Instrument Data Analysis

Working notes for the hands-on course at Procter & Gamble

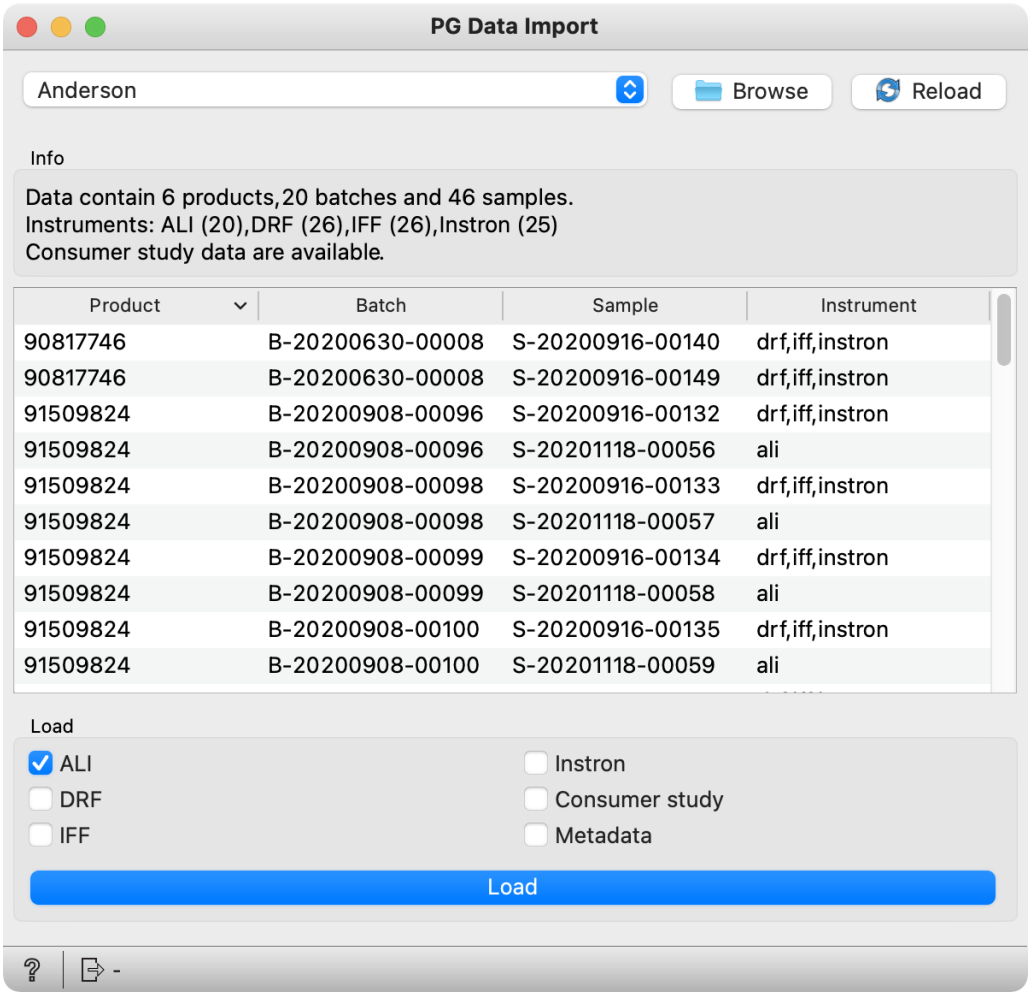
In this tutorial, we will start to work with the Orange widgets designed for Procter & Gamble. We will begin with data reading and walk through data visualizations, filtering, and aggregations. We will conclude the first day with several data analyses.

Handouts prepared by:
Bruna Zupan, Lan Žagar, and Primož Godec

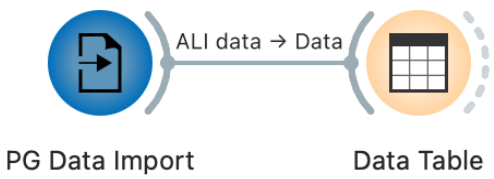
Lesson 1: Data loading

Let us now start exploring the data from P&G shampoo experiments.

Use the *PG Data Import* widget to open a structured data directory with several types of experimental data (ALI, DRF, IFF, Instron) along with Consumer study data and metadata.

