SPSS Modeler Tutorial 1

– The Drug Project

Data Warehousing and Data Mining

March 2014

SPSS Modeler (formerly Clementine) is the SPSS enterprise-strength data mining workbench. It helps organizations to improve customer and citizen relationships through an in-depth understanding of data. Organizations use the insight gained from SPSS Modeler to retain profitable customers, identify cross-selling opportunities, attract new customers, detect fraud, reduce risk, and improve government service delivery. The current version is "SPSS Modeler 15".

1 The Drug Project Exercise

Briefing: 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.

1.1 Launch the SPSS Modeler:

Open the SPSS Modeler by going to the Start menu → All Programs → IBM SPSS Modeler 15.0 → IBM SPSS Modeler 15.0. Select "Open an existing project" and double-click on "More files…". In the Open dialog window, goto the path of "N:\DWDM\SPSSModeler\Demos" and double-click on the "drug.cpj" file to open it. The SPSS Modeler should open and displays as Figure 1.

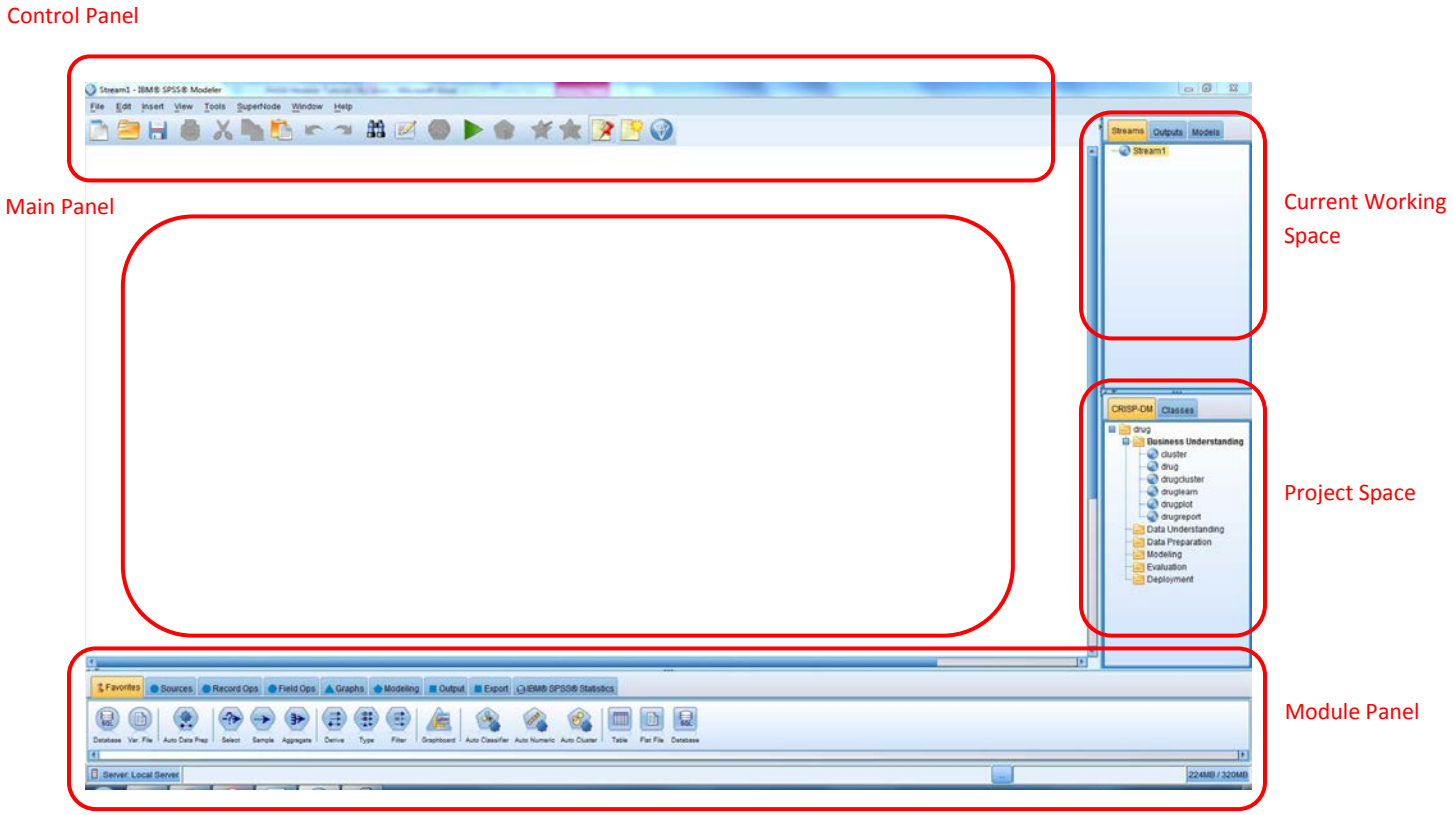

Figure 1: The Drug Project

1.2 Displaying the Properties of the Data

To open a data source, the SPSS Modeler provides many options listed in the "Sources" tab from the "Module Panel". Here, we will use the "Var. File" node.

