Purge Domain

Removes unused attribute values and useless attributes, sorts the remaining values.

Inputs

- Data: input dataset

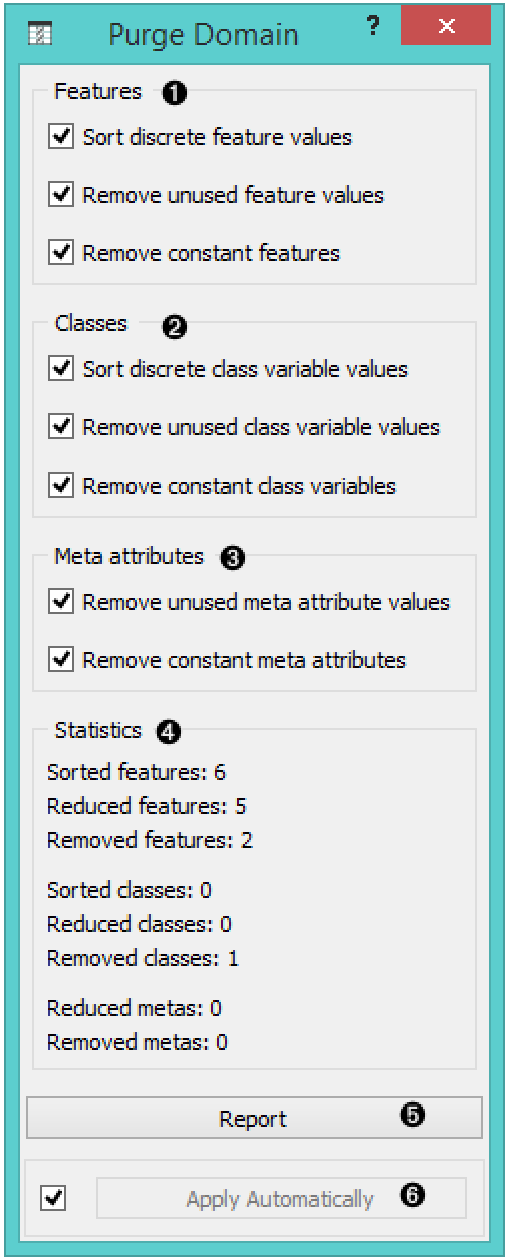
Outputs

- Data: filtered dataset

Definitions of nominal attributes sometimes contain values which don’t appear in the data. Even if this does not happen in the original data, filtering the data, selecting exemplary subsets and alike can remove all examples for which the attribute has some particular value. Such values clutter data presentation, especially various visualizations, and should be removed.

After purging an attribute, it may become single-valued or, in extreme case, have no values at all (if the value of this attribute was undefined for all examples). In such cases, the attribute can be removed.

A different issue is the order of attribute values: if the data is read from a file in a format in which values are not declared in advance, they are sorted “in order of appearance”. Sometimes we would prefer to have them sorted alphabetically.



Purge Domain

Features 1

☑ Sort discrete feature values

☑ Remove unused feature values

☑ Remove constant features

Classes 2

☑ Sort discrete class variable values

☑ Remove unused class variable values

☑ Remove constant class variables

Meta attributes 3

☑ Remove unused meta attribute values

☑ Remove constant meta attributes

Statistics 4

Sorted features: 6
Reduced features: 5
Removed features: 2

Sorted classes: 0
Reduced classes: 0
Removed classes: 1

Reduced metas: 0
Removed metas: 0

☑

1. Purge attributes.
2. Purge classes.
3. Purge meta attributes.
4. Information on the filtering process.

5. Produce a report.
6. If *Apply automatically* is ticked, the widget will output data at each change of widget settings.

