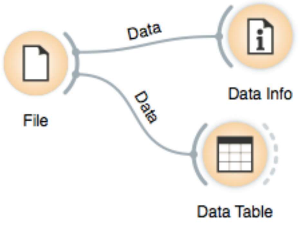
Lesson 2: Basic Data Exploration

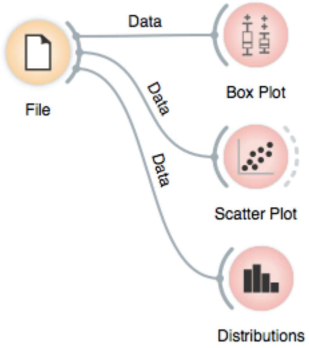
Let us consider another problem, this time from clinical medicine. We will dig for something interesting in the data and explore it a bit with various widgets. You will get to know Orange better and also learn about several interesting visualizations.

We will start with an empty canvas; to clean it from our previous lesson, use either File→New or select all the widgets and remove them (use the backspace/delete key, or Cmd-backspace if you are on Mac).

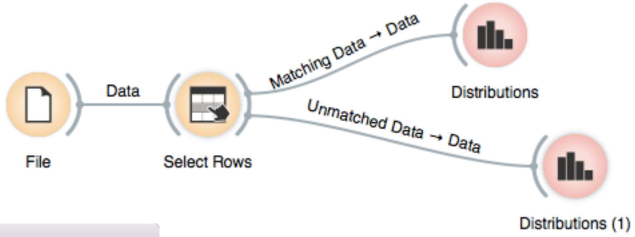
Now again, add the File widget and open another documentation data set: heart_disease. How does the data look?



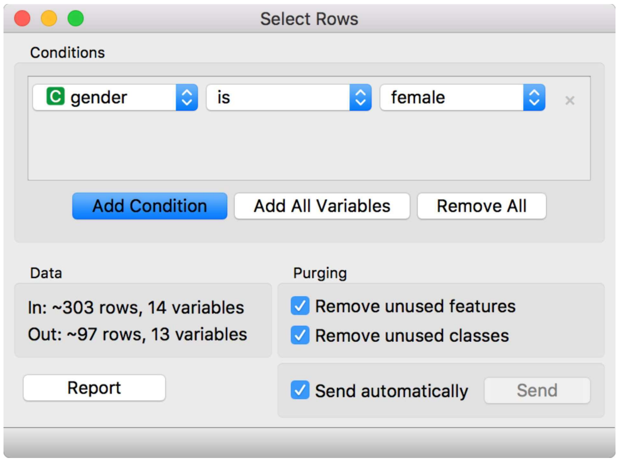
Let us check whether standard visualizations tell us anything interesting. (Hint: look for gender differences. These are always interesting and occasionally even real.)



Data can also be split by the value of features — in this case, gender — and analyze it separately.



The two Distributions widgets get different data: the upper gets the selected rows, and the lower gets the others. Double-click the connection between the widgets to access setup dialog, as you've learned in the previous lesson.



Select Rows

Conditions

C gender | is | female ×

Add Condition | Add All Variables | Remove All

Data

In: ~303 rows, 14 variables
Out: ~97 rows, 13 variables

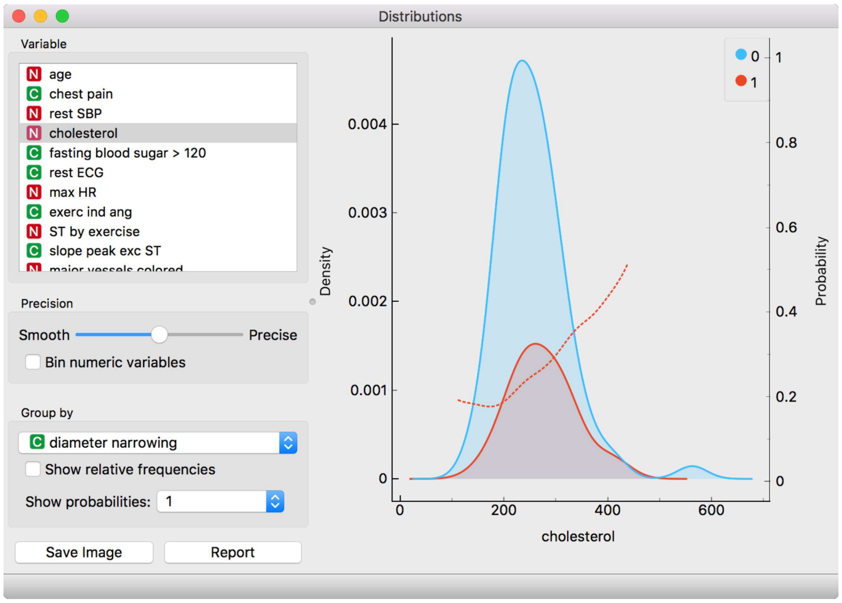
Purging

☑ Remove unused features
☑ Remove unused classes

Report

☑ Send automatically | Send

In the Select Rows widget, we choose the female patients. You can also add other conditions. Selection of data instances works well with visualization of data distribution. Try having at least two widgets open at the same time and explore the data.



