Imputation Component

This component automatically fills missing values in datasets. For each column in the source dataset, select the most suitable method for processing missing values. Missing values are null values.

> **Note:** The node does not process fields with the `variant` data type (see data types).

Ports

Input

- **Input data source**: Accepts an input dataset (a data table).

Output

- **Output dataset**: Provides the resulting dataset (a data table).

Configuration wizard

Configure column mapping in the table or links interface.

Imputation settings

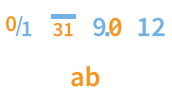
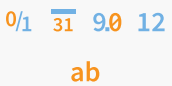
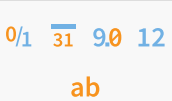
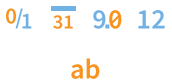
- **Source data ordered**: Tick this checkbox if you know the data is ordered. It could be a time series or another sequence ordered ascending or descending (e.g., by date or time). Use different missing data processing methods for ordered and unordered data.
- **Allowable percentage of nulls**: Enter a percentage value to set the threshold beyond which the node will not fill missing values. For example, if you set this parameter to 50, the node does not fill in fields that contain more than 50% missing values.
- **Random seed**: Specify the initial positive integer to initialise the pseudo-random number generator. The seed determines the generator's output sequence. If you reinitialise the generator with the same seed, you get the same number sequence. This parameter affects the random value replacement: When using the same input and seed, the node outputs the same results. Use the following commands for this parameter:

    - **Always randomize**: Always use a random seed.
    - **Generate**: Generate a new seed.
    - **Copy**: Copy the specified value to the clipboard.
- **Input fields**: View the list of fields available for processing. For each field, tick the checkbox to enable processing, and then select the method in the **Processing method** column:
    - **Replace with the previous value**: Replace detected missing values with the previous known (non-empty) value from the same column. If the dataset starts with missing values, the node leaves them empty until the first non-empty value.
    - **Replace with average**: Replace detected missing values with the column average.
    - **Replace with median**: Replace detected missing values with the column median.
    - **Replace with most frequent**: Replace detected missing values with the most probable value of the column, i.e., the *mean value of the most frequent interval*. The number of intervals varies depending on the sample size: the larger it is, the more intervals the node uses. The system employs different processing methods for discrete and continuous data:
        - **Discrete:** The node fills missing values with the most frequent value. If there are several such values, the *last* value is used.
        - **Continuous:** The node fills missing values with the mean value of the most frequent interval. If there are several such intervals, the system uses the *first* interval of those with the same maximum frequency.
    - **Replace with 0**: Replace detected missing values with `0`.
    - **Replace with random values**: Replace detected missing values with random values that the node generates in the range from the minimum to the maximum column value.
    - **Linear interpolation**: Replace detected missing values with intermediate values of a linear function that the node builds from known values (as if you drew a straight line between them).
    - **Cubic interpolation**: Replace detected missing values with intermediate values of a cubic spline (third-degree splines with a continuous first derivative) that the node builds from known values.
    - **Spline interpolation**: Replace detected missing values with intermediate values of a spline function that the node builds from known values.
    - **Leave unchanged**: Do not fill detected missing values.
    - **Delete records**: Exclude rows with detected missing values from the output dataset.
    - **Replace with set value**: Replace detected missing values with the default value `Unspecified` or with a custom value. To specify a value, click **... More**.

For each field, the set of available methods depends on the following three data characteristics at the same time (see data):

- ordered/unordered
- discrete/continuous
- data type

See the applicability table:

Method	Unordered dataset		Ordered dataset	
	Discrete	Continuous	Discrete	Continuous
Replace with the previous value				
Replace with average				
Replace with median				
Replace with most frequent				
Replace with 0				
Replace with random values				
Linear interpolation				
Cubic interpolation				
Spline interpolation				
Leave unchanged				
Delete records				
Replace with set value				

Read on: Binning Component

See also: Eliminate Outliers