Data Quality Visualizer

Use this visualizer to perform a comprehensive data quality assessment for each field.

Views and options

Here is the list of all panes and commands available in the **Data Quality** visualizer:

- **Summary**: This is the first screen that opens when you enter the visualizer. Select the fields for analysis. After processing, the **Summary** screen shows the general data quality indicators and the problems found in the selected data fields.
- **Discrete**: Review data quality for discrete fields.
- **Continuous**: Review data quality for continuous fields.
- **Parameters**: Select and configure indicators:
  - **Selection of Indicators**: Select the indicators that you need to assess.
  - **Configure Parameters**: Select the identification method, configure the method parameters, and set the **Maximum percentage of null data** value.
- **Absolute or relative values**: Toggle the display mode of quantitative values between absolute and relative (percent).
- **Export**: Export a report file. The system exports only the currently open view (**Summary**, **Discrete**, or **Continuous**). On the toolbar, you can see the selected export format:
  - **Export to XLS (html)**
  - **Export to HTML**
- **Drill-through**: Show the input dataset filtered by the selected indicator for the selected field. For example, if you select a **Nulls** cell, the system shows all records of the input dataset that have nulls in the selected field. The **Drill-through** command is available only for the **Discrete** and **Continuous** tables.
- **Histogram**: Show a pane with a histogram for the selected field. The **Histogram** option is available only for the **Discrete** and **Continuous** tables.

Dataset statistics

Here is the list of available statistics that appears on the right of the **Summary** screen:

- **Determination of outliers and extreme values**: Shows the technique used to identify outliers and extreme values.. By default, *standard deviation*.
- **Columns**: The total number of columns in the dataset.
- **Rows**: The total number of records in the dataset.

- **Fully filled fields**: The number of fields that do not contain missing values, as a percentage.
- **Fully filled records**: The number of records that do not contain missing values, as a percentage.
- **Suitable columns**: The number of columns used for analysis. The system treats a column as unsuitable if:
  - The percentage of **Nulls** is higher than **Maximum percentage of null data**. (See Configure parameters.)
  - The **Quality indicator** of a discrete field is *0* (that is, the field contains a constant).
- **RPV index**: The ratio of the number of records to the number of columns.

Configuring tables

When you open the **Data Quality** visualizer, the system shows the **Summary** table with the list of fields. Use the checkboxes to select the fields for processing. The **Discrete** and **Continuous** tables are not available at this step. Click the **Parameters** button to configure the required quality indicators.

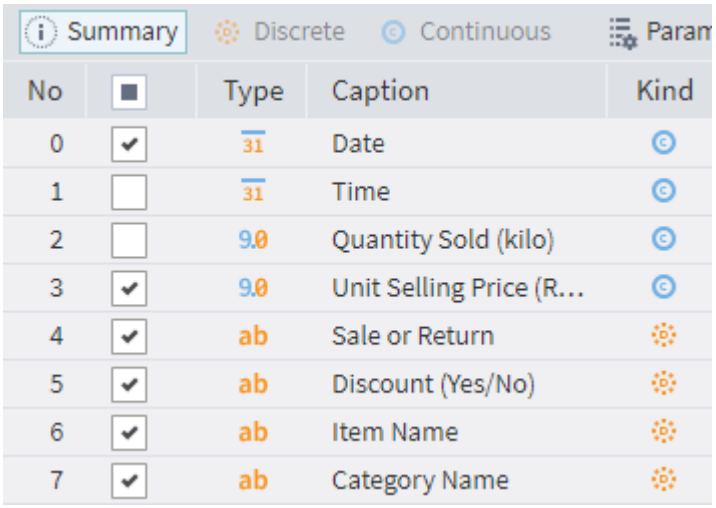


Figure 1. Selecting fields for processing

Selection of indicators

The table shows the list of quality indicators that may be calculated for the selected fields.

