Data Mining – LAB 1

PASW Modeler 13 – Getting Started

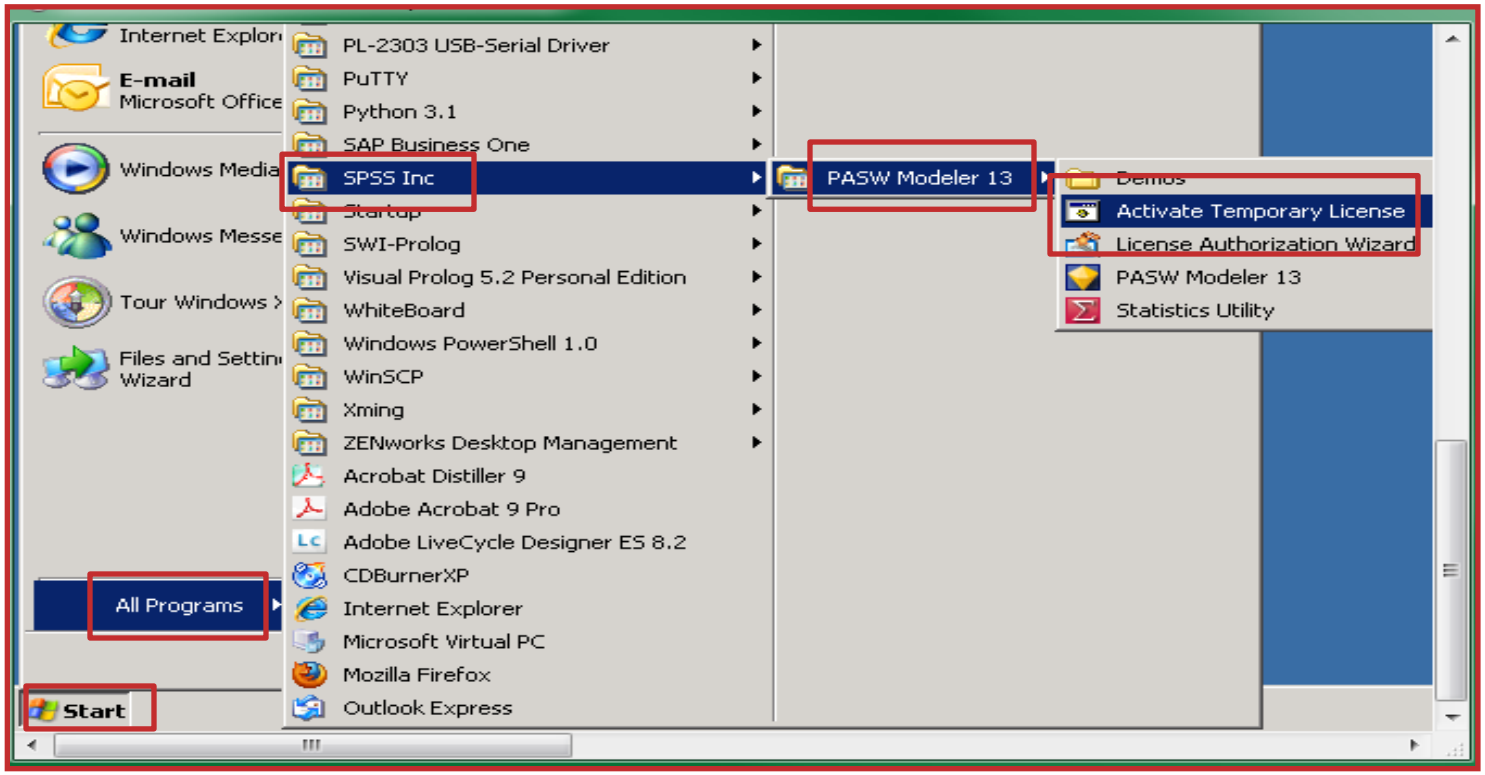
Association – Apriori

Getting Started – 1

- In department's labs, you may use the software directly by activating a temporary license in the PC.
- At home, you may download and use a virtual machine image. We use "Virtual Box", so you may download the tool from the official site (http://www.virtualbox.org/wiki/Downloads), and install it in your machine.

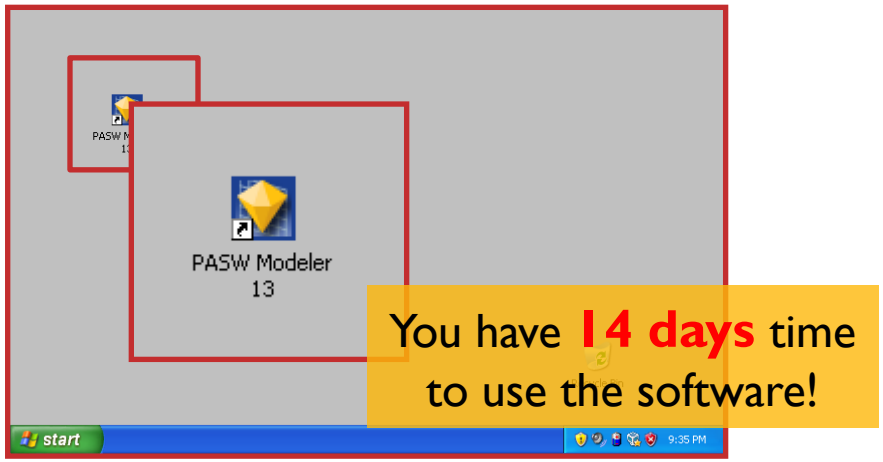
Getting Started – 2

- Run PASW in lab
  - [Start] ➔ [Programs] ➔ [SPSS Inc] ➔ [PASW Modeler 13] ➔ [Activate Temporary License]



Getting Started – 3

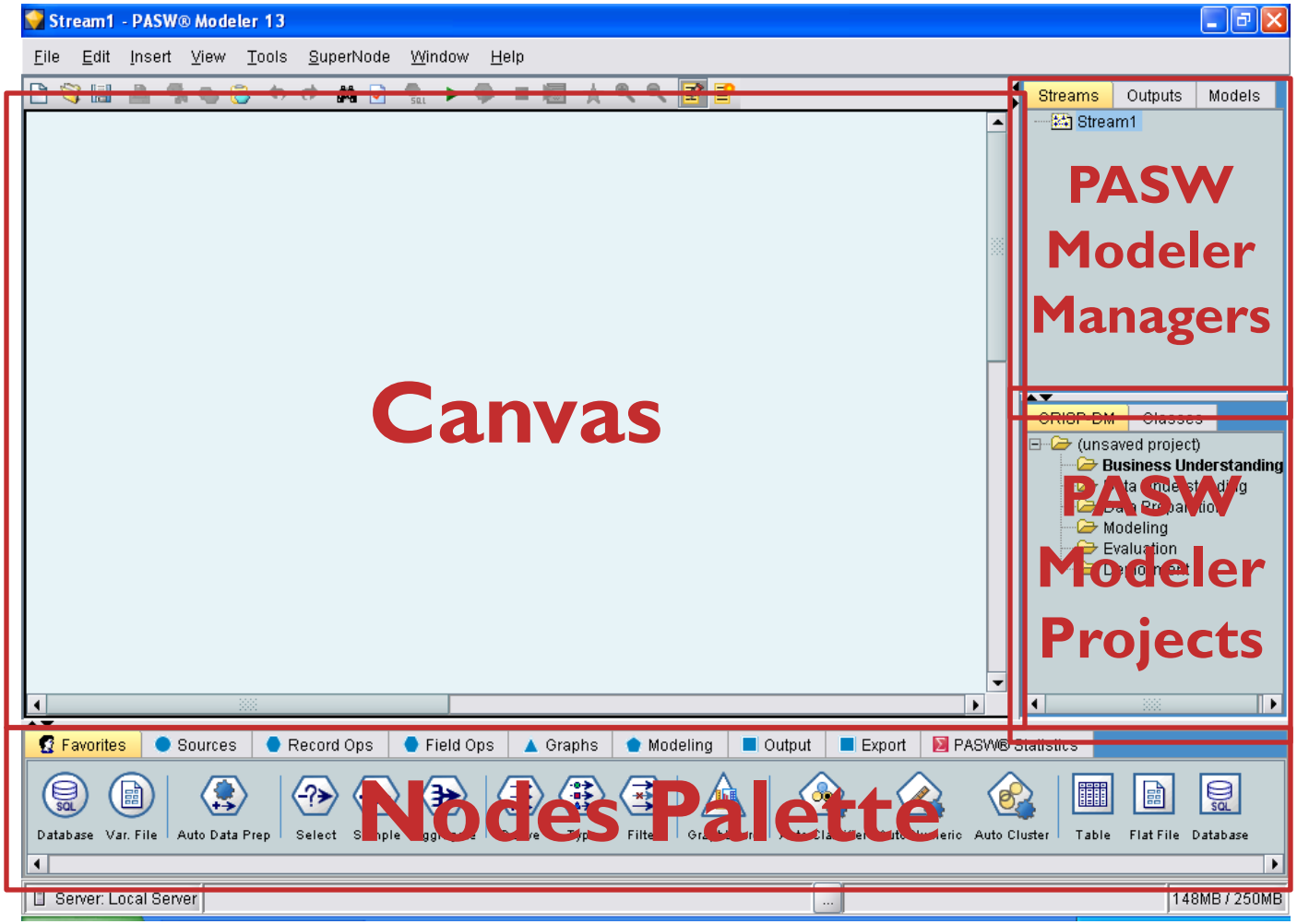
- Use the software at home.
  - Download "**vb_pasw.vdi**" from department network drive and start running the machine.
  - You should find the "**PASW Modeler 13**" icon on the desktop.
  - Double click the icon to run the application.



Getting Started – 4

- For lab's copy, the "activate" command needs to be run for each new start (or reboot) of the machine.
- For Virtualbox image, as it is a trial version, it lasts for only two weeks' time. A new copy should be replaced after the trial period.

PASW – First Look



PASW - Canvas

- The **stream canvas** is the largest area of the PASW Modeler window and is where you will build and manipulate data streams.
- **Streams** are created by drawing of data operations relevant to your business on the main canvas in the interface. Each operation is represented by an icon or **node**, and the nodes are linked together in a **stream** representing the flow of data through each operation.

PASW – Nodes Palette

- Most of the data and modeling tools in PASW Modeler reside in the **Nodes Palette**, across the bottom of the window below the stream canvas.
- For example, the **Record Ops** palette tab contains nodes that you can use to perform operations on the data **records**, such as selecting, merging, and appending.
- To add nodes to the canvas, double-click icons from the Nodes Palette or drag and drop them onto the canvas. You than connect them to create a **stream**, representing the flow of data.

PASW – Some Other Nodes (1)

- **Sources** – Nodes brings data into PASW Modeler.
- **Records Ops** – Nodes perform operations on data records, such as selecting, merging, and appending.
- **Field Ops** – Nodes perform operations on data fields, such as filtering, derving new fields, and determining the data type for given fields.

PASW – Some Other Nodes (2)

- **Graphs** – Nodes graphically display data before and after modeling. Graphs include plots, histograms, web nodes, and evaluation charts.
- **Modeling** – Nodes use the modeling algorithms available in PASW Modeler, such as neural nets, decision trees, clustering algorithms, and data sequencing.
- **Output** – Nodes produce a variety of output of data, charts, and model results, which can be viewed in PASW Modeler or sent directly to another application, such as PASW Statistics or Excel.

PASW Modeler Manager

- You can use the **Streams** tab to open, rename, save and delete the streams created in a session.
- The **Outputs** tab contains a variety of files, such as graphs and tables, produced by stream operations in PASW Modeler. You can display, save, rename, and close the tables, graphs, and reports listed on this tab.
- The **Models** tab contains all models nuggets, which are models generated in PASW Modeler, for the current session. These models can be browsed directly from the Models tab or added to the stream in the canvas.

PASW Modeler Projects (1)

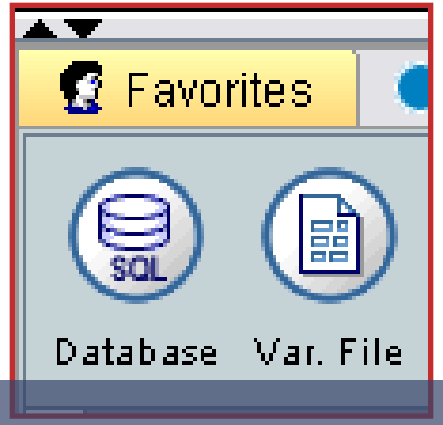
- On the lower right side of the window is the projects tool, used to create and manage data mining projects (groups of files related to a data mining task).
- There are two ways to view projects you create in PASW Modeler – in the Classes view and the CRISP-DM view.

