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

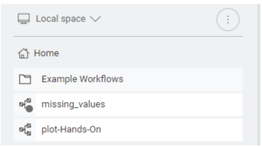
Recall that in the exercise on Data Exploration, we observed some missing values in the dataset in daily_weather.csv. In this exercise, we will look at some techniques to address those missing values.

Steps

Start a New Workflow

Let's start a new workflow for this exercise.

1. Open KNIME, and below the example projects click on *Create workflow in your local space*
2. Name the workflow something descriptive, e.g. "missing_values" 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

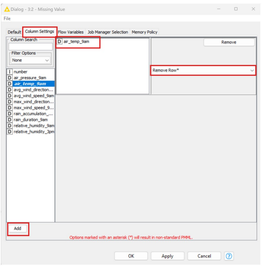
The first step in the workflow is to read in your dataset.

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/data/, 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.

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. 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 CSV Reader node.
2. 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. Click OK
3. Execute the workflow



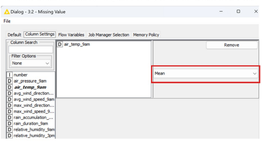
If you compare the outputs from both nodes, you will notice that after removing the missing values, our table has lost 5 samples. You can click first on the *CSV Reader* node to display the original table.

Now click on the *Missing Value* node to display the output.

Impute Missing Values with Mean

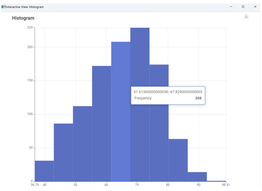
Another method to handle missing values is to replace the missing values with the mean or median of the column. This can be accomplished with the same **Missing Value** node pipeline we have already. For this exercise, we will use **mean**.

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.



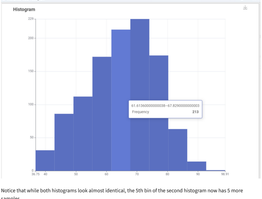
To see the effect in our analysis, we are going to compare two histograms. The first one will use the output of *CSV Reader* (which is the default table), and the other will use the output of *Missing Value*.

Open the Configure Dialog in each of the *Histogram* nodes, and make sure to specify to select **air_temp_9am** as the dimension and 10 bins per histogram. No other change is needed. The first Histogram will look as the following:



Notice that KNIME removes the missing values by default when creating the histogram. In the previous activity, we explicitly asked for the missing values to be displayed.

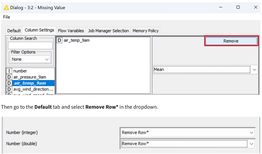
Now we can look at the second histogram, the one with the imputed means.



Notice that while both histograms look almost identical, the 5th bin of the second histogram now has 5 more samples.

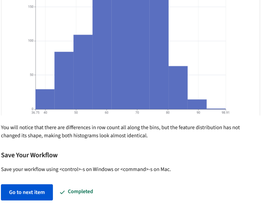
Remove All Rows

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**.



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

Run the workflow again. And compare the histograms. Remember that the first histogram will remove by default the samples with missing values for **air_temp_9am**.



You will notice that there are differences in row count at along the bins, but the feature distribution has not changed its shape, making both histograms look almost identical.

Save Your Workflow

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

Go to next Item ✓ Completed