (note that homeowner is also coded numerically)

	A	B	C	D	E	F
1	**Credit Approval Decisions**					
2						
3	**Homeowner**	**Credit Score**	**Years of Credit History**	**Revolving Balance**	**Revolving Utilization**	**Decision**
4	Y	725	20	$ 11,320	25%	Approve
5	Y	573	9	$ 7,200	70%	Reject
6	Y	677	11	$ 20,000	55%	Approve
7	N	625	15	$ 12,800	65%	Reject
8	N	527	12	$ 5,700	75%	Reject
9	Y	795	22	$ 9,000	12%	Approve
10	N	733	7	$ 35,200	20%	Approve
11	N	620	5	$ 22,800	62%	Reject
12	Y	591	17	$ 16,500	50%	Reject
13	Y	660	24	$ 9,200	35%	Approve

Example 10.7: Classifying Credit-Approval Decisions Intuitively

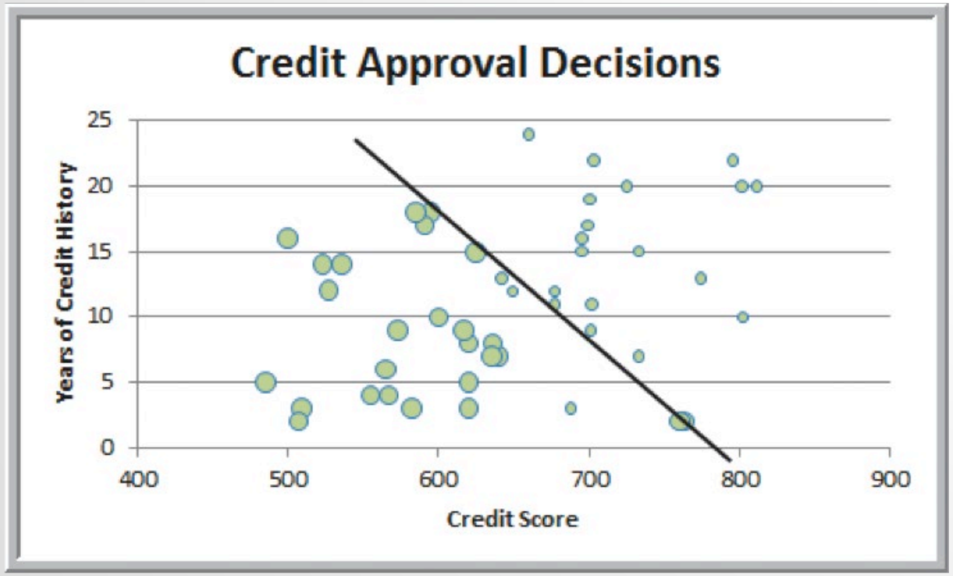
- Large bubbles correspond to rejected applications
- When the credit score is > 640, most applications were approved
  - Classification rule: Reject if credit score ≤ 640



2 misclassifications out of 50 = 4%

Example 10.7 Continued

- Alternate classification rule using visualization
  - Reject if years + 0.095(credit score) ≤ 74.66



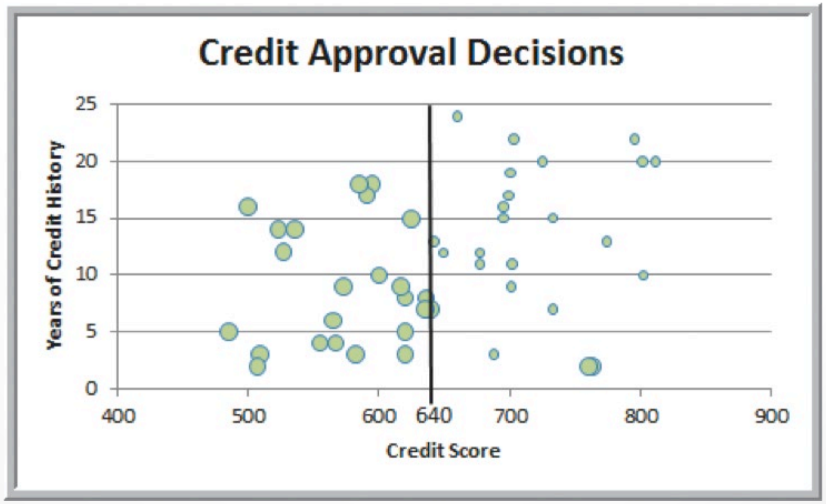
3 misclassifications out of 50 = 6%

Measuring Classification Performance

- Find the probability of making a misclassification error and summarize the results in a **classification matrix**, which shows the number of cases that were classified either correctly or incorrectly.

Example 10.8: Classification matrix for Credit-Approval Classification Rules

	Predicted Classification	
Actual Classification	Decision = 1	Decision = 0
Decision = 1	23	2
Decision = 0	0	25



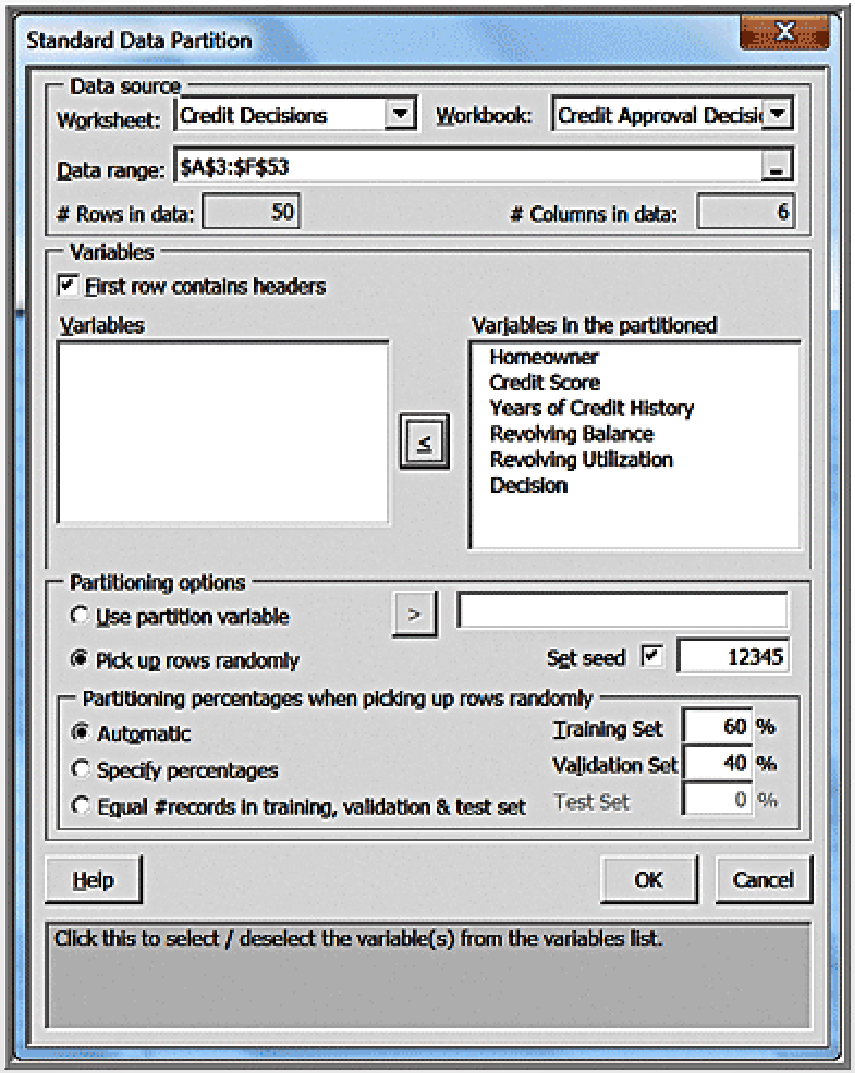
- Off-diagonal elements are the misclassifications
- Probability of a misclassification = 2/50 = 0.04

Using Training and Validation Data

- Data mining projects typically involve large volumes of data.
- The data can be partitioned into:
  - training data set – has known outcomes and is used to "teach" the data-mining algorithm
  - validation data set – used to fine-tune a model
  - test data set – tests the accuracy of the model
- In *XLMiner*, partitioning can be random or user-specified.