PASW Modeler Projects (2)

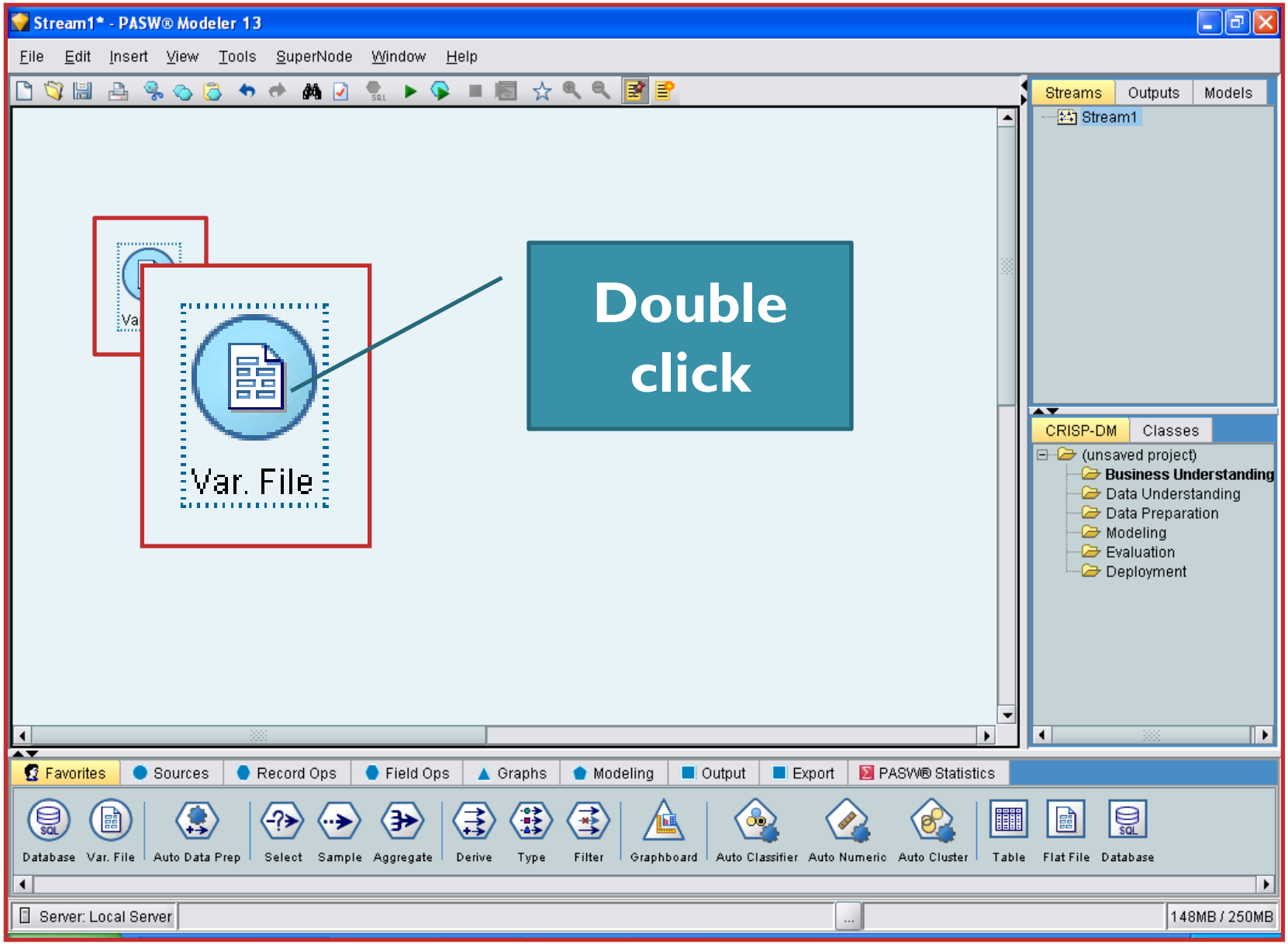
- The **CRISP-DM** tab provides a way to organize projects according to the *Cross-Industry Standard Process for Data Mining*, an industry-proven, nonproprietary methodology. For both experienced and first-time data miners, using the CRISP-DM tool will help you to better organize and communicate your efforts.
- The **Classes** tab provides a way to organize your work in PASW Modeler categorically – by the types of objects you create. This view is useful when taking inventory of data, streams and models.

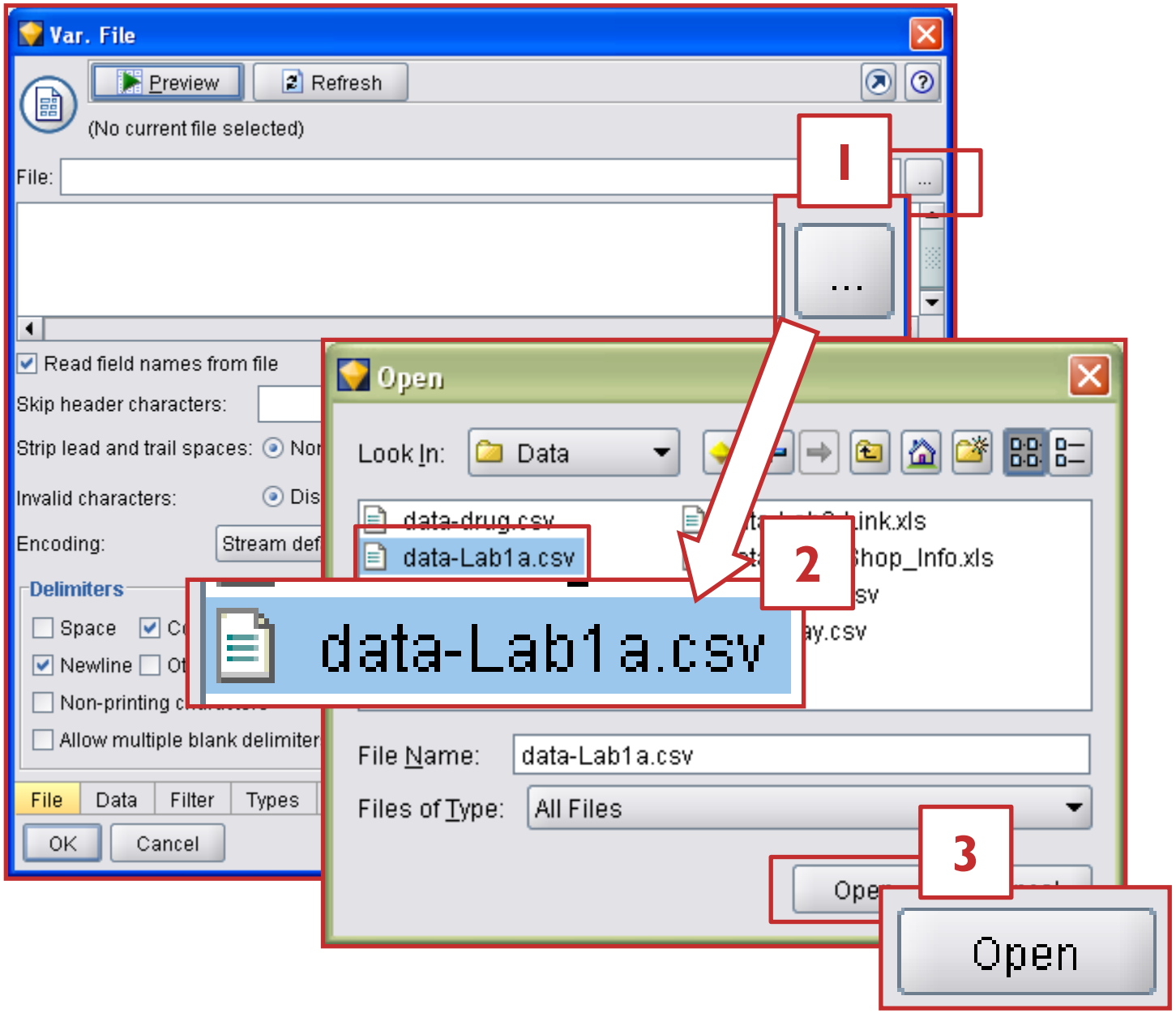
PASW – First Use

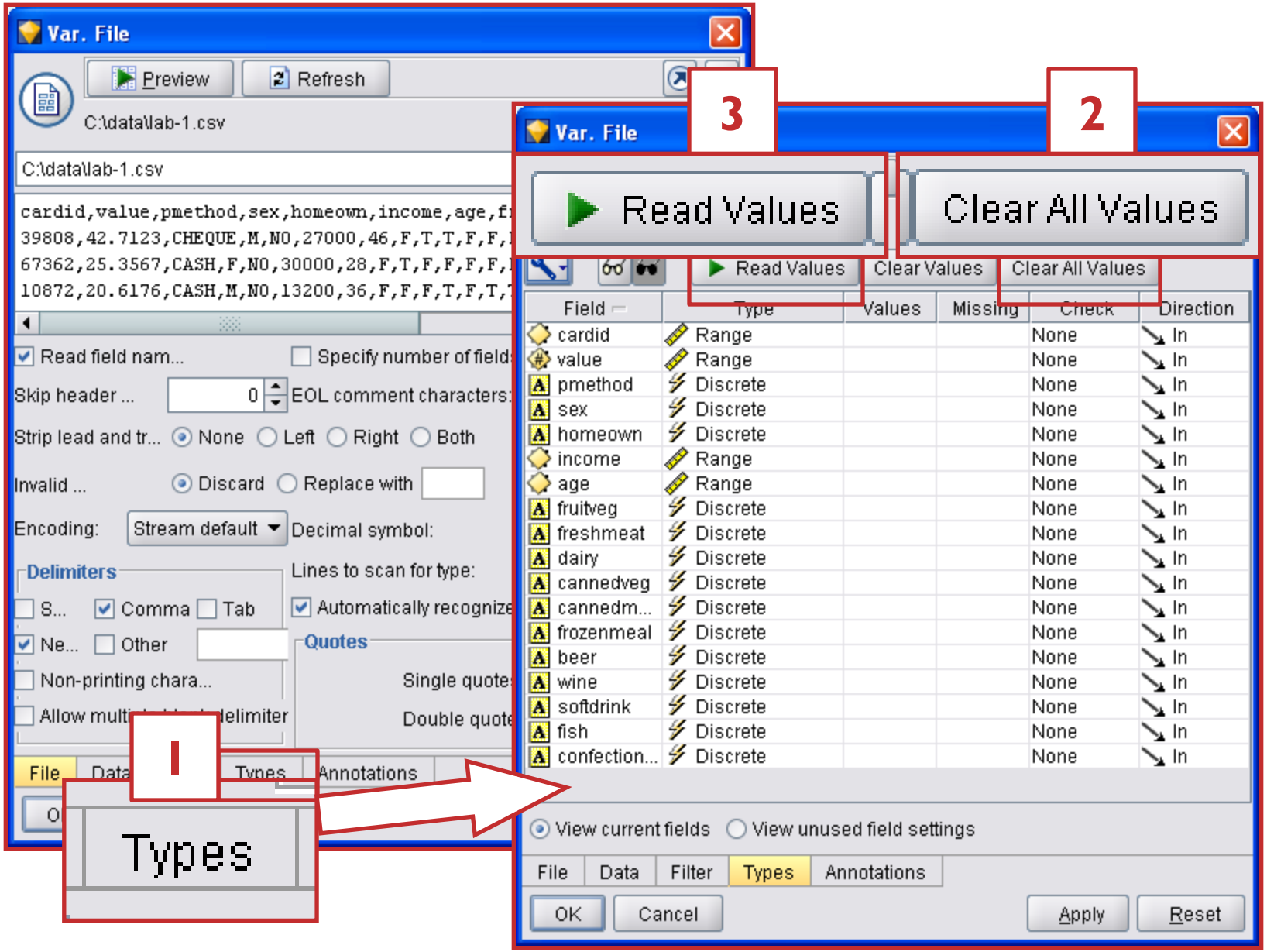
- Create a stream by applying an algorithm model, **Apriori**.
- Start PASW Modeler 13.
- Double click or "drag and drop" the icon, "**Var. File**".

