Exploring Data with KNIME Plots

Learning Objectives

By the end of this activity, you will be able to perform the following operations in KNIME:

1. Create a workflow
2. Import a dataset
3. Explore a dataset using plots

Problem Description

Wildfires have caused significant damage in southern California in recent years, making the threat of wildfires very real in San Diego and other southern California regions. One of the weather conditions that increases the risks of wildfires is low relative humidity. Low relative humidity leads to dry conditions, which can hasten the spread of wildfires. Having a way to predict these conditions would be very helpful in avoiding the dangers of wildfires. With this in mind, the problem that we want to address is to predict days with low relative humidity. The hands-on exercises in this and subsequent modules are to build a model for this predictive task.

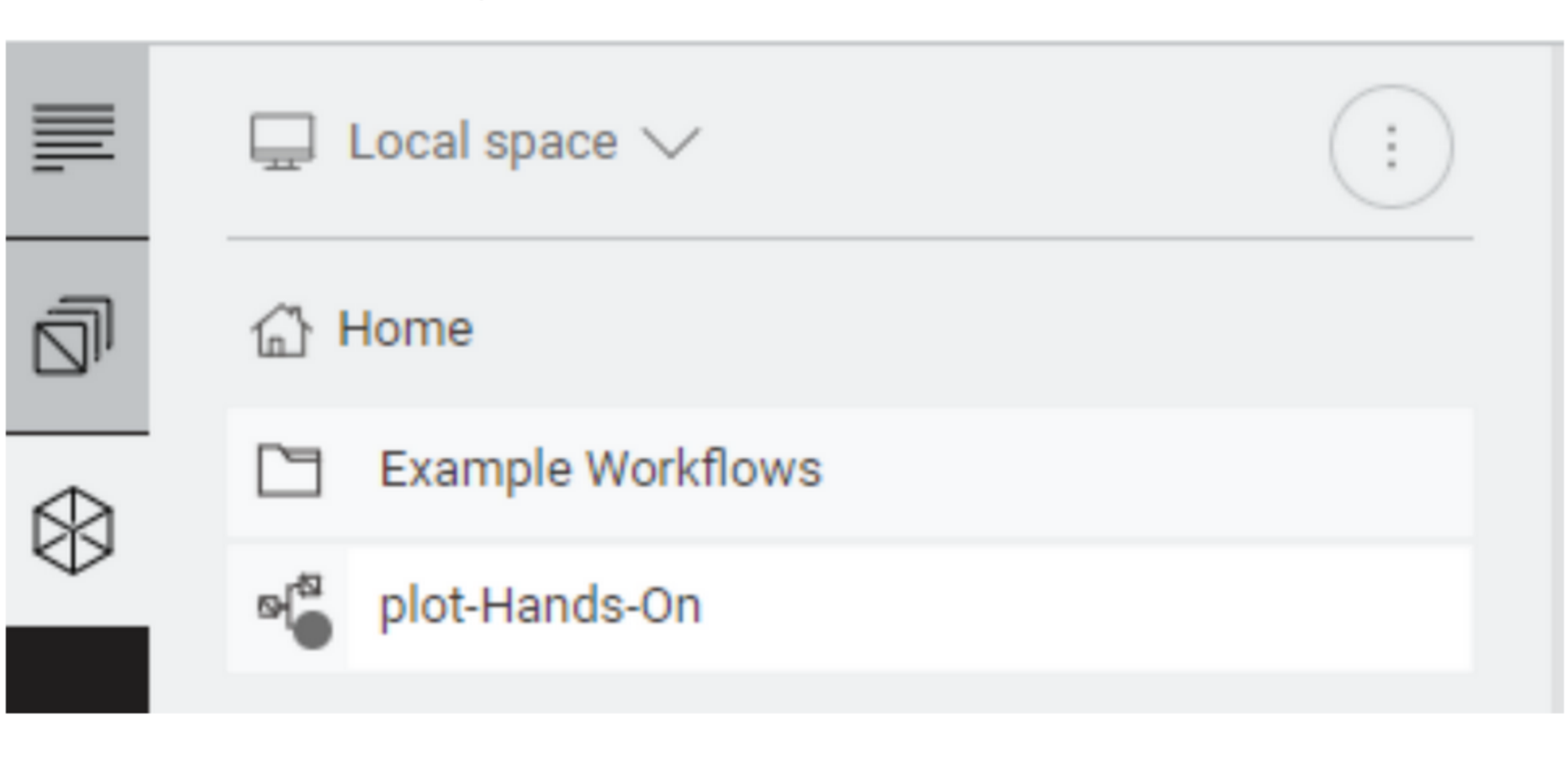
Before we create a model, we need to explore the dataset that we have to work with. This dataset contains weather data collected over three years from a weather station in San Diego. For a detailed description of the data, see the Reading *Description of Daily Weather Dataset.*

Steps

Start a KNIME Workflow

Let's start a new KNIME workflow.

