Data Mining

SPSS Clementine 12.0

3. Exploratory Graphs

Spring 2010

Instructor: Dr. Masoud Yaghini

Outline

- Overview
- Reading in Text Data
- Adding a Table
- Creating a Distribution Graph
- Creating a Scatterplot
- Creating a Web Graph
- References

Overview

Drug Treatments

- For this section, imagine that you are a medical researcher compiling data for a study.
- You have collected data about a set of patients, all of whom suffered from the same illness.
- During their course of treatment, each patient responded to one of five medications.
- Part of your job is to use data mining to find out which drug might be appropriate for a future patient with the same illness.
- This example uses the data file named DRUG1n.

Drug Treatments

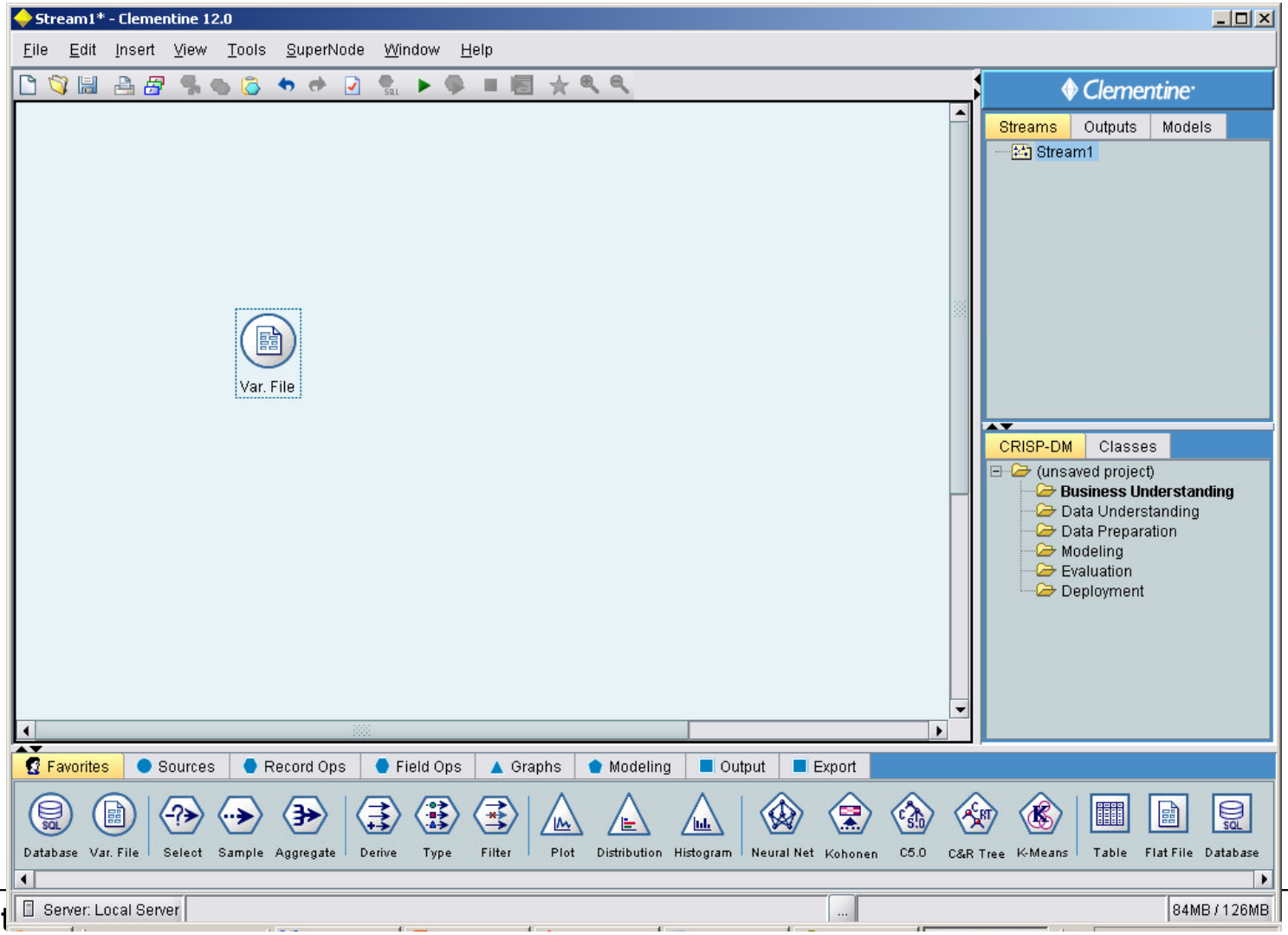
- The data fields used in the demo are:

Data field	Description
Age	(Number)
Sex	*M* or *F*
BP	Blood pressure: *HIGH*, *NORMAL*, or *LOW*
Cholesterol	Blood cholesterol: *NORMAL* or *HIGH*
Na	Blood sodium concentration
K	Blood potassium concentration
Drug	Prescription drug to which a patient responded

Reading in Text Data

Reading in Text Data

- Adding a Variable File node



Reading in Text Data

- Select the file called DRUG1n in Variable File dialog box

Var. File

Refresh

C:\Program Files\SPSSInc\Clementine12.0\Demos\DRUG1n

File: C:\Program Files\SPSSInc\Clementine12.0\Demos\DRUG1n

```
Age,Sex,BP,Cholesterol,Na,K,Drug
23,F,HIGH,HIGH,0.792535,0.031258,drugY
47,M,LOW,HIGH,0.739309,0.056468,drugC
47,M,LOW,HIGH,0.697269,0.068944,drugC
```

Read field names from file | Specify number of fields 1

Skip header characters: 0 | EOL comment characters:

Strip lead and trail spaces: None Left Right Both

Invalid characters: Discard Replace with

Encoding: Stream default | Decimal symbol: Stream default

Delimiters

Space Comma Tab

Newline Other

Non-printing characters

Allow multiple blank delimiters

Lines to scan for type: 50

Quotes

Single quotes: Discard

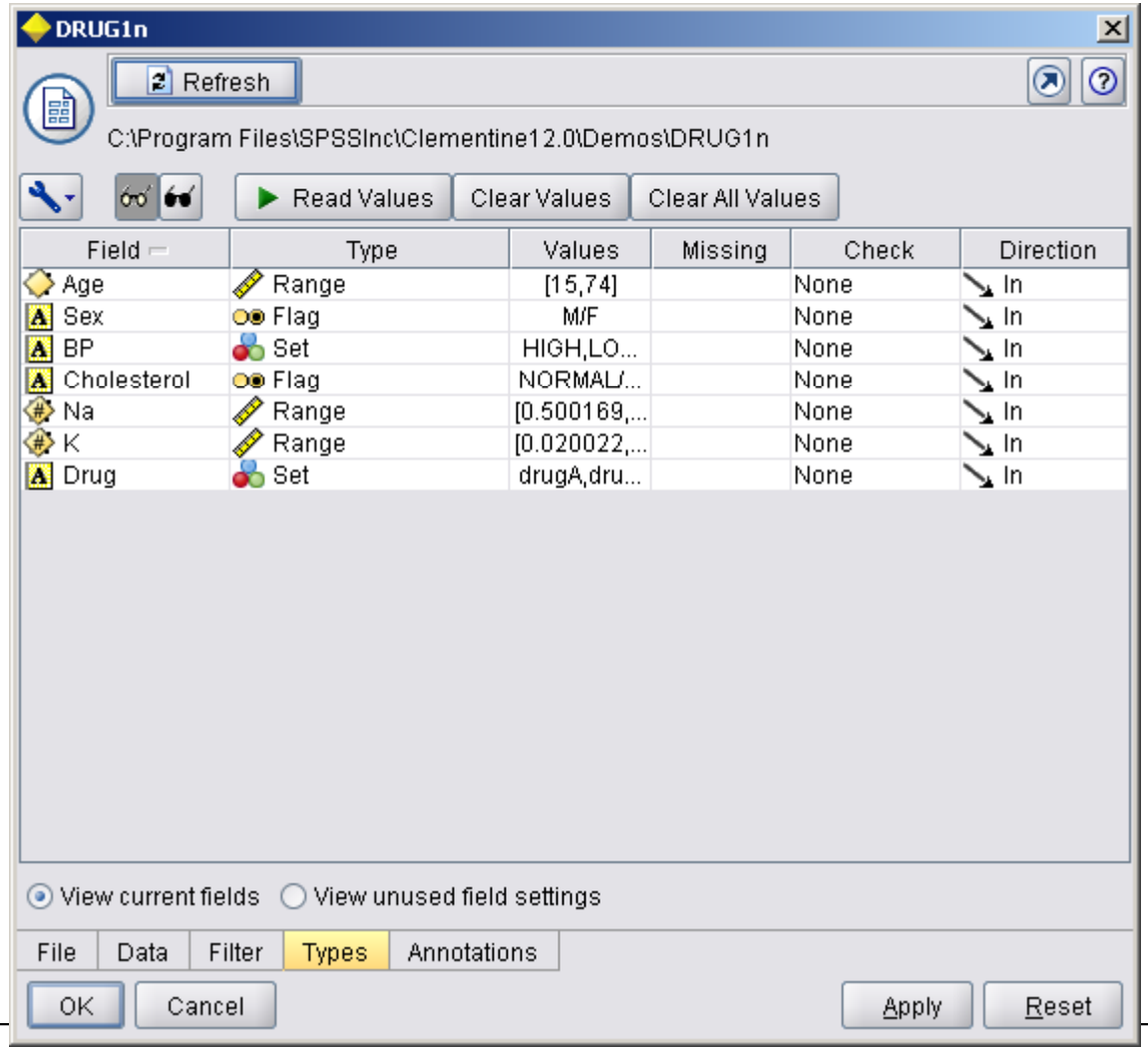
Double quotes: Discard

File Data Filter Types Annotations

OK Cancel Apply Reset

Reading in Text Data

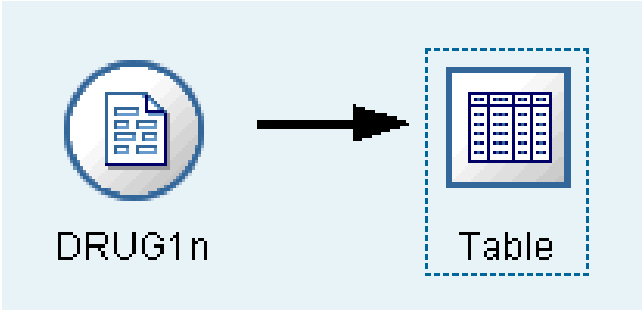
- Read Values to view the actual values in Types tab



Adding a Table

Adding a Table

- Add Table node to glance at the values for some of the records by Table node.
- To place a Table node in the stream, double-click the icon in the Output palette.



Adding a Table

- To view the table, right-click the Table node and choose Execute. Sort columns by clicking on the column header, or reorder columns using drag and drop.