Example 10.9: Partitioning Data Sets in *XLMiner*

- *Modified Credit Approval Decisions* data
- *XLMiner > Partition Data > Standard Partition*
- *Select the variables*
- *Choose partitioning options and percentages*



Example 10.9 Continued

- Results

XLMiner : Data Partition Sheet

Output Navigator		
Training Data	Validation Data	Test Data

Data						
Data source	Credit Decisions!$A$4:$F$53					
Selected variables	Homeowner	Credit Score	Years of Credit	Revolving Balar	Revolving Utiliz	Decision
Partitioning Method	Randomly chosen					
Random Seed	12345					
# training rows	30					
# validation rows	20					

	Selected variables					
Row Id.	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
1	1	725	20	$ 11,320	25%	1
4	0	625	15	$ 12,800	65%	0
5	0	527	12	$ 5,700	75%	0
6	1	795	22	$ 9,000	12%	1
9	1	591	17	$ 16,500	50%	0
10	1	660	24	$ 9,200	35%	1

Classifying New Data

- After a classification scheme is chosen and the best model is developed based on existing data, we use the predictor variables as inputs to the model to predict the output.

Example 10.9: Classifying New Data for Credit Decisions Using Credit Scores and Years of Credit History

- Classify new data using the prior rules developed

	A	B	C	D	E	F
1						
2	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
3	1	700	8	$21,000	15%	
4	0	520	1	$4,000	90%	
5	1	650	10	$8,500.00	25%	
6	0	602	7	$16,300.00	70%	
7	0	549	2	$2,500.00	90%	
8	1	742	15	$16,700.00	18%	

- Using the second rule, if years + 0.095 × credit score ≤ 74.66, then only the last record would be approved for credit

Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Years + 0.095*Credit Score	Decision
1	700	8	$21,000.00	15%	74.50	0
0	520	1	$4,000.00	90%	50.40	0
1	650	10	$8,500.00	25%	71.75	0
0	602	7	$16,300.00	70%	64.19	0
0	549	2	$2,500.00	90%	54.16	0
1	742	15	$16,700.00	18%	85.49	1

Classification Techniques

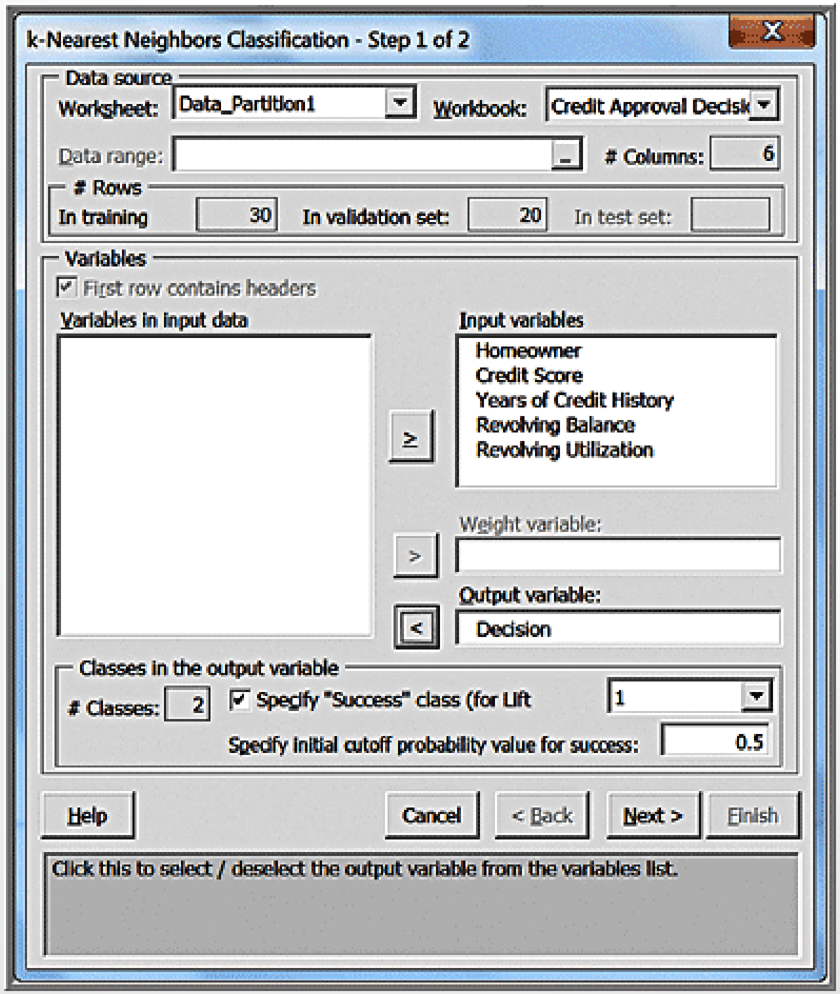
- *k-Nearest Neighbors (k-NN) Algorithm*
  - Finds records in a database that have similar numerical values of a set of predictor variables
- *Discriminant Analysis*
  - Uses predefined classes based on a set of linear discriminant functions of the predictor variables
- *Logistic Regression*
  - Estimates the probability of belonging to a category using a regression on the predictor variables

k-Nearest Neighbors (*k*-NN)

- Measure the Euclidean distance between records in the training data set.
- The nearest neighbor to a record in the training data set is the one that that has the smallest distance from it.
  - If $k = 1$, then the 1-NN rule classifies a record in the same category as its nearest neighbor.
  - *k*-NN rule finds the *k*-Nearest Neighbors in the training data set to each record we want to classify and then assigns the classification as the classification of majority of the *k* nearest neighbors
- Typically, various values of *k* are used and then results inspected to determine which is best.

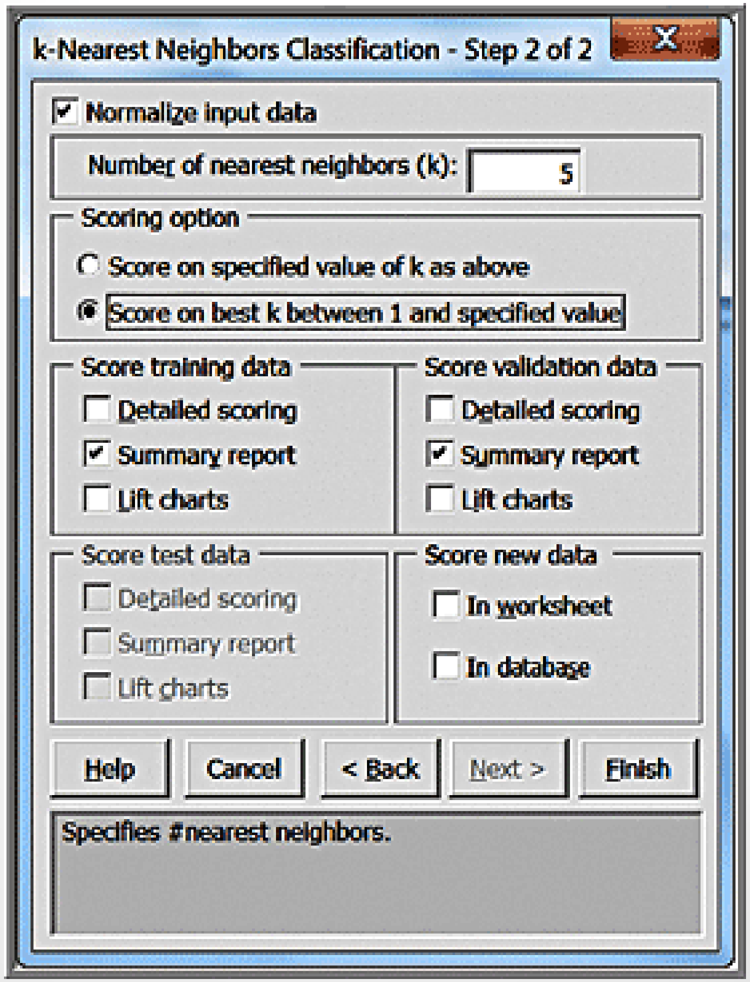
Example 10.10: Classifying Credit Decisions Using the *k*-NN Algorithm

