Handling Missing Values in KNIME

Learning Objectives

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

1. Remove samples with missing values for a variable
2. Impute missing values with the column mean
3. Remove samples with any missing values

Problem Description

Recall that in the exercise on Data Exploration, we observed some missing values in the dataset in daily_weather.csv. In this exercise, we will look at some techniques to address those missing values.

Steps

Start a New Workflow

Let's start a new workflow for this exercise.

1. In the upper-menu bar, choose **File > New...**
2. In the window, choose **New KNIME Workflow** and click **Next >**
3. Name the workflow something descriptive such as "Missing Values Hands-On".
4. The destination LOCAL is fine. Click **Finish**. Check that your workflow shows up under LOCAL in the KNIME Explorer view.

Import the Dataset

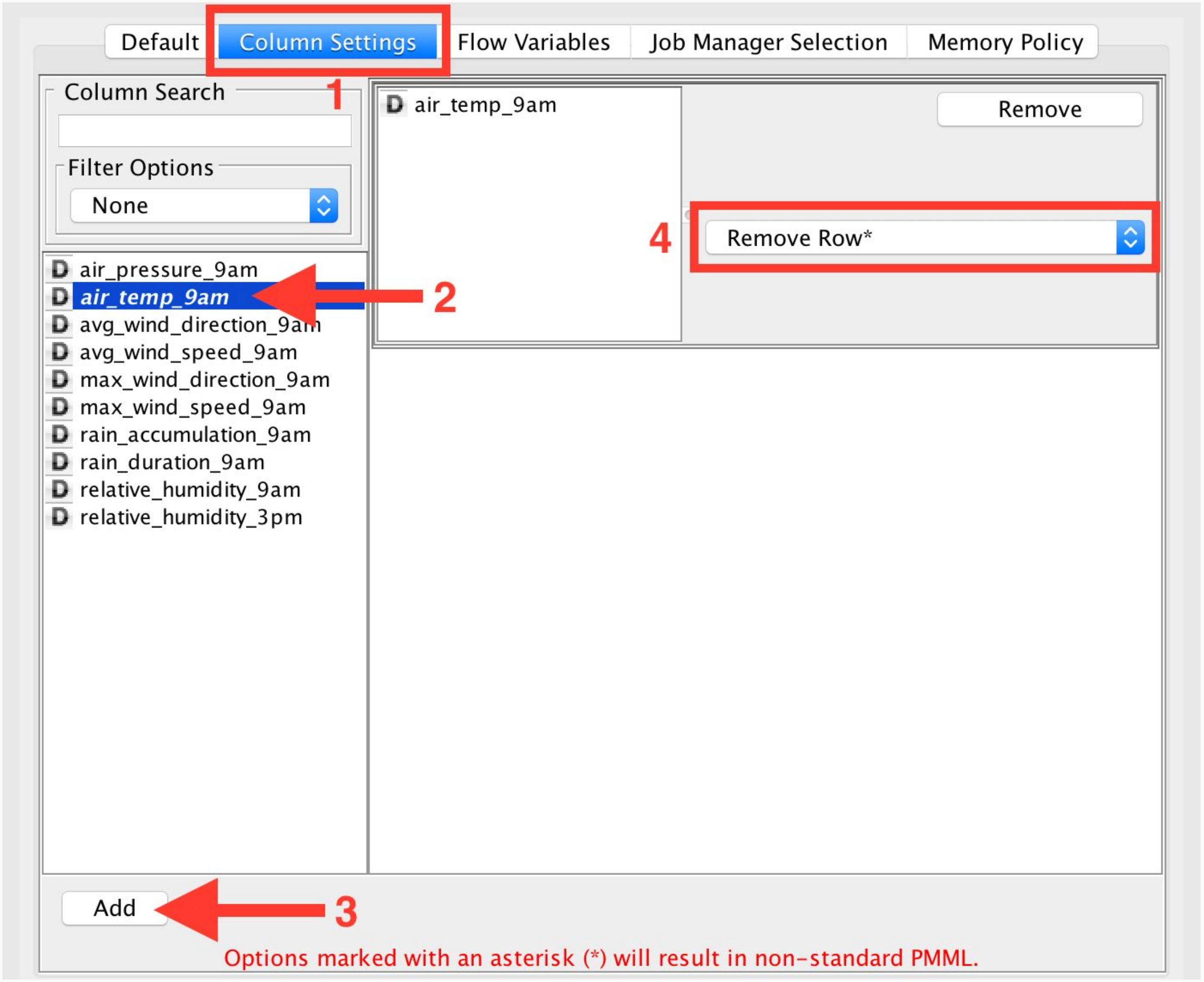
The first step in the workflow is to read in your dataset.