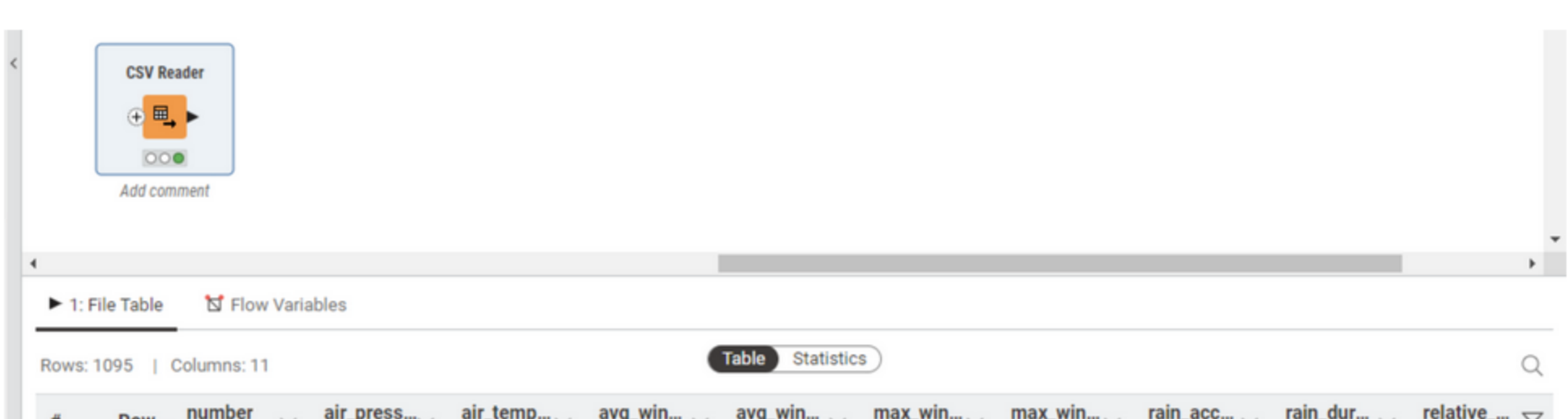
1. Open KNIME, and below the example projects click on *Create workflow in your local space*
2. Name the workflow something descriptive, e.g. "plot Hands-On" and click on *Create.*
3. The workflow will be saved in the Local space of your KNIME installation. You should see the new workflow under LOCAL in the KNIME Explorer view.



Import the Dataset

The first step in the workflow is to read in your dataset.

1. Drag the **CSV Reader** node onto the Workflow Editor. You can search for the File Reader node by typing "CSV Reader" in the search box in the Node Repository, or find it under IO.
2. Open the Configure Dialog.
3. Click **Browse** and select the location of the dataset file *big-data-4/knime/daily_weather.csv*, which you should have downloaded already. If not, go to "Instructions for Downloading Hands On Datasets" under Week 1 section**.**
4. Click **OK** to close the Configure Dialog.
5. **Execute** the node.



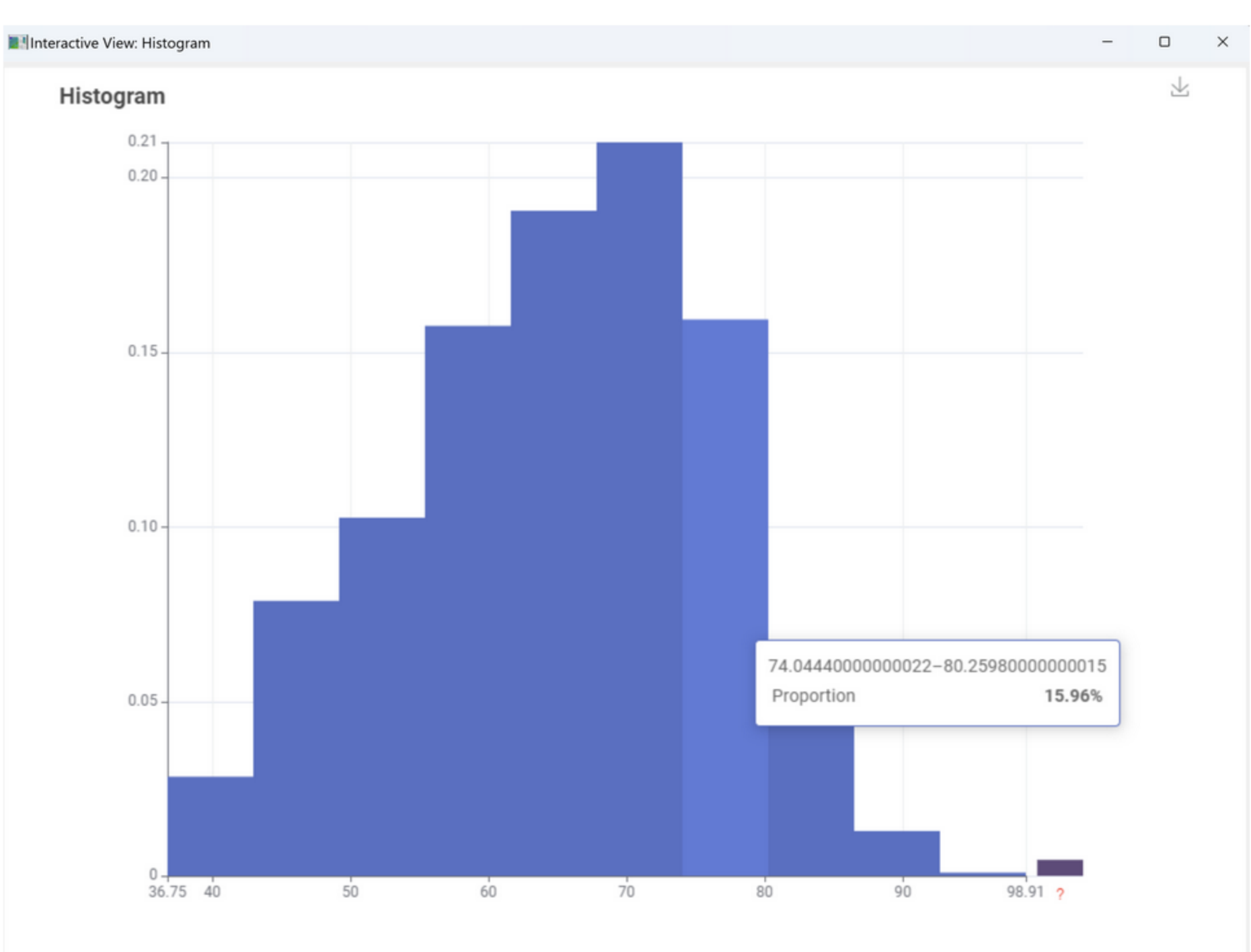
#	Row...	number	air_press...	air_temp...	avg_win...	avg_win...	max_win...	max_win...	rain_acc...	rain_dur...	relative_...
1	Row0	0	918.06	74.822	271.1	2.08	295.4	2.863	0	0	42.42
2	Row1	1	917.348	71.404	101.935	2.443	140.472	3.533	0	0	24.329
3	Row2	2	923.04	60.638	51	17.068	63.7	22.101	0	20	8.9
4	Row3	3	920.503	70.139	198.832	4.337	211.203	5.19	0	0	12.189

Create a Histogram

A histogram is used to examine the distribution of a continuous variable. It divides the data into bins, and plots the frequency of occurrences within the range of each bin.