Table (7 fields, 200 records)

File Edit Generate

	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

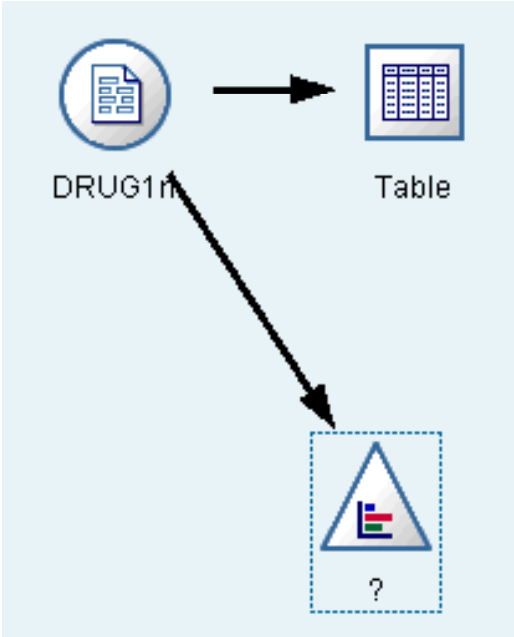
Table Annotations

OK

Creating a Distribution Graph

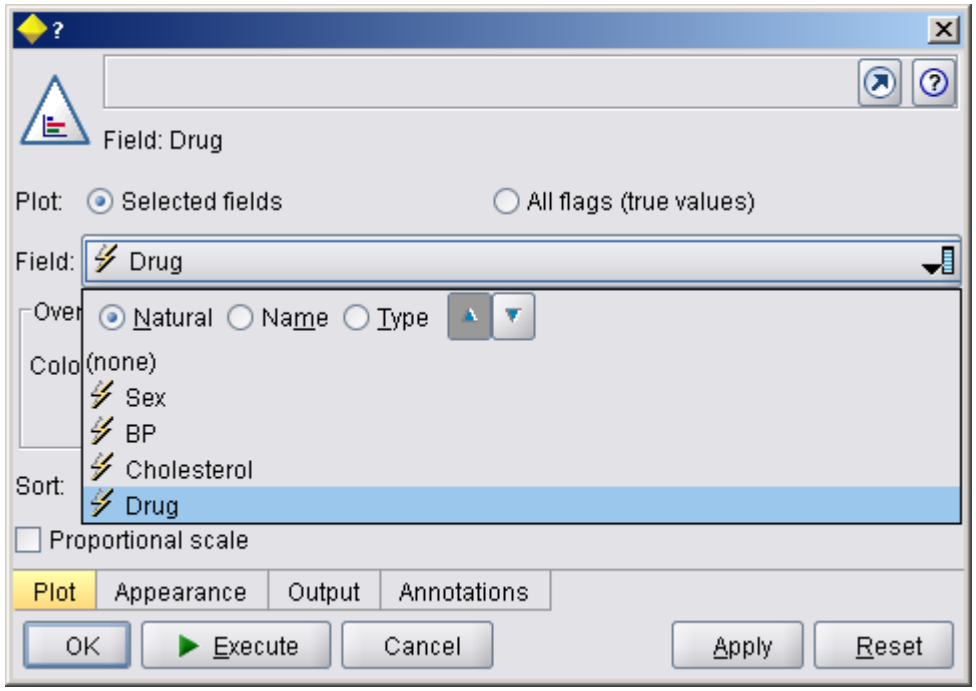
Creating a Distribution Graph

- For example, to find out what proportion of the patients responded to each drug, use a Distribution node.
- Add a Distribution node to the stream and connect it to the Source node, then double-click the node to edit options for display.