1. Drag the **File Reader** node onto the Workflow Editor.
2. Double-click it to open the Configure Dialog.
3. Click **Browse** and select the location of the dataset file **daily_weather.csv**, which you should have downloaded already.

Remove Samples

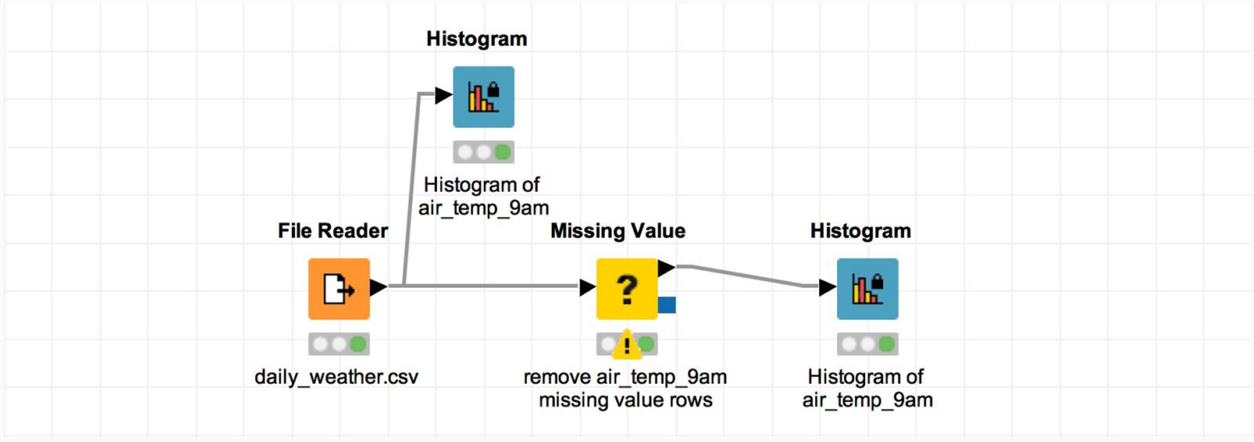
One method to handle missing values is to simply remove the rows that contain them. This can be accomplished with the **Missing Value** node. We will look at the variable **air_temp_9am** and remove any samples with a missing value for this variable.

1. First, let's look at the histogram of air_temp_9am to see if there are any missing values for this variable. Locate the **Histogram** node, and drag it to the Workflow Editor. Connect the Histogram node to the File Reader node. Execute the File Reader node. Open up the Histogram node's Configure Dialog, and set the **Binning column** and **Aggregation column** to **air_temp_9am**.
2. From the Node Repository, search for "Missing Value" and drag the **Missing Value** node onto the Workflow Editor. Connect the Missing Value node to the File Reader node.
3. In the Configure Dialog of the Missing Value node, go to the Column Settings tab. Select the air_temp_9am column, click on Add, and choose the **Remove Row*** in the combo-box. This means that any sample with a missing value for air_temp_9am will be removed.