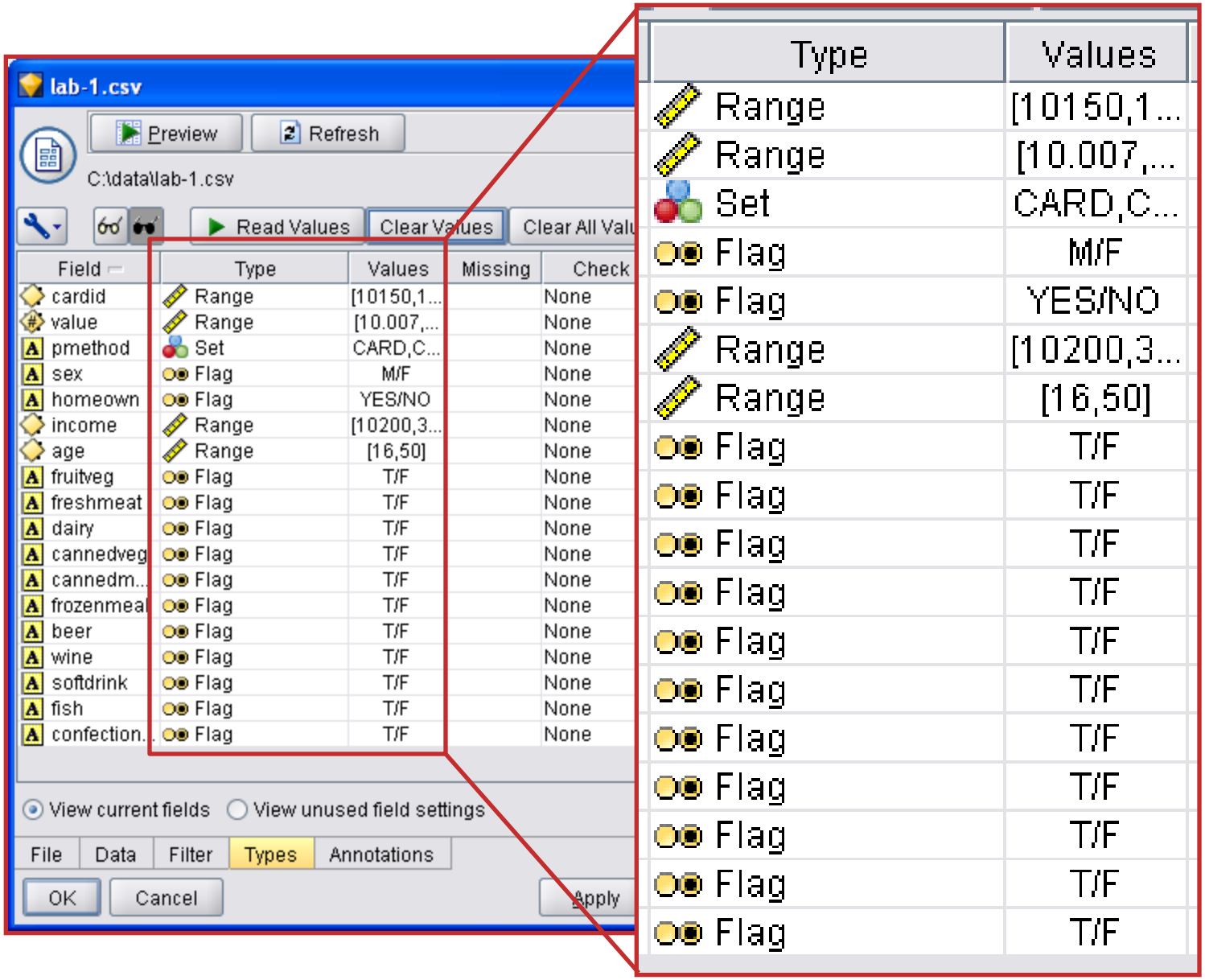


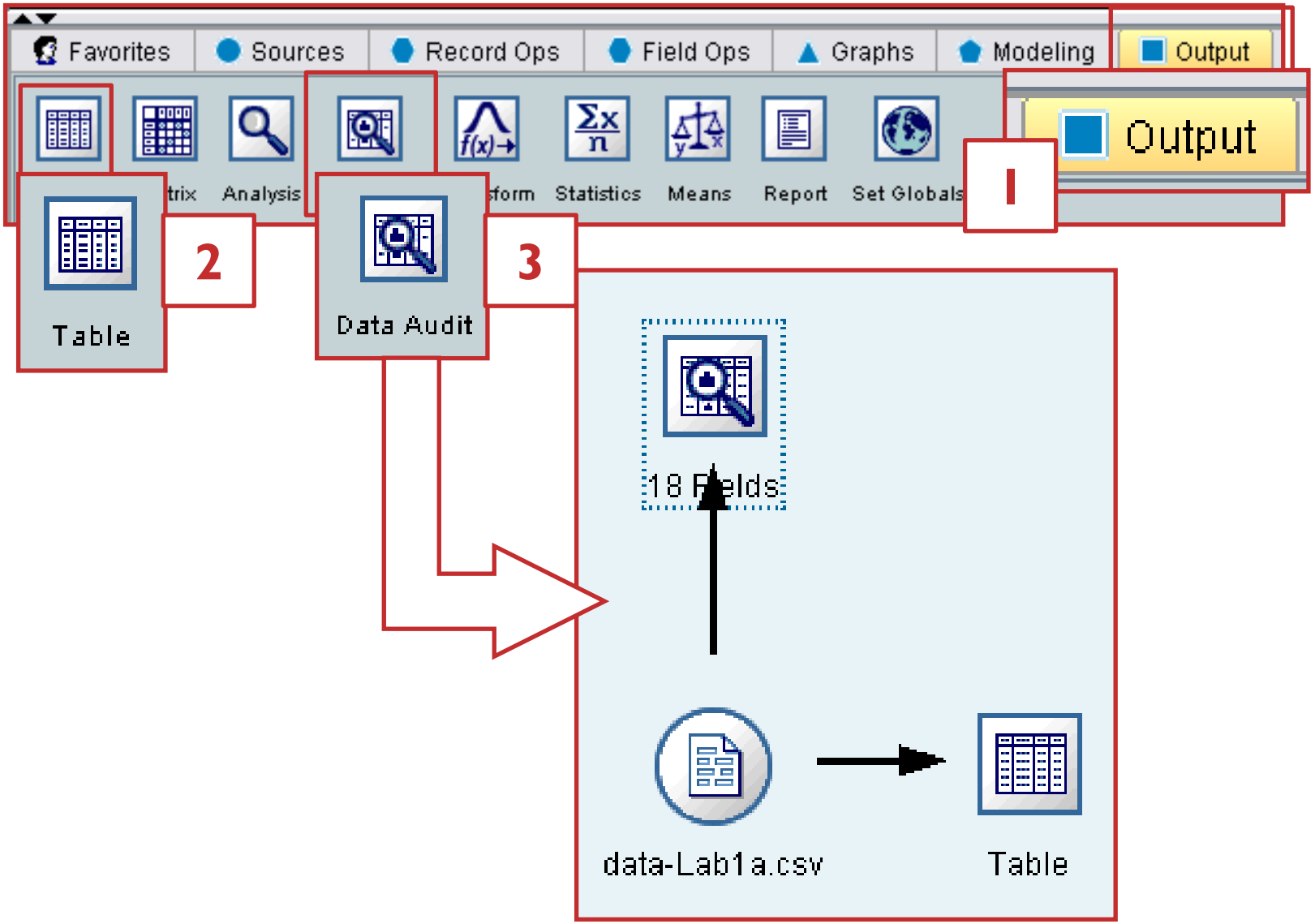
Download the data file from:
http://www4.comp.polyu.edu.hk/~csamak/data/data-Lab1a.csv


Stream1* - PASW® Modeler 13
File Edit Insert View Tools SuperNode Window Help
Var. File
Double click
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export PASW® Statistics
Database Var. File Auto Data Prep Select Sample Aggregate Derive Type Filter Graphboard Auto Classifier Auto Numeric Auto Cluster Table Flat File Database
Server: Local Server
148MB / 250MB


Var. File
Preview
Refresh
(No current file selected)
File:
1
Read field names from file
Skip header characters:
Strip lead and trail spaces:
Invalid characters:
Encoding:
Delimiters
Space
Newline
Allow multiple blank delimiter
File
Data
Filter
Types
OK
Cancel
Open
Look In:
Data
data-drug.csv
data-Lab1a.csv
2
data-Lab1a.csv
File Name:
data-Lab1a.csv
Files of Type:
All Files
3
Open


Var. File
Preview
Refresh
C:\data\lab-1.csv
C:\data\lab-1.csv
cardid,value,pmethod,sex,homeown,income,age,f
39808,42.7123,CHEQUE,M,NO,27000,46,F,T,T,F,F,
67362,25.3567,CASH,F,NO,30000,28,F,T,F,F,F,F,
10872,20.6176,CASH,M,NO,13200,36,F,F,F,T,F,T,
Read field nam...
Specify number of field
Skip header ...
0
EOL comment characters:
Strip lead and tr...
None
Left
Right
Both
Invalid ...
Discard
Replace with
Encoding:
Stream default
Decimal symbol:
Delimiters
Lines to scan for type:
S...
Comma
Tab
Automatically recognize
Ne...
Other
Quotes
Non-printing chara...
Single quote
Allow multi delimiter
Double quote
File
Data
Types
Annotations
I
Types
Var. File
3
2
Read Values
Clear All Values
Read Values
Clear Values
Clear All Values
Field
Type
Values
Missing
Check
Direction
cardid Range None In
value Range None In
pmethod Discrete None In
sex Discrete None In
homeown Discrete None In
income Range None In
age Range None In
fruitveg Discrete None In
freshmeat Discrete None In
dairy Discrete None In
cannedveg Discrete None In
cannedm... Discrete None In
frozenmeal Discrete None In
beer Discrete None In
wine Discrete None In
softdrink Discrete None In
fish Discrete None In
confection... Discrete None In
View current fields
View unused field settings
File
Data
Filter
Types
Annotations
OK
Cancel
Apply
Reset


lab-1.csv
Preview
Refresh
C:\data\lab-1.csv
Read Values
Clear Values
Clear All Valu
Field
Type
Values
Missing
Check
cardid Range [10150,1... None
value Range [10.007,... None
pmethod Set CARD,C... None
sex Flag M/F None
homeown Flag YES/NO None
income Range [10200,3... None
age Range [16,50] None
fruitveg Flag T/F None
freshmeat Flag T/F None
dairy Flag T/F None
cannedveg Flag T/F None
cannedm... Flag T/F None
frozenmeal Flag T/F None
beer Flag T/F None
wine Flag T/F None
softdrink Flag T/F None
fish Flag T/F None
confection... Flag T/F None
View current fields
View unused field settings
File
Data
Filter
Types
Annotations
OK
Cancel
Apply
Type
Values
Range [10150,1...
Range [10.007,...
Set CARD,C...
Flag M/F
Flag YES/NO
Range [10200,3...
Range [16,50]
Flag T/F
Flag T/F
Flag T/F
Flag T/F
Flag T/F
Flag T/F
Flag T/F
Flag T/F
Flag T/F
Flag T/F


Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Analysis
Statistics
Means
Report
Set Globals
Output
1
2
Table
3
Data Audit
18 Fields
data-Lab1a.csv
Table

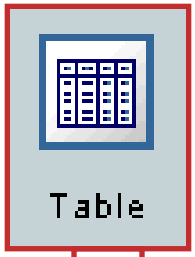


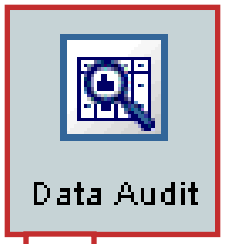
Table (18 fields, 1,000 records)

File Edit Generate

	cardid	value	pmethod	sex	homeown	income	age	fruitveg	freshmeat	
1	39808	42.712	CHEQUE	M	NO	27000	46	F	T	T
2	67362	25.357	CASH	F	NO	30000	28	F	T	F
3	10872	20.618	CASH	M	NO	13200	36	F	F	F
4	26748	23.688	CARD	F	NO	12200	26	F	F	T
5	91609	18.813	CARD	M	YES	11000	24	F	F	F
6	26630	46.487	CARD	F	NO	15000	35	F	T	F
	62995	14.047	CASH	F	YES	20800	30	T	F	F
	38765	22.203	CASH	M	YES	24400	22	F	F	F
9	28935	22.975	CHEQUE	F	NO	29500	46	T	F	F
10	41792	14.569	CASH	M	NO	29600	22	T	F	F
11	59480	10.328	CASH	F	NO	27100	18	T	T	T
12	60755	13.780	CASH	F	YES	20000	48	T	F	F
13	70998	36.509	CARD	M	YES	27300	43	F	F	T
14	80617	10.201	CHEQUE	F	YES	28000	43	F	F	F
15	61144	10.374	CASH	F	NO	27400	24	T	F	T
16	36405	34.822	CHEQUE	F	YES	18400	19	F	F	F
17	76567	42.248	CARD	M	YES	23100	31	T	F	F
18	85699	18.169	CASH	F	YES	27000	29	F	F	F
19	11357	10.753	CASH	F	YES	23100	26	F	F	F
20	97761	32.318	CARD	F	YES	25800	38	T	F	F

Table Annotations

OK



Data Audit of [18 fields]

File Edit Generate

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
ordid		Range	10150	109884	60405.704	29494.202	-0.010	--	1000
value		Range	10.007	49.886	29.398	11.555	0.027	--	1000
pmethod		Set	--	--	--	--	--	3	1000
sex		Flag	--	--	--	--	--	2	1000
homeown		Flag	--	--	--	--	--	2	1000
income		Range	10200	30000	20171.800	5767.558	0.005	--	1000
age		Range	16	50	32.840	10.343	0.040	--	1000
fruitveg		Flag	--	--	--	--	--	2	1000

[1] Indicates a multimode result [2] Indicates a sampled result

Audit Quality Annotations

OK

Quality

Data Audit of [18 fields]