Distributions

Variable

- N age
- C chest pain
- N rest SBP
- N cholesterol
- C fasting blood sugar > 120
- C rest ECG
- N max HR
- C exerc ind ang
- N ST by exercise
- C slope peak exc ST
- N major vessels colored

Precision

Smooth ——○—— Precise

□ Bin numeric variables

Group by

C diameter narrowing

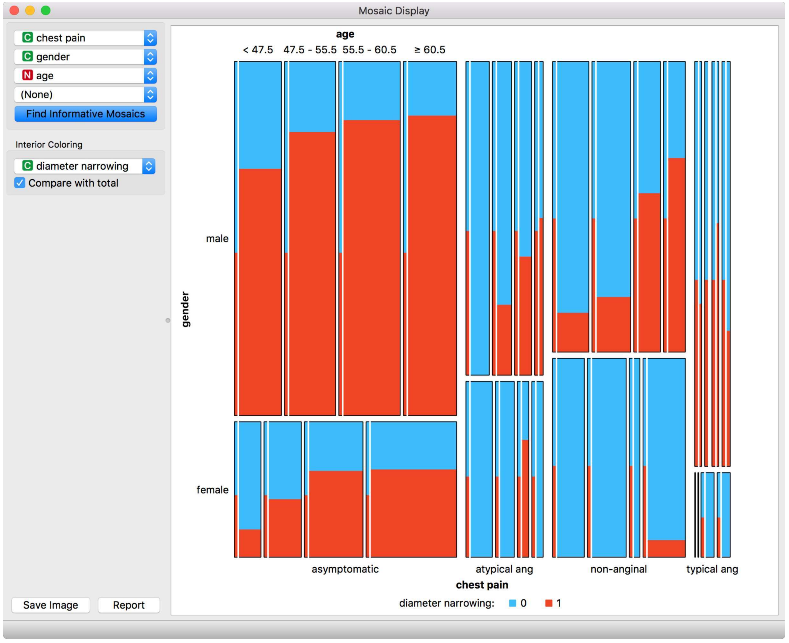
□ Show relative frequencies

Show probabilities: 1

Save Image | Report

There are two less known — but great — visualizations for observing interactions between features.

You can play with the widget by trying different combinations of 1-4 features.



Mosaic Display

C chest pain
C gender
N age
(None)

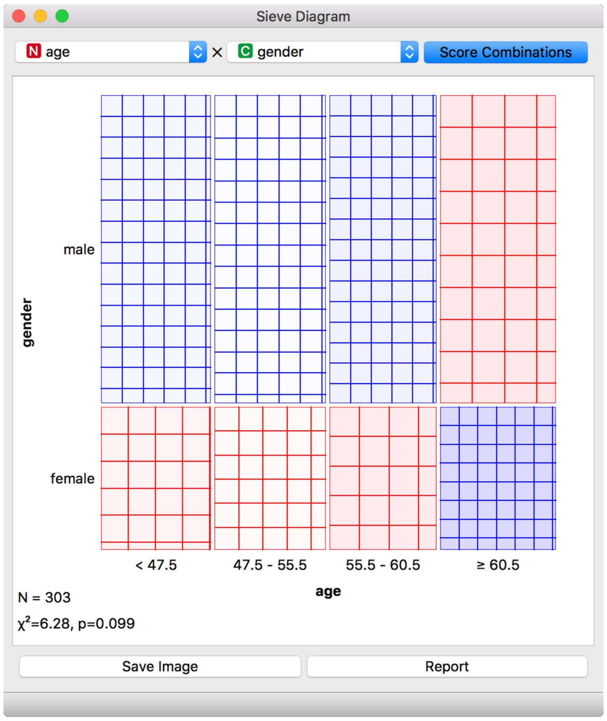
Interior Coloring
C diameter narrowing
☑ Compare with total

The mosaic display shows a rectangle split into columns with widths reflecting the prevalence of different types of chest pain. Each column is then further split vertically according to gender distributions within the column. The resulting rectangles are divided again horizontally according to age group sizes. Within the resulting bars, the red and blue areas represent the outcome distribution for each group and the tiny strip to the left of each shows the overall distribution.

What can we conclude from this diagram?

Another visualization, Sieve diagram, also splits a rectangle horizontally and vertically, but with independent cuts, so the areas correspond to the expected number of data instances assuming the observed variables are independent. For example, 1/4 of patients are older than 60, and 1/3 of patients are female, so the area of the bottom right rectangle is 1/12 of the total area. With roughly 300 patients, we would expect 1/12 × 300 = 25 older women in our data. There are 34. Sieve diagram shows the difference between the expected and the observed frequencies by the grid density and the color of the field.

See the Score Combinations button? Guess what it does? And how it scores the combinations? (Hint: there are some Greek letters at the bottom of the widget.)



N = 303

χ^2=6.28, p=0.099