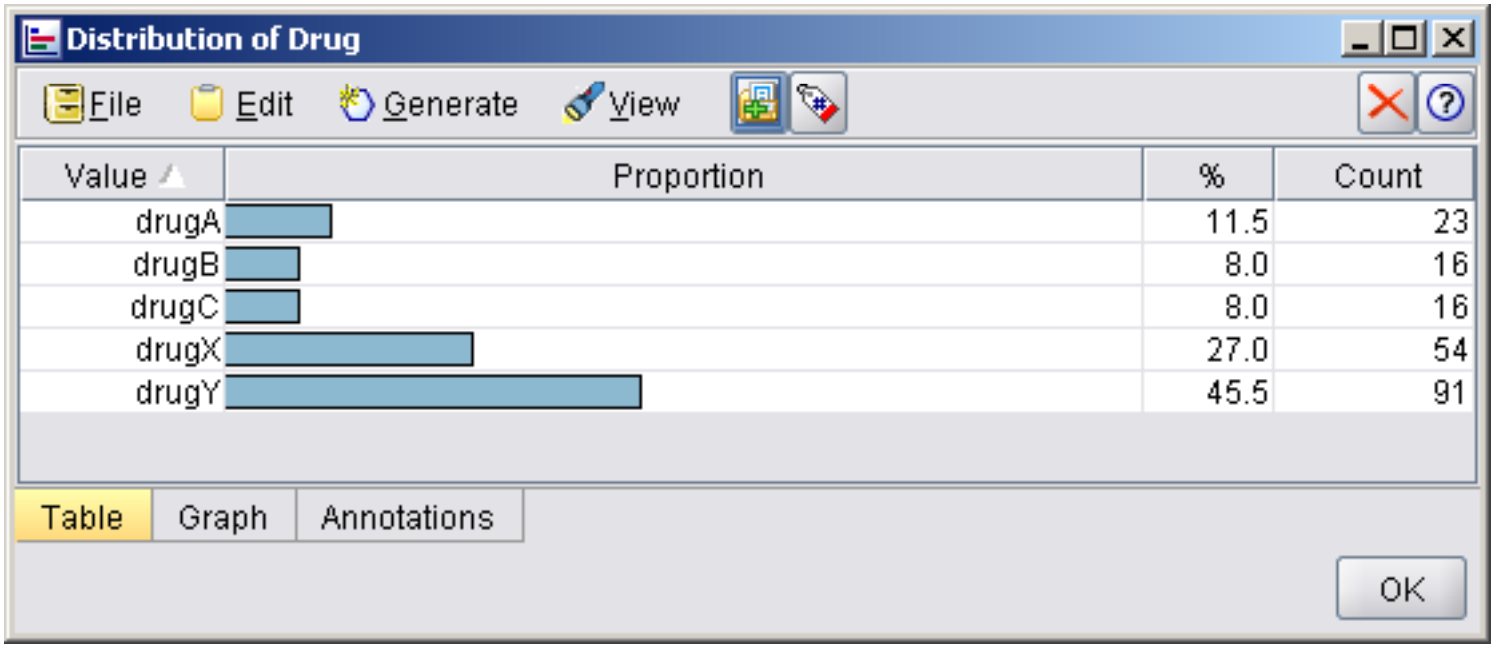
Creating a Distribution Graph

- Select Drug as the target field whose distribution you want to show.
- Then, click Execute from the dialog box.



Creating a Distribution Graph

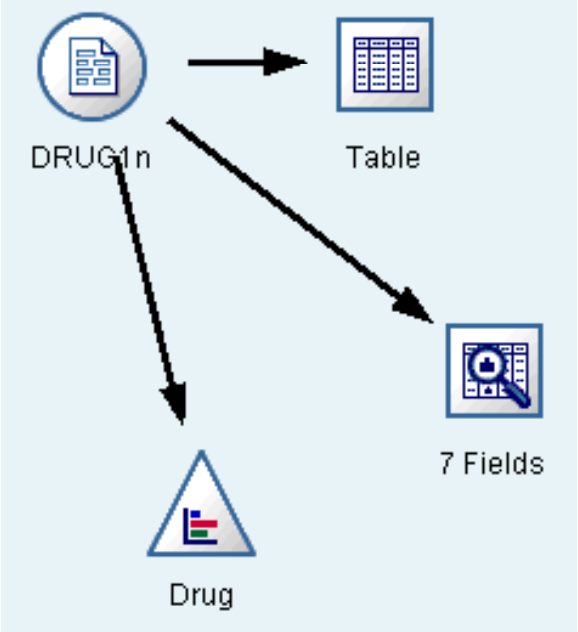
- The resulting graph helps you see the "shape" of the data.
- It shows that patients responded to drug Y most often and to drugs B and C least often.



Value	Proportion	%	Count
drugA		11.5	23
drugB		8.0	16
drugC		8.0	16
drugX		27.0	54
drugY		45.5	91

Creating a Data Audit node

- Alternatively, you can attach and execute a Data Audit node for a quick glance at distributions and histograms for all fields at once.
- The Data Audit node is available on the Output tab.



Creating a Distribution Graph

Data Audit of [7 fields]

File Edit Generate

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Age		Range	15	74	44.315	16.544	0.030	--	200
Sex		Discrete	--	--	--	--	--	2	200
BP		Discrete	--	--	--	--	--	3	200
Choleste...		Discrete	--	--	--	--	--	2	200
Na		Range	0.500	0.896	0.697	0.119	-0.074	--	200
K		Range	0.020	0.080	0.050	0.018	-0.039	--	200
Drug		Discrete	--	--	--	--	--	5	200

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

Audit Quality Annotations

OK

Clementine

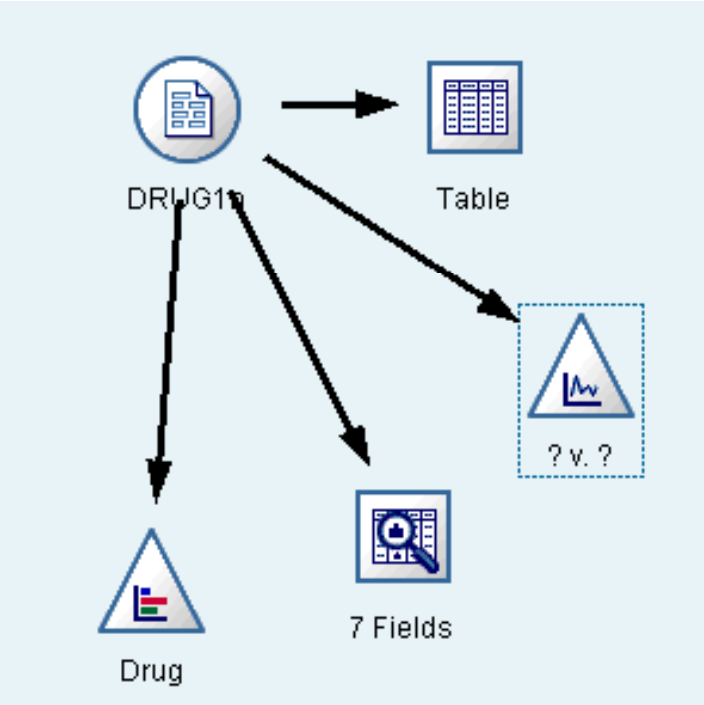
Creating a Scatterplot

Creating a Scatterplot

- Now let's take a look at what factors might influence Drug, the target variable.
- As a researcher, you know that the concentrations of sodium and potassium in the blood are important factors.
- Since these are both numeric values, you can create a scatterplot of sodium versus potassium, using the drug categories as a color overlay.