1. Select the "Sources" tab from the "Module Panel"
2. Double click on the "Var.File" node and it will appear in the "Main Panel". You can also add a node by single left-click on the node in the "Module Panel", then single left-click at the place where you want to place that node in the "Main Panel".
3. Double click the "Var.File" node in the "Main Panel" to open its property window (Figure 2), and Click the "…" button next to the "File" field. In the "Open" dialog window, select to open the "DRUG1n" file that contains records of drug information. The "Var.File" node now should have properties as in Figure 2. The DRUG1n file contains records for 7 attributes, termed "Age", "Sex", "BP", "Cholesterol", "Na", "K", and "Drug".
4. Click "OK" to close the "Var.File" property window.

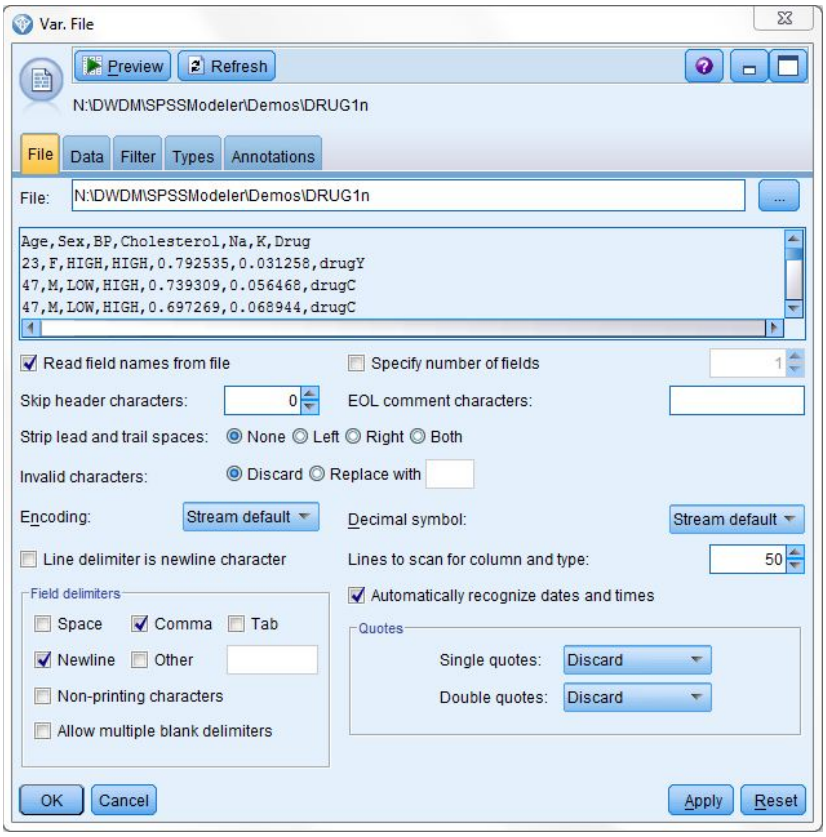

Figure 2: Var.File Property

To display the properties of the data, we use a "Distribution" node.

1. Select the "Distribution" node listed in the "Graphs" tab from the "Module Panel", and add it to the "Main Panel".
2. Establish a link between the "DRUG1n" node and the "Distribution" node by right-clicking on the "DRUG1n" node and select the "Connect…" option, then left-clicking on the "Distribution" node (Figure 3).

Figure 3: Link between two nodes

3. Double-click the "Distribution" node to open its property window.
4. Select "Drug" for the "Field" option (Figure 4) to display the distribution of drugs. Click "Run"

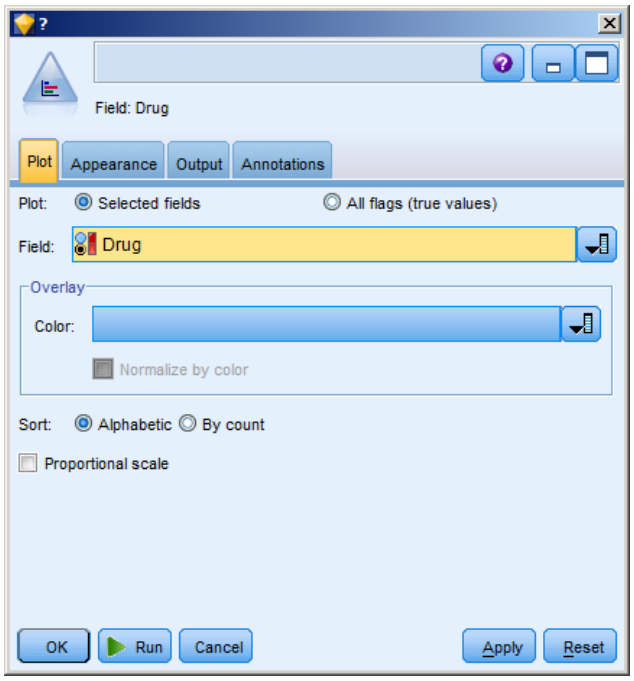

Figure 4: Distribution Node Property

5. You should see a distribution window for the Drug attribute in the DRUG1n file (Figure 5). This window illustrates the count of different drugs and their percentages.