- Partition the data into training and validation sets.
- *XLMiner > Classify < k-Nearest Neighbors*



Example 10.10 Continued

- Step 2
- Check the box *Normalize input data*
- Enter the value for *k*
- Choose scoring option



Example 10.10 Continued

- Results

XLMiner : k-Nearest Neighbors Classification

Output Navigator				
Inputs	Train. Score - Summary	Valid. Score - Summary	Test Score - Summary	Database Score
Elapsed Time	Train. Score - Detailed Rep.	Valid. Score - Detailed Rep.	Test Score - Detailed Rep.	New Score - Detailed Rep.
Prior Class Pr	Training Lift Charts	Validation Lift Charts	Test Lift Charts	Validation error log

Prior class probabilities

According to relative occurrences in training data

Class	Prob.	
1	0.366666667	<-- Success Class
0	0.633333333	

Validation error log for different k

Value of k	% Error Training	% Error Validation	
1	0.00	15.00	
2	6.67	10.00	<--- Best k
3	0.00	10.00	
4	3.33	10.00	
5	3.33	10.00	

Training Data scoring - Summary Report (for k=2)

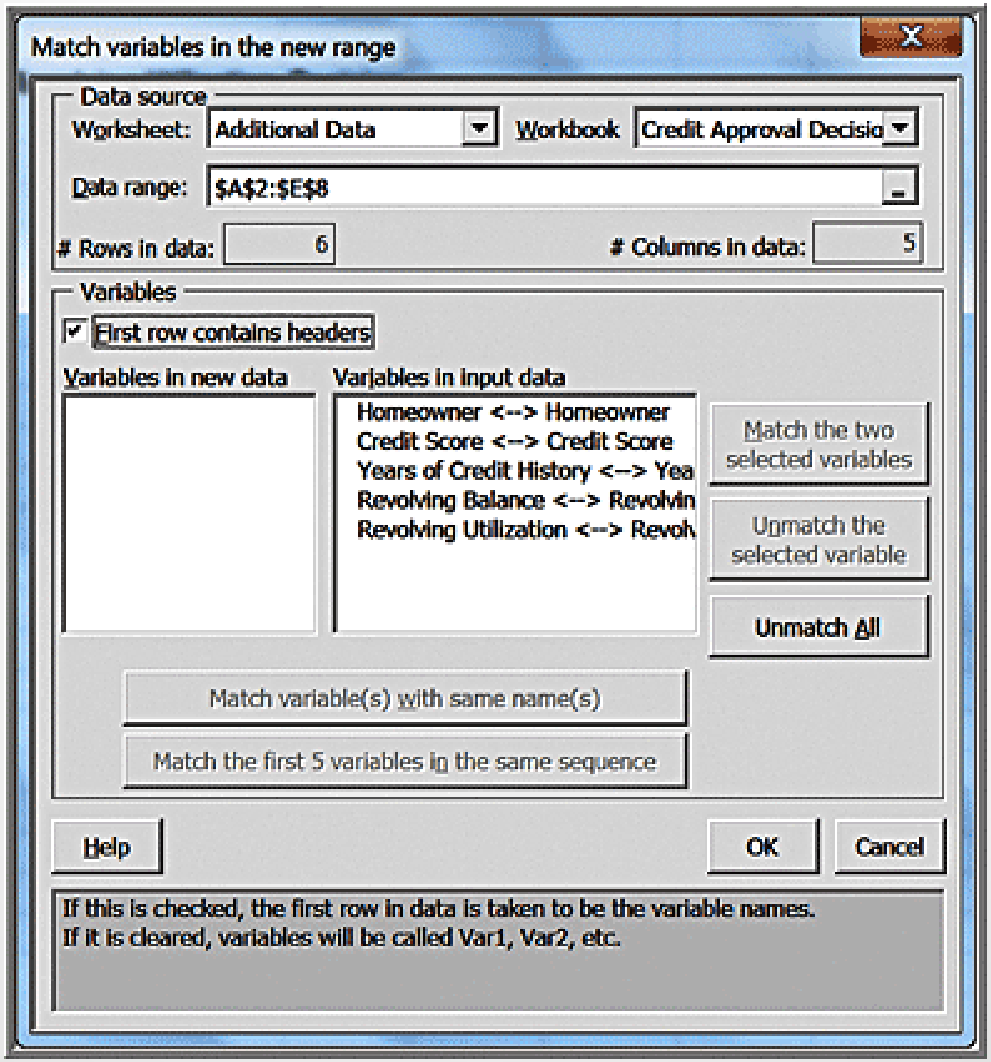
Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	11	0
0	2	17

Error Report			
Class	# Cases	# Errors	% Error
1	11	0	0.00
0	19	2	10.53
Overall	30	2	6.67

Example 10.11 Classifying New Data using *k*-NN

- Partition the data
- In Step 2 of k-NN, normalize the input data and set the number of nearest neighbors (*k*) to 2, the best value.
- Click on *In worksheet* in the *Score new data pane* of the dialog to open the *Match variables in the new range* dialog



Example 10.11 Continued

- Results

	A	B	C	D	E	F	G	H	I	J	K
1		XLMiner : k-Nearest Neighbors - Classification of New Data (for k=2)									
2											
3		Data range		['Credit Approval Decisions Coded.xlsx]'Additional Data'!$A$3:$E$8						Back to Navigator	
4											
5		Cut off Prob.Val. for Success (Updatable)				0.5	(Updating the value here will NOT update value in summary report)				
6											
7		Row Id.		Predicted Class	Prob. for 1 (success)	Actual #Nearest Neighbors	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization
8		1		1	1	2	1	700	8	$21,000.00	15%
9		2		0	0	2	0	520	1	$4,000.00	90%
10		3		1	1	2	1	650	10	$8,500.00	25%
11		4		0	0	2	0	602	7	$16,300.00	70%
12		5		0	0	2	0	549	2	$2,500.00	90%
13		6		1	1	2	1	742	15	$16,700.00	18%

