Feature Statistics

Show basic statistics for data features.

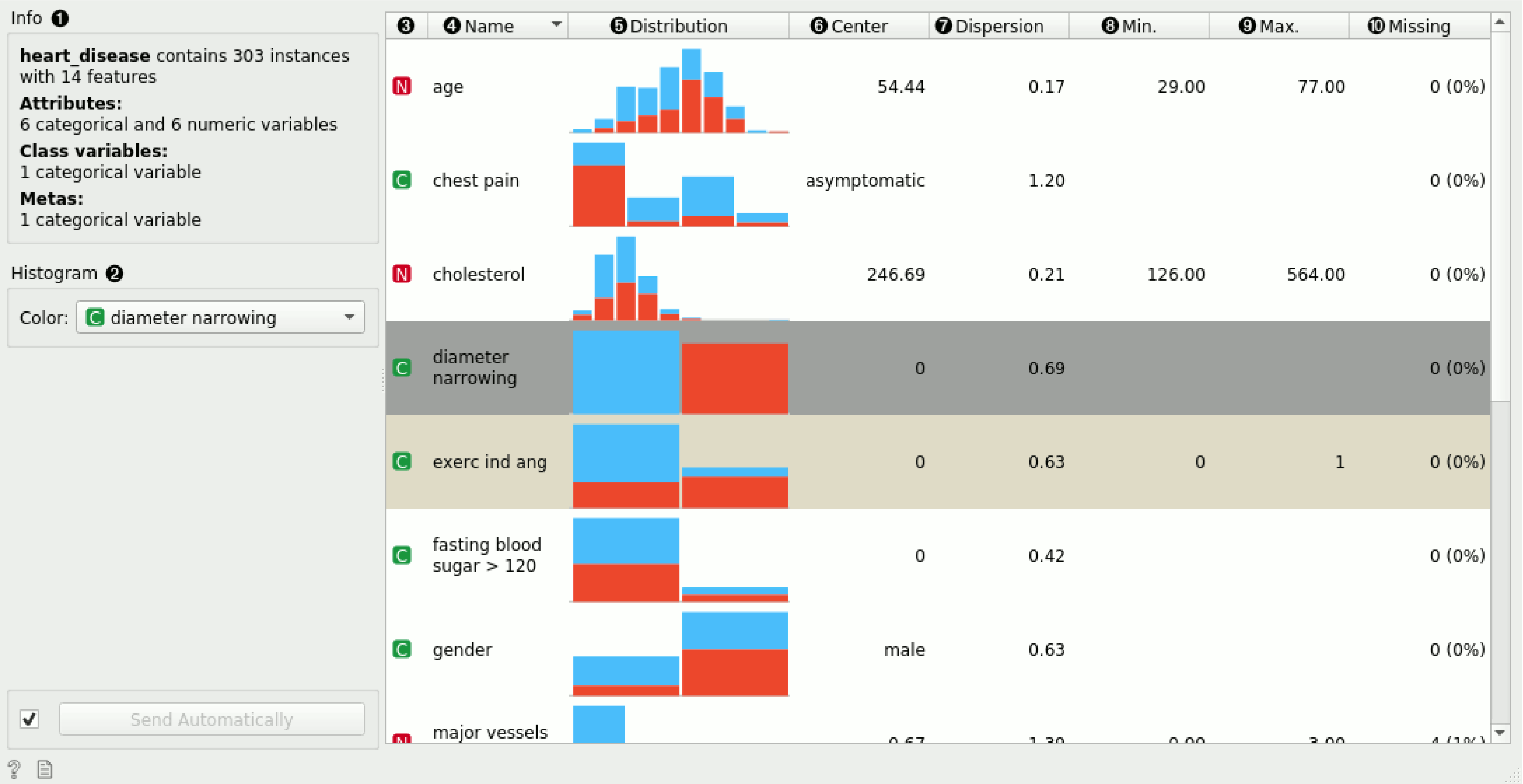
Inputs

- Data: input data

Outputs

- Reduced data: table containing only selected features
- Statistics: table containing statistics of the selected features

The **Feature Statistics** widget provides a quick way to inspect and find interesting features in a given data set.



Info ❶

heart_disease contains 303 instances with 14 features
Attributes:
6 categorical and 6 numeric variables
Class variables:
1 categorical variable
Metas:
1 categorical variable

Histogram ❷

Color: C diameter narrowing

☑ Send Automatically

❸	❹ Name	❺ Distribution	❻ Center	❼ Dispersion	❽ Min.	❾ Max.	❿ Missing
N	age		54.44	0.17	29.00	77.00	0 (0%)
C	chest pain		asymptomatic	1.20			0 (0%)
N	cholesterol		246.69	0.21	126.00	564.00	0 (0%)
C	diameter narrowing		0	0.69			0 (0%)
C	exerc ind ang		0	0.63	0	1	0 (0%)
C	fasting blood sugar > 120		0	0.42			0 (0%)
C	gender		male	0.63			0 (0%)
N	major vessels		0.67	1.30	0.00	3.00	4 (1%)

The Feature Statistics widget on the *heart-disease* data set. The feature *exerc ind ang* was manually changed to a meta variable for illustration purposes.