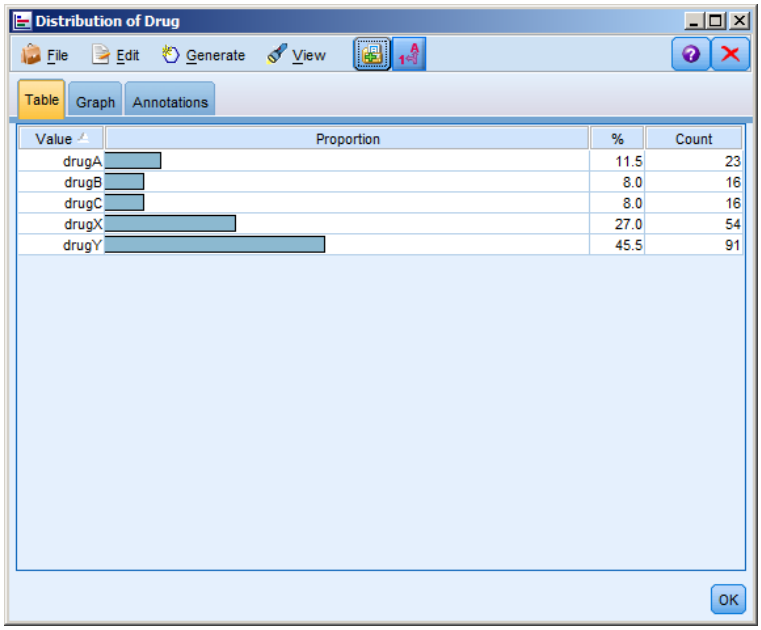

Figure 5: Distribution of Drugs

6. Click OK to close the window.

1.3 Finding a Relationship in Numeric Data

To investigate a relationship between sodium (Na) and potassium (K) levels, the most natural way would be to produce a point plot. To do this, we create a "Plot" node and connect it to the "Var.File" node.

1. Select the "Plot" node listed in the "Graphs" tab from the "Module Panel", and add it to the "Main Panel".
2. Establish a link between the "DRUG1n" node and the "Distribution" node by right-clicking on the "DRUG1n" node and select the "Connect…" option, then left-clicking on the "Plot" node (Figure 6).

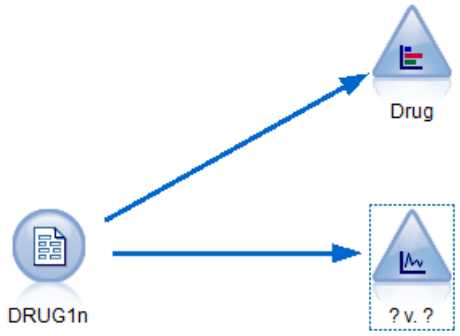

Figure 6: Link between DRUG1n and Plot

3. Double-click the “Plot” node to open its property window.
4. Select “K” (Potasium) for the “X Field” option and select “Na” (Sodium) for the “Y Field” option (Figure 7).

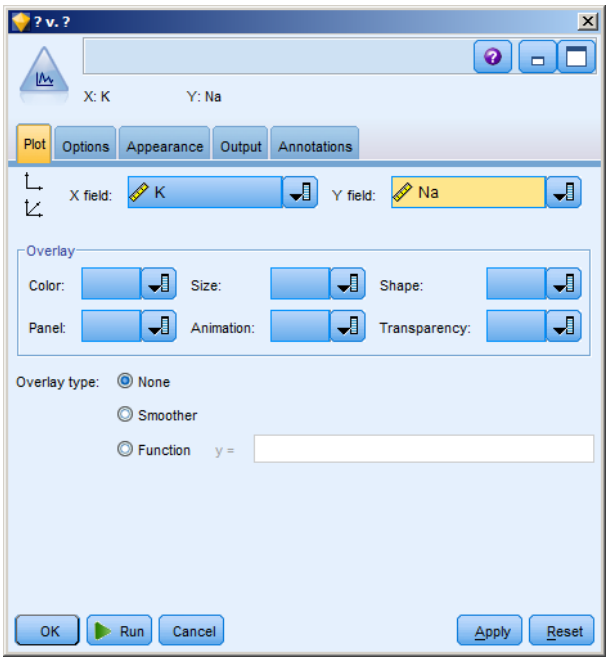

Figure 7: Plot Node Property

5. Click Run. The plot window of the K attribute and Na attribute will be displayed (Figure 8). This appears to be a random scattering, with no obviously apparent relationship between the Na and K attributes. However, this graph takes no account of which drug was used in each case. Therefore, we need to modify the property of the “Plot” node in order to display the correlations between Na and K with respect to different drugs.