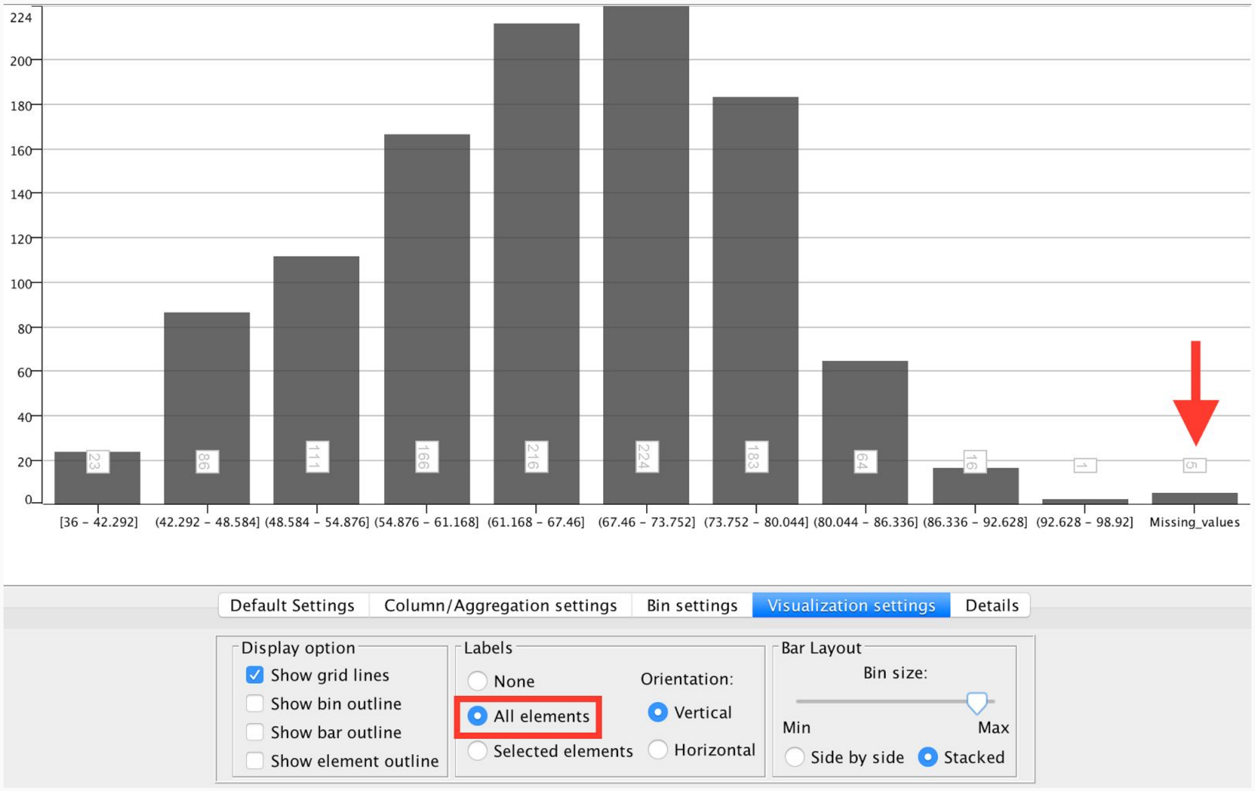


Options marked with an asterisk (*) will result in non-standard PMML.

4. Right-click and copy the Histogram node, and paste it to the right of the Missing Value node. Connect the black triangle output of the Missing Value node to the input of the 2nd Histogram node.


Histogram
Histogram of air_temp_9am
File Reader
daily_weather.csv
Missing Value
remove air_temp_9am missing value rows
Histogram
Histogram of air_temp_9am

We can now compare the 2 histograms to check the distribution of air_temp_9am before and after missing values are removed. We need to check this to ensure that removing samples with missing values do not significantly change the distribution of the variable. In the Histogram View **Visualization Settings** tab, you can select **Labels > All elements** to see the exact row count in each bin.



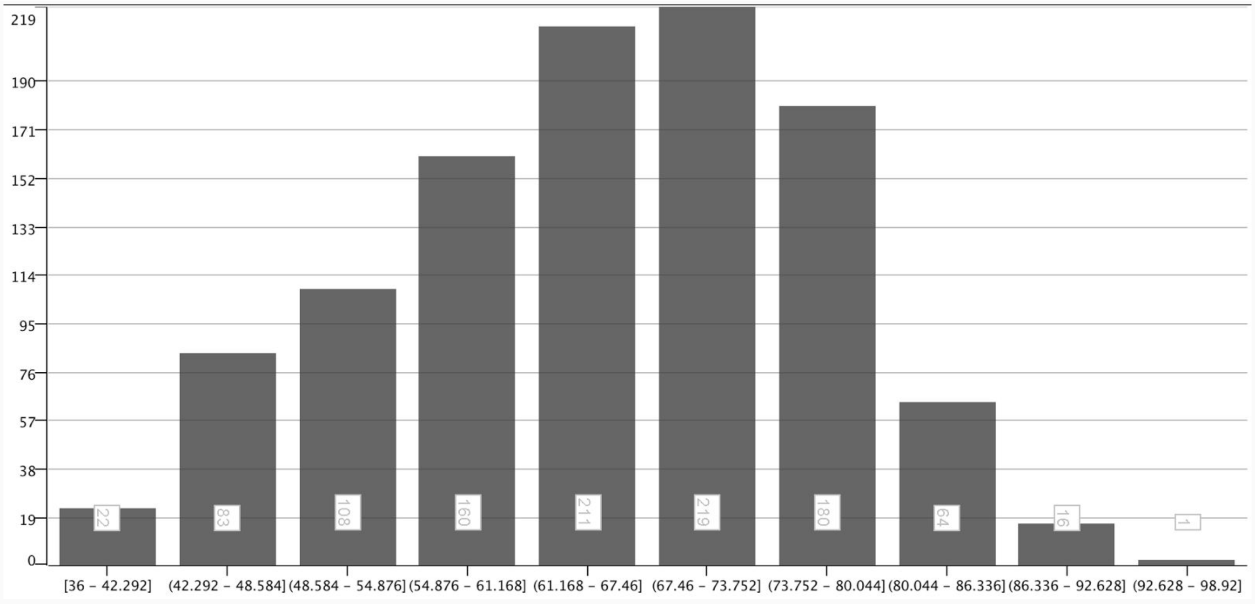
Display option
☑ Show grid lines
□ Show bin outline
□ Show bar outline
□ Show element outline