Credit for records 1, 3 and 6 are approved

Discriminant Analysis

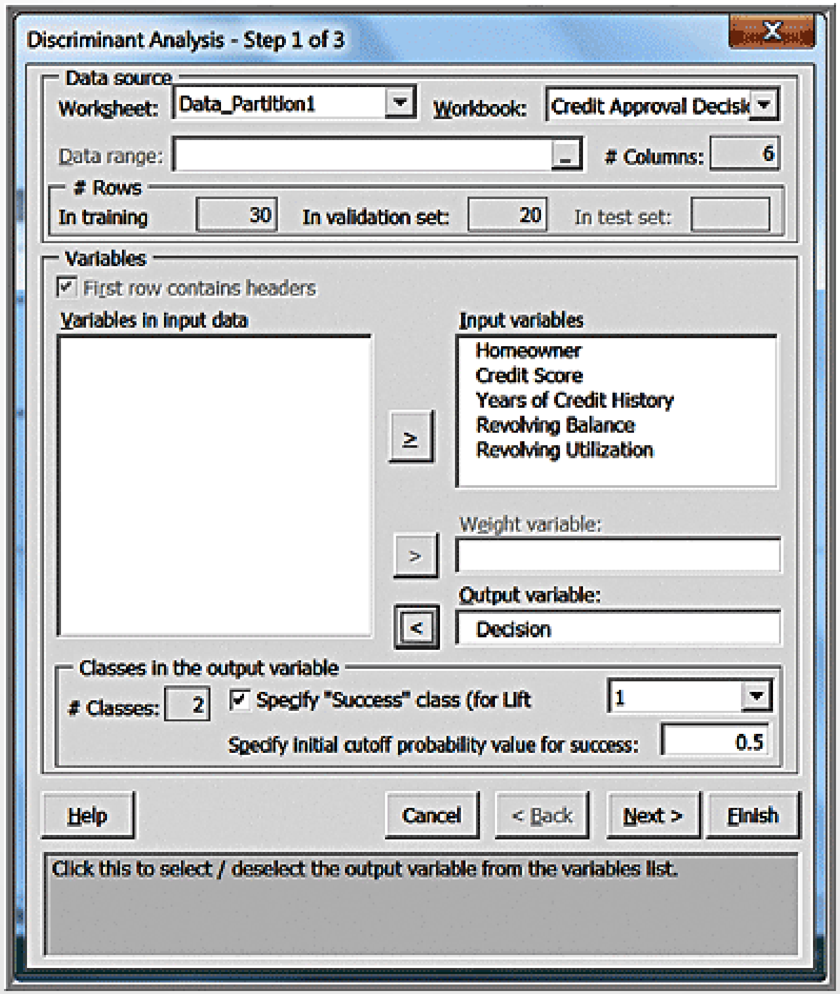
- **Discriminant analysis** is a technique for classifying a set of observations into predefined classes.
- Based on the training data set, the technique constructs a set of linear functions of the predictors, known as **discriminant functions**:

$$L = b_1X_1 + b_2X_2 + \cdots + b_nX_n + c \qquad \textbf{(10.2)}$$

  - b_i are the discriminant coefficients (weights), X_i are the input variables (predictors), c is a constant (intercept)
- For k categories, k discriminant functions are constructed. For a new observation, each of the k discriminant functions is evaluated, and the observation is assigned to class i if the i^{th} discriminant function has the highest value.

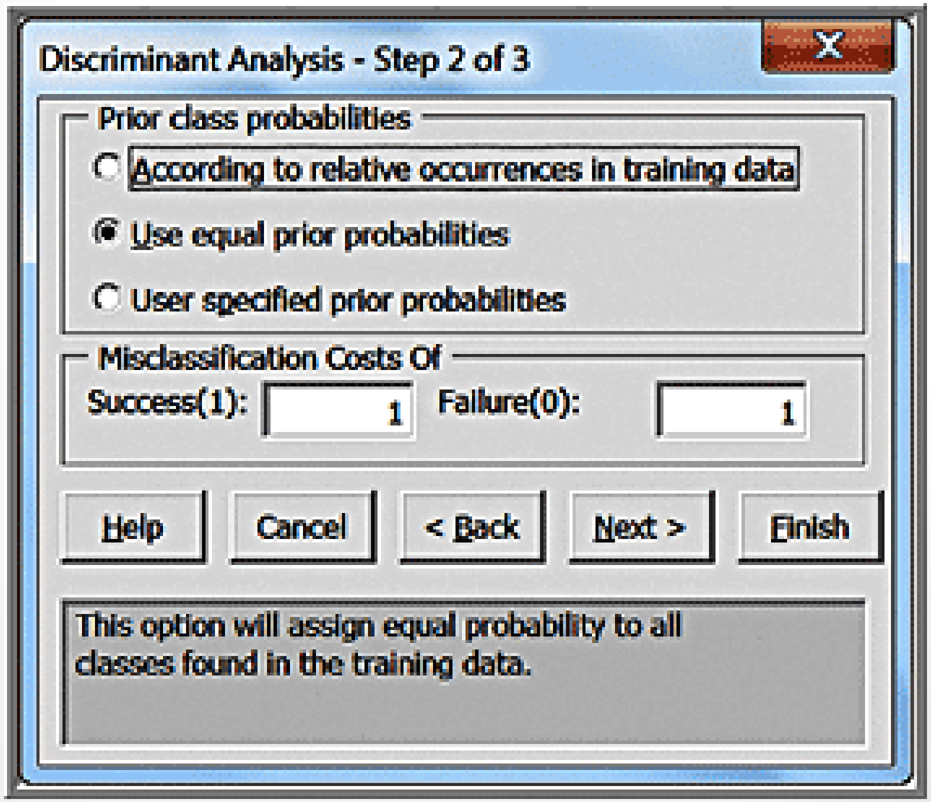
Example 10.12: Classifying Credit Decisions Using Discriminant Analysis

- *XLMiner > Classify > Discriminant Analysis*



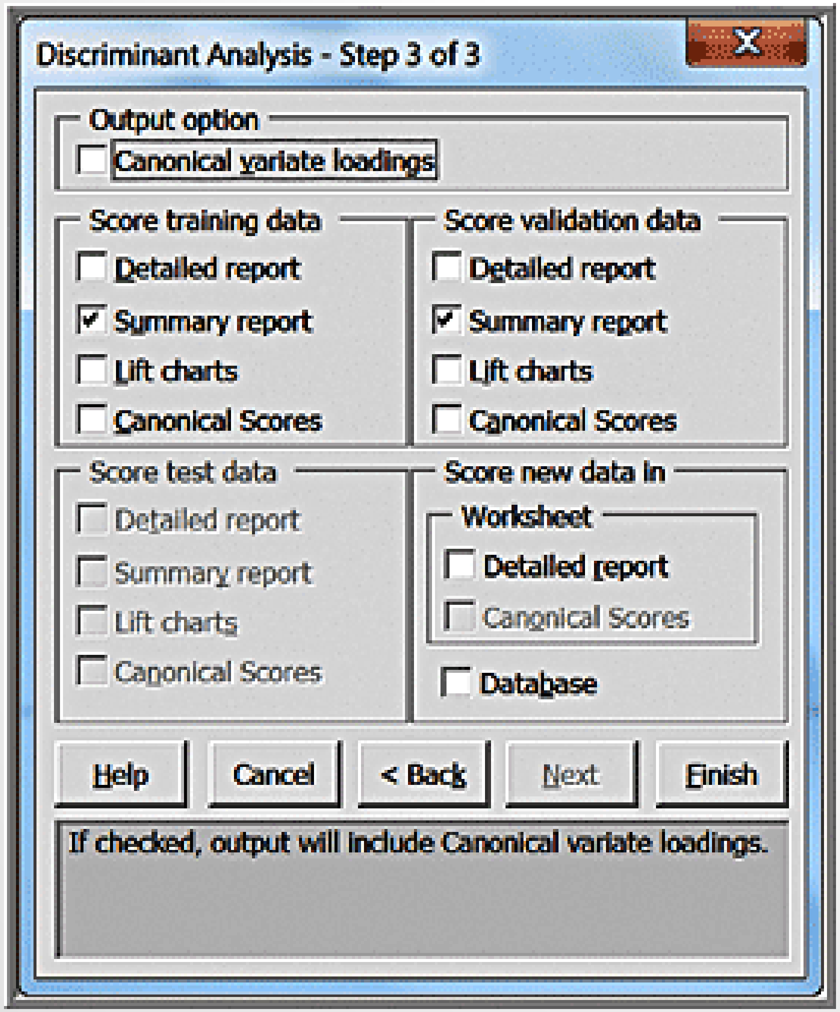
Example 10.12 Continued

- Step 2
- Select options for prior assumptions about how frequently the different classes occur.