Indicator	Data kind	Value
Quality indicator		Shows the diversity of values in the field

Indicator	Data kind	Value
Histogram	P C	Histogram preview
Extremes	P C	The number of records with extreme values
Empty	P	The number of records with empty values
Trailing spaces	P	The number of records with values that end with a whitespace character. (This indicator does not include records that consist only of whitespace characters.)
Range of values	P	The range of values in the `<first value>...<last value>` format in the sorted list of values
Zeros	C	The number of records in which the values are equal to zero
Box plot	C	Calculates the following indicators: **Minimum outlier**, **Lower quartile**, **Median**, **Upper quartile**, **Maximum outlier**, **Sample maximum**
Minimum	C	The minimum value
Mean	C	The mean value
Mode)	P C	The most frequently occurring value[1]
Nulls	P C	The number of records in which the value is missing for the field
Values	P	Displays a chart that shows the diversity of values by row
Outliers	P C	The number of records in which the value is an outlier
Whitespaces	P	The number of records with values that consist only of whitespace characters
String lengths	P	The range of string lengths in the format: `<minimum length> – <maximum length>`
Negative	C	The number of records with negative values

Indicator	Data kind	Value
Infinities		The number of records in which infinity values
Unique		The number of unique values
Maximum		The maximum value
Median		The median for the field
Monotony		Shows how uniform the field values are (**Descending**, **Increasing**, or **Non-monotonic**)

> [1]. If the field has several mode values, the system shows the first of these values and the multimodal icon. ↩

Configure parameters

Switch to the **Configure parameters** tab in the **Parameters** window and select the detection technique for outliers and extreme values, marking the radio button for one of the two **Identification method** options:

- **Standard deviation**
- **Interquartile width**

Set the required values for the selected method:

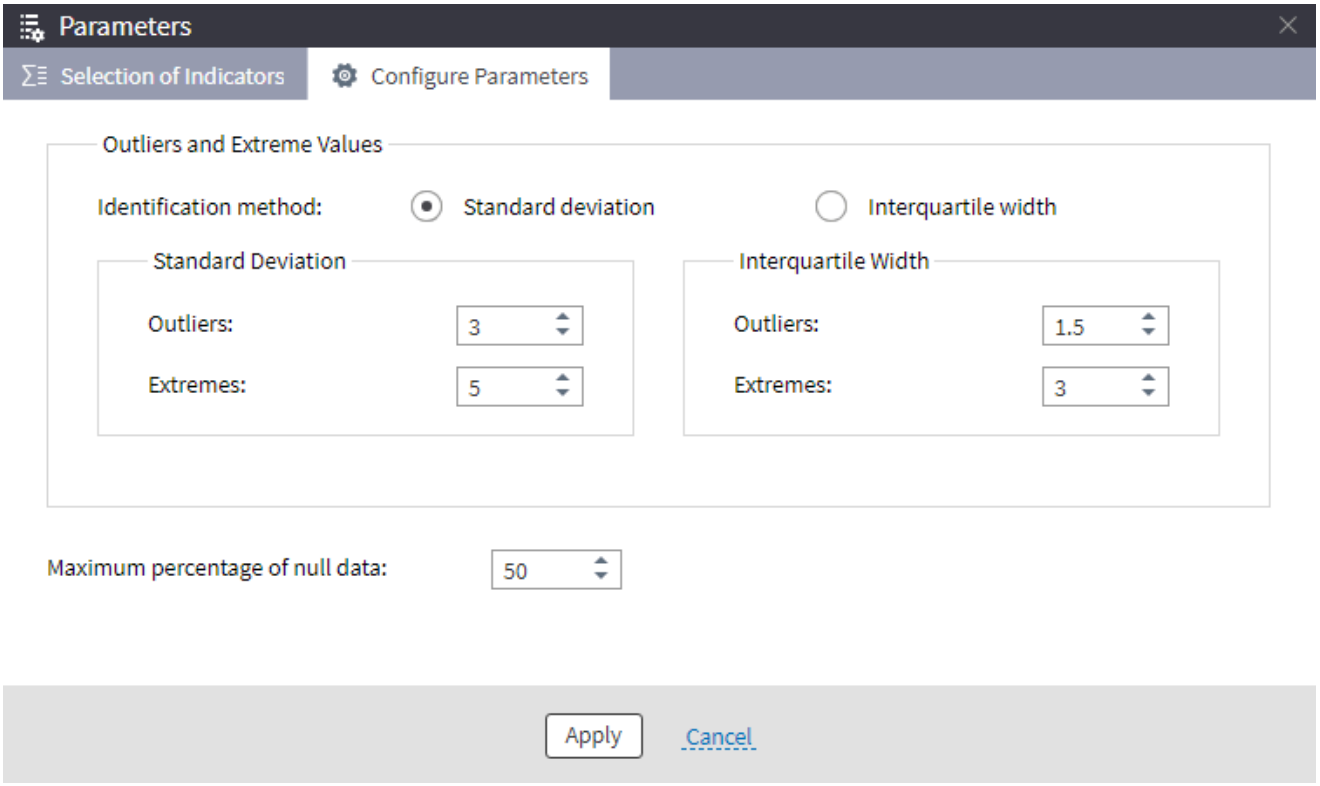


Figure 2. Configure Parameters

Calculating quality statistics

After setting up the fields and indicators for the data quality analysis, click the **Calculate statistics** button in the center of the visualizer area.

The system processes the selected fields, showing the general results in the central **Problems** and **Kinds or problems** columns.

Problems and problem types

The **Problems** column shows the percentage of problem rows for each field.

The **Kinds of problems** columnn displays the quality issues the system encountered in each field. Possible quality issues are:

- **Nulls**: The number of records that contain missing values.
- **Empty**: The number of records that contain empty values.
- **Zeros**: The number of records that contain zero values.
- **Whitespaces**: The number of records that contain only whitespace characters (space, non-breaking space, tab, etc.).
- **Infinities**: The number of records that contain an infinity value.
- **Constant**: Indicates that the field contains only one unique value. (The field may also contain nulls.)

- **Unique**: Indicates that every value in this field is unique.
- **Conversion error**: Appears if the field contains values that you cannot convert to the data type of the field.
- **Not calculated**: Marks fields that do not contain calculated indicators.
- **No data**: This problem appears when you analyze an empty dataset.
- **Extreme**: The number of records in which the value is extreme.
- **Outliers**: The number of records in which the value is an outlier.
- **Negative**: The number of records with negative values.
- **Trailing spaces**: The number of records with values that end with a whitespace character (this indicator does not include records that consist only of whitespace characters).

Next to the each issue value, the system shows the percentage of problem rows relative to the total number of rows, and in parentheses the number of rows with that issue.

Problems	Kinds of problems
100,00%	Extreme - 2,96% (114) Zeros - 97,04% (3 735)
100,00%	Extreme - 0,62% (24) Zeros - 99,38% (3 825)
100,00%	Extreme - 2,73% (105) Zeros - 97,27% (3 744)
84,59%	Zeros - 84,59% (3 256)
68,75%	Zeros - 68,75% (2 646)
55,75%	Zeros - 55,75% (2 146)
51,44%	Extreme - 0,03% (1) Outliers - 0,47% (18) Zeros - 50,95% (1 961)
50,95%	Zeros - 50,95% (1 961)
1,40%	Extreme - 0,05% (2) Outliers - 1,35% (52)
1,30%	Extreme - 0,88% (34) Outliers - 0,42% (16)
1,22%	Extreme - 0,08% (3) Outliers - 1,14% (44)
0,96%	Outliers - 0,96% (37)
0,60%	Extreme - 0,08% (3) Outliers - 0,52% (20)
0,55%	Extreme - 0,05% (2) Outliers - 0,49% (19)

Figure 3. Types of problems

Detailed statistics for discrete and continuos fields

To view the detailed statistics, switch to the **Discrete** or **Continuous** table. (These buttons become active on the toolbar.)

In the table that opens, click the ⌄ **Options** button next to each column header to sort fields by the selected indicator and to hide or show indicator columns.

