Eliminate Outliers Component

The **Eliminate Outliers** component automatically corrects outliers and extreme values in your datasets. Define criteria for each field by specifying the allowable standard deviation or interquartile range.

Outliers deviate significantly from the mean. *Extreme values* deviate so far from typical data that they no longer align with the logic of the processes or phenomena you are studying.

Ports

Input

- **Input data source:** The data table you want to process.

Output

- **Output dataset:** The processed data table.
- **Outliers:** A table containing rows where the component detected outliers.
- **Extremes:** A table containing rows where the component detected extreme values.

Configuration wizard

- **Source data ordered**: Check this box if your numeric series follows a specific order (e.g., sorted by date or time). This setting affects specific processing methods but does not change how the system handles logical or string fields.
- **Input fields**: Select fields for processing from this list. Enable the checkbox for a field and click it to configure its specific processing method.
- **Determination of outliers and extreme values**: Choose one of two identification methods:
  - **Standard deviation**: Best for data that follows a normal distribution. The system flags values that deviate from the mean by more than your specified number of standard deviations. You can set different thresholds for outliers and extreme values.
  - **Interquartile width**: This method calculates the distance between the 1st and 3rd quartiles. The system treats values as anomalies if they deviate from the median by more than your specified interquartile width. Use this for non-normal distributions (see Interquartile width).

Apply these processing methods to both outliers and extreme values:

- **Leave unchanged**: Keep the values as they are.
- **Delete records**: Remove rows containing anomalies from the dataset.
- **Replace with average**: Replace anomalies with the column's mean value.
- **Replace with median**: Replace anomalies with the column's median value.
- **Replace with most frequent**: Replace anomalies with the most common value. The system uses the mean value from the most likely interval, adjusting the number of intervals based on your sample size.
- **Replace with set value**: Enter a specific manual value to replace anomalies.
- **Limit**: Replace anomalies with the nearest boundary value of the threshold.

The available methods for each field depend on three characteristics (see data):

- Data type
- Ordered/unordered
- Discrete/continuous

The following table shows the applicability of different methods depending on these data properties:

Method	Unordered dataset		Ordered dataset	
	Discrete	Continuous	Discrete	Continuous
Leave unchanged	0/1 31 9.0 12 ab	31 9.0 12	0/1 31 9.0 12 ab	31 9.0 12
Delete records	0/1 31 9.0 12 ab	31 9.0 12		
Replace with average	31	31 9.0 12	31	31 9.0 12
Replace with median	31 9.0 12	31 9.0 12	31 9.0 12	31 9.0 12
Replace with most frequent	0/1 31 9.0 12 ab	31 9.0 12	0/1 31 9.0 12 ab	31 9.0 12
Replace with set value	0/1 31 9.0 12 ab	31 9.0 12	0/1 31 9.0 12 ab	
Limit	0/1 31 9.0 12 ab	31 9.0 12	0/1 31 9.0 12 ab	31 9.0 12

Read on: Smoothing Component

See also:

- Imputation

- Outliers