Labels
○ None
● All elements
○ Selected elements

Orientation:
● Vertical
○ Horizontal

Bar Layout
Bin size:
Min Max
○ Side by side ● Stacked

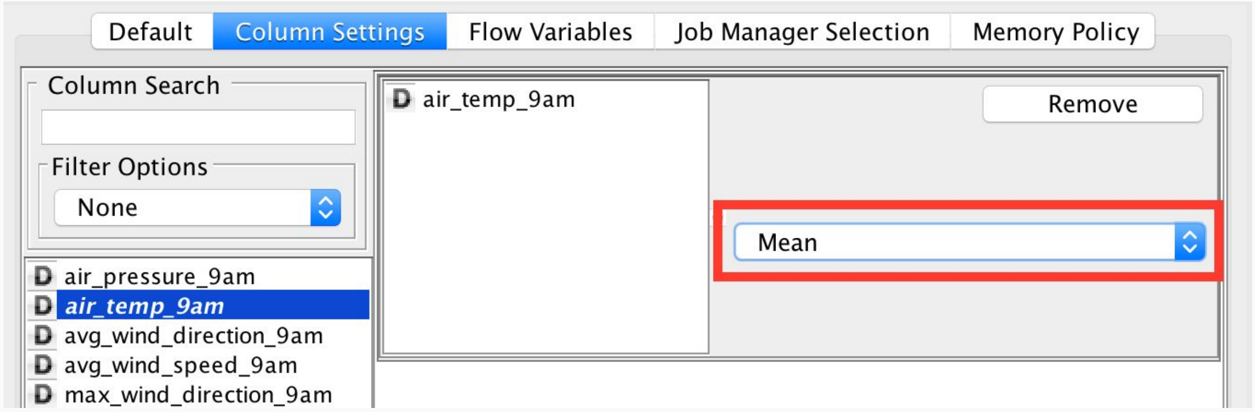
In this case, you should see that the first histogram has a Missing_values bin with some samples, whereas the second histogram has no Missing_values bin since they have been removed by the Missing Values node. Note also that the histograms are similar, indicating that removing missing values did not have a significant impact on the distribution of air_temp_9am.