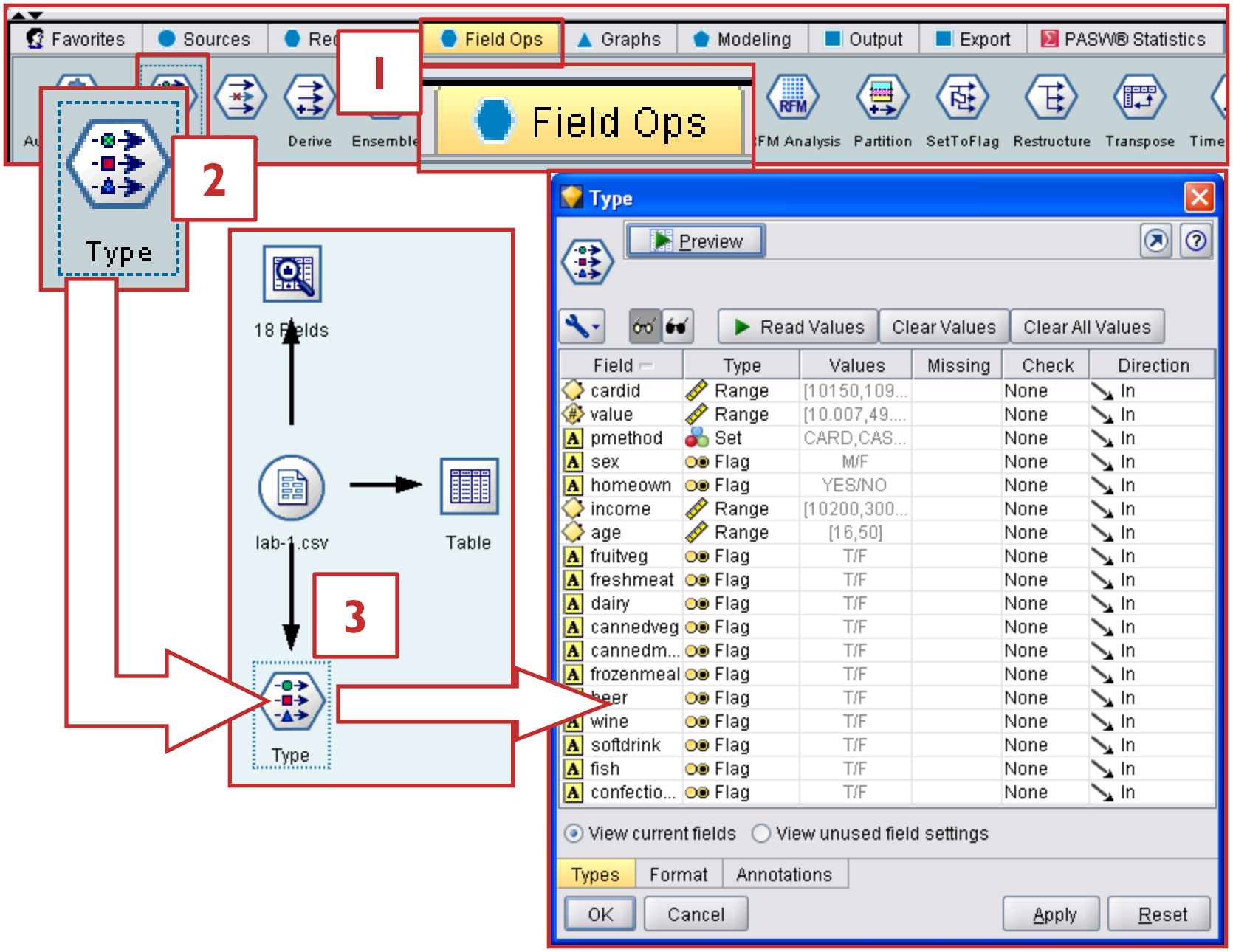
File Edit Generate

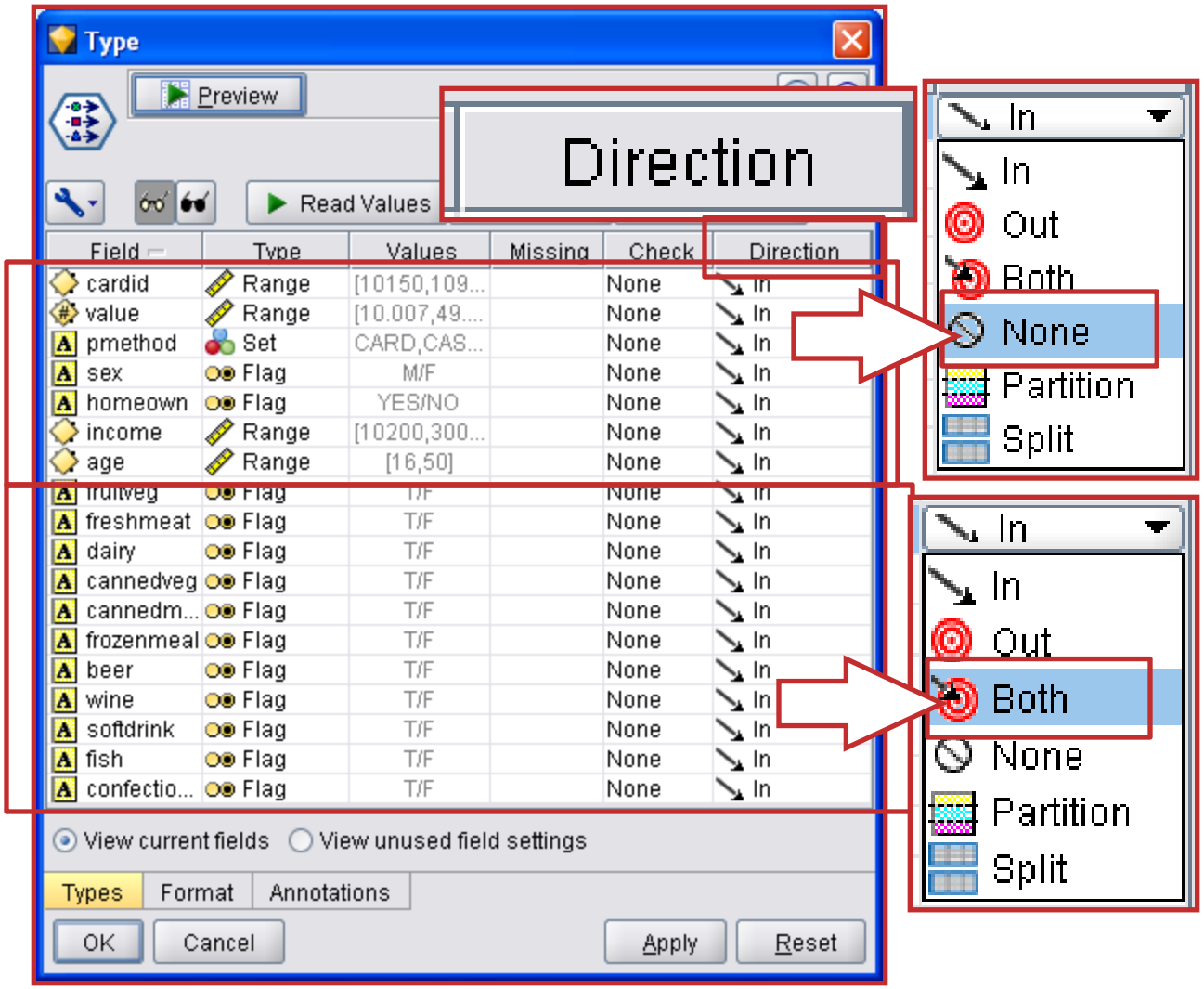
Complete fields (%): 1.0 Complete records (%): 1.0

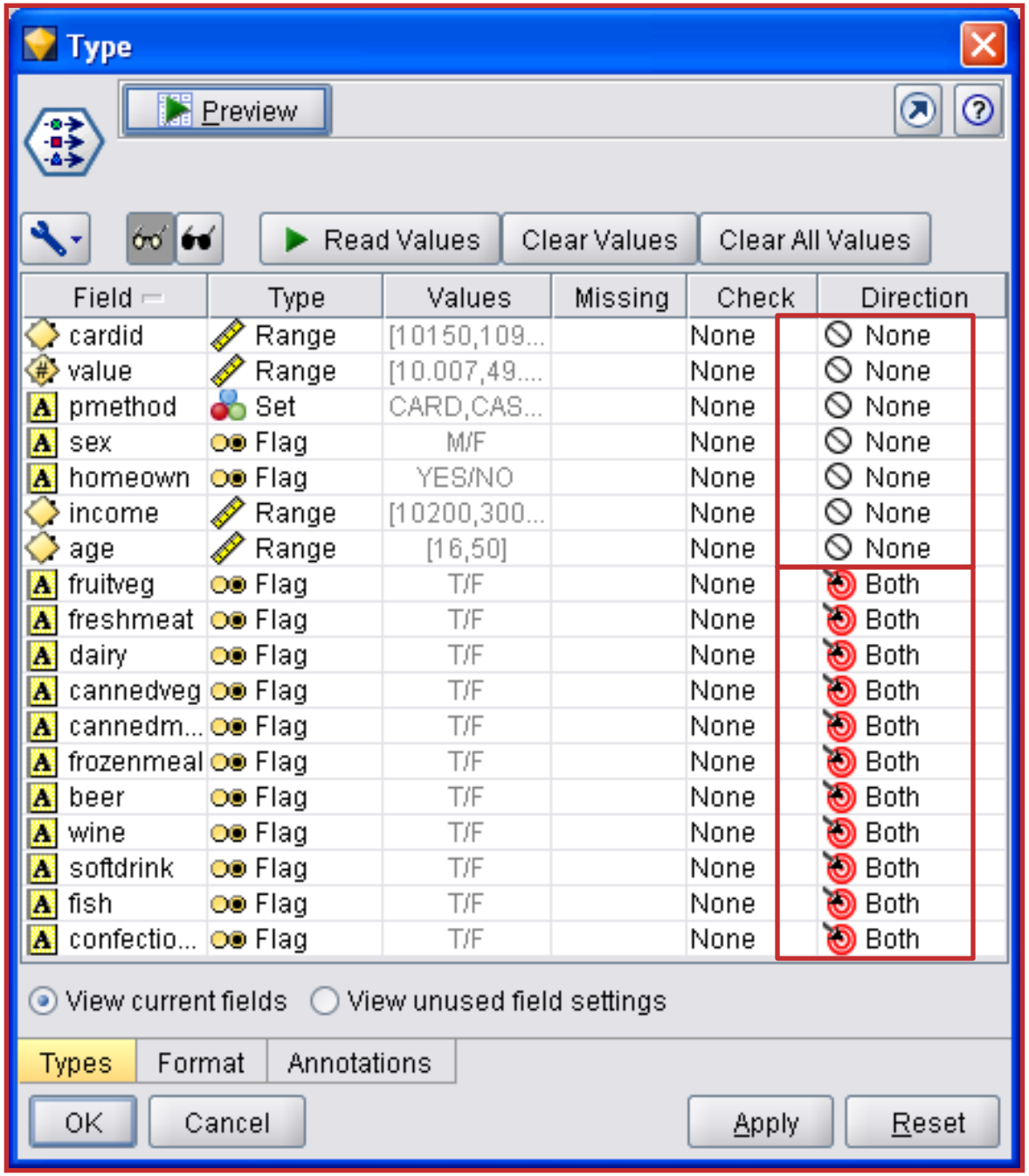
Field	Type	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value
cardid	Range	0	0	None	Never	Fixed	100	1000	0
value	Range	0	0	None	Never	Fixed	100	1000	0
pmethod	Set	--	--	--	Never	Fixed	100	1000	0
sex	Flag	--	--	--	Never	Fixed	100	1000	0
homeown	Flag	--	--	--	Never	Fixed	100	1000	0
income	Range	0	0	None	Never	Fixed	100	1000	0
age	Range	0	0	None	Never	Fixed	100	1000	0
fruitveg	Flag	--	--	--	Never	Fixed	100	1000	0
freshmeat	Flag	--	--	--	Never	Fixed	100	1000	0
dairy	Flag	--	--	--	Never	Fixed	100	1000	0
cannedveg	Flag	--	--	--	Never	Fixed	100	1000	0
cannedmeat	Flag	--	--	--	Never	Fixed	100	1000	0
frozenmeal	Flag	--	--	--	Never	Fixed	100	1000	0
beer	Flag	--	--	--	Never	Fixed	100	1000	0
wine	Flag	--	--	--	Never	Fixed	100	1000	0
softdrink	Flag	--	--	--	Never	Fixed	100	1000	0
fish	Flag	--	--	--	Never	Fixed	100	1000	0
confectionery	Flag	--	--	--	Never	Fixed	100	1000	0