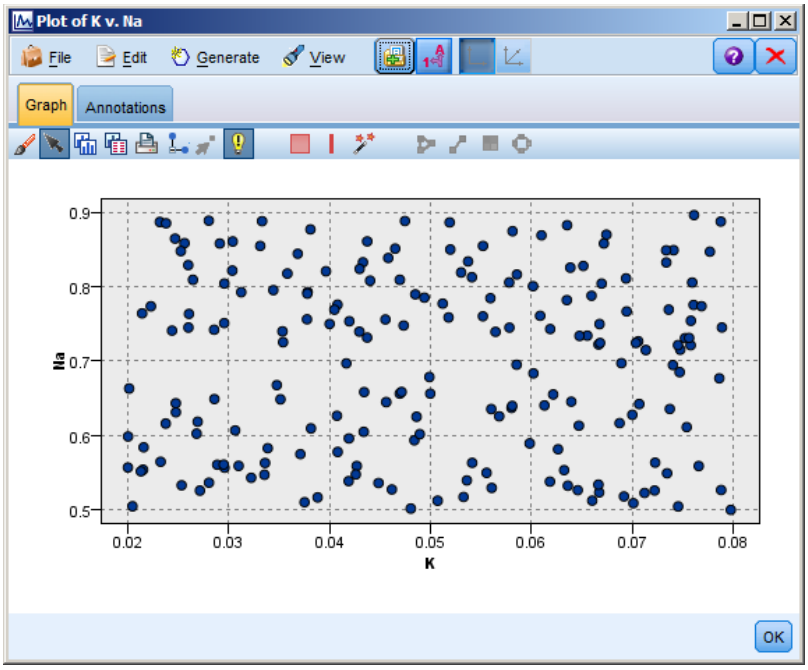

Figure 8: Plot of K v. Na

6. Double-click the "Plot" node to open its property window.
7. Select "Drug" for the "Color" option in the "Overlay" group (Figure 9).

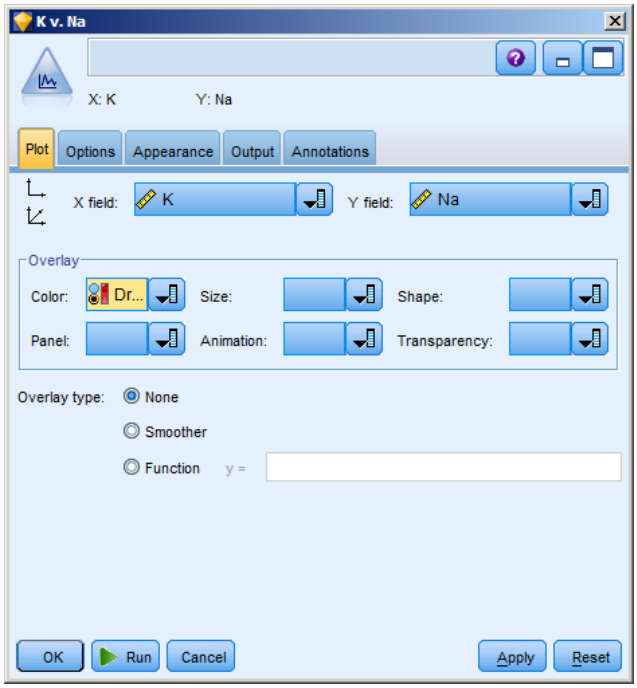

Figure 9: Plot Node Property

8. Click Run. The plot window of the K attribute and Na attribute with respect to different drugs will be displayed (Figure 10). We can observe that a clear pattern emerges in the overlaid plot. The threshold is neither the Na nor K field, but in a ratio between them.
9. Click OK to close the window.