The continuous data we are going to visualize is the **air_temp_9am** column.

1. Locate the **Histogram** node, which is located in the "Views" category of the Node Repository, and drag it to the Workflow Editor.
2. Connect the Histogram node to the CSV Reader node.
3. Open the Configure dialog. Choose **air_temp_9am** as the **Dimension**. The Number of Bins can be anything that makes sense for the data. Select 10. Select **Show separate bar for each type of special values,** so we can display missing values. Select **Proportion(%)** as the frequency axis. Click **OK** to save this configuration and close the dialog.
4. Execute the workflow.
5. To view the histogram of **air_temp_9am**, hover the Histogram node, and select 'Open View'. The x-axis labels show the temperature, and the y-axis shows the % frequency of samples that in each bin. Hovering over each bin will show the range



5. The histogram of **air_temp_9am** shows that the distribution of the temperature at 9am has almost a normal (bell-shaped) curve. The temperature values range from 36.75 to 98.91 degrees Fahrenheit, with the most frequent values around 70 degrees, which is typical of weather in San Diego, California. There are no outliers of concern. Note that the last bin on the right (?) specifies the number of samples with missing values for **air_temp_9am**. We will address these missing values in the hands-on on Data Preparation.

Create a Scatter Plot

A scatter plot shows points on a graph to show the relationship between two variables, and can be used to visually inspect the correlation between the variables.

Let's create a scatter plot to visualize the relationship between **relative_humidity_9am** vs. **air_temp_9am**.

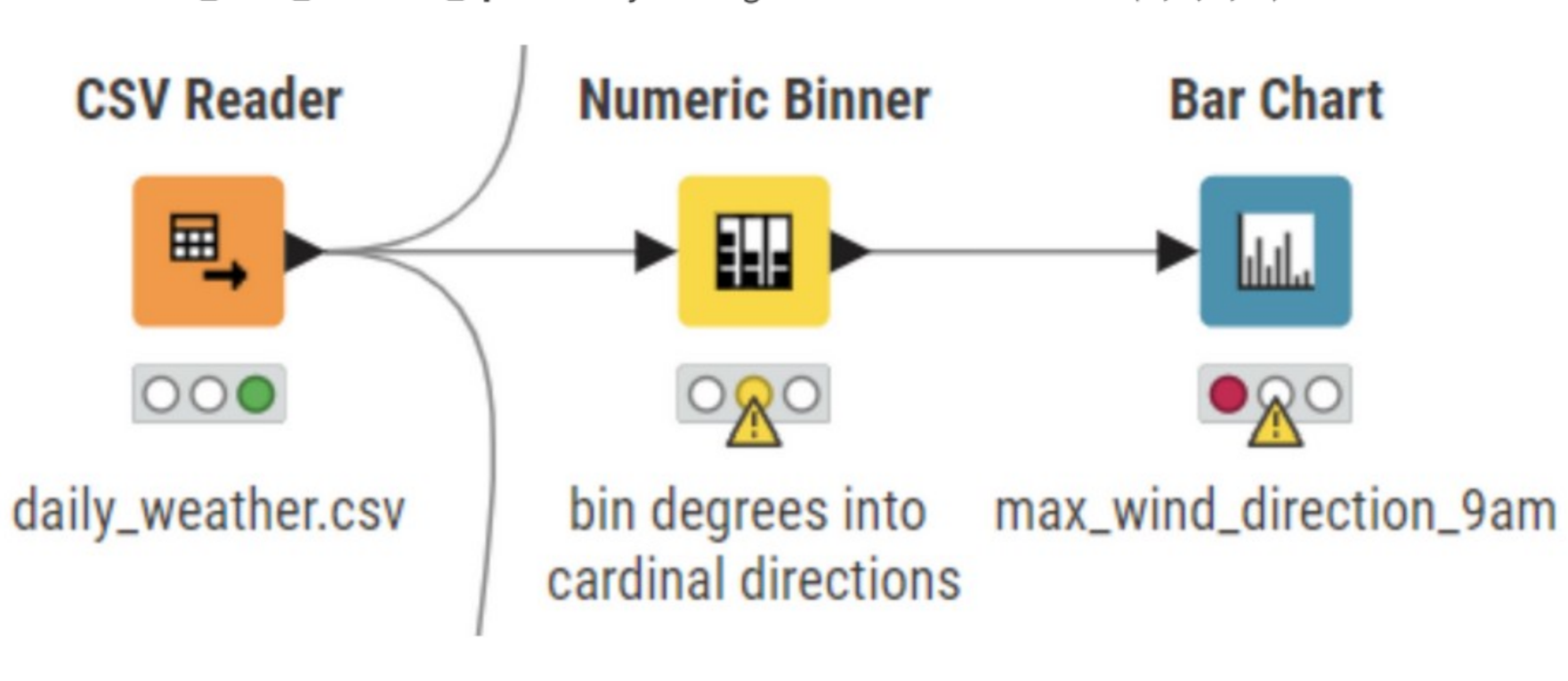
1. Locate the **Scatter Plot** node, which is in the Views category. Drag it to the Workflow Editor, and connect it to the CSV Reader node.
2. Open the configure dialog. Select **air_temp_9am** as the horizontal dimension, and **relative_humidity_9am** as the vertical dimension. Click on **OK.**
3. Execute the workflow, and view the Scatter Plot.