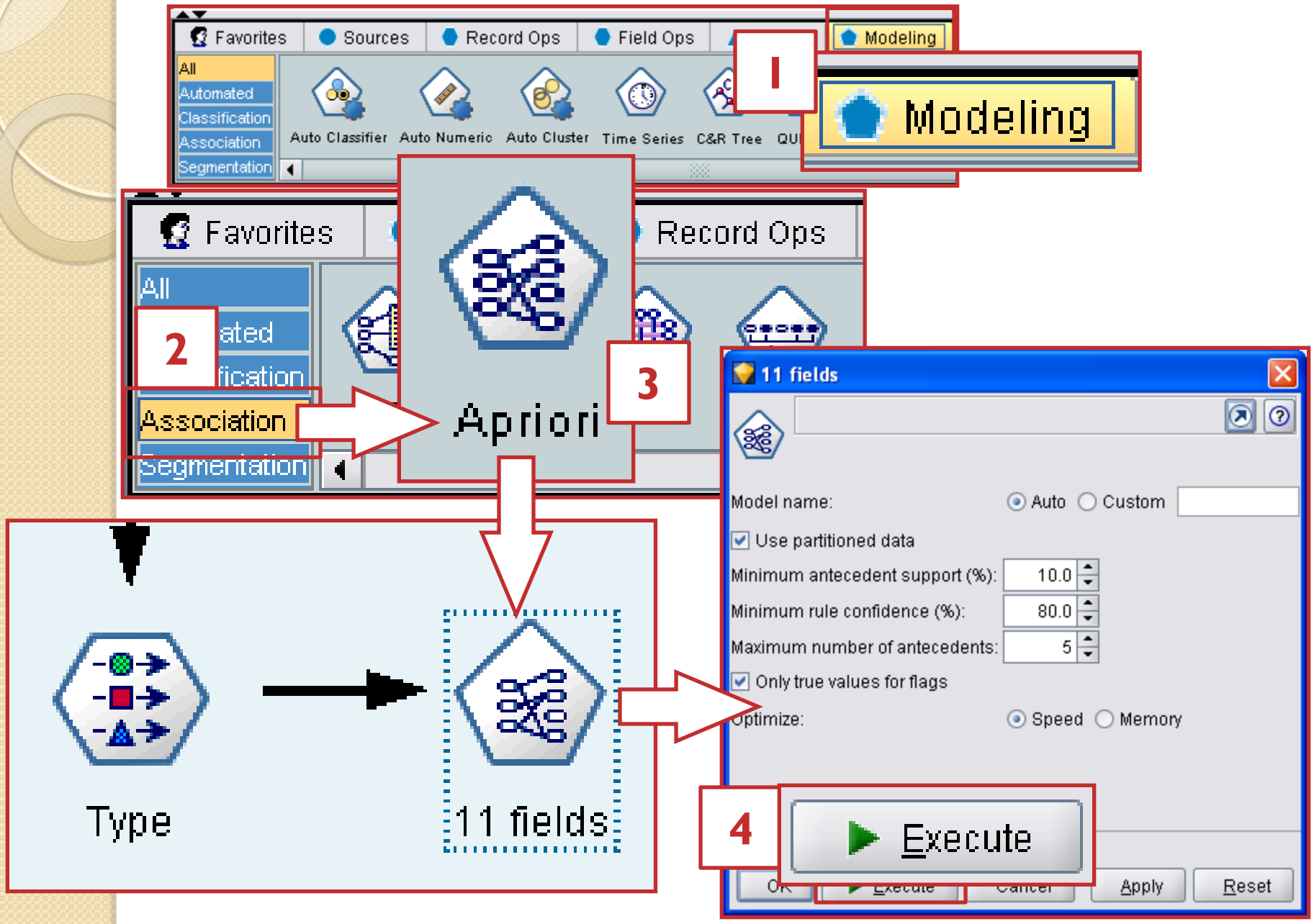
Audit Quality Annotations

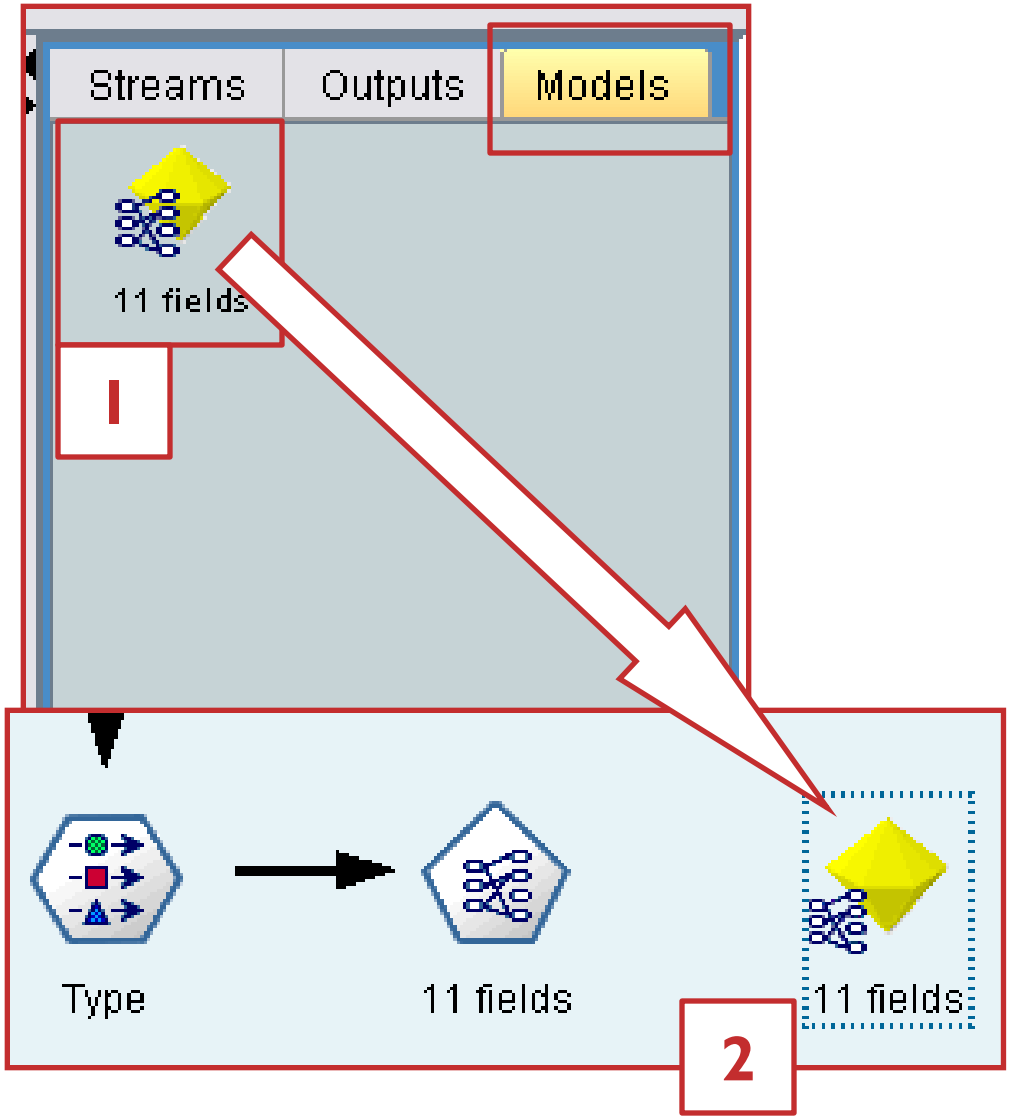
OK

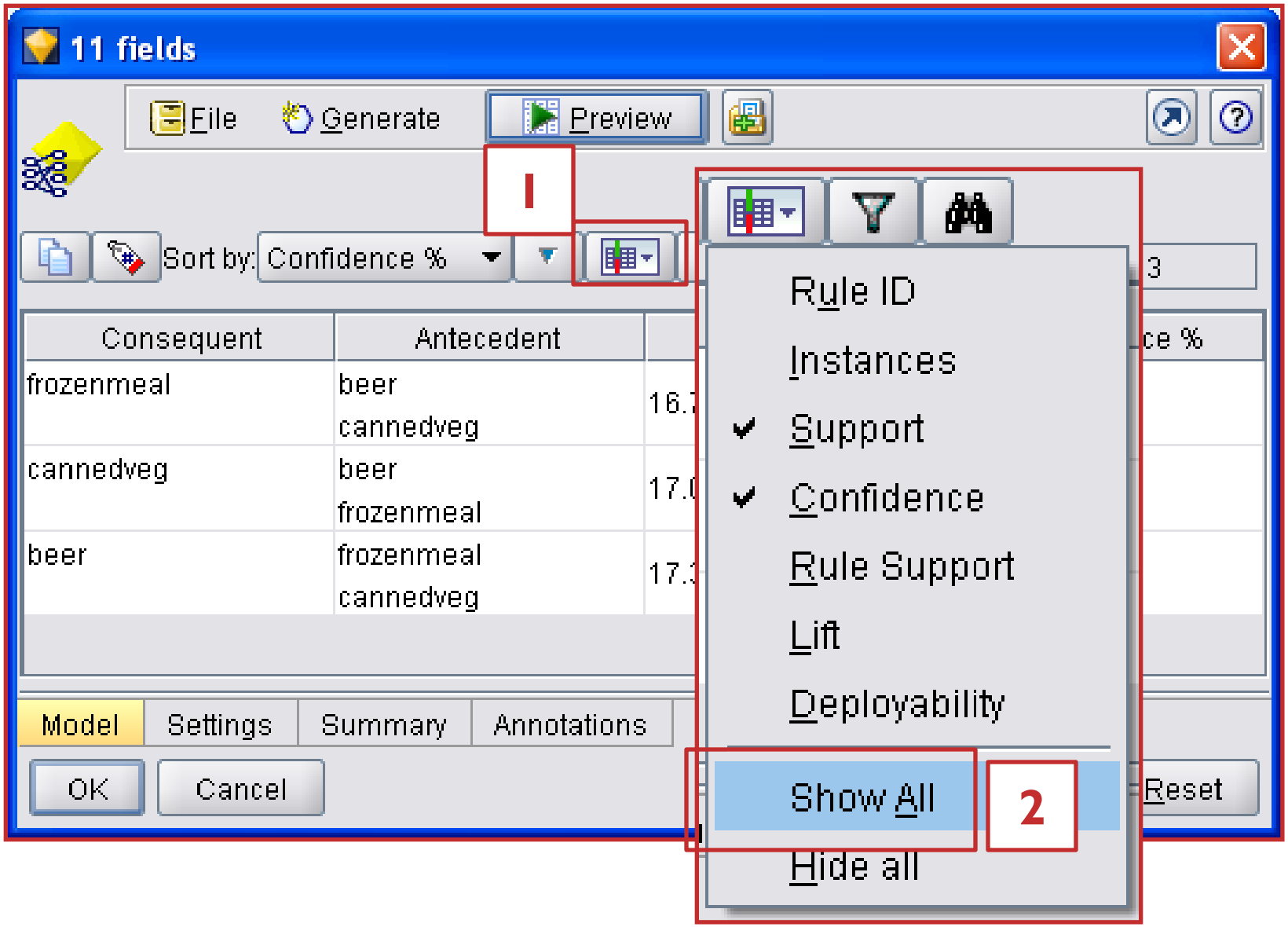

Favorites
Sources
Field Ops
Graphs
Modeling
Output
Export
PASW® Statistics
1
Field Ops
Derive
Ensemble
RFM Analysis
Partition
SetToFlag
Restructure
Transpose
Time
2
Type
18 Fields
lab-1.csv
Table
3
Type
Type
Preview
Read Values
Clear Values
Clear All Values
Field
Type
Values
Missing
Check
Direction
cardid Range [10150,109... None In
value Range [10.007,49.... None In
pmethod Set CARD,CAS... None In
sex Flag M/F None In
homeown Flag YES/NO None In
income Range [10200,300... None In
age Range [16,50] None In
fruitveg Flag T/F None In
freshmeat Flag T/F None In
dairy Flag T/F None In
cannedveg Flag T/F None In
cannedm... Flag T/F None In
frozenmeal Flag T/F None In
beer Flag T/F None In
wine Flag T/F None In
softdrink Flag T/F None In
fish Flag T/F None In
confectio... Flag T/F None In
View current fields
View unused field settings
Types
Format
Annotations
OK
Cancel
Apply
Reset

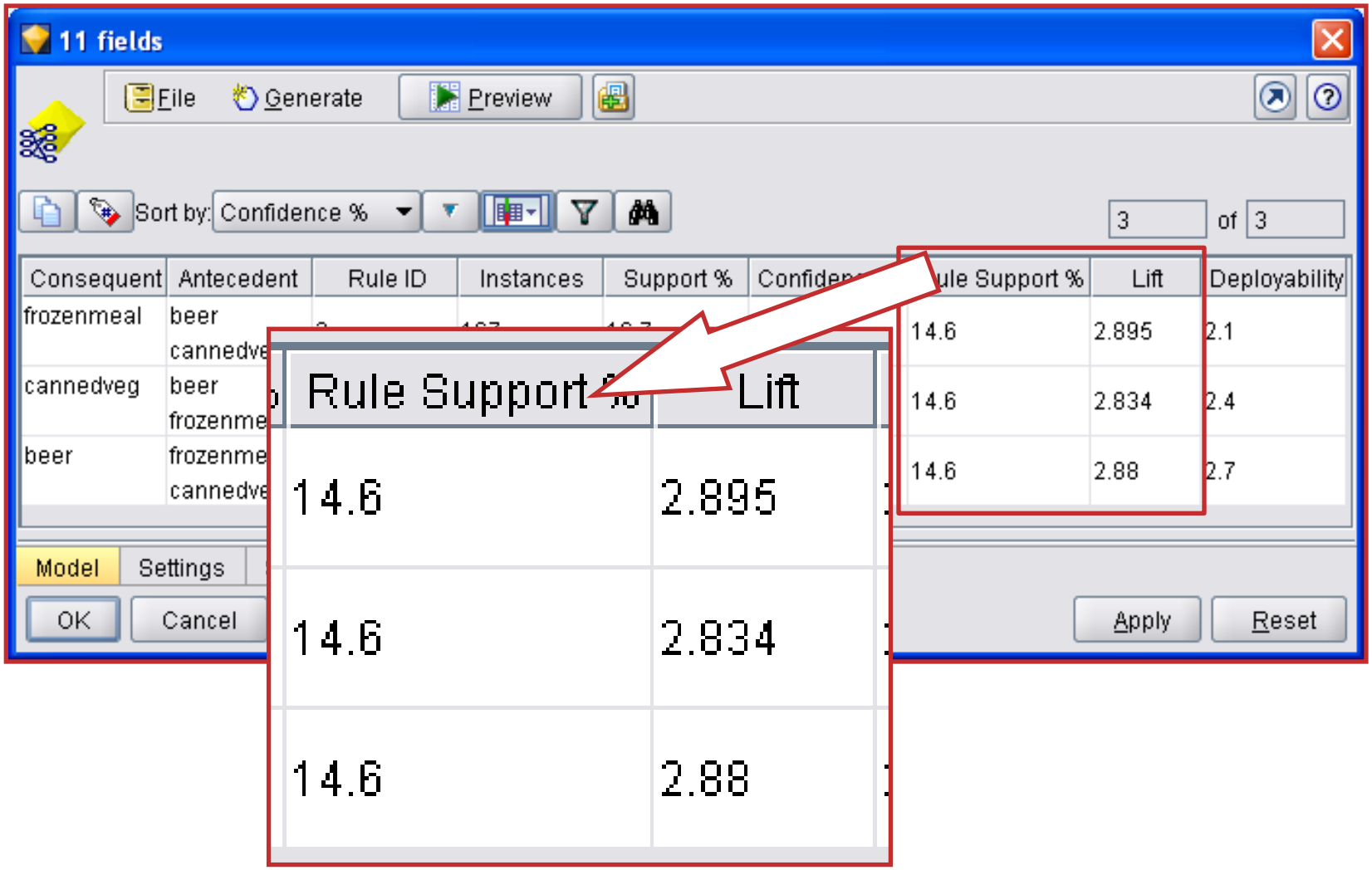

Type
Preview
Read Values
Direction
Field
Type
Values
Missing
Check
Direction
cardid Range [10150,109... None In
value Range [10.007,49.... None In
pmethod Set CARD,CAS... None In
sex Flag M/F None In
homeown Flag YES/NO None In
income Range [10200,300... None In
age Range [16,50] None In
fruitveg Flag T/F None In
freshmeat Flag T/F None In
dairy Flag T/F None In
cannedveg Flag T/F None In
cannedm... Flag T/F None In
frozenmeal Flag T/F None In
beer Flag T/F None In
wine Flag T/F None In
softdrink Flag T/F None In
fish Flag T/F None In
confectio... Flag T/F None In
View current fields
View unused field settings
Types
Format
Annotations
OK
Cancel
Apply
Reset
In
In
Out
Both
None
Partition
Split
In
In
Out
Both
None
Partition
Split