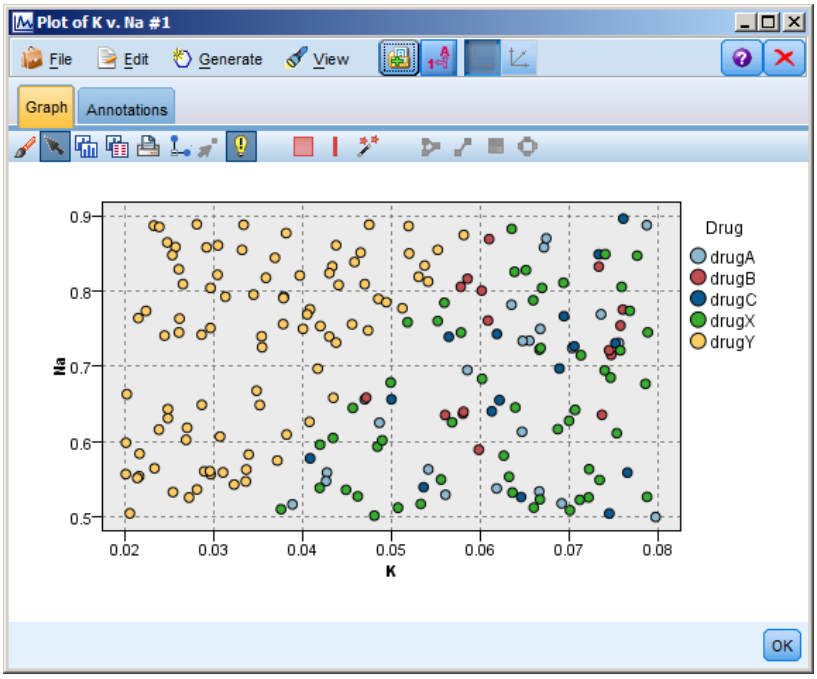

Figure 10: Plot of K v. Na

1.4 Finding the Threshold

We can find the threshold by calculating the ratio and examining its distribution. To do so, we need to create a "Derive" node and connect it to the "Var.file" node.

1. Select the "Derive" node listed in the "Field Ops" tab from the "Module Panel", and add it to the "Main Panel".
2. Establish a link between the "DRUG1n" node and the "Derive" node by right-clicking on the "DRUG1n" node and select the "Connect…" option, then left-clicking on the "Derive" node (Figure 11).

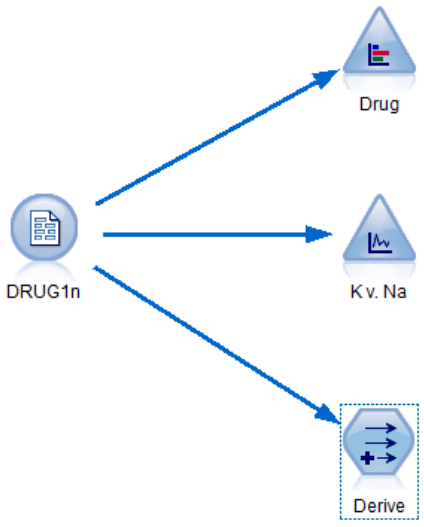

Figure 11: Link between Var.File and Derive Nodes

3. Double-click the "Derive" node to open its property window.
4. Type string "Na_to_K" in the "Derive field", and formula "Na/K" in the "Formula" area (Figure 12). This will create a new field named "Na_to_K" containing numbers calculated as "Na/K".

5. Click “OK” to close this property window. The “Derive” node will be renamed to “Na_to_K”.

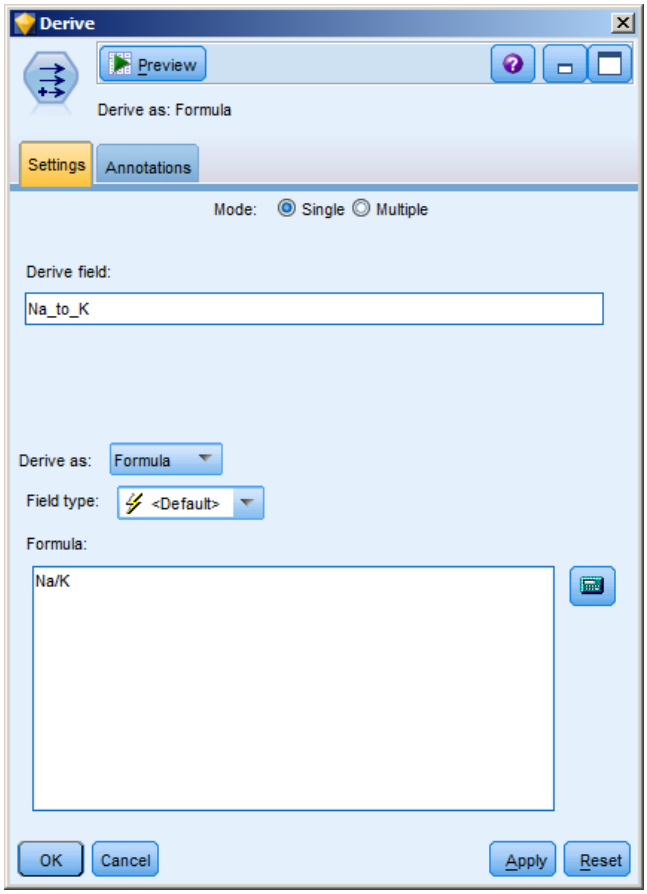

Figure 12: Derive Node Property

Next, we need to create a “Histogram” node to display the output from the “Derive” node.