3. The scatter plot of **air_temp_9am** and **relative_humidity_9am** shows that there is a negative correlation between temperature and relative humidity. This means that an increase in temperature corresponds to a decrease in relative humidity. Since warm air can hold more water vapor than cool air, relative humidity falls when the temperature rises. This relationship is captured by the scatter plot.

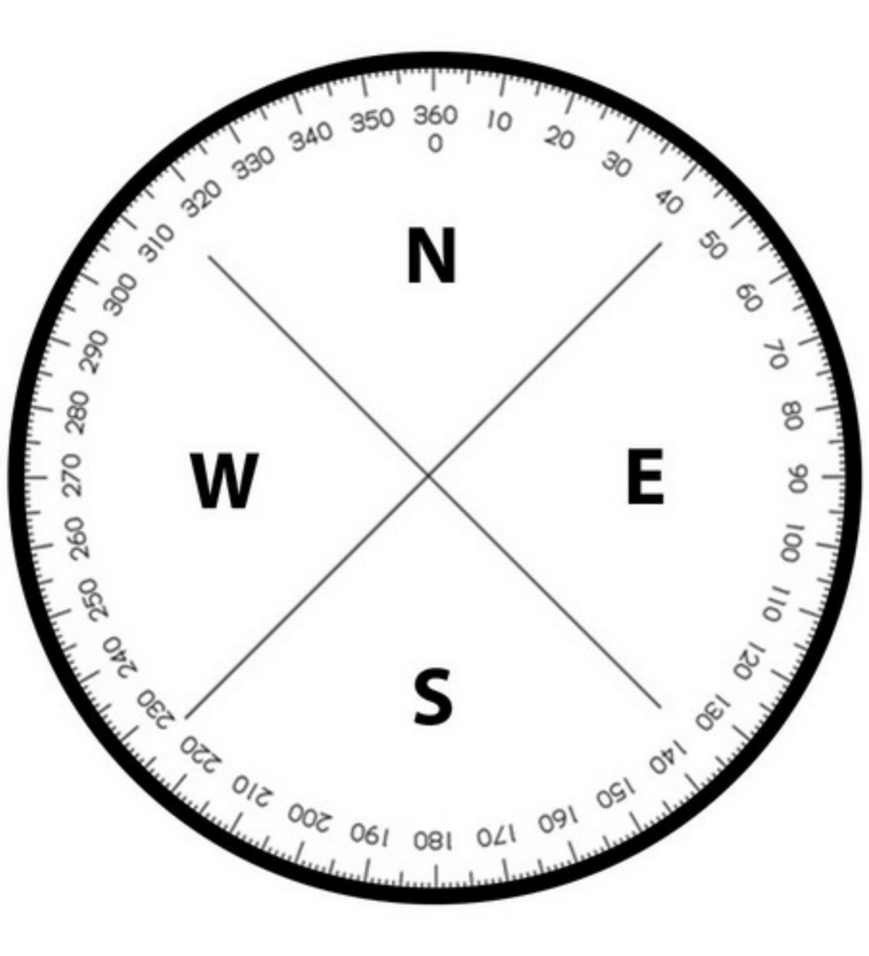
Create a Bar Chart

A bar chart is used to show the distribution of a categorical variable. To make a bar chart in KNIME, we can use the **Bar Chart** node.

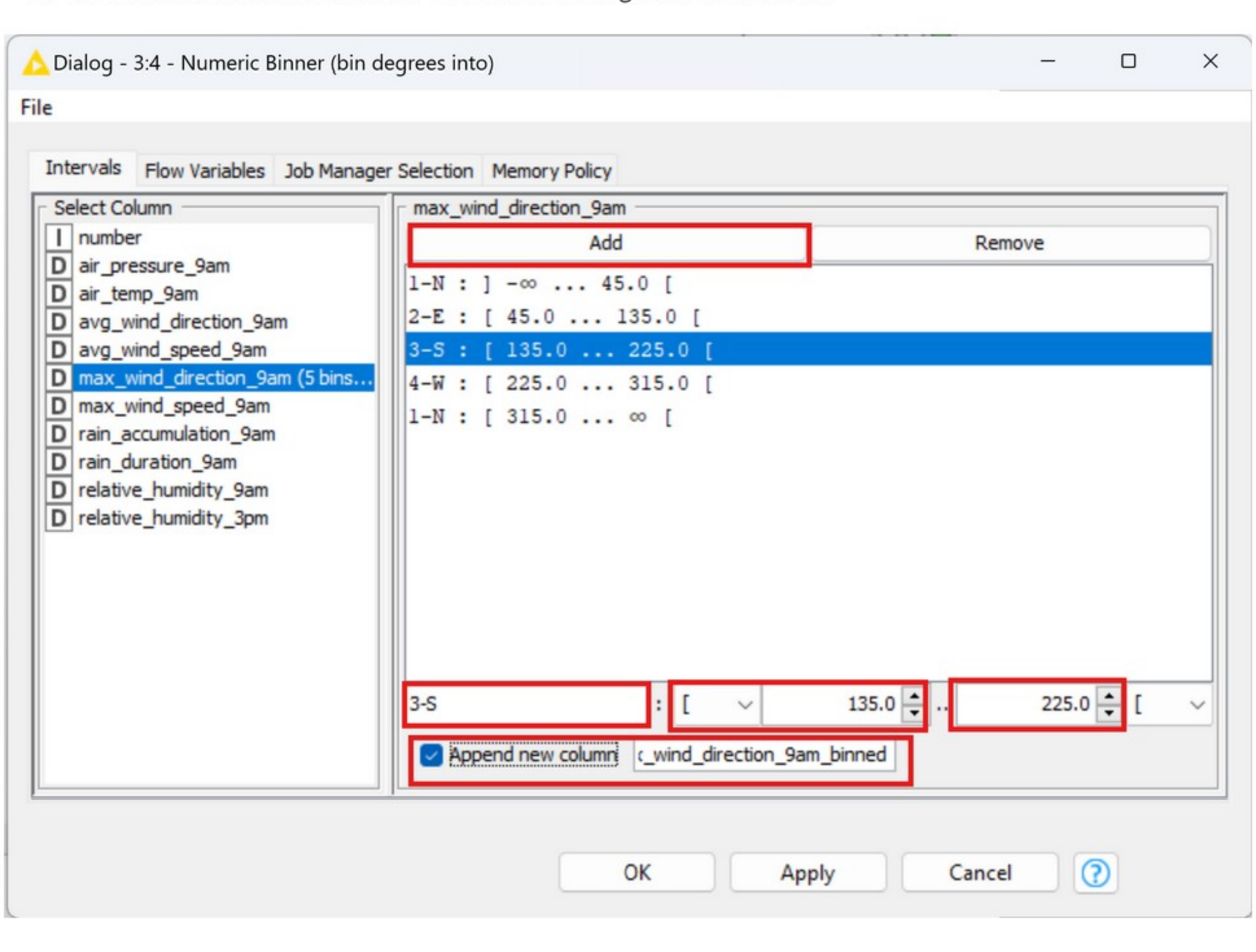
We will use a bar chart to visualize the distribution of the **max_wind_direction_9am** variable. Notice that this variable specifies the wind direction in degrees, and so is a continuous variable. We will generate a categorical version of the **max_wind_direction_3pm** data by binning it into cardinal directions (N, E, S, W).