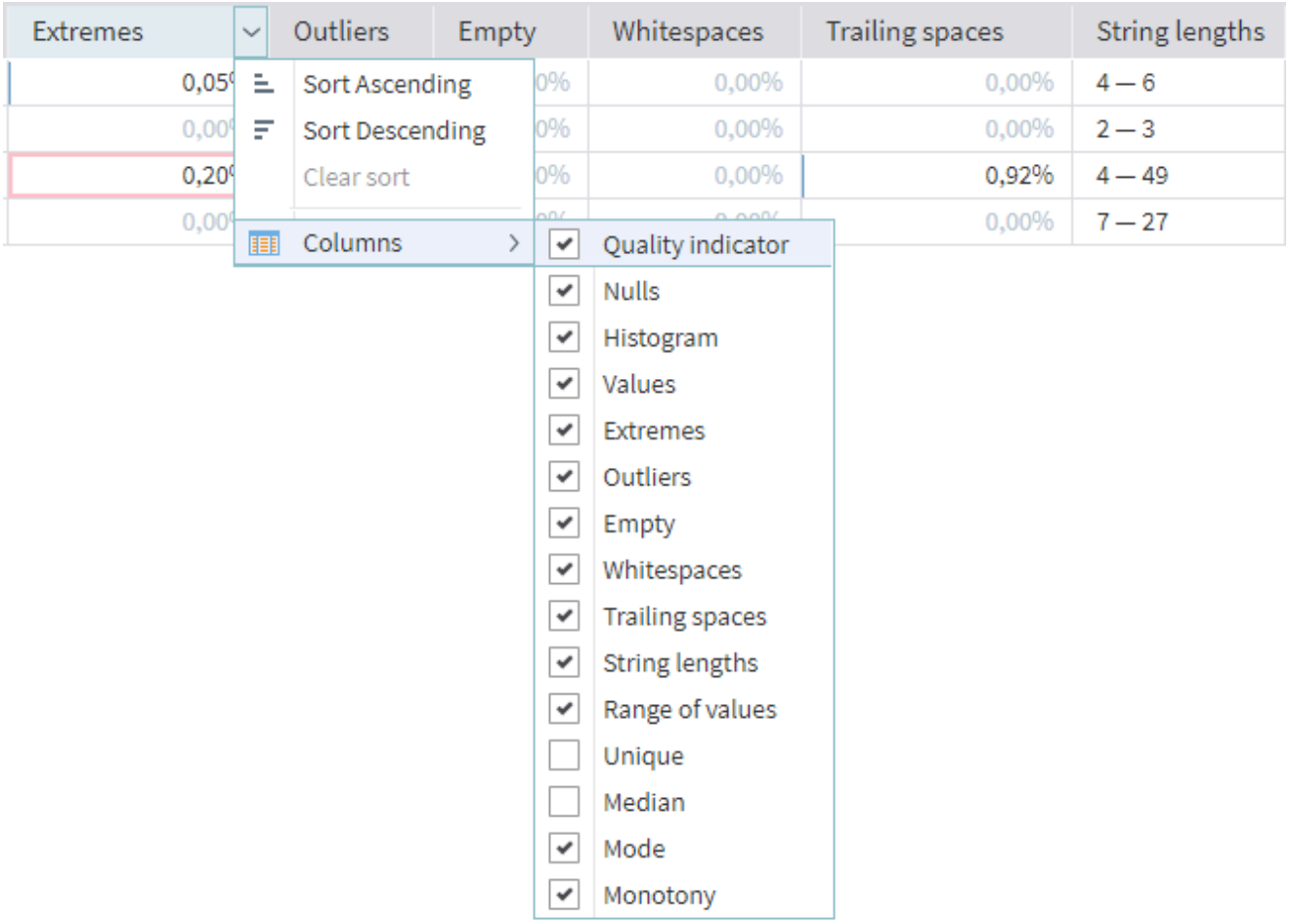
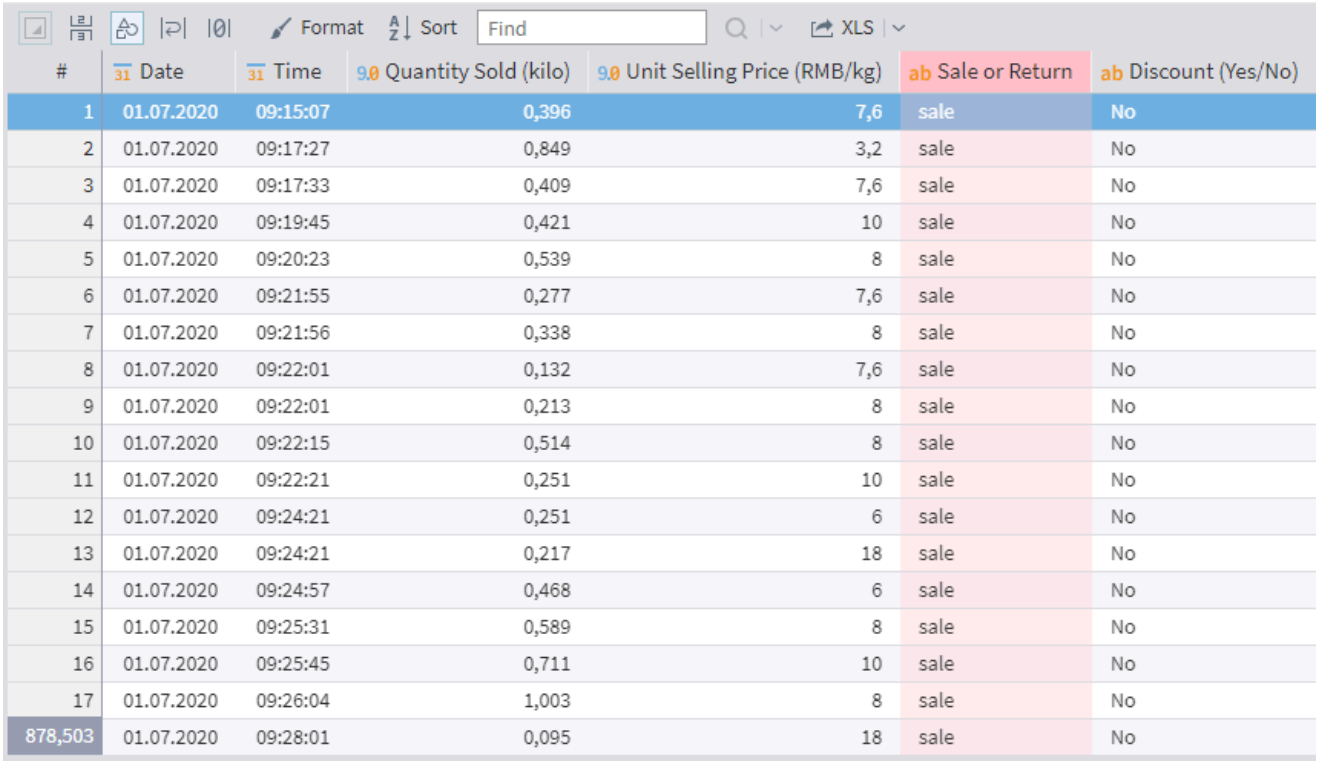