1. Select the “Histogram” node listed in the “Graphs” tab from the “Module Panel”, and add it to the “Main Panel”.
2. Establish a link between the “Na_to_K” node and the “Histogram” node by right-clicking on the “Na_to_K” node and select the “Connect…” option, then left-clicking on the “Histogram” node (Figure 13).

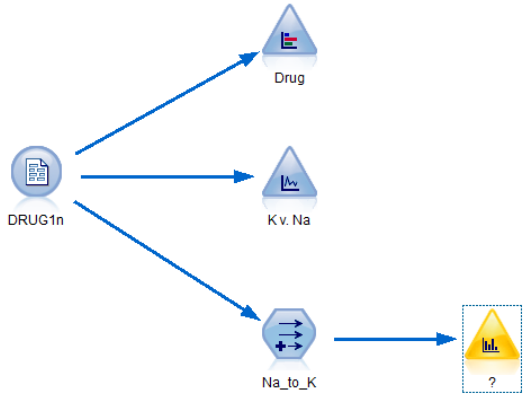

Figure 13: Link between Na_to_K and Histogram Nodes

3. Double-click the “Histogram” node to open its property window.
4. Select “Na_to_K” for the “Field” option, and “Drug” for the “Color” option in the “Overlay” group (Figure 14).

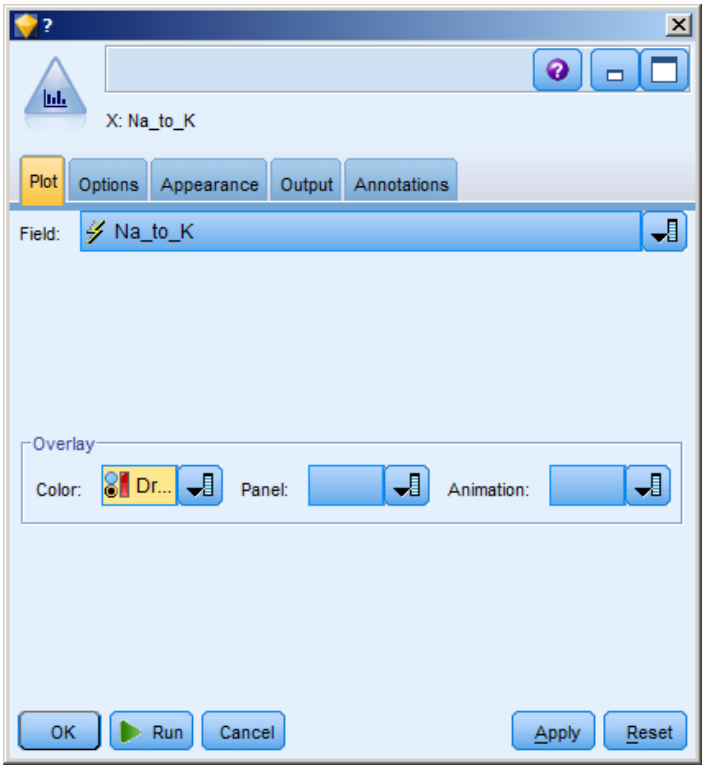

Figure 14: Histogram Node Property

5. Click "Run". The histogram window will be display as in Figure 15.

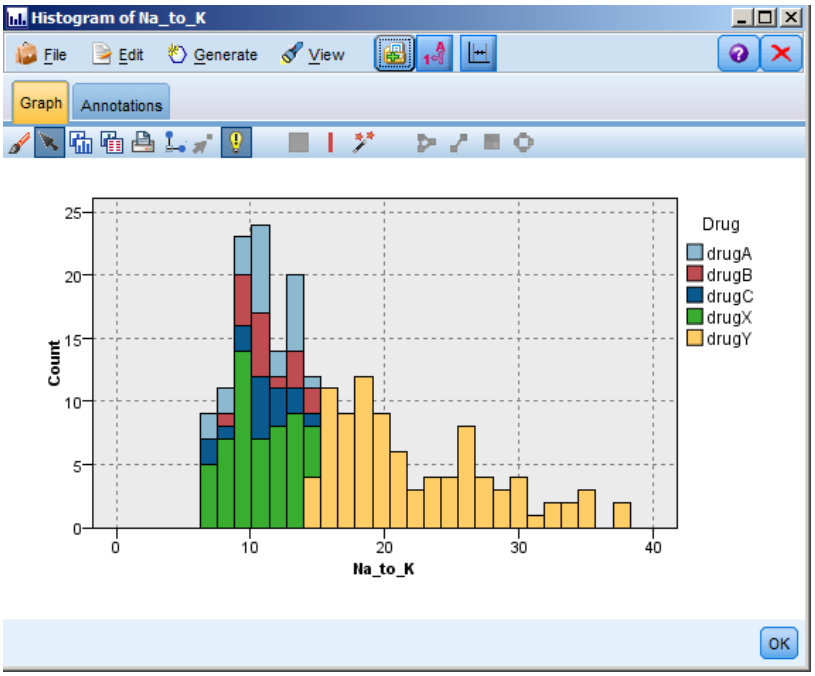

Figure 15: Histogram of Na_to_K

The histogram shows that the distribution of the ratio of Na and K. In addition, the threshold is clear as the column in the bars change from multi-coloured to the pure yellow colour at the critical value.

We can now add a band selection line to this histogram to separate the records before and after the threshold.

1. Tick the "Interactions" option from the "View" menu (Figure 16).
2. Left-click the "Activates band selection" option (Figure 17).

3. Place the RED colour line as close as possible to the point at which the bars of the histogram change colour (the threshold point). (Figure 18).
4. Right-click at the right side of the threshold line, and select “Generate Derive Node for Band” option (Figure 19).
5. A new “Derive” node will then be added to the “Main Panel”. Open its property window, and observe the selection condition. Rename this node as “band’’.
6. Connect this “band” node to “Na_to_K” derive node and also add a new histogram node to connect to it (Figure 20).
7. Double-click the “Histogram” node to open its property window.
8. Select “Na_to_K” for the “Field” option, and “Band” for the “Color” option in the “Overlay” group
9. Run the new histogram node and observe the result (Figure 21).