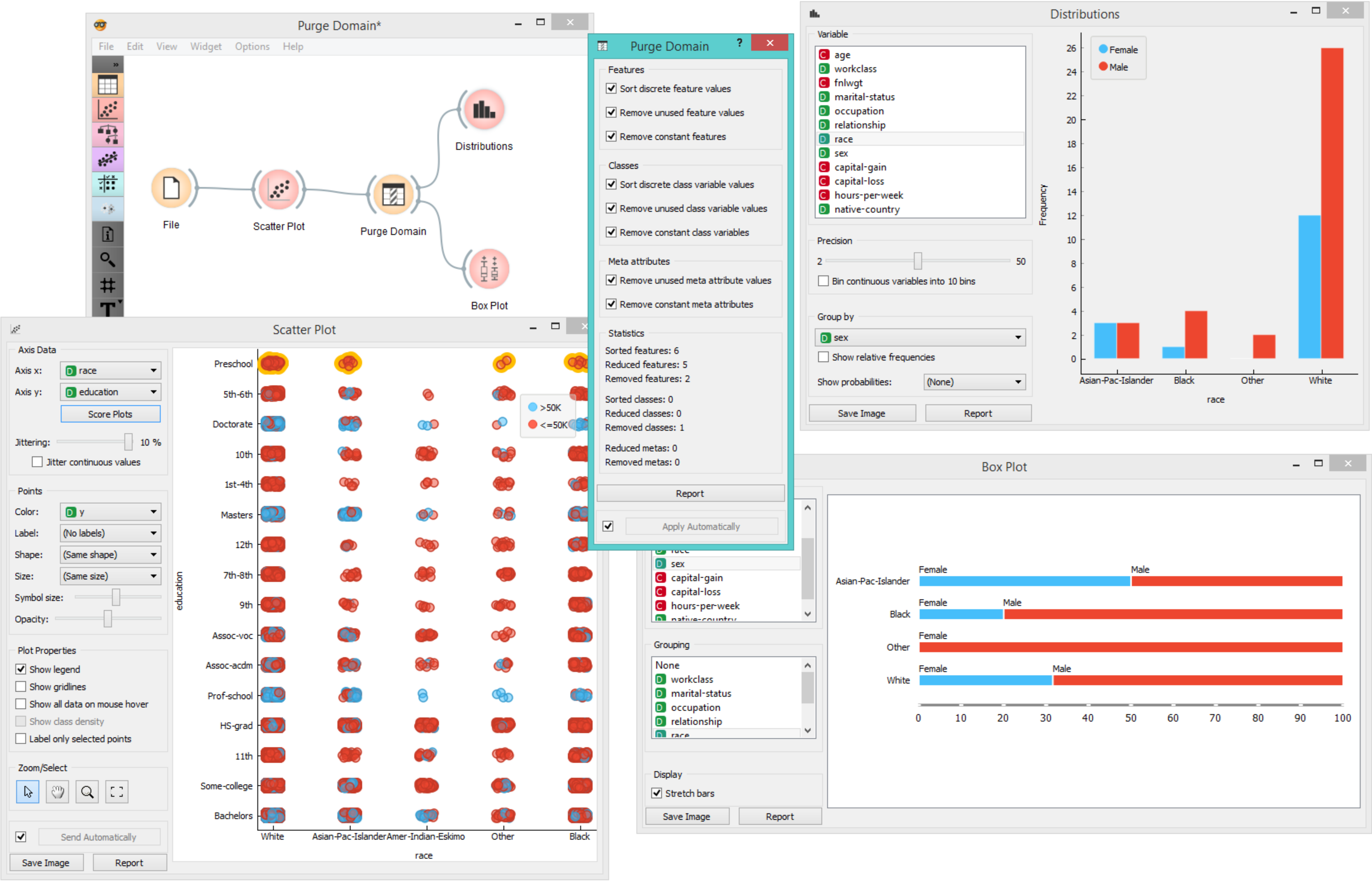
Such purification is done by the widget **Purge Domain**. Ordinary attributes and class attributes are treated separately. For each, we can decide if we want the values sorted or not. Next, we may allow the widget to remove attributes with less than two values or remove the class attribute if there are less than two classes. Finally, we can instruct the widget to check which values of attributes actually appear in the data and remove the unused values. The widget cannot remove values if it is not allowed to remove the attributes, since having attributes without values makes no sense.

The new, reduced attributes get the prefix “R”, which distinguishes them from the original ones. The values of new attributes can be computed from the old ones, but not the other way around. This means that if you construct a classifier from the new attributes, you can use it to classify the examples described by the original attributes. But not the opposite: constructing a classifier from the old attributes and using it on examples described by the reduced ones won’t work. Fortunately, the latter is seldom the case. In a typical setup, one would explore the data, visualize it, filter it, purify it… and then test the final model on the original data.

Example

The **Purge Domain** widget would typically appear after data filtering, for instance when selecting a subset of visualized examples.

In the above schema, we play with the *adult.tab* dataset: we visualize it and select a portion of the data, which contains only four out of the five original classes. To get rid of the empty class, we put the data through **Purge Domain** before going on to the Box Plot widget. The latter shows only the four classes which are in the **Purge Data** output. To see the effect of data purification, uncheck *Remove unused class variable values* and observe the effect this has on Box Plot.



Purge Domain

Features

☑ Sort discrete feature values

☑ Remove unused feature values

☑ Remove constant features

Classes

☑ Sort discrete class variable values

☑ Remove unused class variable values

☑ Remove constant class variables

Meta attributes

☑ Remove unused meta attribute values

☑ Remove constant meta attributes

Statistics

Sorted features: 6
Reduced features: 5
Removed features: 2

Sorted classes: 0
Reduced classes: 0
Removed classes: 1

Reduced metas: 0
Removed metas: 0

Report

☑ Apply Automatically

Scatter Plot

Axis Data

Axis x: race

Axis y: education

Score Plots

Jittering: 10 %

☐ Jitter continuous values

Points

Color: y

Label: (No labels)

Shape: (Same shape)

Size: (Same size)

Symbol size:

Opacity:

Plot Properties

☑ Show legend

☐ Show gridlines

☐ Show all data on mouse hover

☐ Show class density

☐ Label only selected points

Zoom/Select

☑ Send Automatically

Save Image Report

Distributions

Variable

- age
- workclass
- fnlwgt
- marital-status
- occupation
- relationship
- race
- sex
- capital-gain
- capital-loss
- hours-per-week
- native-country

Precision 2 50

☐ Bin continuous variables into 10 bins

Group by: sex

☐ Show relative frequencies

Show probabilities: (None)

Save Image Report

Box Plot

- race
- sex
- capital-gain
- capital-loss
- hours-per-week
- native-country

Grouping

- None
- workclass
- marital-status
- occupation
- relationship
- race

Display

☑ Stretch bars

Save Image Report