To do this, use the **Numeric Binner** node. If you look at a compass, the conversion of degrees to direction looks like this:



1. Locate the **Numeric Binner** node, by searching it and then clicking on *More advanced nodes*. Drag it to the Workflow Editor, and connect it to the CSV Reader node.
2. In the Configure Dialog of the Numeric Binner node, select the **max_wind_direction_9am** column and add 5 bins.
3. The first AND last bin will be named exactly the same because north has two ranges of degrees (The node status will show a warning. This is expected.).
4. In order to display the directions in correct order in the chart, enter the names of each direction as follows: "1-N", "2- E", "3- S", "4- W", etc.
5. Fill in the min and max values for each bin according to the screenshot.

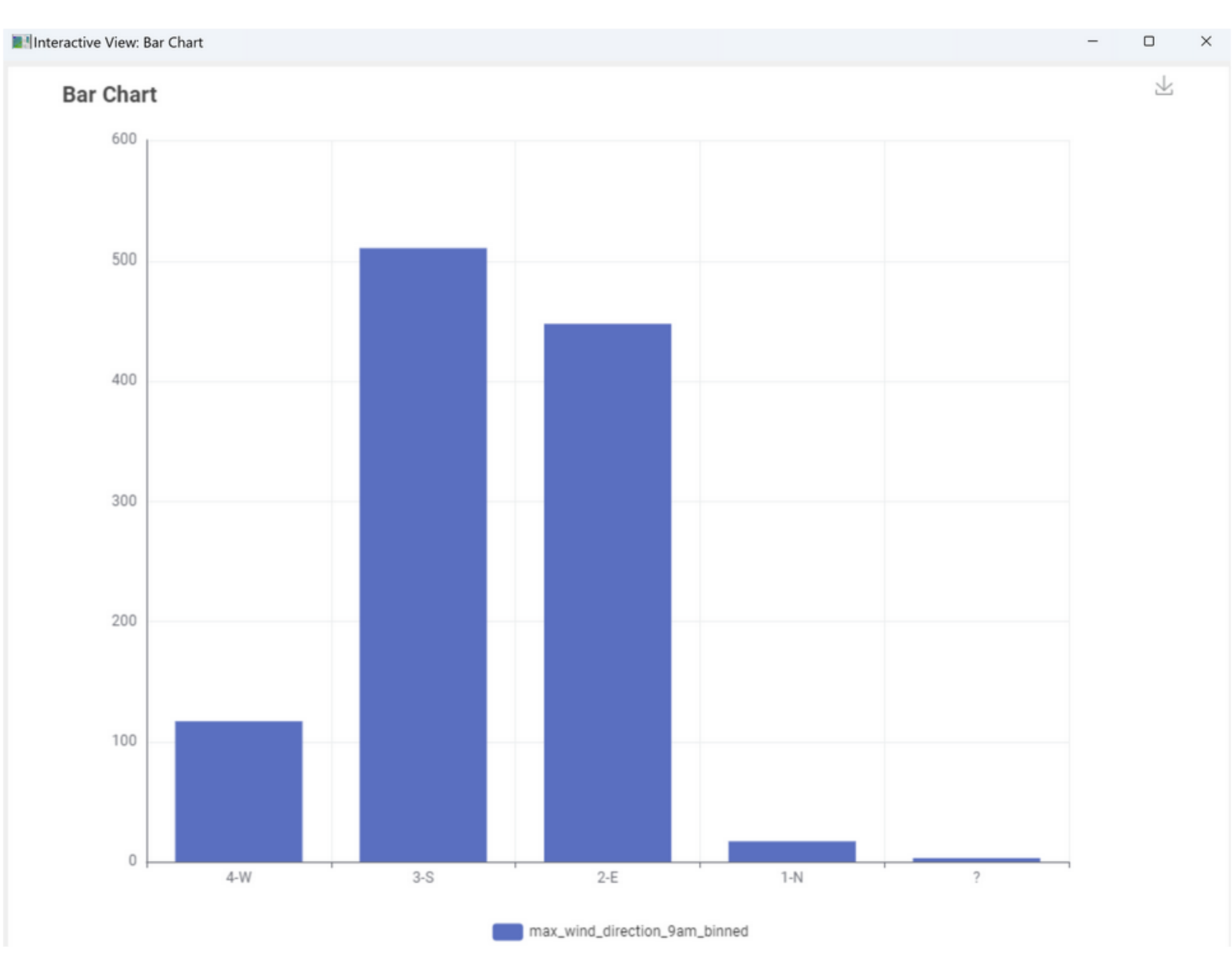


6. Make sure "Append new column" is checked; let's call it **categorical_max_wind_direction_9am.**

7. Locate the **Bar Chart** node (in the Views category), and drag it to the Workflow Editor. Then connect the Numeric Binner node to the Bar Chart node.