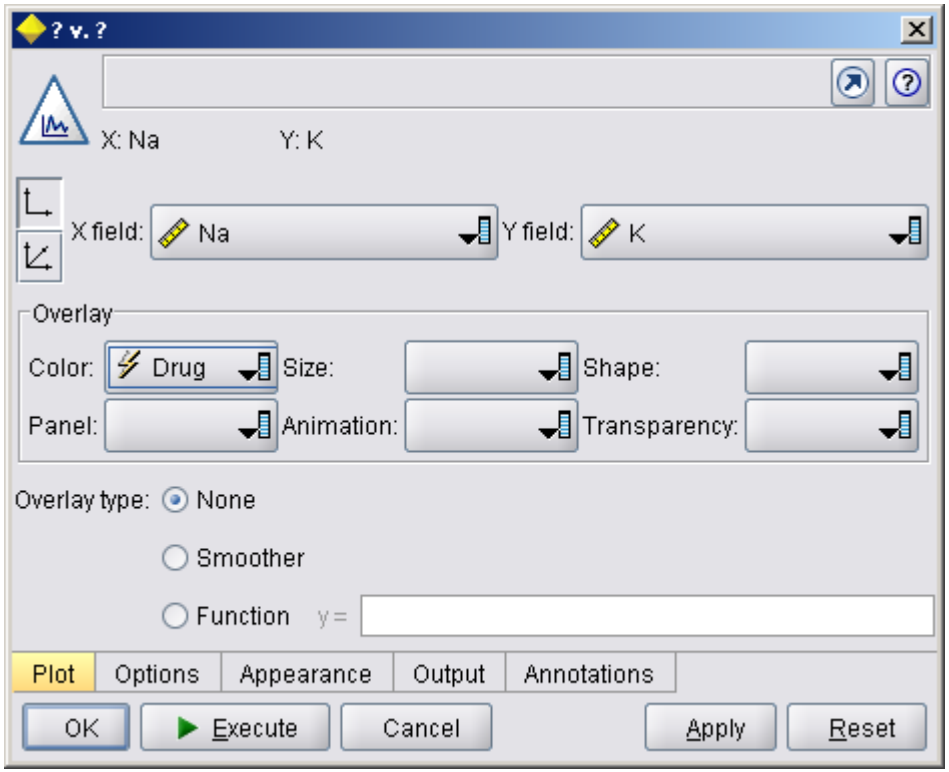
Creating a Scatterplot

- Place a Plot node in the workspace and connect it to the Source node, and double-click to edit the node.



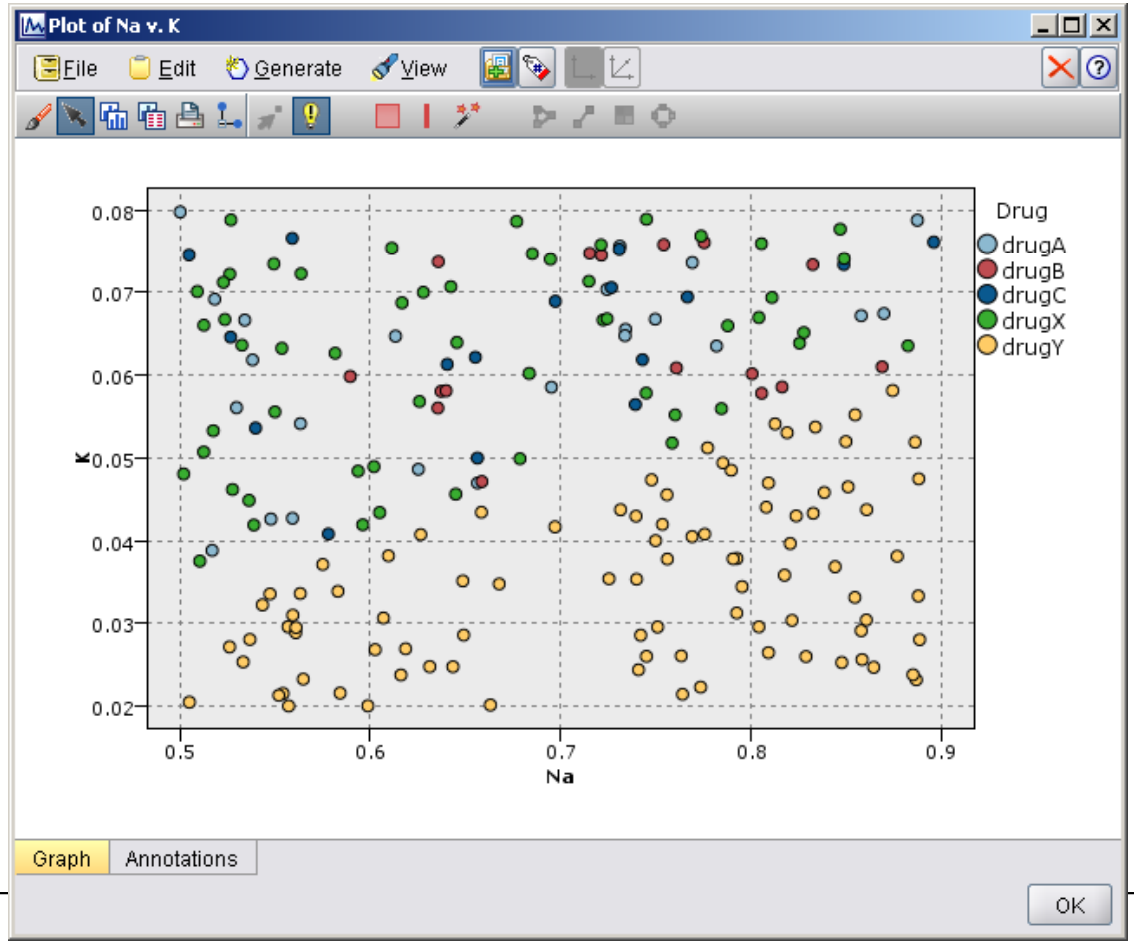
Creating a Scatterplot

- On the Plot tab, select Na as the X field, K as the Y field, and Drug as the overlay field.
- Then, click Execute.