Type
Preview
Read Values
Clear Values
Clear All Values
Field
Type
Values
Missing
Check
Direction
cardid Range [10150,109... None None
value Range [10.007,49.... None None
pmethod Set CARD,CAS... None None
sex Flag M/F None None
homeown Flag YES/NO None None
income Range [10200,300... None None
age Range [16,50] None None
fruitveg Flag T/F None Both
freshmeat Flag T/F None Both
dairy Flag T/F None Both
cannedveg Flag T/F None Both
cannedm... Flag T/F None Both
frozenmeal Flag T/F None Both
beer Flag T/F None Both
wine Flag T/F None Both
softdrink Flag T/F None Both
fish Flag T/F None Both
confectio... Flag T/F None Both
View current fields
View unused field settings
Types
Format
Annotations
OK
Cancel
Apply
Reset


Favorites
Sources
Record Ops
Field Ops
Modeling
All
Automated
Classification
Association
Segmentation
Auto Classifier
Auto Numeric
Auto Cluster
Time Series
C&R Tree
1
Modeling
Favorites
Record Ops
All
2
Association
Segmentation
Apriori
3
11 fields
Model name:
Auto
Custom
Use partitioned data
Minimum antecedent support (%):
10.0
Minimum rule confidence (%):
80.0
Maximum number of antecedents:
5
Only true values for flags
Optimize:
Speed
Memory
4
Execute
Apply
Reset
Type
11 fields


Streams
Outputs
Models
11 fields
1
Type
11 fields
11 fields
2


11 fields
File
Generate
Preview
1
Sort by:
Confidence %
Consequent
Antecedent
frozenmeal
beer
cannedveg
cannedveg
beer
frozenmeal
beer
frozenmeal
cannedveg
Rule ID
Instances
Support
Confidence
Rule Support
Lift
Deployability
Show All
Hide all
2
Model
Settings
Summary
Annotations
OK
Cancel
Reset


11 fields
File
Generate
Preview
Sort by: Confidence %
3
of 3
Consequent
Antecedent
Rule ID
Instances
Support %
Confiden
ule Support %
Lift
Deployability
frozenmeal
beer
cannedve
14.6
2.895
2.1
cannedveg
beer
frozenme
14.6
2.834
2.4
beer
frozenme
cannedve
14.6
2.88
2.7
Model
Settings
OK
Cancel
Apply
Reset
Rule Support %
Lift
14.6
2.895
14.6
2.834
14.6
2.88

Try Yourself

- Can you see there are some differences from a normal rule set (Apriori)?
- Change the "Support" and "Confidence" settings.
- See what happens to the rule set.

Example 2 – data-Lab1b.csv

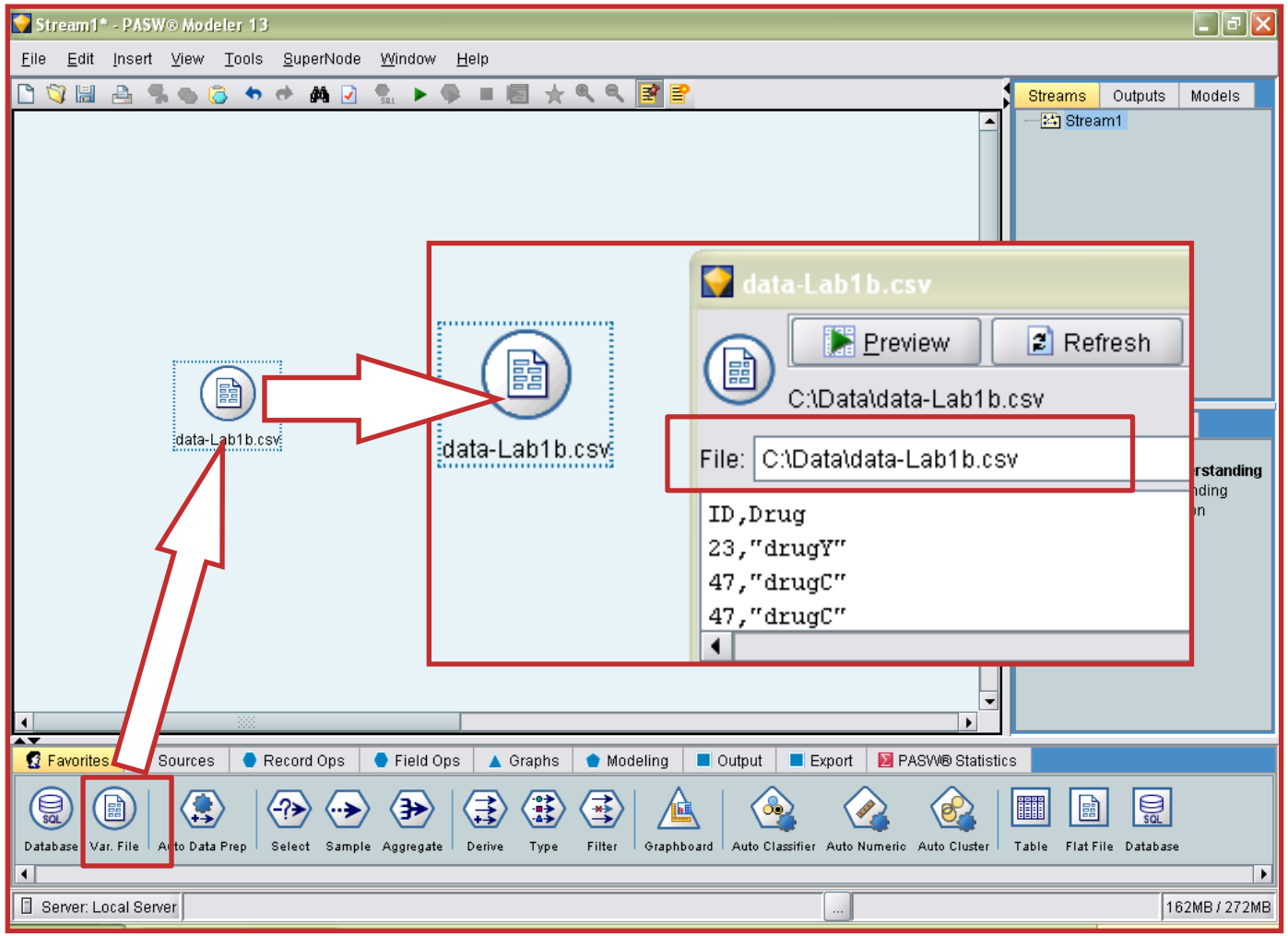
- Data Description:

Attributes	Description
ID	Transaction ID
Drug	Drug Type

- Total number of records: 200
- Total number of drugs: 5

Data Understanding

- Import data – Add node, "Var. File"

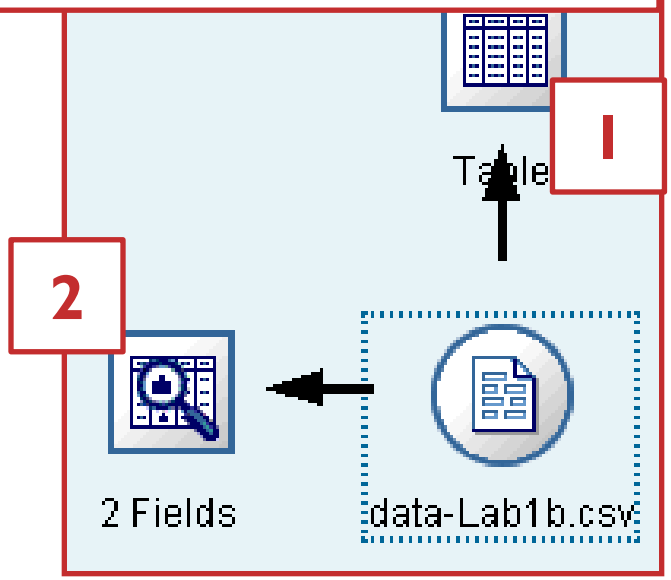


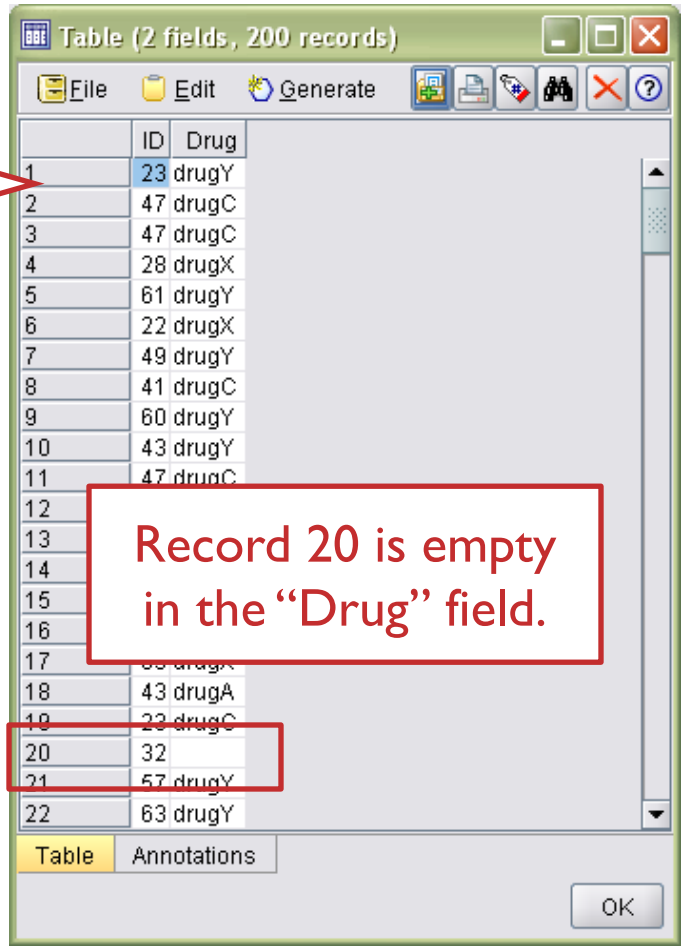
Data Understanding

- Analyze data distribution – add nodes: "Table" + "Data Audit"

3

Right click [Table] ➔ [Execute].





Record 20 is empty in the "Drug" field.

Data Understanding

- For "Data Audit",

Not all records are [Valid].

Data Audit of [ID Drug]

File Edit Generate

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
ID		Range	15	74	44.315	16.544	0.030	--	200
Drug		Set	--	--	--	--	--	13	192

[1] Indicates a multimode result [2] Indicates a sampled result

Audit Annotations

1

OK

File Edit Generate

Complete fields (%): 0.5 Complete records (%): 0.96

Field	Type	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value	Outliers	Extremes	Action
Drug	Set	96	192	0	8	8	0	--	--	--
ID	Range	100	200	0	0	0	0	0	0	None

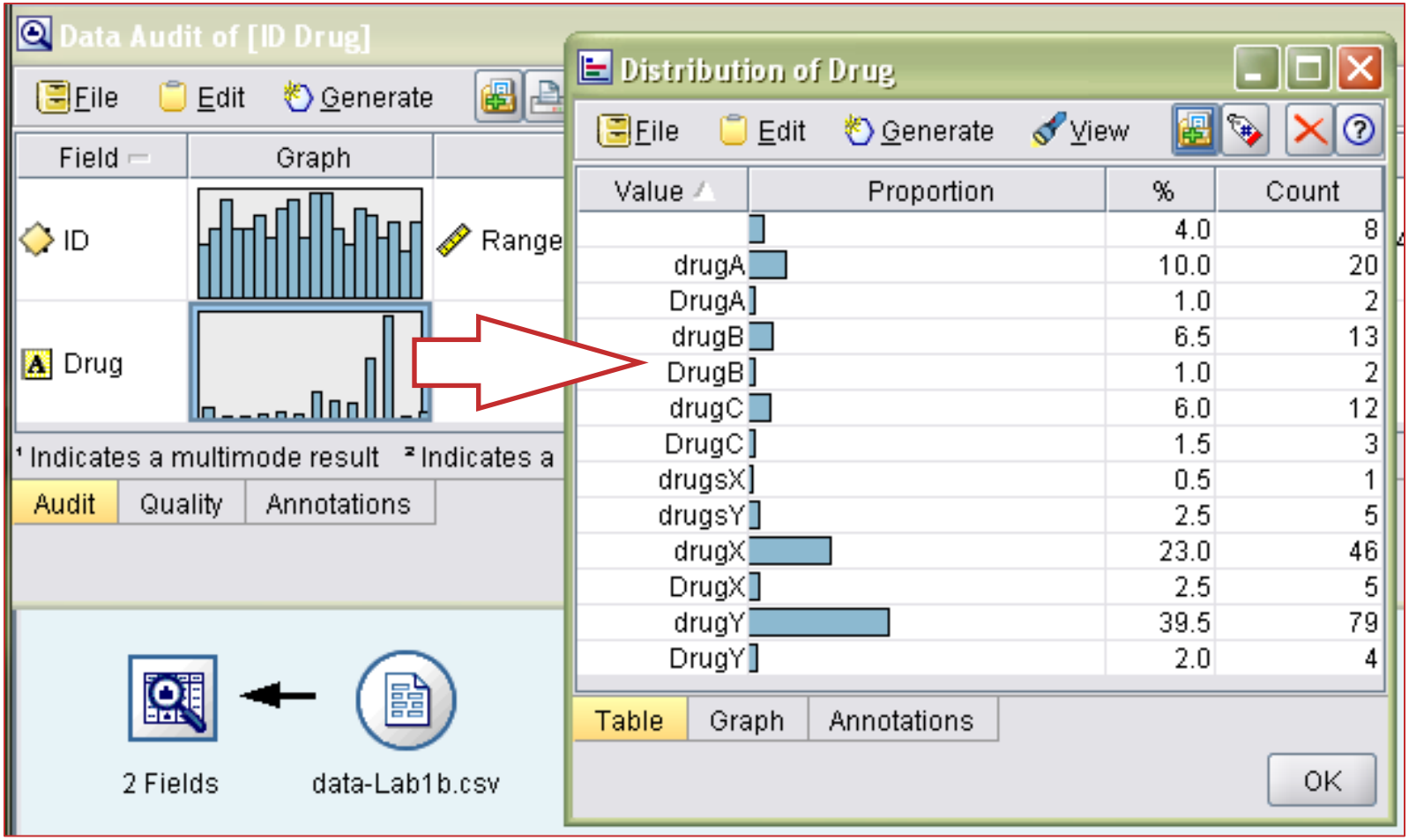
2

For the field, Drug, it is only 96% complete.