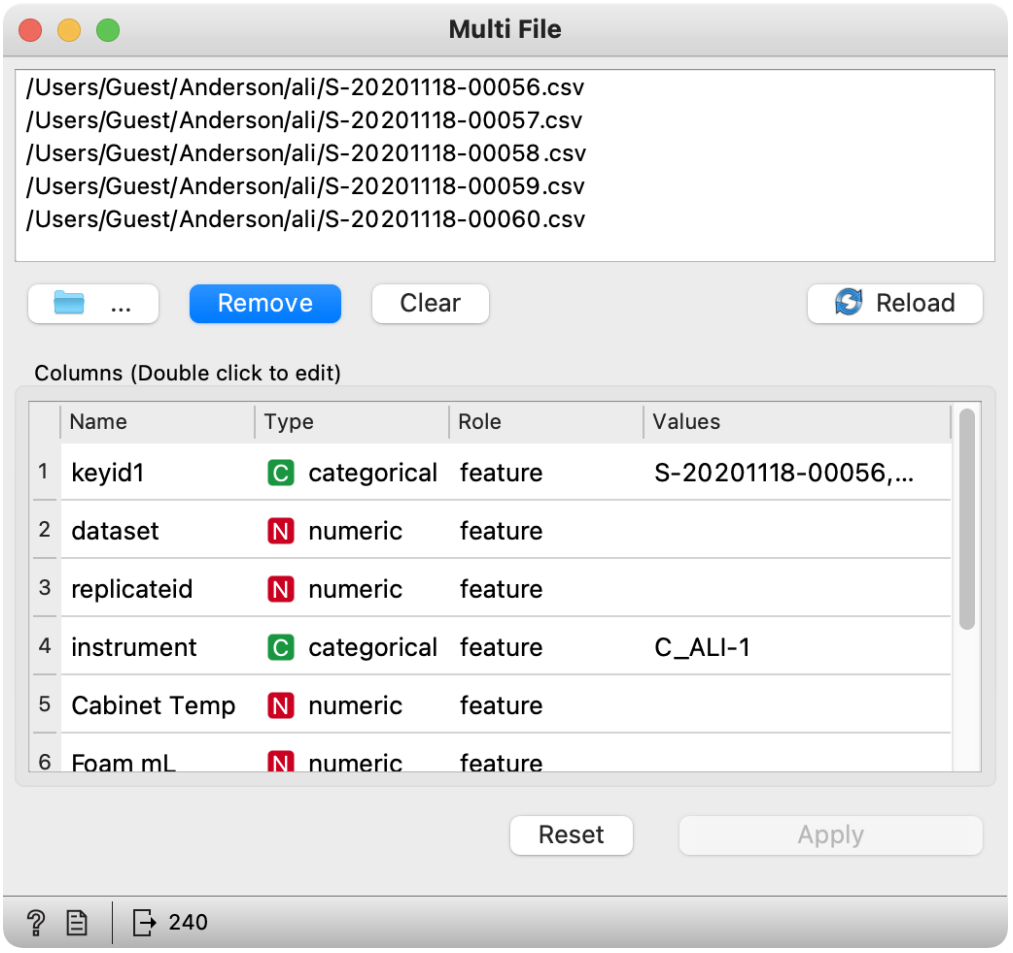


The widget will show the product ID, batch ID, sample ID, and the instrument data available for that sample.



Connect the *PG Data Import* widget to the *Data Table* widget to see the contents of the loaded data set.

Alternatively, you can load the data using the *Multi File* widget. Go to the ALI subfolder and select a subset of files to load. You can use this widget to load data files from other types of experiments that the *PG Data Import* widget would not recognize.