Example 10.12 Continued

- Step 3



Example 10.12 Continued

- Results
  - Approve the application: L(1) = -137.48 + 32.295 × homeowner + 0.286 × credit score + 0.833 × years of credit history + 0.00010274 × revolving balance + 128.248 × revolving utilization
  - Reject the application: L(0) = -157.2 + 30.747 3 homeowner = 0.289 × credit score + 0.473 3 years of credit history + 0.0004716 × revolving balance + 167.7 × revolving utilization

For record 1, L(1) = 152.05; L(0) = 139.8. Assign to category 1

Classification Function

Variables	Classification Function	
	1	0
Constant	-137.4815521	-157.2017517
Homeowner	32.2950325	30.74663162
Credit Score	0.285761	0.28945312
Years of Credit History	0.83345157	0.47282016
Revolving Balance	0.00010274	0.0004716
Revolving Utilization	128.2484283	167.7003479

Example 10.12 Continued

- Scoring Reports

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	11	0
0	0	19

Error Report			
Class	# Cases	# Errors	% Error
1	11	0	0.00
0	19	0	0.00
Overall	30	0	0.00

Validation Data scoring - Summary Report

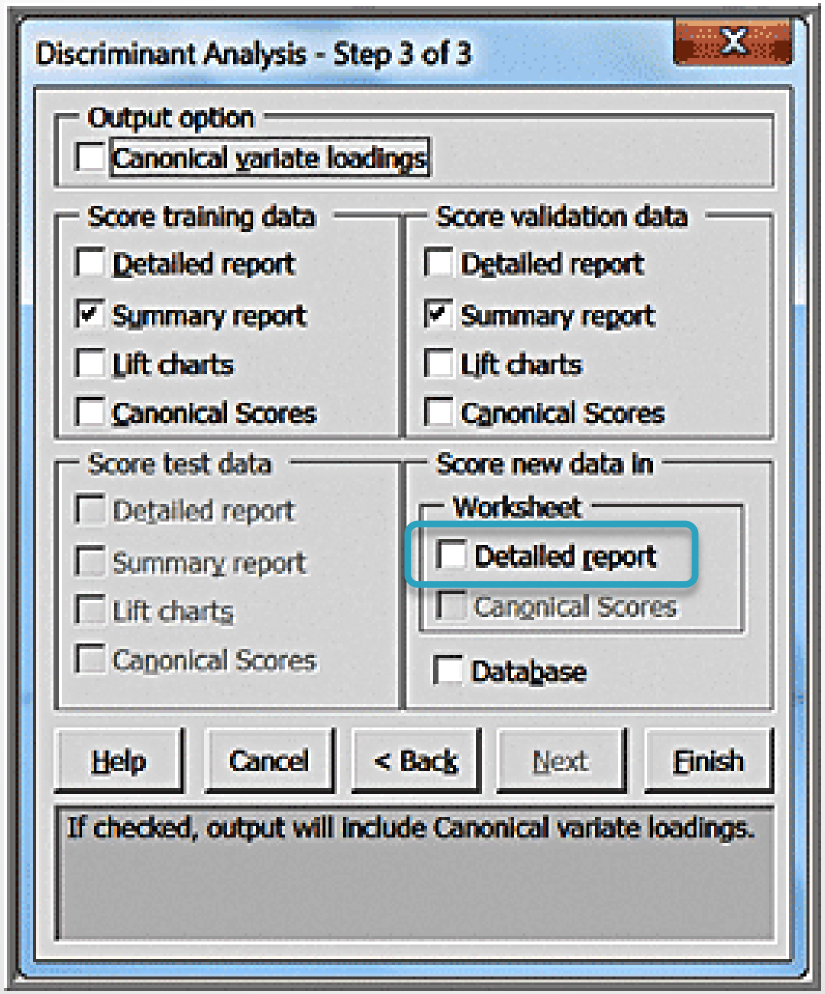
Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	10	2
0	1	7

Error Report			
Class	# Cases	# Errors	% Error
1	12	2	16.67
0	8	1	12.50
Overall	20	3	15.00

Example 10.13: Using Discriminant Analysis to Classify New Data

- In Step 3, click *Detailed report* in *Score new data in Worksheet* pane.



Example 10.13 Continued

- Results

XLMiner : Discriminant Analysis - Classification of New Data

Data range	['Credit Approval Decisions Coded.xlsx']'Additional Data'!$A$3:$E$8

Back to Navigator

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in summary report)

Row Id.	Predicted Class	Prob. for 1 (success)	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization
1	1	0.999631359	1	700	8	$21,000.00	15%
2	0	6.69946E-09	0	520	1	$4,000.00	90%
3	1	0.999923393	1	650	10	$8,500.00	25%
4	0	1.23209E-06	0	602	7	$16,300.00	70%
5	0	1.50124E-08	0	549	2	$2,500.00	90%
6	1	0.999976936	1	742	15	$16,700.00	18%

Logistic Regression

- **Logistic regression** is variation of linear regression in which the dependent variable is categorical.
  - Seeks to predict the probability that the output variable will fall into a category based on the values of the independent (predictor) variables. This probability is used to classify an observation into a category.
- Generally used when the dependent variable is binary—that is, takes on two values, 0 or 1.

Classification Using Logistic Regression

- Estimate the probability p that an observation belongs to category 1, $P(Y = 1)$, and, consequently, the probability $1 - p$ that it belongs to category 0, $P(Y = 0)$.
- Then use a *cutoff value*, typically 0.5, with which to compare p and classify the observation into one of the two categories.
- The dependent variable is called the **logit**, which is the natural logarithm of $p/(1 – p)$ – called the **odds** of belonging to category 1.
- The form of a logistic regression model is

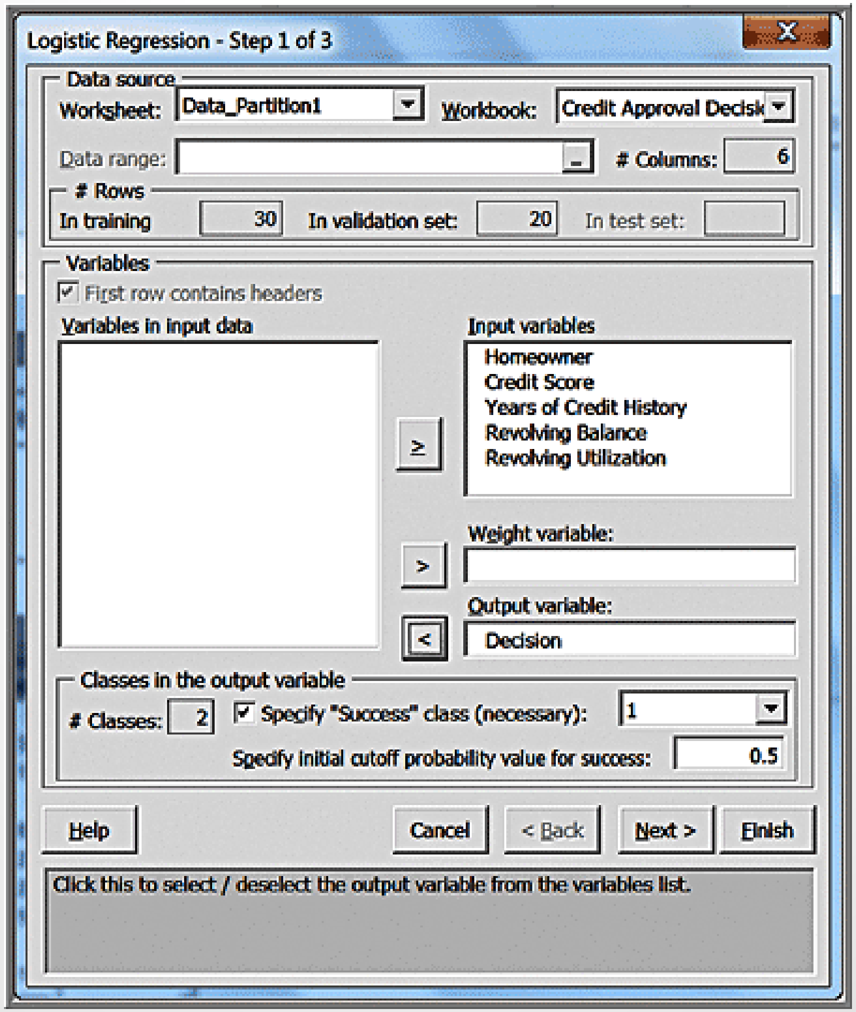
$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad \textbf{(10.3)}$$