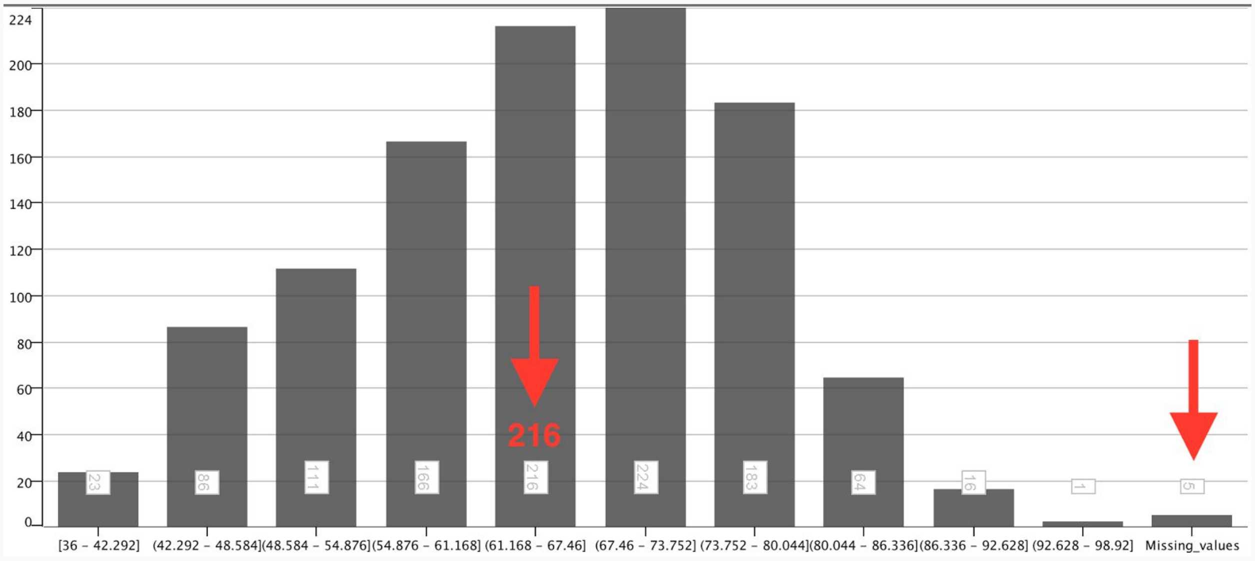
Impute Missing Values with Mean

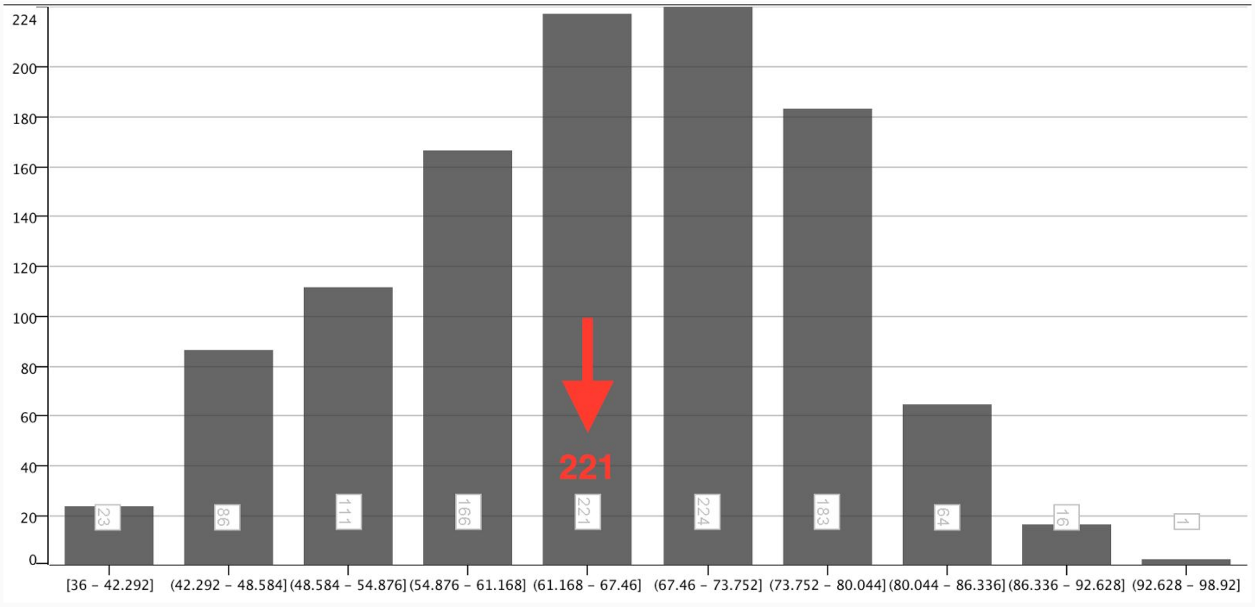
Another method to handle missing values is to replace the missing values with the mean or median of the column. This can be accomplished with the same **Missing Value** node pipeline we have already. For this exercise, we will use **mean**.

1. In the Configure Dialog of the Missing Value node, change the Column Setting for **air_temp_9am** from **Remove*** to **Mean**. This replaces any missing value for air_temp_9am with the mean value for that variable.