Figure 4. Sorting columns and selecting columns to display

Select a cell and click **Drill-through** to see all records from that cell. (The drill-through pane appears at the bottom.)



Format Sort Find XLS

#	Date	Time	Quantity Sold (kilo)	Unit Selling Price (RMB/kg)	Sale or Return	Discount (Yes/No)
1	01.07.2020	09:15:07	0,396	7,6	sale	No
2	01.07.2020	09:17:27	0,849	3,2	sale	No
3	01.07.2020	09:17:33	0,409	7,6	sale	No
4	01.07.2020	09:19:45	0,421	10	sale	No
5	01.07.2020	09:20:23	0,539	8	sale	No
6	01.07.2020	09:21:55	0,277	7,6	sale	No
7	01.07.2020	09:21:56	0,338	8	sale	No
8	01.07.2020	09:22:01	0,132	7,6	sale	No
9	01.07.2020	09:22:01	0,213	8	sale	No
10	01.07.2020	09:22:15	0,514	8	sale	No
11	01.07.2020	09:22:21	0,251	10	sale	No
12	01.07.2020	09:24:21	0,251	6	sale	No
13	01.07.2020	09:24:21	0,217	18	sale	No
14	01.07.2020	09:24:57	0,468	6	sale	No
15	01.07.2020	09:25:31	0,589	8	sale	No
16	01.07.2020	09:25:45	0,711	10	sale	No
17	01.07.2020	09:26:04	1,003	8	sale	No
878,503	01.07.2020	09:28:01	0,095	18	sale	No

Figure 5. Cell drill-through

Select a cell and click **Histogram** to view a detailed histogram. (This opens on the right side of the visualizer window.) For discrete fields, you can select which values to display:

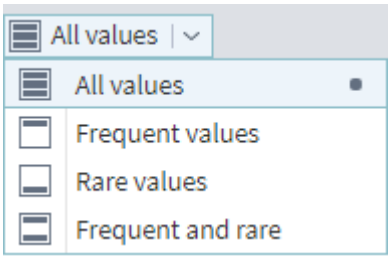


Figure 6. Display options for the histogram

Read on: Cube Visualizer