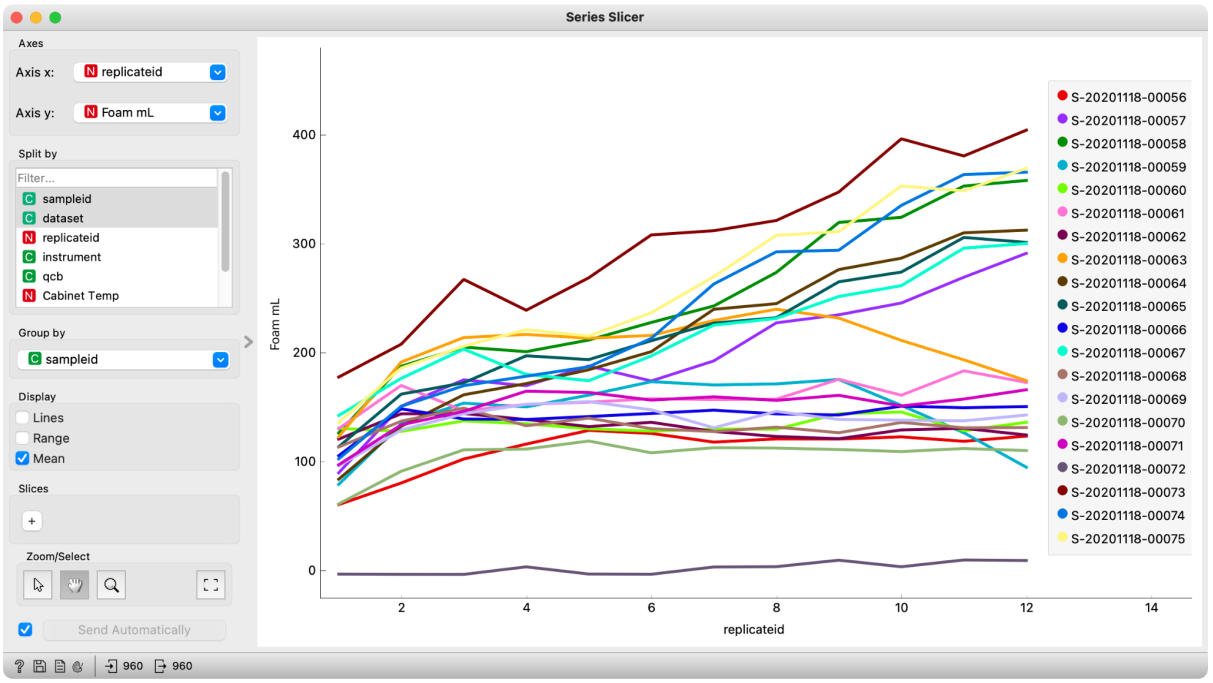


Lesson 2: Visualize time-series data

All four experimental data sets contain time-series data, and examining it in tabular form (with *Data Table*) can be challenging. We will use the *Series Slicer* widget to plot the series.

Let us start by trying to visualize the ALI data. Connect PG Data Import to the Series Slicer widget and set the following settings.

- Axis x - Select the time variable (e.g. `replicateid`)
- Axis y - Select the variable for the series of choice (e.g. `Foam mL`)
- Split by - Select all variables that uniquely define one time-series (e.g., `sampleid` and `dataset`)
- Group by - Select a variable if you wish to group multiple series lines (e.g., `sampleid`)