Creating a Scatterplot

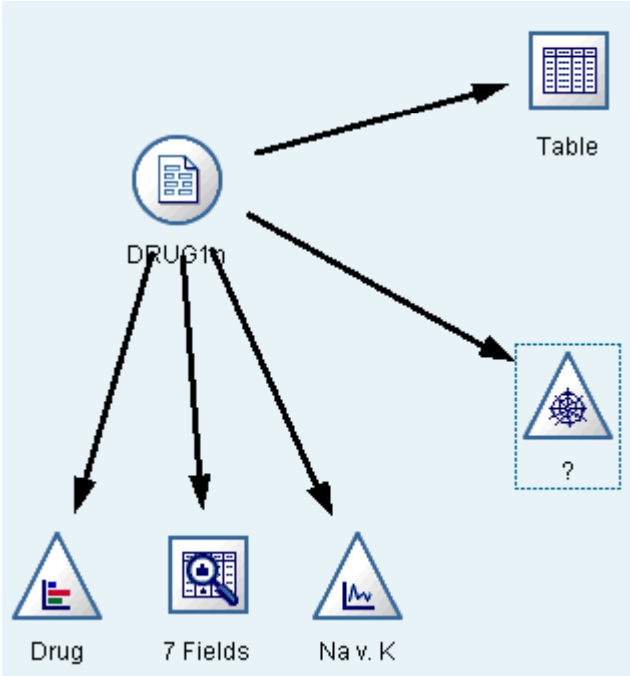
- The plot clearly shows a threshold above which the correct drug is always drug Y and below which the correct drug is never drug Y. This threshold is a ratio—the ratio of sodium (Na) to potassium (K).



Creating a Web Graph

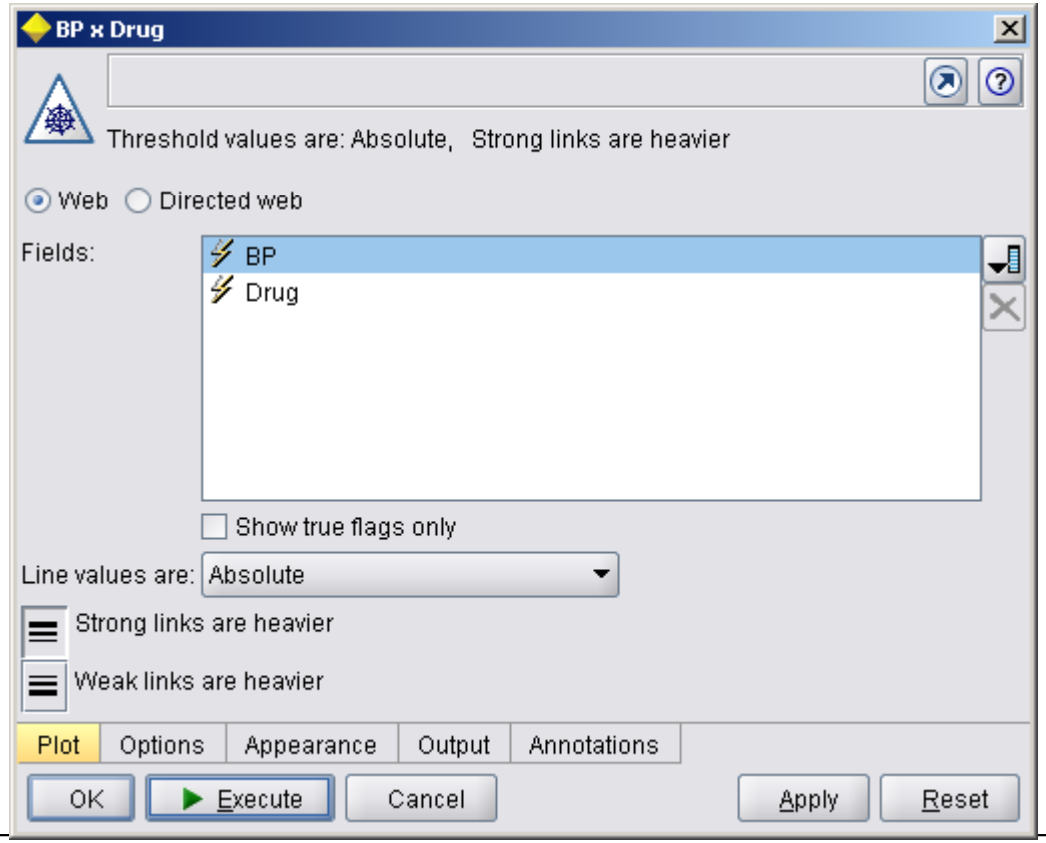
Creating a Web Graph

- Since many of the data fields are categorical, you can also try plotting a web graph, which maps associations between different categories.
- Start by connecting a Web node to the Source node in your workspace.