- The logit function can be solved for p:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}} \qquad \textbf{(10.4)}$$

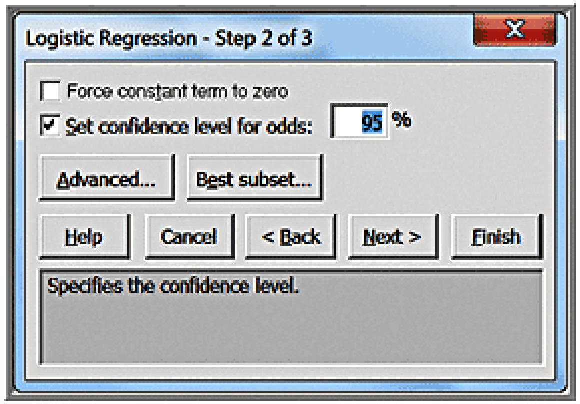
Example 10.14: Classifying Credit Approval Decisions Using Logistic Regression

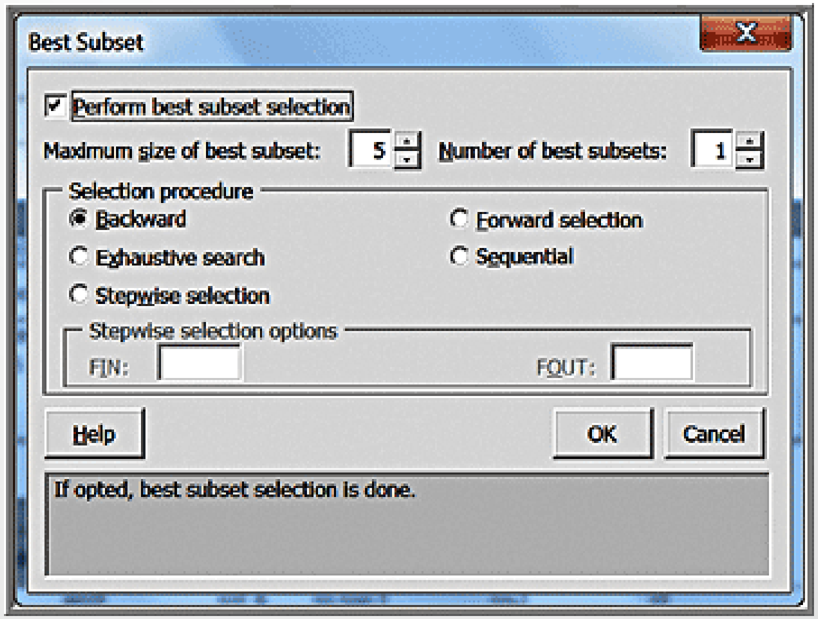
- *XLMiner > Classify > Logistic Regression*
- Partition the data
- Specify the data range, the input variables, and the output variable.



Example 10.14 Continued

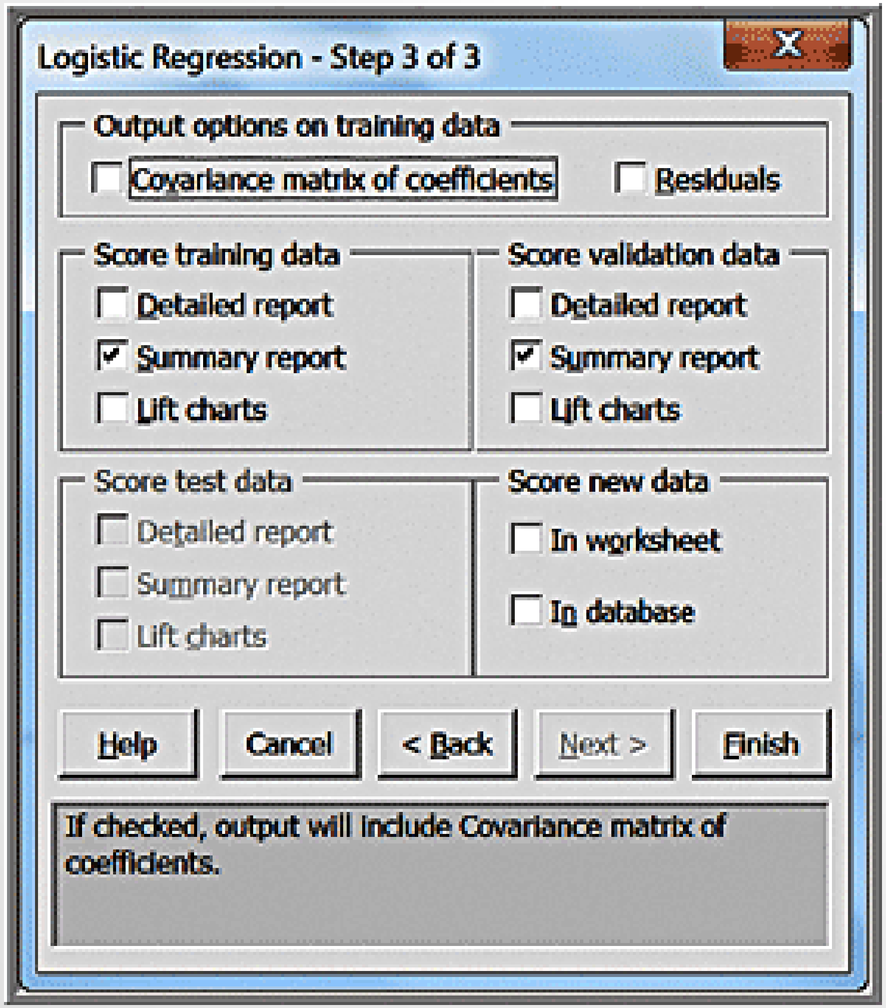
- Step 2
- The *Best Subsets* button allows *XLMiner* to evaluate all possible models with subsets of the independent variables.
  - This is useful in choosing models that eliminate insignificant independent variables.