8. In the Configure Dialog of the Bar chart node, choose the new column as the **Binning column.** Then, in **Aggregation** select **Ocurrence count**. Click **OK.**

9. Now execute the workflow and view the bar chart.

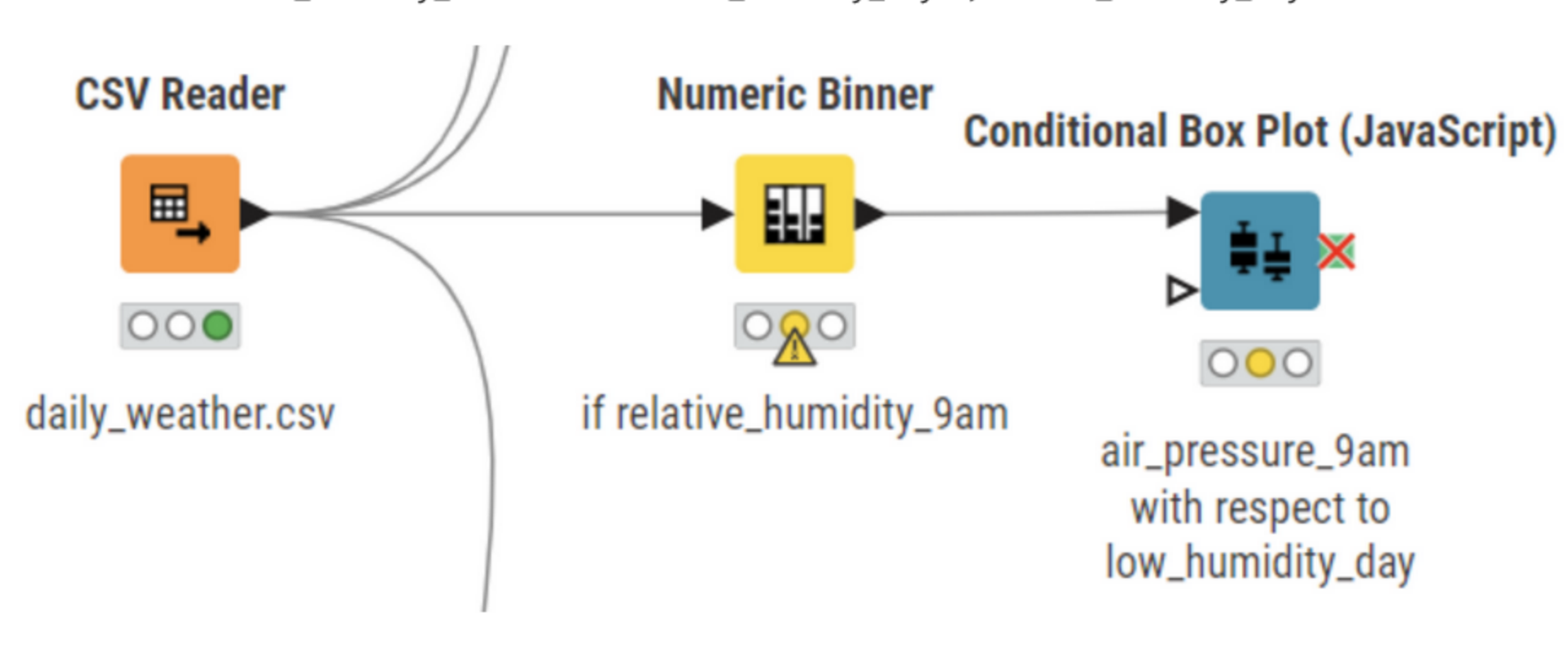


10. The bar chart of categorical_max_wind_direction_9am shows that the most frequent value for max wind direction is South, and there are very few from the North. There are also a few samples with missing values.