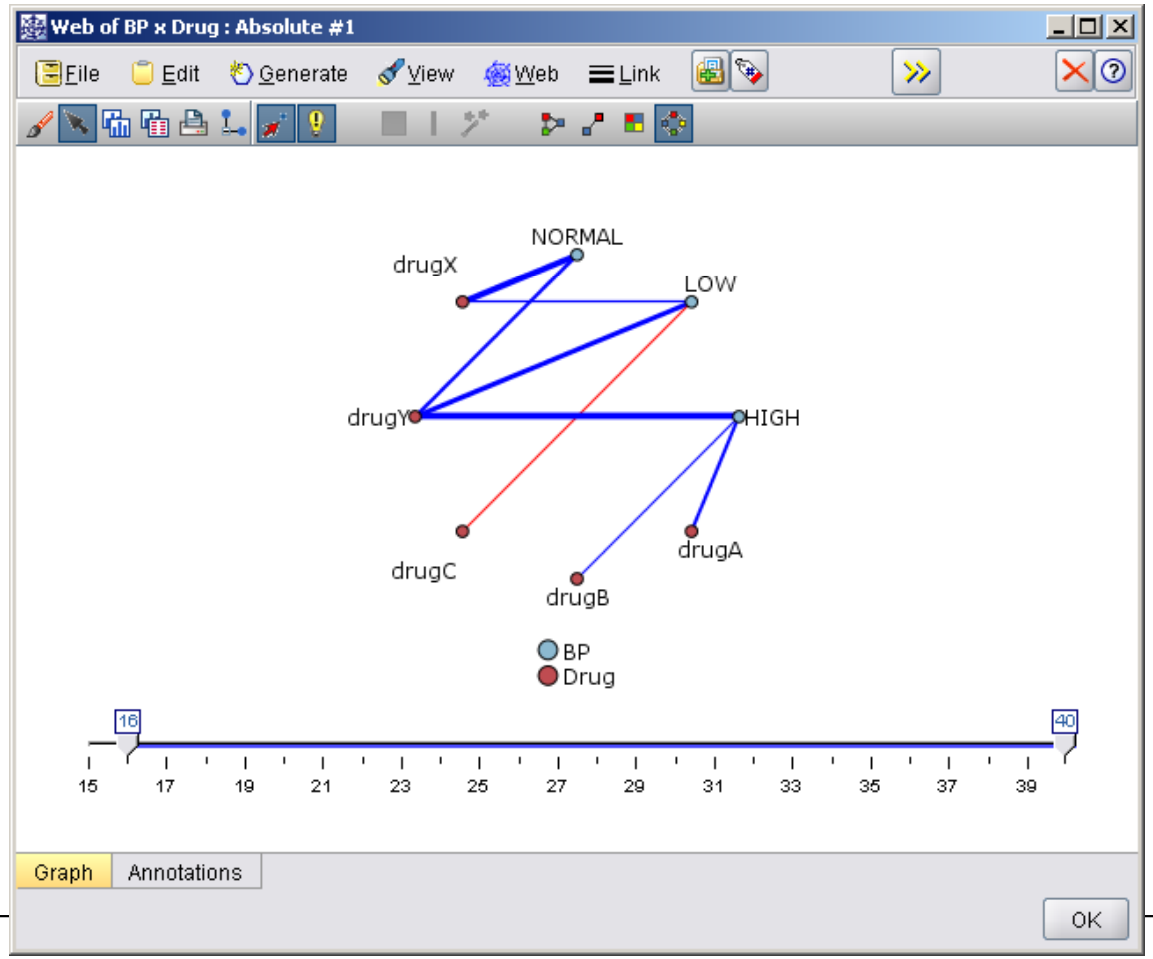
Creating a Web Graph

- In the Web node dialog box, select BP (for blood pressure) and Drug.
- Then, click Execute.