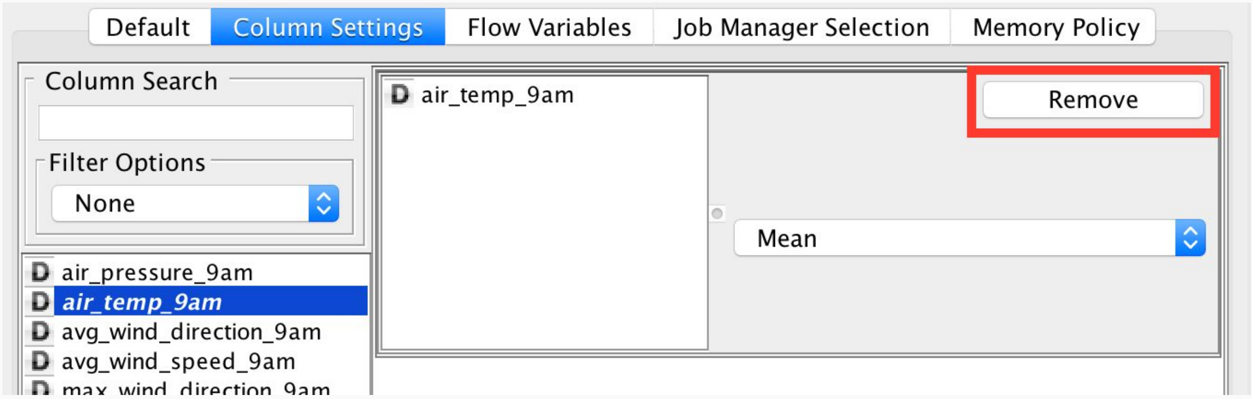
Note the difference in the number of samples in the 5th bin, but the distribution of air_temp_9am remains essentially the same.





Remove All Rows

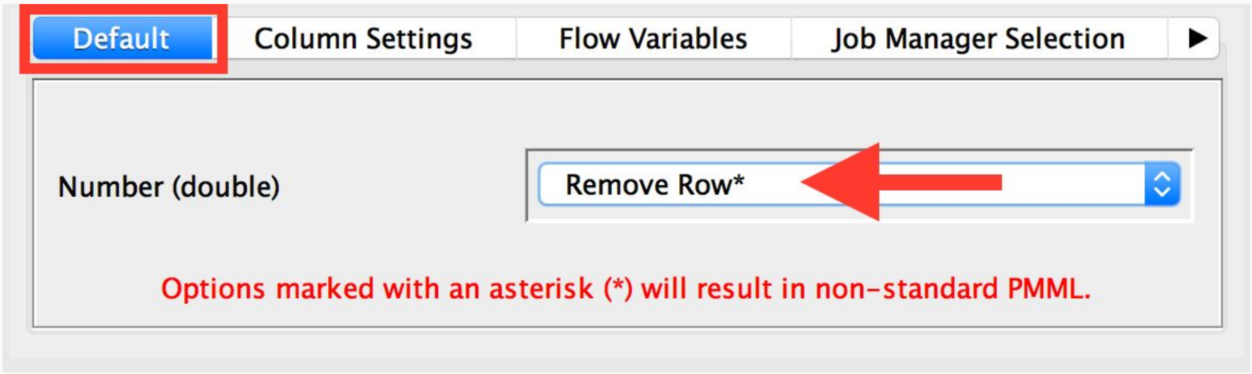
Thus far we have been removing missing values for specific variables, but we can also remove rows with any missing value for any variable. This is done by clicking the **Remove** button in the **Column Settings** tab for **air_temp_9am**.