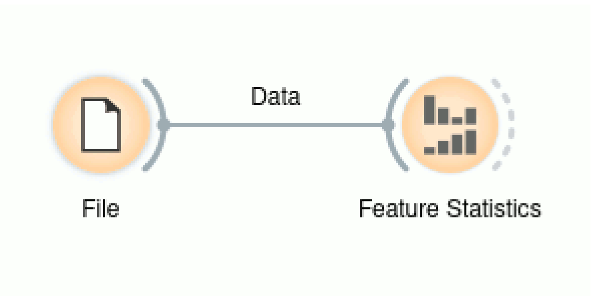
1. Info on the current data set size and number and types of features
2. The histograms on the right can be colored by any feature. If the selected feature is categorical, a discrete color palette is used (as shown in the example). If the selected feature is numerical, a continuous color palette is used. The table on the right contains statistics about each feature in the data set. The features can be sorted by each statistic, which we now describe.
3. The feature type - can be one of categorical, numeric, time and string.
4. The name of the feature.

5. A histogram of feature values. If the feature is numeric, we appropriately discretize the values into bins. If the feature is categorical, each value is assigned its own bar in the histogram.
6. The central tendency of the feature values. For categorical features, this is the mode. For numeric features, this is mean value.
7. The dispersion of the feature values. For categorical features, this is the entropy of the value distribution. For numeric features, this is the coefficient of variation.
8. The minimum value. This is computed for numerical and ordinal categorical features.
9. The maximum value. This is computed for numerical and ordinal categorical features.
10. The number of missing values in the data.

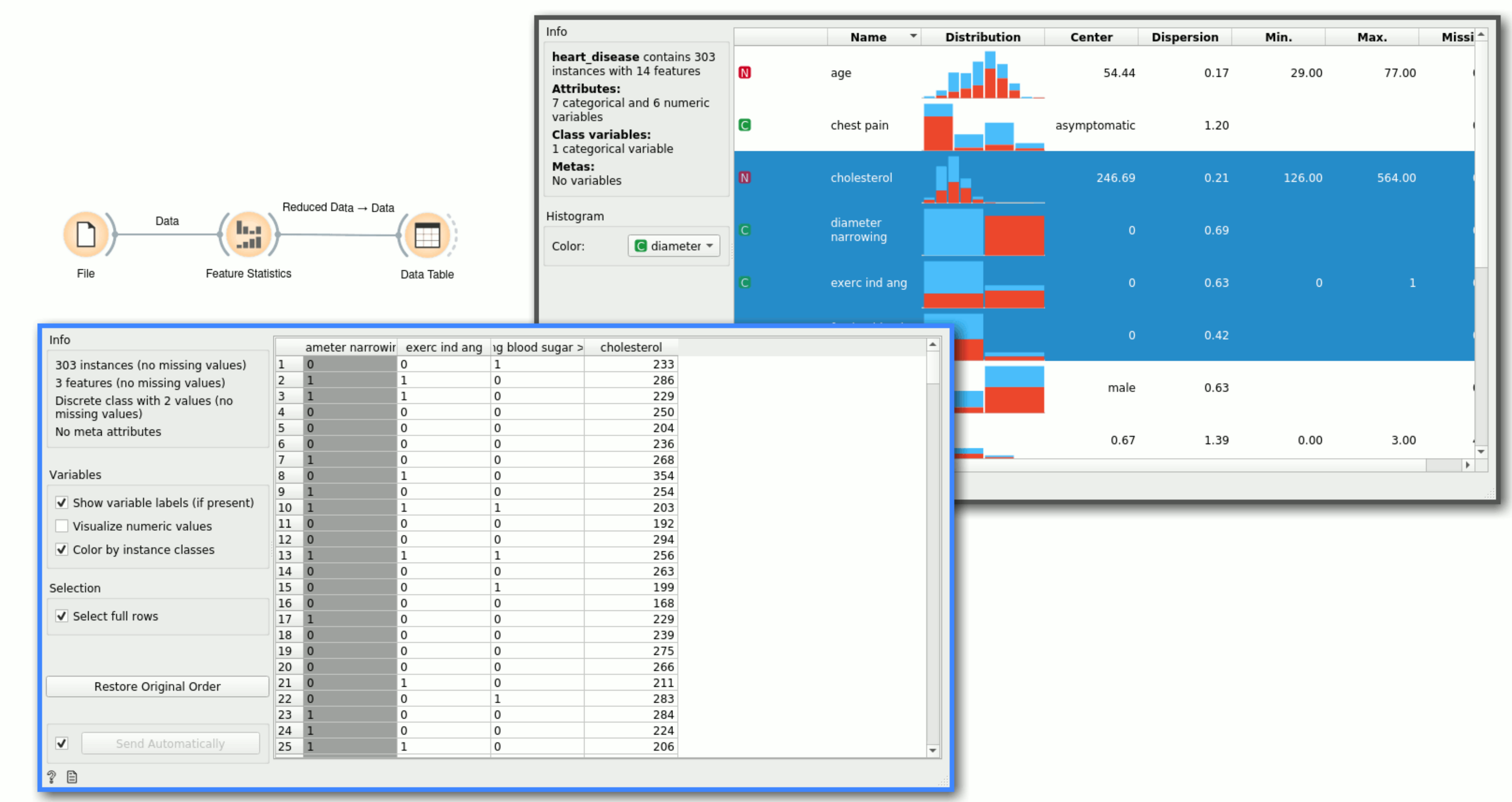
Notice also that some rows are colored differently. White rows indicate regular features, gray rows indicate class variables and the lighter gray indicates meta variables.