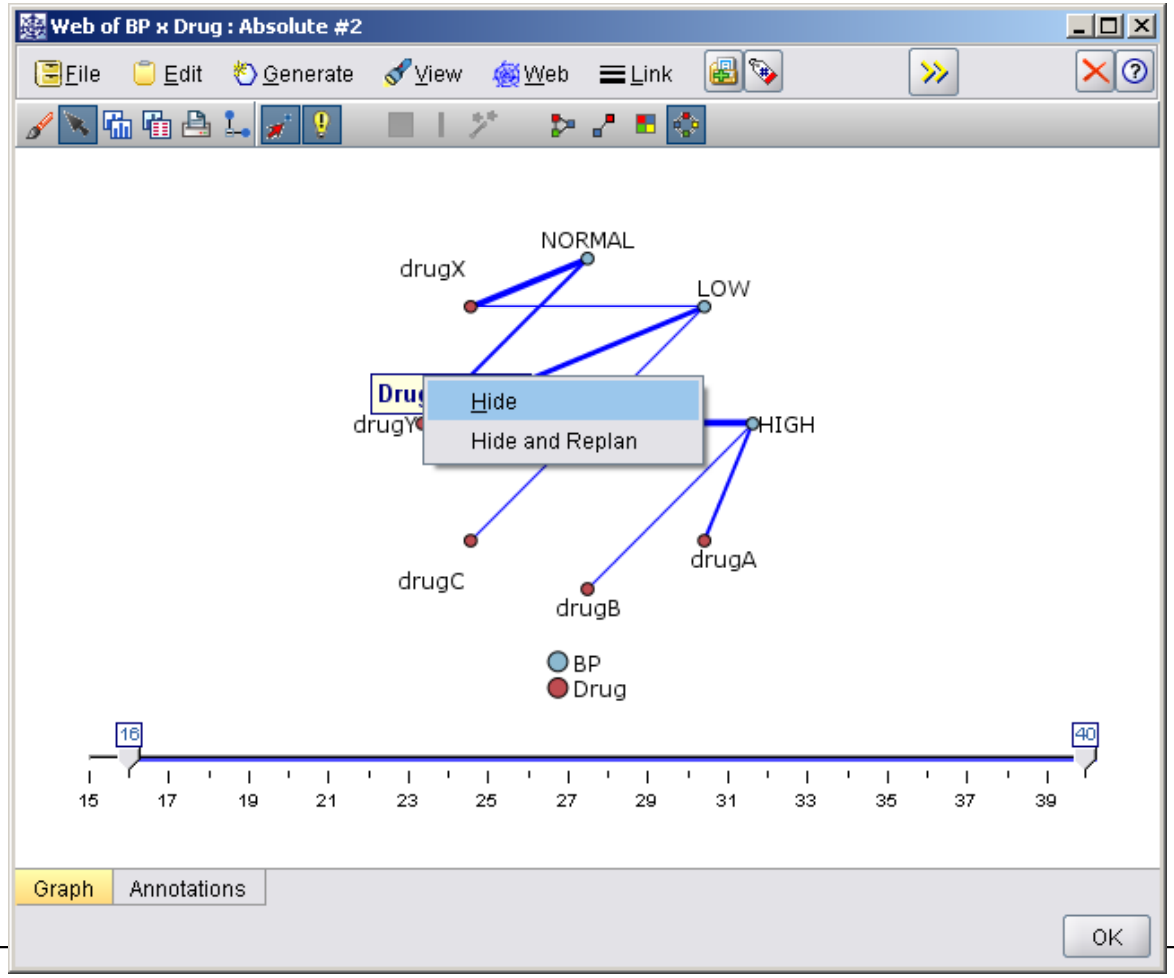
Creating a Web Graph

- From the plot, it appears that drug Y is associated with all three levels of blood pressure. This is no surprise—you have already determined the situation in which drug Y is best.



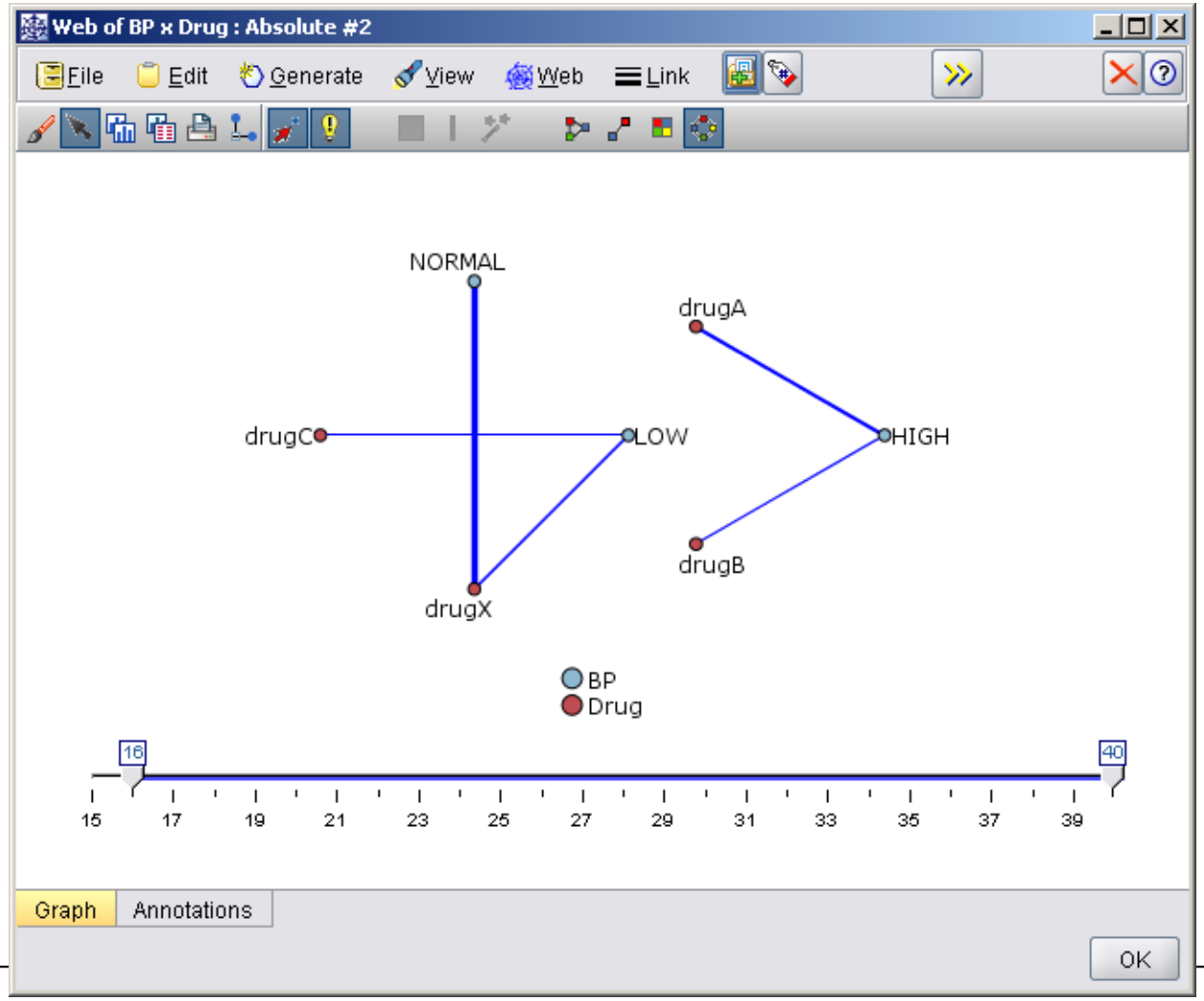
Creating a Web Graph

- To focus on the other drugs, you can hide it. Right-click over the drug Y point and choose Hide and Replan.



Creating a Web Graph

- In the simplified plot, drug Y and all of its links are hidden.



Creating a Web Graph

- Now, you can clearly see that :
  - Only drugs A and B are associated with high blood pressure
  - Only drugs C and X are associated with low blood pressure
  - And normal blood pressure is associated only with drug X

References

References

- Integral Solutions Limited., **Clementine® 12.0 Applications Guide**, 2007. (chapter 8)

The end