air_temp_9am

Mean

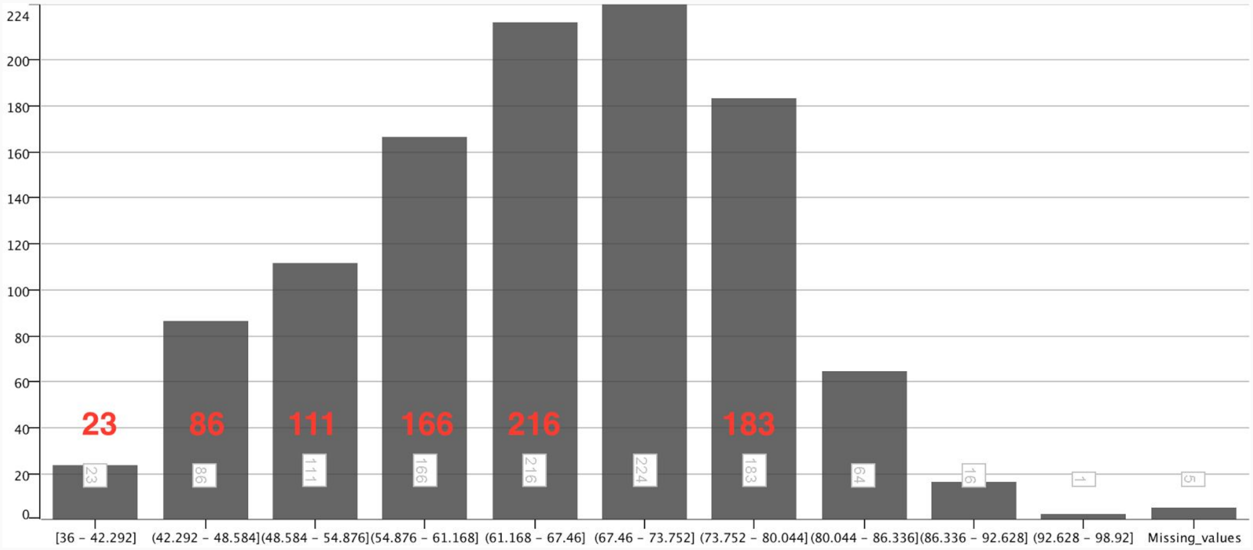
Then go to the **Default** tab and select **Remove Row*** in the dropdown.

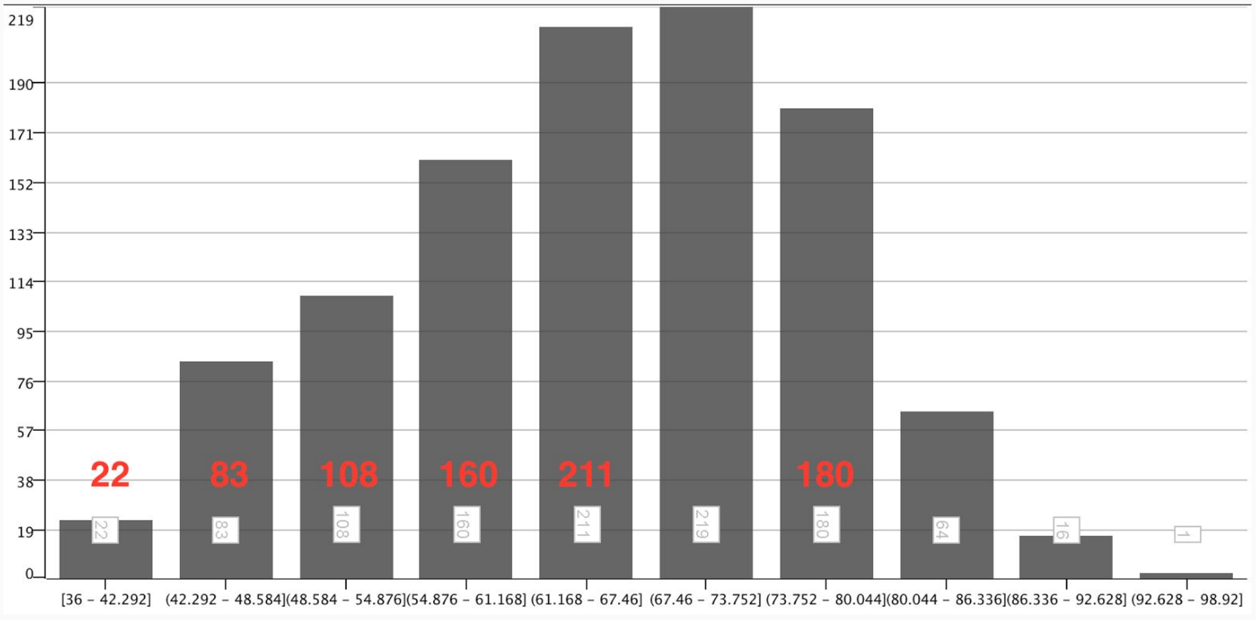


Number (double) Remove Row*

Options marked with an asterisk (*) will result in non-standard PMML.

Note there are differences in row count all along the bins, but the feature distribution has not changed its shape.





Save Your Workflow

Save your workflow using <control>-s on Windows or <command>-s on Mac, or selecting File>Save or File>Save As.