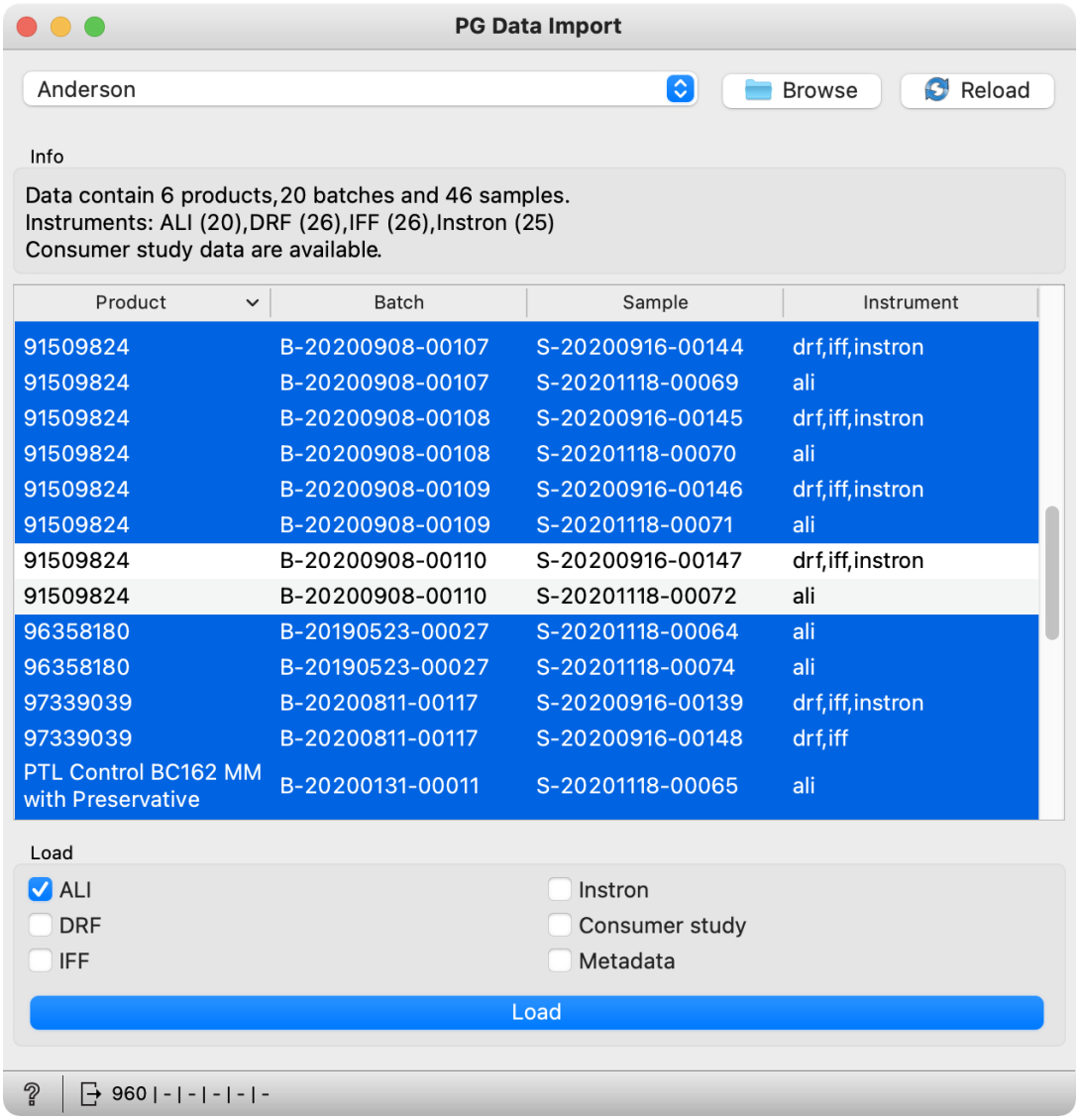


Before you correctly set the Split by variables and x and y axes, you might get an error message "Non-unique series values" informing you that the data contains multiple values of y (`Foam mL`) for the same x (`replicateid`).

Lesson 3: Filtering

Looking at the ALI data, we notice that one of the samples is not like the others. It has very low foam levels, and cross-referencing the actual values in the *Data Table* reveals some are negative. Let's assume something went wrong during that experiment, and we want to remove it to avoid corrupting the statistics and models that we will fit.

We can exclude the experiment by selecting a subset of samples in the *PG Data Import* widget or do it at any later point in the workflow by selecting a subset of rows based on some criteria using *Select Rows*. In our case, we will use `sampleid` is not S-20201118-00072.



Lesson 4: Aggregate

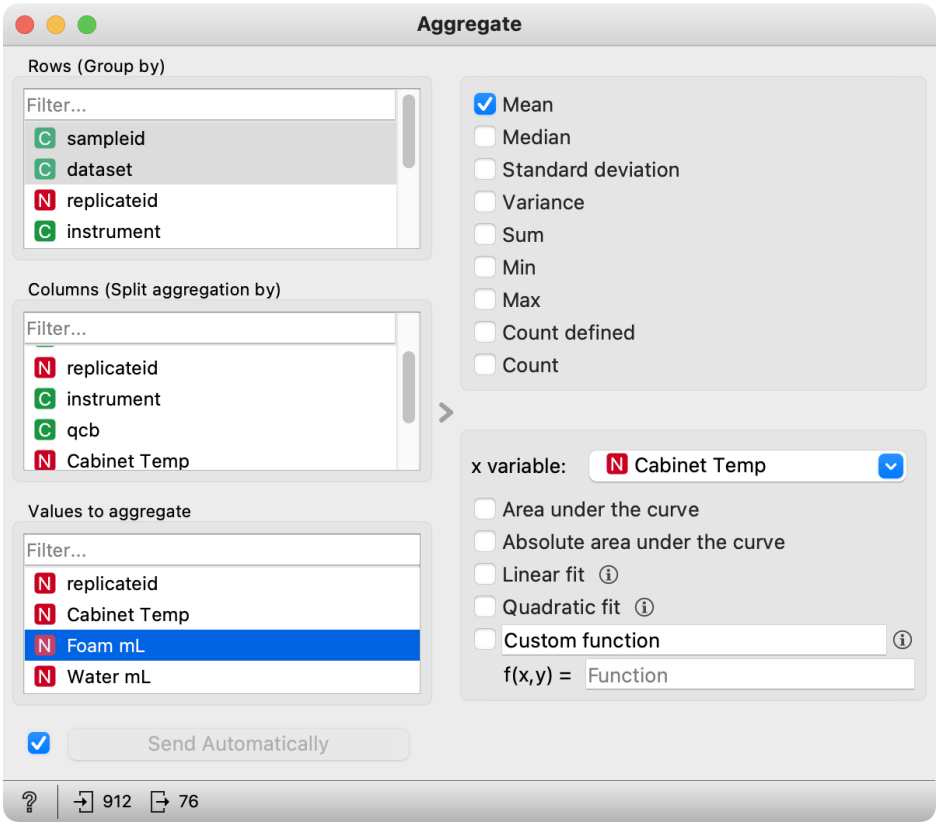
Now that we got familiar with the data, it is time to prepare it for other data mining techniques which we have learned during the previous week. For that, we would like to profile the time-series data to obtain the standard tabular format with the objects of interest as rows and describe them with several features (columns).