Example 10.14 Continued

- Step 3



Example 10.14 Continued

- Results

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds	95% Confidence Interval	
Constant term	8.70898151	177.1350403	0.96078718	6057.07028	3889.615345	8224.525214
Homeowner	-1.89079905	31.67862511	0.95240498	0.15095115	0	1.3923E+26
Credit Score	0.01126203	0.21146901	0.95752782	1.01132572	0.668172	1.53071308
Years of Credit History	0.18884063	1.65134251	0.90895575	1.20784843	0.04746649	30.73532104
Revolving Balance	-0.00022931	0.0020333	0.91020685	0.9997707	0.99579436	1.00376296
Revolving Utilization	-33.73615646	70.85647583	0.63398921	0	0	*

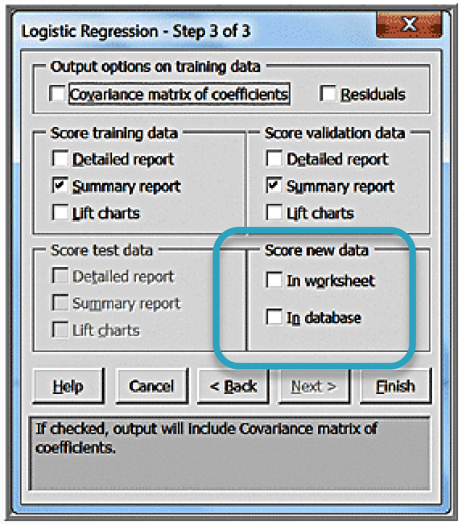
Residual df	24
Residual Dev.	0.09734347
% Success in training data	36.66666667
# Iterations used	9
Multiple R-squared	0.99753118

Best subset selection

	#Coeffs	RSS	Cp	Probability	Model (Constant present in all models)					
					1	2	3	4	5	6
Choose Subset	2	23.09636879	-1.89944279	0.99869645	Constant	Revolving Utilization	*	*	*	*
Choose Subset	3	23.01775742	0.0185279	0.99931508	Constant	Revolving Balance	Revolving Utilization	*	*	*
Choose Subset	4	23.01327133	2.01384687	0.99310237	Constant	Years of Credit History	Revolving Balance	Revolving Utilization	*	*
Choose Subset	5	23.00244331	4.00254774	0.96014309	Constant	Homeowner	Years of Credit History	Revolving Balance	Revolving Utilization	*
Choose Subset	6	23.00000191	6.00000048	1	Constant	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization

Example 10.15: Using Logistic Regression to Classify New Data

- In Step 3 click on *In worksheet* in the *Score new data* pane of the dialog.



XLMiner : Logistic Regression - Classification of New Data

Data range	['Credit Approval Decisions Coded.xlsx']'Additional Data'!$A$3:$E$8

Back to Navigator

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in summary report)

Row Id.	Predicted Class	Prob. for 1 (success)	Log odds	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization
1	1	0.998232456	6.336395031	1	700	8	$21,000.00	15%
2	0	6.6524E-08	-16.52570307	0	520	1	$4,000.00	90%
3	1	0.996472859	5.643734145	1	650	10	$8,500.00	25%
4	0	2.63912E-05	-10.54245454	0	602	7	$16,300.00	70%
5	0	1.57113E-07	-15.66629857	0	549	2	$2,500.00	90%
6	1	0.999698136	8.105233007	1	742	15	$16,700.00	18%

Association Rule Mining

- **Association rule mining**, often called *affinity analysis*, seeks to uncover associations and/or correlation relationships in large data sets
  - Association rules identify attributes that occur together frequently in a given data set.
  - **Market basket analysis**, for example, is used determine groups of items consumers tend to purchase together.
- Association rules provide information in the form of if-then (antecedent-consequent) statements.

Example 10.16: Custom Computer Configuration

- *PC Purchase Data*
- We might want to know which components are often ordered together.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**PC Purchase Data**											
2												
3		**Processor**			**Screen Size**			**Memory**			**Hard Drive**	
4												
5	**Intel Core i3**	**Intel Core i5**	**Intel Core i7**	**10 inch screen**	**12 inch screen**	**15 inch screen**	**2 GB**	**4 GB**	**8 GB**	**320 GB**	**500 GB**	**750 GB**
6	0	1	0	0	1	0	0	1	0	0	1	0
7	0	1	0	0	0	1	0	0	1	0	0	1
8	0	1	0	0	1	0	0	1	0	1	0	0
9	1	0	0	0	1	0	0	0	1	0	1	0
10	0	0	1	0	0	1	0	0	1	0	0	1
11	0	0	1	0	1	0	0	1	0	0	0	1
12	0	0	1	0	0	1	0	0	1	0	0	1
13	1	0	0	0	1	0	0	1	0	0	1	0
14	0	1	0	1	0	0	1	0	0	0	1	0

Measuring Strength of Association

- **Support for the (association) rule** is the percentage (or number) of transactions that include all items both antecedent and consequent.
- **Confidence of the (association) rule** is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

$$\text{confidence} = P(\text{consequent} \mid \text{antecedent}) = \frac{P(\text{antecedent and consequent})}{P(\text{antecedent})} \qquad (10.5)$$

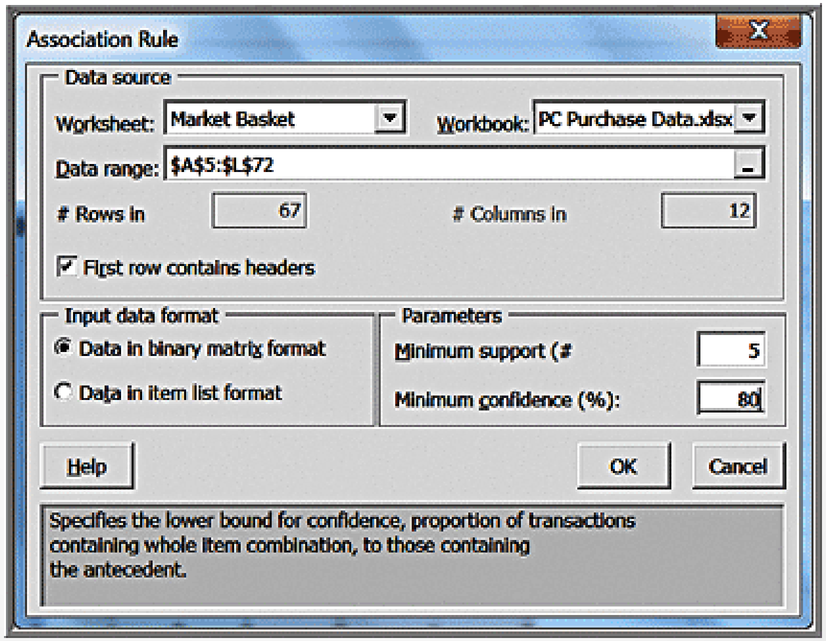
- **Lift** is a ratio of confidence to expected confidence.
  - Expected confidence is the number of transactions that include the consequent divided by the total number of transactions.
  - The higher the lift ratio, the stronger the association rule; a value greater than 1.0 is usually a good minimum.

Example 10.17: Measuring Strength of Association

- A supermarket database has 100,000 point-of-sale transactions; 2000 include both A and B items; 5000 include C; and 800 include A, B, and C
- Association rule: "If A and B are purchased, then C is also purchased."
  - Support = 800/100,000 = 0.008
  - Confidence = 800/2000 = 0.40
  - Expected confidence = 5000/100,000 = 0.05
  - Lift = 0.40/0.05 = 8

Example 10.18: Identifying Association Rules for *PC Purchase Data*

- *XLMiner > Associate > Association Rules*
- *Input options:*
  - Data in binary matrix format: Choose this option if each column in the data represents a distinct item and the data are expressed as 0s and 1s.
  - Data in item list format: Choose this option if each row of data consists of item codes or names that are present in that transaction.
- Specify minimum support and confidence parameters



Example 10.18 Continued

- Results

XLMiner : Association Rules

Data	
Input Data	Market Basket!$A$5:$L$72
Data Format	Binary Matrix
Minimum Support	5
Minimum Confidence %	80
No. of Rules	10
Overall Time (secs)	5

Place the cursor on a cell in the rules table to read a rule.
Use up / down arrow keys to browse through the rules.