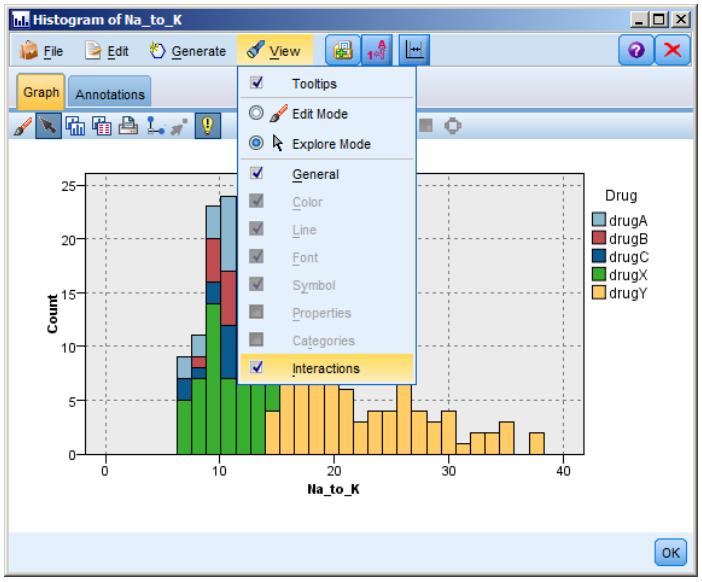

Figure 16: Histogram Interactions

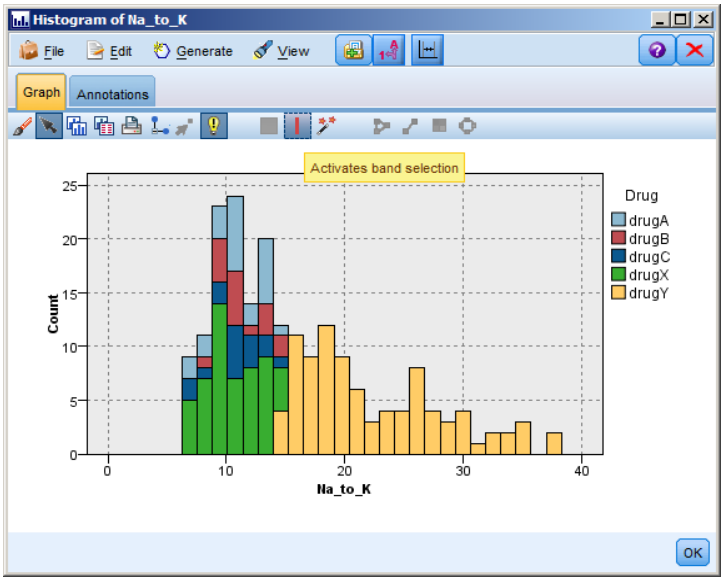

Figure 17: Activates band selections

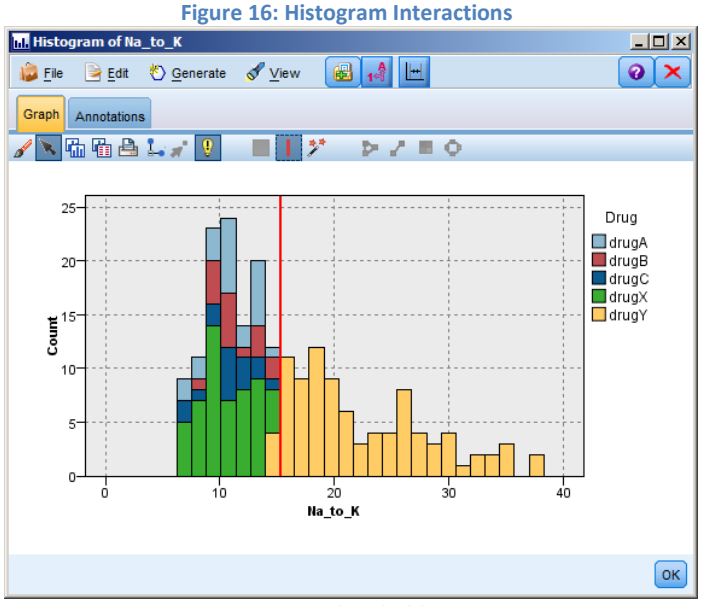

Figure 18: Threshold Line