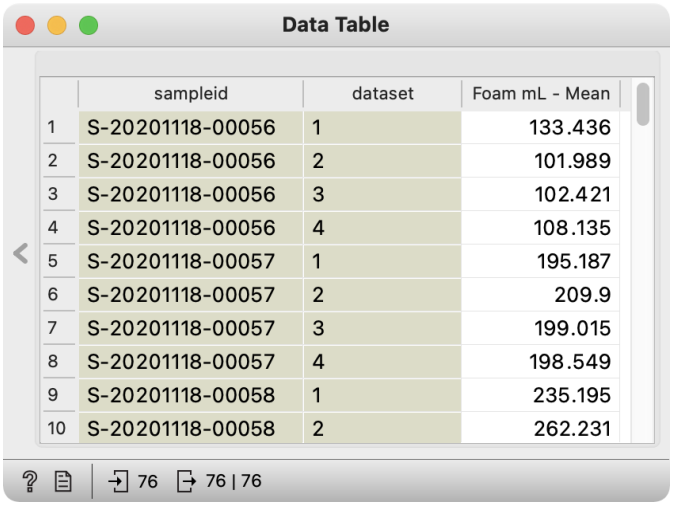
Use the *Aggregate* widget to first compute the simplest aggregation and represent each time-series with a single feature — the mean value.

The important settings in this widget include:

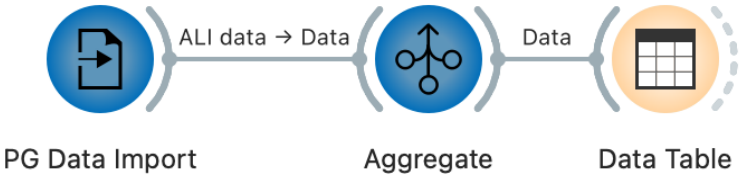
- Rows (Group by) — Select variables which will uniquely identify rows of the aggregated output table (e.g. `sampleid`; or `sampleid` and `dataset`)
- Columns (Split aggregations by) — Select a variable (or several) for which you wish to compute independent aggregate values (e.g. none selected; or `dataset`).
- Values to aggregate — Select all variables that you wish to aggregate (e.g., `Foam mL`; or `Foam mL` and `Water mL`)

Afterward, check the aggregated table by connecting the output from *Aggregate* to *Data Table*.



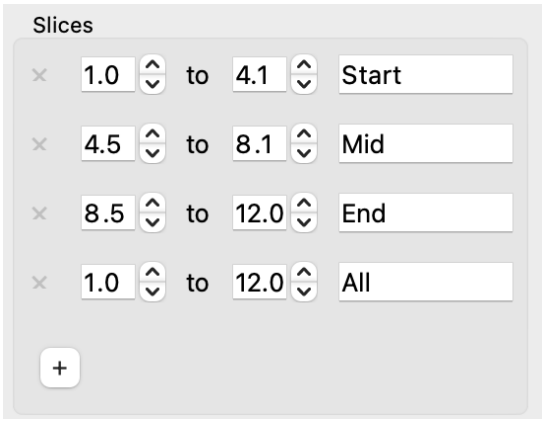


Now play with the settings and check how the output changes when selecting different options for Rows/Columns/Values and multiple aggregations.