Rule No.	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	15 inch screen, Intel Core i7=>	750 GB	5	17	5	3.941176
2	83.33	15 inch screen, 8 GB=>	750 GB	6	17	5	3.284314
3	100	15 inch screen, 500 GB=>	Intel Core i5	5	33	5	2.030303
4	83.33	12 inch screen, 8 GB=>	500 GB	6	31	5	1.801075
5	83.33	12 inch screen, 4 GB, Intel Core i5=>	500 GB	6	31	5	1.801075
6	100	15 inch screen, 320 GB=>	4 GB	6	38	6	1.763158
7	83.33	4 GB, Intel Core i7=>	12 inch screen	6	32	5	1.744792
8	83.33	500 GB, 8 GB=>	12 inch screen	6	32	5	1.744792
9	85.71	10 inch screen, 320 GB=>	4 GB	7	38	6	1.511278
10	80	320 GB, Intel Core i5=>	4 GB	10	38	8	1.410526

Rule 1 states that if a customer purchased a 15-inch screen with an Intel Core i7 processor, then a 750 GB hard drive was also purchased.

Example 10.18 Continued

- Display of Rule #1

Rule 1: If item(s) 15 inch screen, Intel Core i7= is / are purchased, then this implies item(s) 750 GB is / are also purchased. This rule has confidence of 100%.

Rule No.	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	15 inch screen, Intel Core i7=>	750 GB	5	17	5	3.941176
2	83.33	15 inch screen, 8 GB=>	750 GB	6	17	5	3.284314
3	100	15 inch screen, 500 GB=>	Intel Core i5	5	33	5	2.030303
4	83.33	12 inch screen, 8 GB=>	500 GB	6	31	5	1.801075
5	83.33	12 inch screen, 4 GB, Intel Core i5=>	500 GB	6	31	5	1.801075
6	100	15 inch screen, 320 GB=>	4 GB	6	38	6	1.763158
7	83.33	4 GB, Intel Core i7=>	12 inch screen	6	32	5	1.744792
8	83.33	500 GB, 8 GB=>	12 inch screen	6	32	5	1.744792
9	85.71	10 inch screen, 320 GB=>	4 GB	7	38	6	1.511278
10	80	320 GB, Intel Core i5=>	4 GB	10	38	8	1.410526

- Confidence (Conf.%) means that of the people who bought a 15-inch screen and a core i7 processor, all (100%) bought 750 GB hard drives as well.
- Support (a) indicates that 5 customers bought a 15-inch screen and a core i7 processor.
- Support (c) indicates the number of transactions involving the purchase of options, total.
- Support (a U c) is the number of transactions in which a 15-inch screen, Intel Core i7, and 750 GB hard drive were ordered.
- Lift Ratio indicates how much more likely we are to encounter a 750 GB transaction if we consider just those transactions where a 15-inch screen and Intel Core i7 are purchased, as compared to the entire population of transactions.

Cause-and-Effect Modeling

- Correlation analysis can help us develop cause-and-effect models that relate lagging and leading measures.
  - **Lagging measures** tell us what has happened and are often external business results such as profit, market share, or customer satisfaction.
  - **Leading measures** predict what will happen and are usually internal metrics such as employee satisfaction, productivity, and turnover.

Example 10.19: Using Correlation for Cause-and-Effect Modeling

- *Ten Year Survey* data
  - Satisfaction was measured on a 1-5 scale.

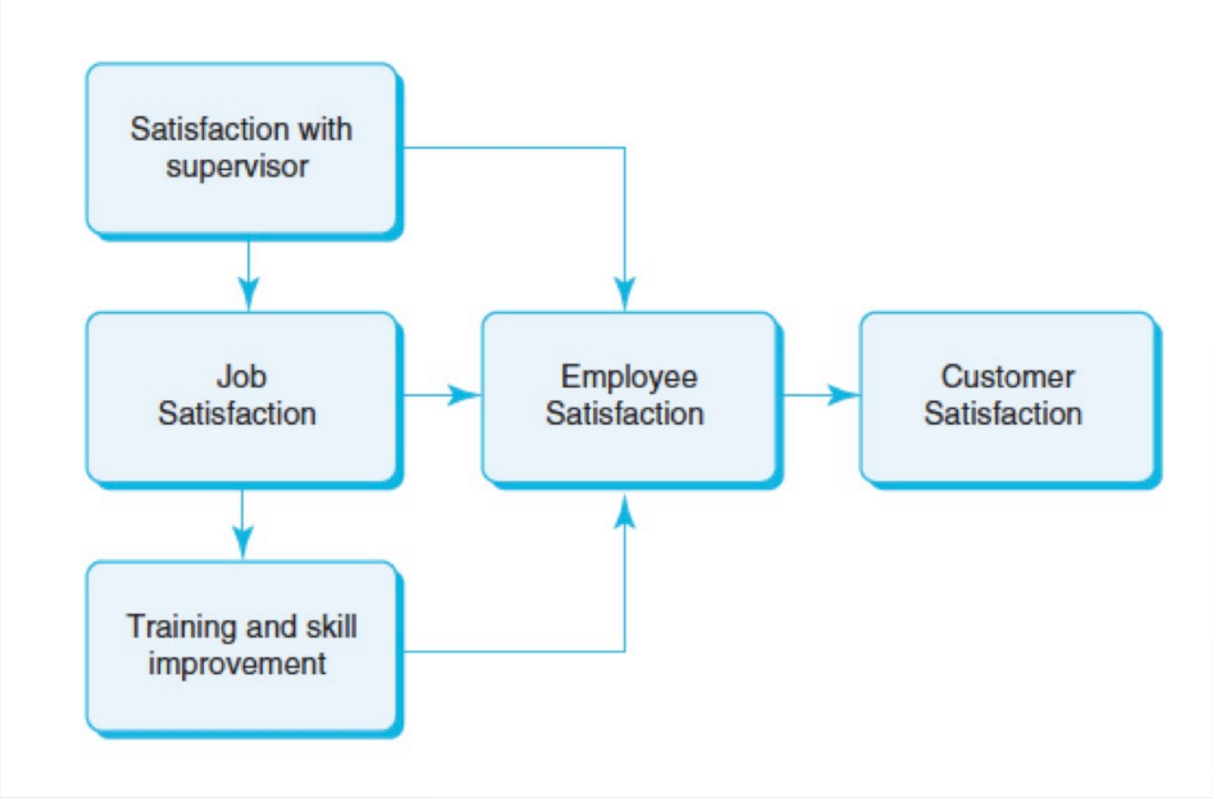
	A	B	C	D	E	F
1	Ten Year Survey					
2						
3	Survey Sample	Customer satisfaction	Employee satisfaction	Job satisfaction	Satisfaction with supervisor	Training and skill improvement
4	1	2.97	3.51	3.92	3.06	3.48
5	2	3.71	3.58	4.13	3.06	2.57
6	3	3.29	3.43	3.62	4.42	3.06
7	4	2.05	3.81	4.12	4.31	3.17
8	5	4.56	4.17	4.25	4.14	4.15
9	6	4.28	4.13	4.13	4.57	3.61
10	7	2.17	2.42	4.19	2.53	2.72
11	8	3.01	2.95	3.95	3.25	2.56

- Correlation matrix

	A	B	C	D	E	F
1		*Customer satisfaction*	*Employee satisfaction*	*Job satisfaction*	*Satisfaction with supervisor*	*Training and skill improvement*
2	Customer satisfaction	1				
3	Employee satisfaction	0.493345395	1			
4	Job satisfaction	0.151693544	0.840444148	1		
5	Satisfaction with supervisor	0.495977225	0.881324581	0.606796166	1	
6	Training and skill improvement	0.532307756	0.828657884	0.710624973	0.769700425	1

Example 10.19 Continued

- Logical model



Group Homework 4 – Email me (albert.kalim@asbury.edu) your GROUP answers and the viz link by Sunday, 6/26,11:59 p.m. ET (10 points total)

Using the data file “Banking Data.xls” (click here to download data)

- 1. Answer Chapter 10, Problem 2 (parts a and b – 2.5 pts each).
- 2. Create a Tableau dashboard with one visualization (viz) based on the data above. You pick the viz type. What observations can you make about this viz based on the picked data? (5 pts)