Example

The Feature Statistics widget is most often used after the File widget to inspect and find potentially interesting features in the given data set. In the following examples, we use the *heart-disease* data set.



Once we have found a subset of potentially interesting features, or we have found features that we would like to exclude, we can simply select the features we want to keep. The widget outputs a new data set with only these features.



Info

heart_disease contains 303 instances with 14 features

Attributes:
7 categorical and 6 numeric variables

Class variables:
1 categorical variable

Metas:
No variables

Histogram

Color: diameter

	Name	Distribution	Center	Dispersion	Min.	Max.	Missi
N	age		54.44	0.17	29.00	77.00	
C	chest pain		asymptomatic	1.20			
N	cholesterol		246.69	0.21	126.00	564.00	
C	diameter narrowing		0	0.69			
C	exerc ind ang		0	0.63	0	1	
			0	0.42			
			male	0.63			
			0.67	1.39	0.00	3.00	

Info

303 instances (no missing values)
3 features (no missing values)
Discrete class with 2 values (no missing values)
No meta attributes

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

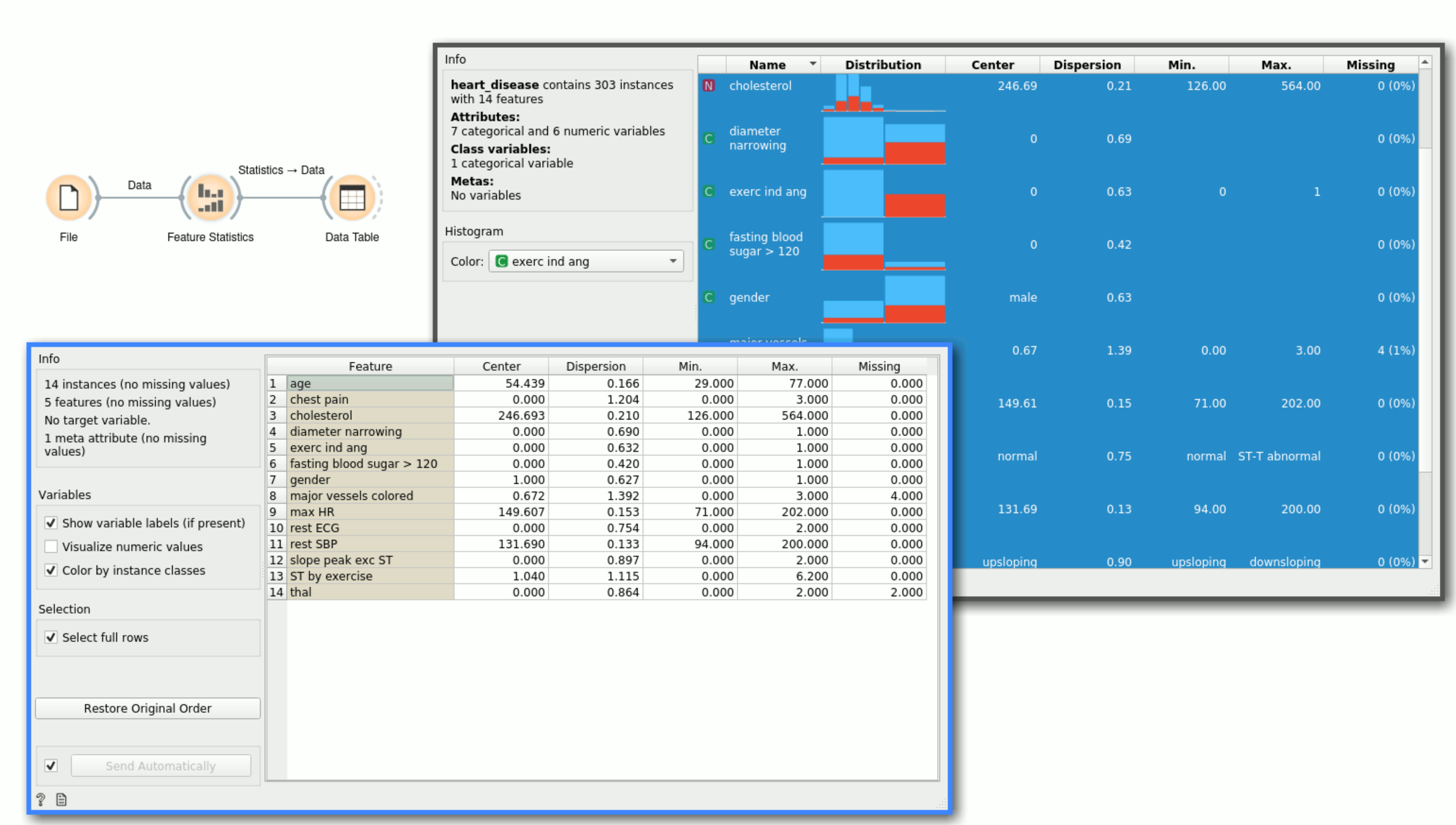
☑ Select full rows

Restore Original Order

☑ Send Automatically

	ameter narrowir	exerc ind ang	ıg blood sugar >	cholesterol
1	0	0	1	233
2	1	1	0	286
3	1	1	0	229
4	0	0	0	250
5	0	0	0	204
6	0	0	0	236
7	1	0	0	268
8	0	1	0	354
9	1	0	0	254
10	1	1	1	203
11	0	0	0	192
12	0	0	0	294
13	1	1	1	256
14	0	0	0	263
15	0	0	1	199
16	0	0	0	168
17	1	0	0	229
18	0	0	0	239
19	0	0	0	275
20	0	0	0	266
21	0	1	0	211
22	0	0	1	283
23	1	0	0	284
24	1	0	0	224
25	1	1	0	206