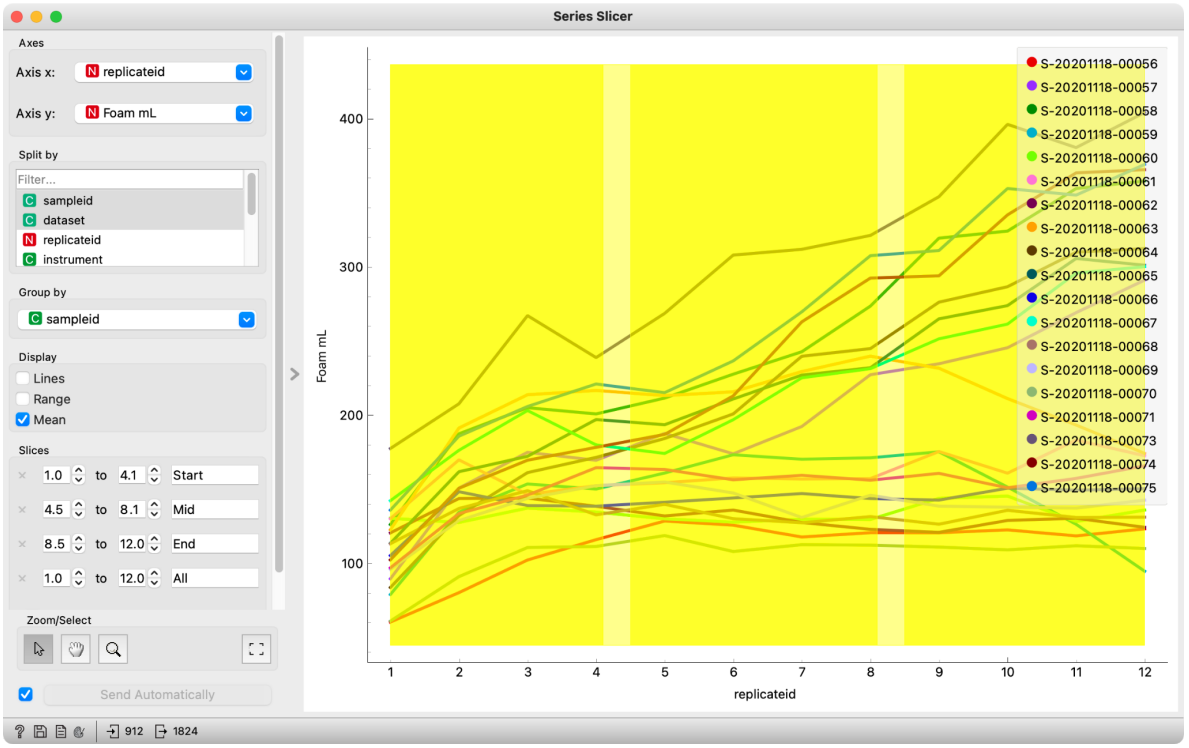
Lesson 5: Time Slices

A combination of *Series Slicer* and *Aggregate* widgets enables powerful analysis of different sections (slices) of the time-series.