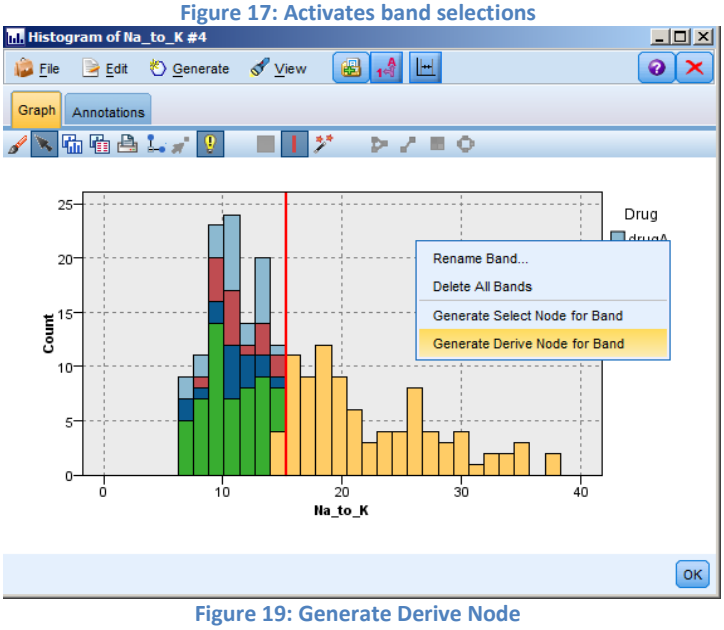

Figure 19: Generate Derive Node

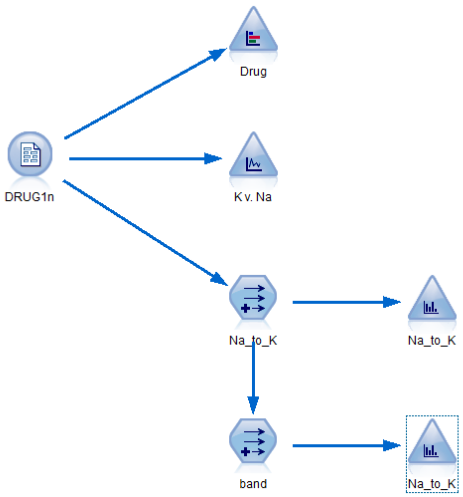

Figure 20: The new band node

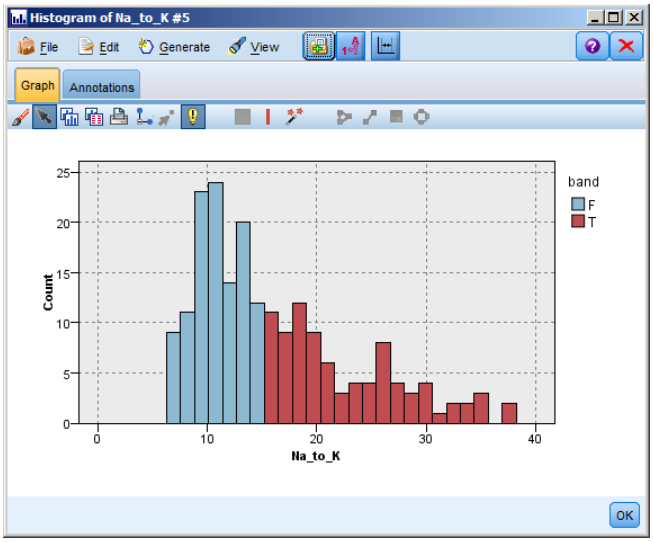

Figure 21: The new band

End of Tutorial 1