Alternatively, if we want to store feature statistics, we can use the *Statistics* output and manipulate those values as needed. In this example, we simply select all the features and display the statistics in a table.



Info

heart_disease contains 303 instances with 14 features

Attributes:
7 categorical and 6 numeric variables

Class variables:
1 categorical variable

Metas:
No variables

Histogram

Color: C exerc ind ang

	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N	cholesterol		246.69	0.21	126.00	564.00	0 (0%)
C	diameter narrowing		0	0.69			0 (0%)
C	exerc ind ang		0	0.63	0	1	0 (0%)
C	fasting blood sugar > 120		0	0.42			0 (0%)
C	gender		male	0.63			0 (0%)
	major vessels		0.67	1.39	0.00	3.00	4 (1%)
			149.61	0.15	71.00	202.00	0 (0%)
			normal	0.75	normal	ST-T abnormal	0 (0%)
			131.69	0.13	94.00	200.00	0 (0%)
			upsloping	0.90	upsloping	downsloping	0 (0%)

Info

14 instances (no missing values)
5 features (no missing values)
No target variable.
1 meta attribute (no missing values)

Variables

☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

☑ Send Automatically

	Feature	Center	Dispersion	Min.	Max.	Missing
1	age	54.439	0.166	29.000	77.000	0.000
2	chest pain	0.000	1.204	0.000	3.000	0.000
3	cholesterol	246.693	0.210	126.000	564.000	0.000
4	diameter narrowing	0.000	0.690	0.000	1.000	0.000
5	exerc ind ang	0.000	0.632	0.000	1.000	0.000
6	fasting blood sugar > 120	0.000	0.420	0.000	1.000	0.000
7	gender	1.000	0.627	0.000	1.000	0.000
8	major vessels colored	0.672	1.392	0.000	3.000	4.000
9	max HR	149.607	0.153	71.000	202.000	0.000
10	rest ECG	0.000	0.754	0.000	2.000	0.000
11	rest SBP	131.690	0.133	94.000	200.000	0.000
12	slope peak exc ST	0.000	0.897	0.000	2.000	0.000
13	ST by exercise	1.040	1.115	0.000	6.200	0.000
14	thal	0.000	0.864	0.000	2.000	2.000