Audit Quality Annotations

OK

Data Understanding



Data Understanding

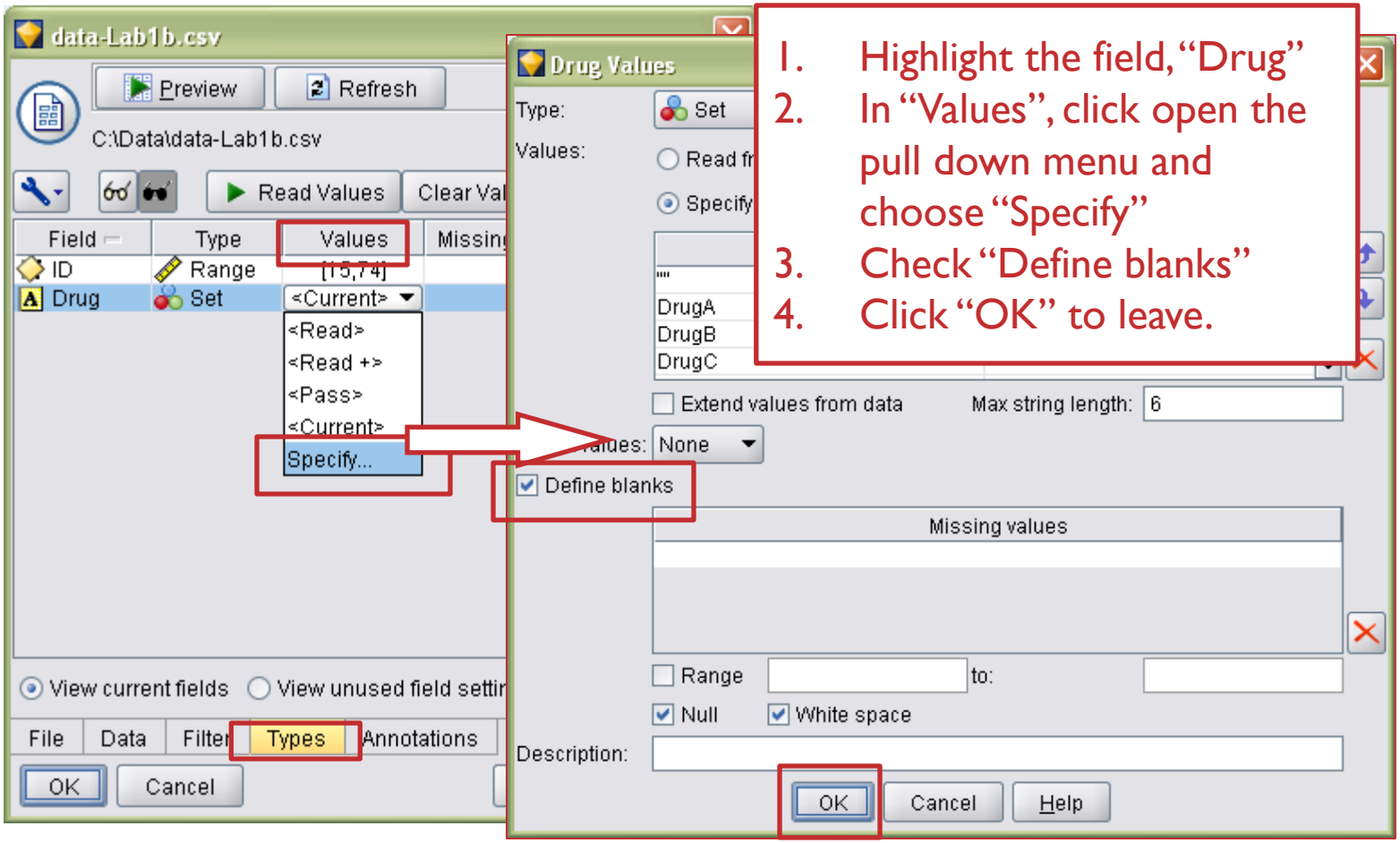
- Anything wrong with the dataset?
- How about the file format?

Data Preparation

- As we have found out there are something wrong with the dataset, it is time to correct them.
  - Null value
  - Inconsistent value
  - Inappropriate file format