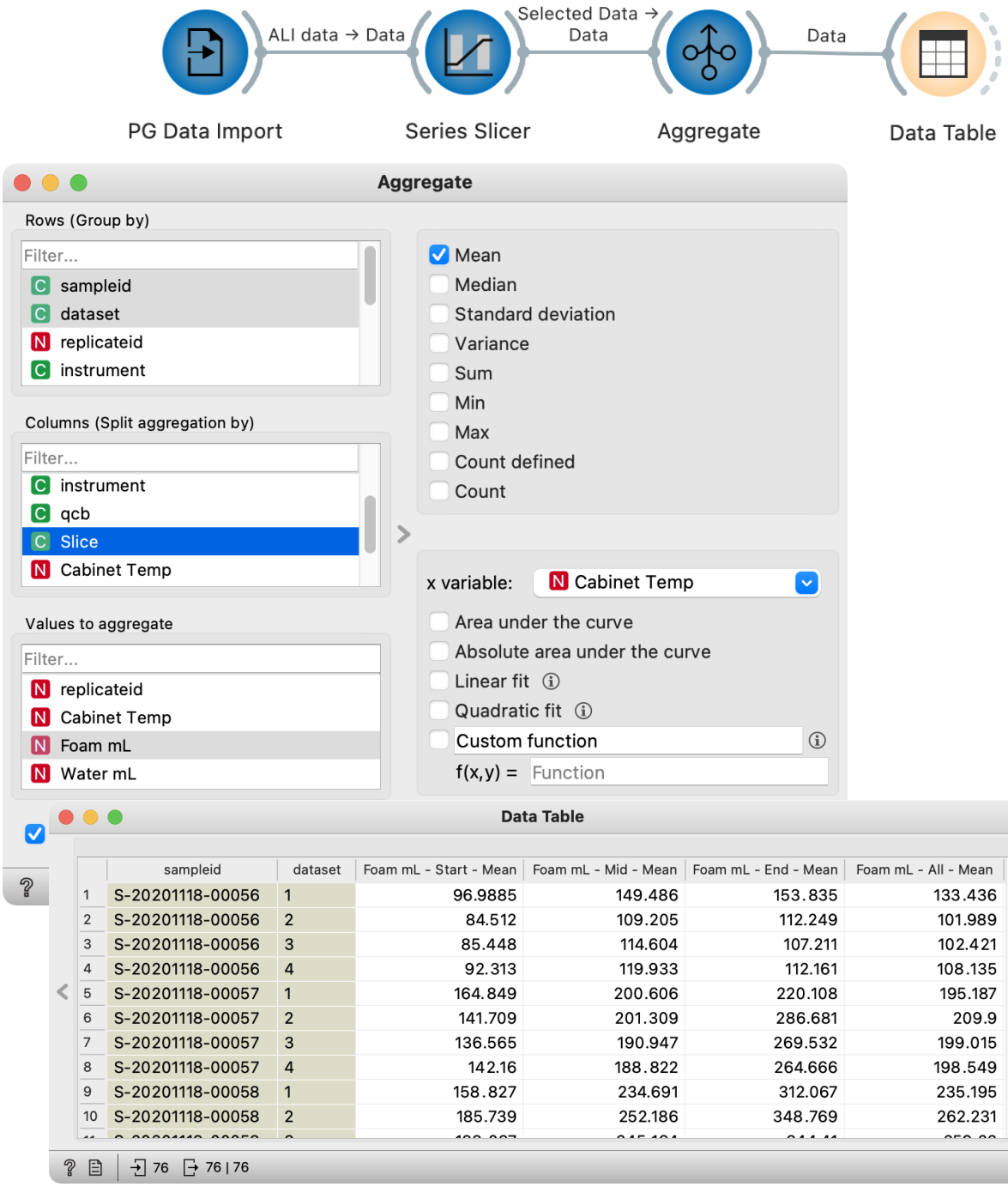


To see a simple example of this, go back to the *Series Slicer* visualization of the ALI data and select multiple slices using click-and-drag actions on the plot and editing the Slices section in the control area.

Check the output of *Series Slicer* in the *Data Table* and notice that the data looks almost the same, but now has an additional column with the Slice information. This can be used in subsequent analysis.



Connect Series Slicer to Aggregate and repeat the aggregation with mean values, but now split the aggregations based on the Slice variable.

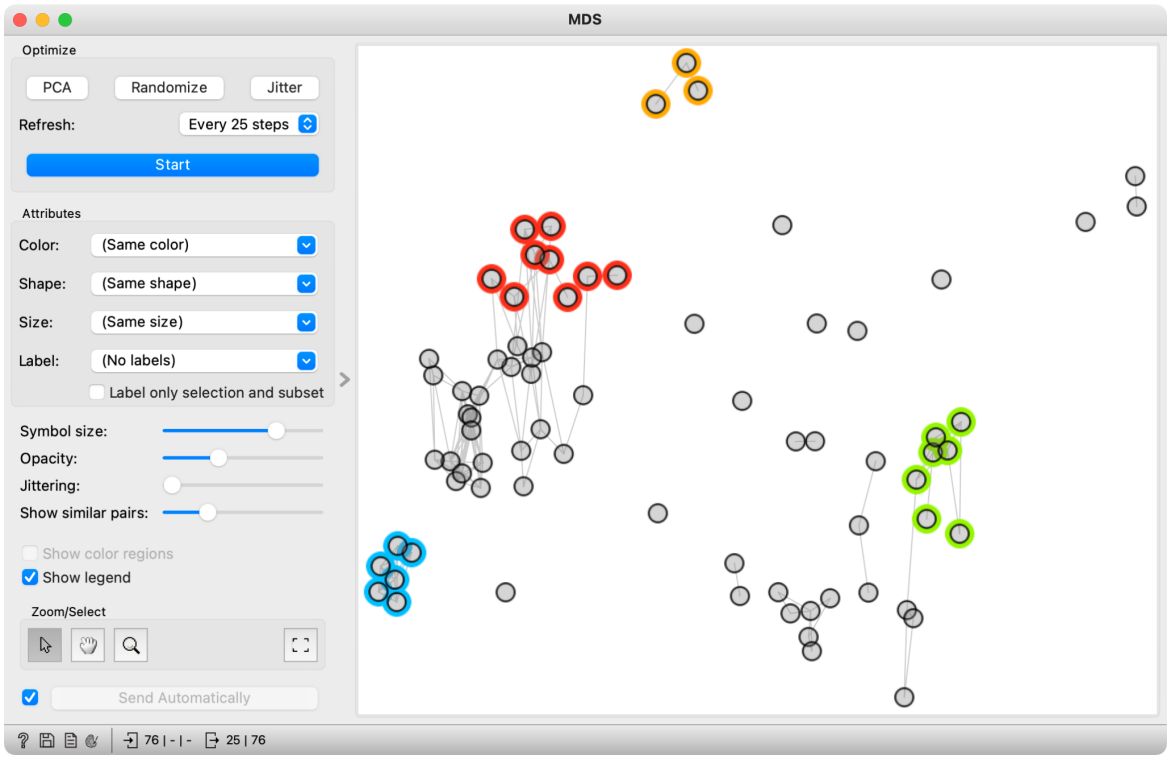


	sampleid	dataset	Foam mL - Start - Mean	Foam mL - Mid - Mean	Foam mL - End - Mean	Foam mL - All - Mean
1	S-20201118-00056	1	96.9885	149.486	153.835	133.436
2	S-20201118-00056	2	84.512	109.205	112.249	101.989
3	S-20201118-00056	3	85.448	114.604	107.211	102.421
4	S-20201118-00056	4	92.313	119.933	112.161	108.135
5	S-20201118-00057	1	164.849	200.606	220.108	195.187
6	S-20201118-00057	2	141.709	201.309	286.681	209.9
7	S-20201118-00057	3	136.565	190.947	269.532	199.015
8	S-20201118-00057	4	142.16	188.822	264.666	198.549
9	S-20201118-00058	1	158.827	234.691	312.067	235.195
10	S-20201118-00058	2	185.739	252.186	348.769	262.231