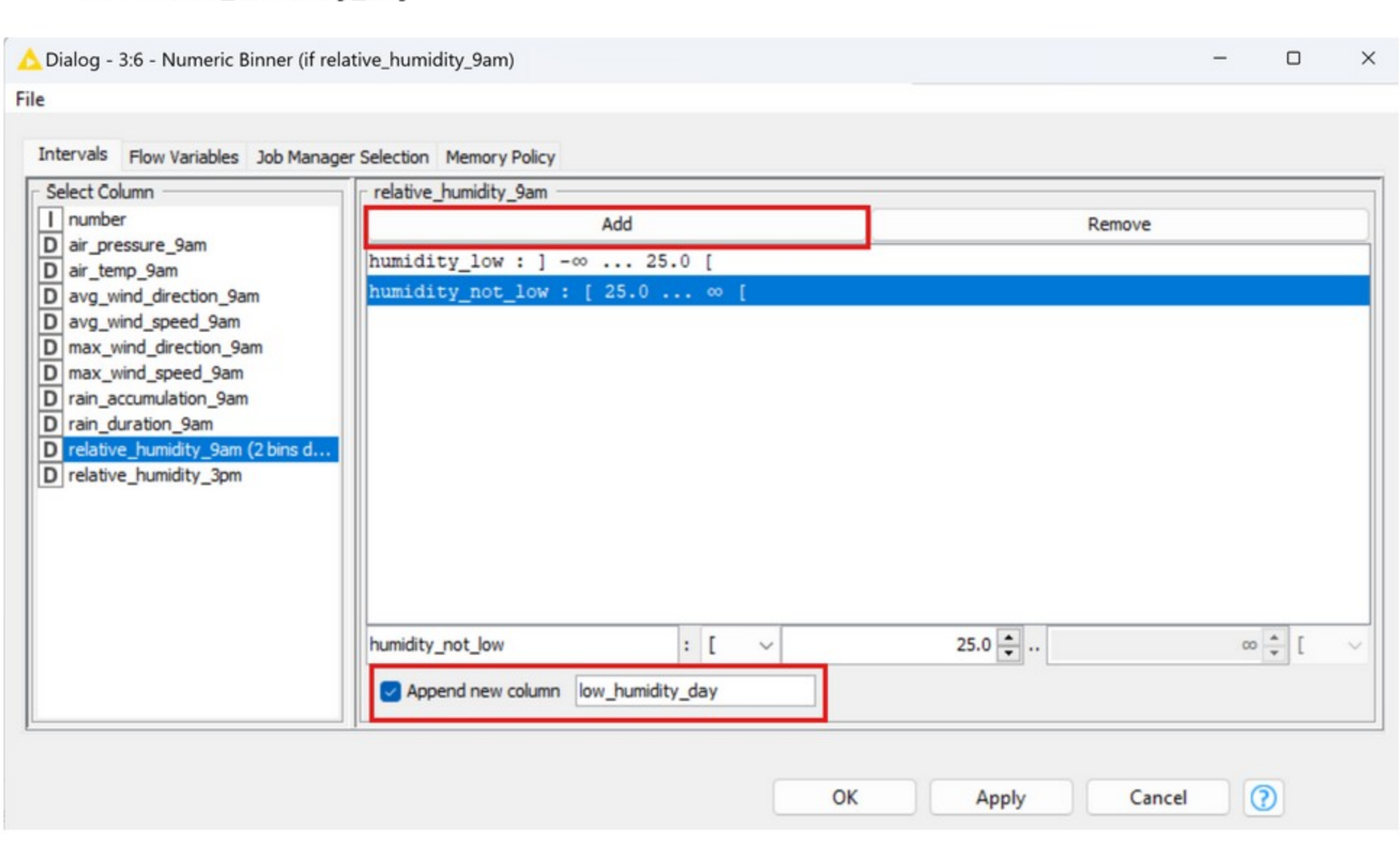
Create a Box Plot

A box plot can be used to compare different distributions. Data for a numeric variable is partitioned into categories, and a box plot is created to show the distribution for each category. The box plots are then shown on a single graph to compare the different categories.

We will create a box plot to examine how the distribution of **air_pressure_9am** differs for low-humidity days vs. days with normal or high humidity. First, we will will create a categorical variable named **low_humidity_day** to specify whether a day is low-humidity or not. We can do this using the **Numeric Binner** node. The condition for this new variable is *"if relative_humidity_9am < 25% then low_humidity_day=1, else low_humidity_day=0"*.