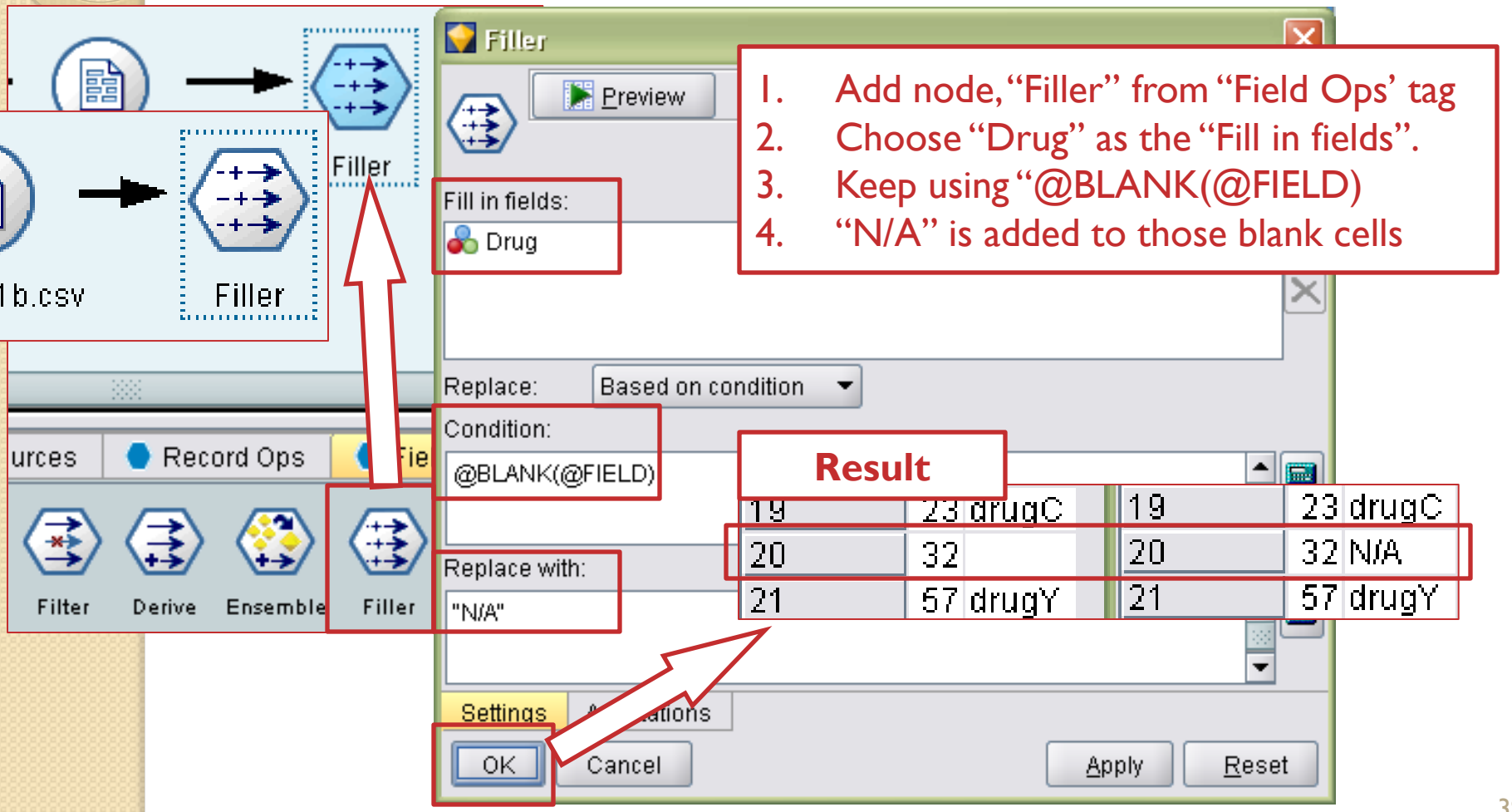
Data Preparation

- Define blanks and/or null values



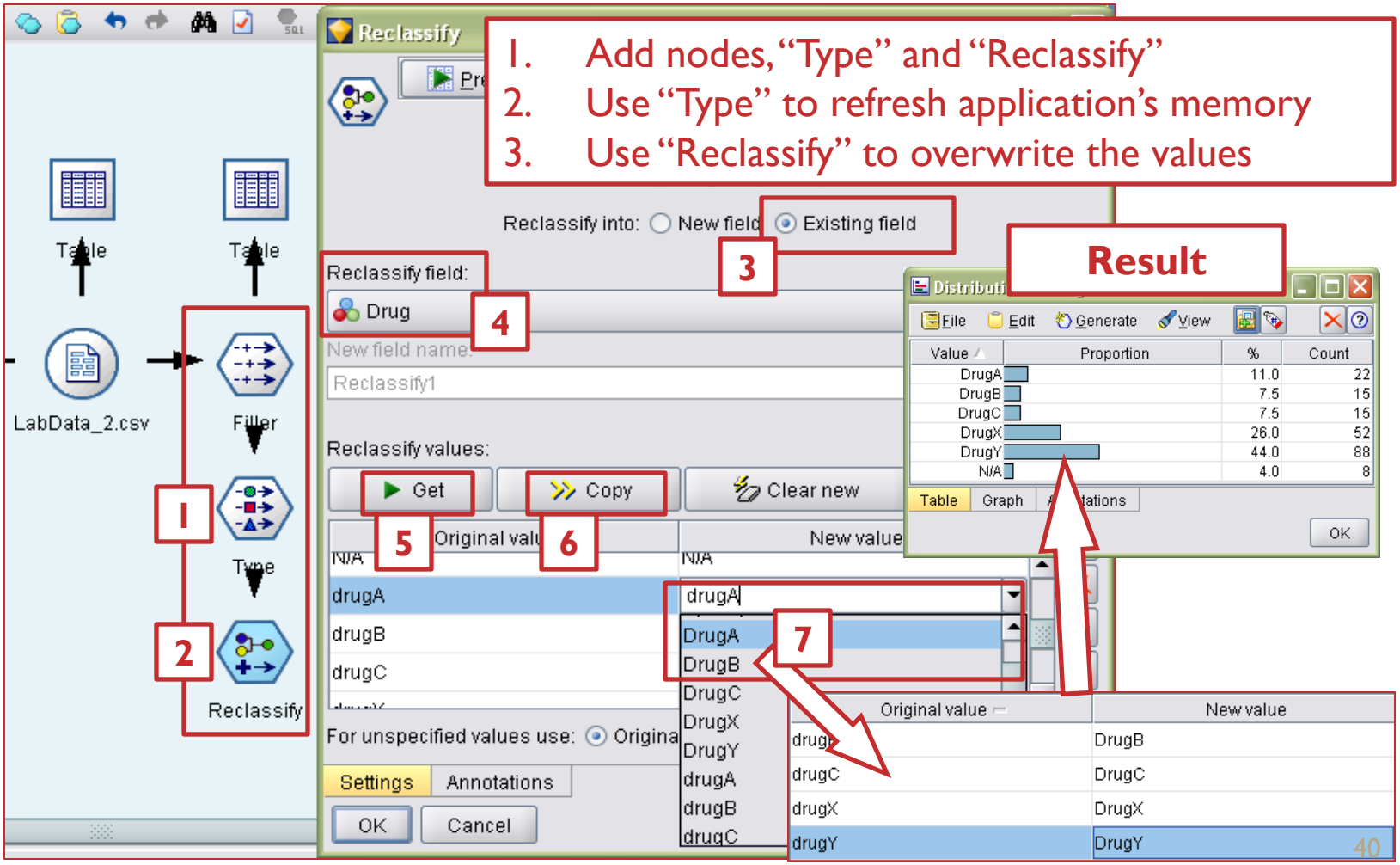
Data Preparation

- Replace all blanks and/or null values



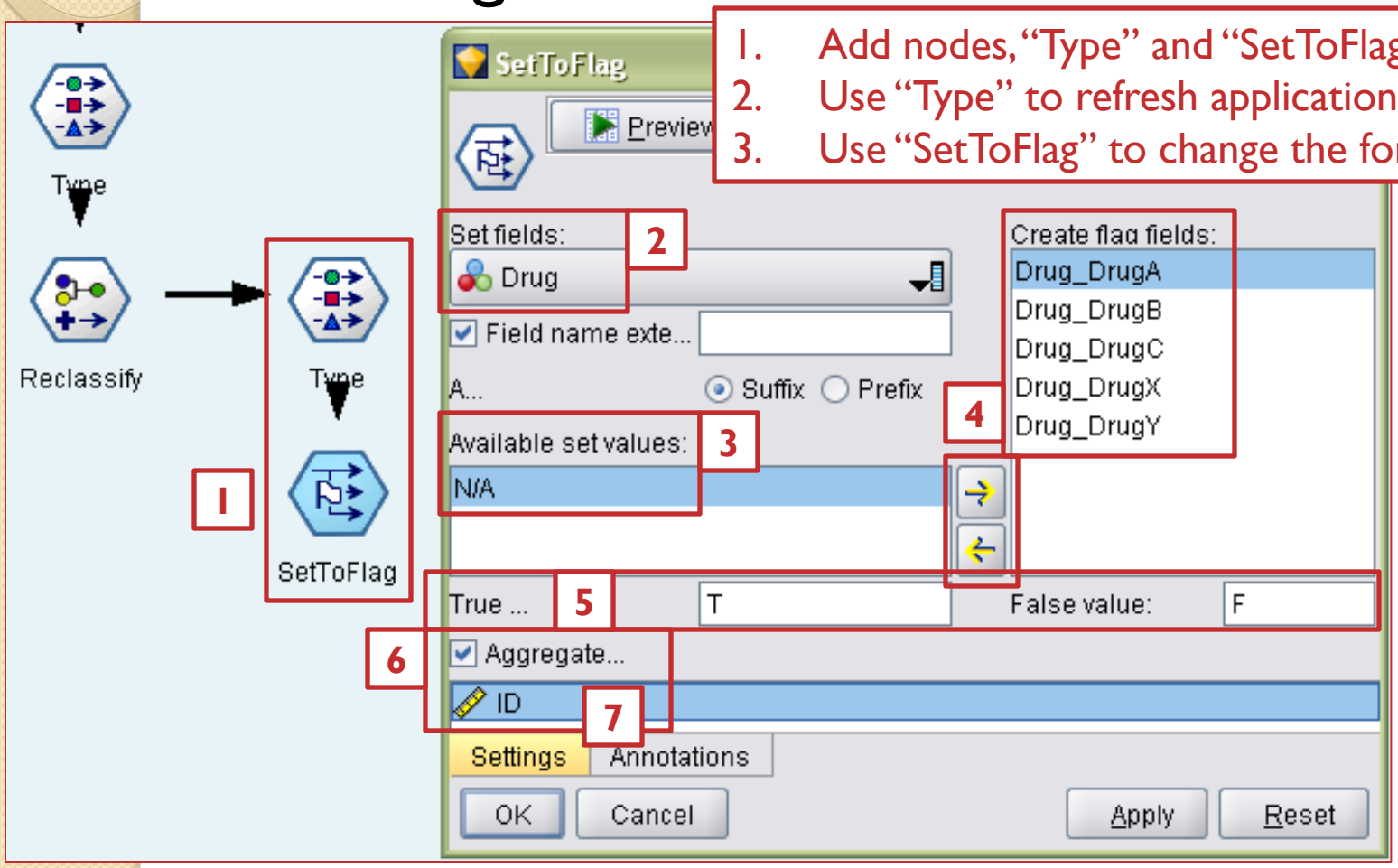
Data Preparation

- Correct inconsistent value



Data Preparation

- Change the file format



1. Add nodes, "Type" and "SetToFlag"
2. Use "Type" to refresh application's memory
3. Use "SetToFlag" to change the format

Data Preparation

- Finally the dataset becomes

Table (6 fields, 57 records)

File Edit Generate

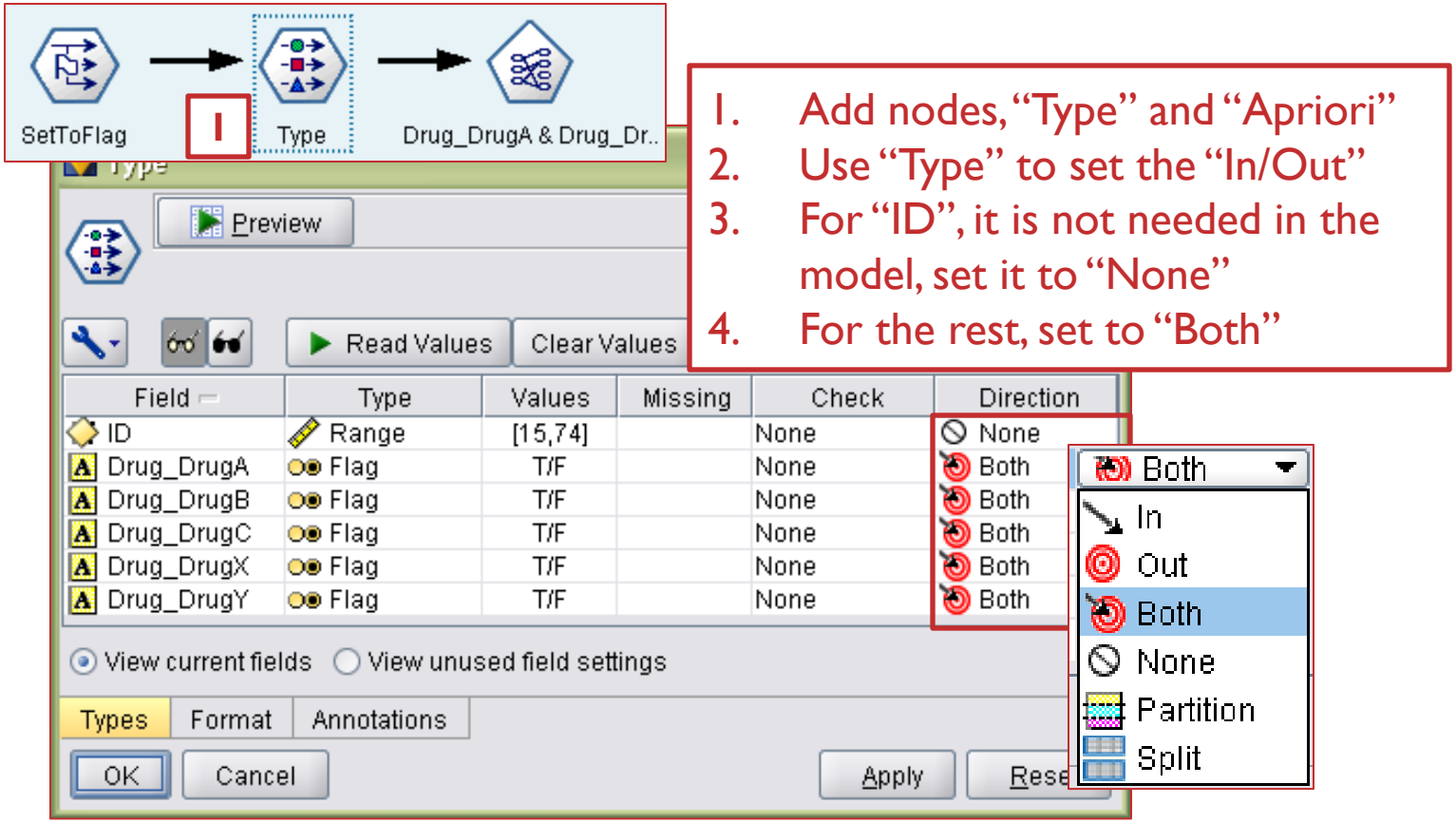
	ID	Drug_DrugA	Drug_DrugB	Drug_DrugC	Drug_DrugX	Drug_DrugY
1	25	F	F	F	F	T
2	21	F	F	F	F	T
3	30	F	F	F	T	F
4	55	F	T	F	F	F
5	52	F	F	F	T	T
6	35	T	F	F	T	F
7	72	F	T	T	T	T
8	20	T	F	F	T	T
9	56	F	F	T	T	T
10	42	T	F	F	F	T
11	51	F	T	F	T	T
12	59	F	T	T	T	F
13	64	F	F	F	T	T
14	19	T	F	F	F	T
15	36	T	F	F	T	T
16	70	F	T	F	F	T
17	54	F	F	F	F	T
18	17	F	F	F	T	F
19	29	T	F	F	F	T
20	38	T	F	F	F	T

Table Annotations

OK

Build Association Model

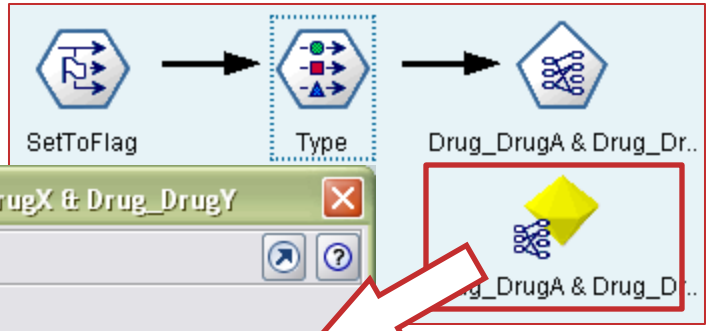
- For "Type" node,

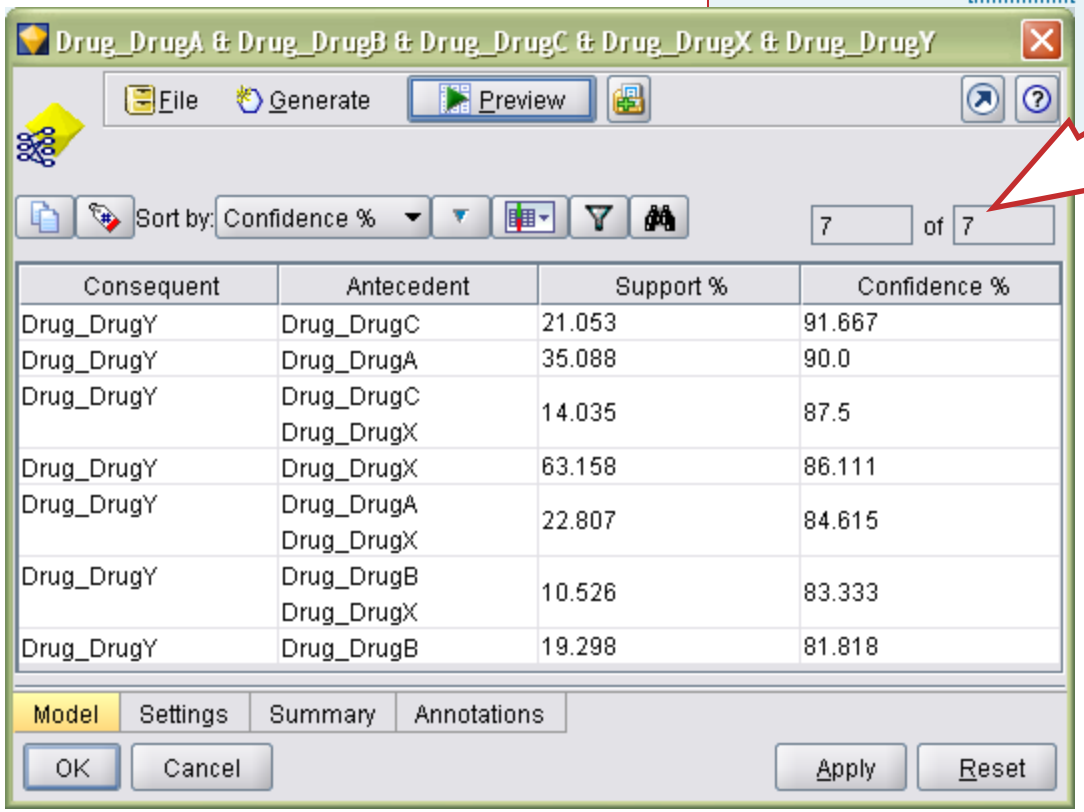


1. Add nodes, "Type" and "Apriori"
2. Use "Type" to set the "In/Out"
3. For "ID", it is not needed in the model, set it to "None"
4. For the rest, set to "Both"

Apriori Model

- For "Apriori", right click it and choose "Execute"





Consequent	Antecedent	Support %	Confidence %
Drug_DrugY	Drug_DrugC	21.053	91.667
Drug_DrugY	Drug_DrugA	35.088	90.0
Drug_DrugY	Drug_DrugC Drug_DrugX	14.035	87.5
Drug_DrugY	Drug_DrugX	63.158	86.111
Drug_DrugY	Drug_DrugA Drug_DrugX	22.807	84.615
Drug_DrugY	Drug_DrugB Drug_DrugX	10.526	83.333
Drug_DrugY	Drug_DrugB	19.298	81.818

Apriori Model

- View all

Drug_DrugA & Drug_DrugB & Drug_DrugC & Drug_DrugX & Drug_DrugY

File Generate Preview

Sort by: Confidence % 7 of 7

Consequent	Antecedent	Rule ID	Instances	Support %	Confidence...	Rule Support %	Lift	Deployability
Drug_DrugY	Drug_DrugC	2	12	21.053	91.667	19.298	1.025	1.754
Drug_DrugY	Drug_DrugA	3	20	35.088	90.0	31.579	1.006	3.509
Drug_DrugY	Drug_DrugC Drug_DrugX	6	8	14.035	87.5	12.281	0.978	1.754
Drug_DrugY	Drug_DrugX	4	36	63.158	86.111	54.386	0.962	8.772
Drug_DrugY	Drug_DrugA Drug_DrugX	7	13	22.807	84.615	19.298	0.946	3.509
Drug_DrugY	Drug_DrugB Drug_DrugX	5	6	10.526	83.333	8.772	0.931	1.754
Drug_DrugY	Drug_DrugB	1	11	19.298	81.818	15.789	0.914	3.509

Model Settings Summary Annotations

OK Cancel Apply Reset

Remarks

- Note that the tool (PASW Modeler) just helps you to generate/build models quickly.
- It does not give you the solution.
- How many rules in a model would be good enough? 3 or 7 or what?