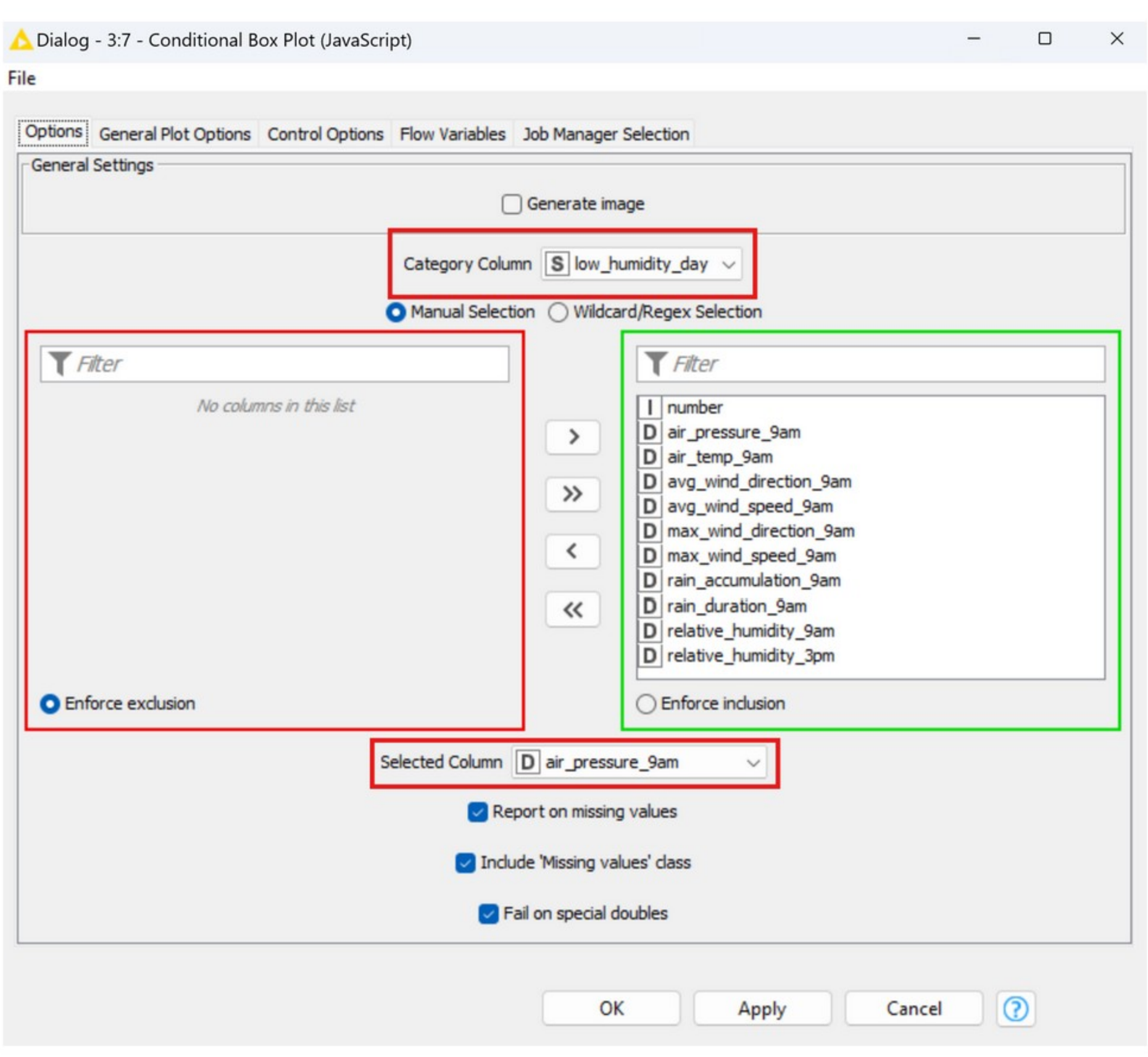


1. Locate the **Numeric Binner** node, drag it to the Workflow Editor, and connect it to the CSV Reader node.
2. Open the Configure Dialog for the Numeric Binner node. Select **relative_humidity_9am**, and add 2 bins. Make one bin called "**humidity_low**" with the range **-∞ to 25 excluding 25**, and another called "**humidity_not_low**" with the range **25 to ∞**. The endpoint brackets specify that humidity_low excludes 25.0, while humidity_not_low includes 25.0. This is necessary to capture the condition "*if relative_humidity_9am < 25% then low_humidity_day=1, else low_humidity_day=0*". Click the checkbox to "**Append new column**" and name it **low_humidity_day**.

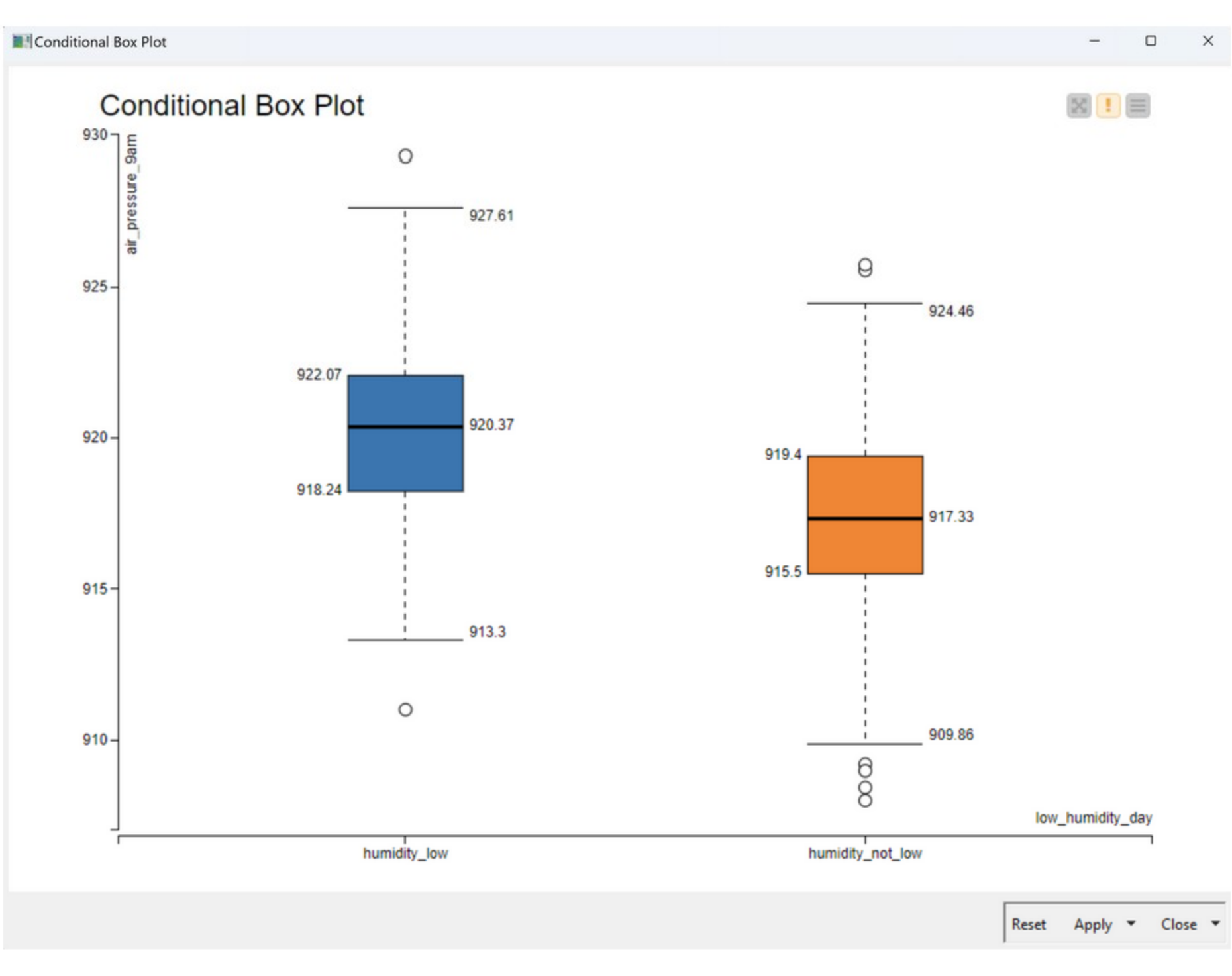


3. Add a **Conditional Box Plot** node to the workflow and connect to the output of Numeric Binner.

4. In the Conditional Box Plot node's Configure Dialog, select low_humidity_day as the nominal column and select air_pressure_9am as the numeric column. This means that data values for air_pressure_9am are to be partitioned according to low_humidity_day.



5. Now execute the workflow, and view the Conditional Box Plot to see the comparison.



The box plot of air_pressure_9am for *humidity_low vs. humidity_not_low* shows that, on average, pressure is higher for low-humidity days. Low-pressure weather systems are associated with stormy & rainy weather (with high humidity), while high-pressure systems bring sunny weather and low humidity. This relationship is captured in the box plot.

Save Your Workflow

Save your workflow using <control>-s on Windows or <command>-s on Mac.