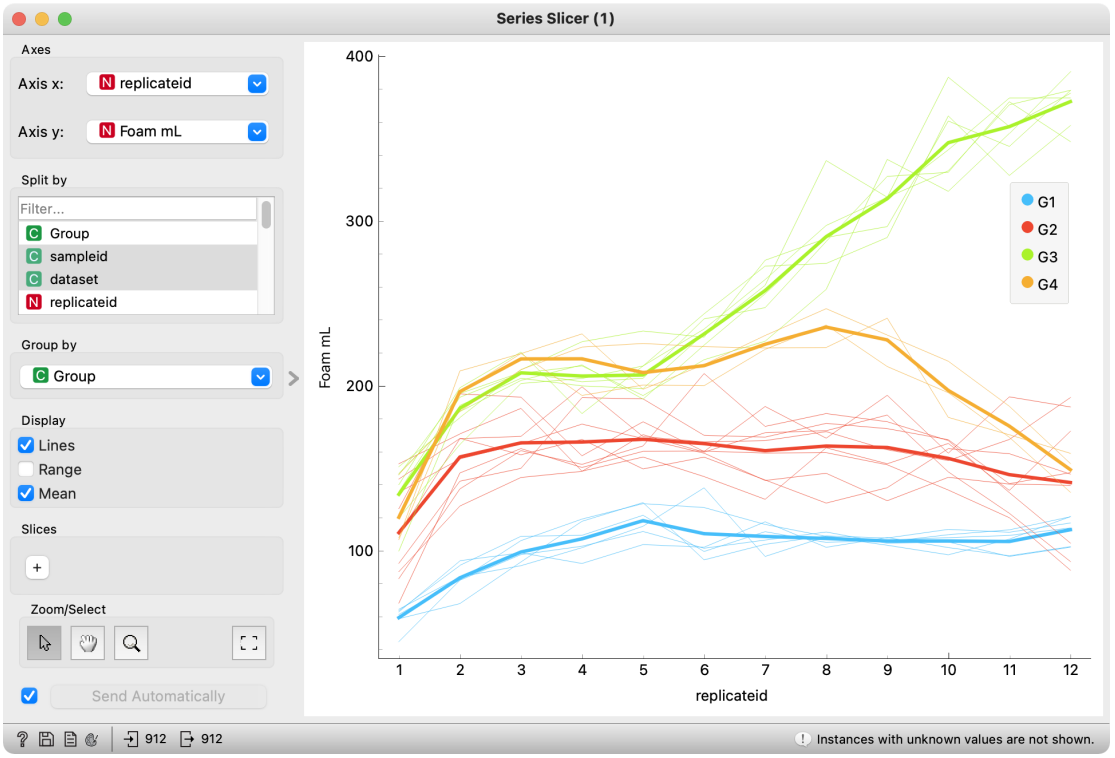
Lesson 6: Analysis of profiled data

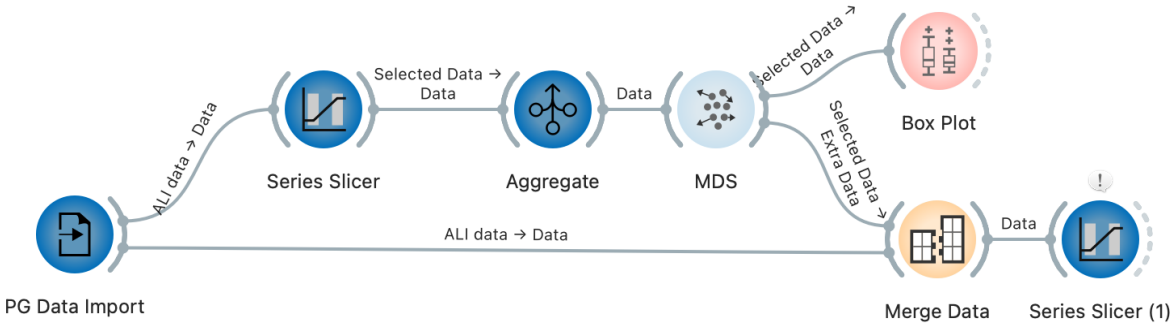
Let us now use some standard Orange widgets to explore the relationships between samples and try to find some patterns.

Using *MDS* we can embed all samples into 2D space based on all the features characterizing them (e.g., slice means or even all 12 measurements).



Try selecting different groups and check the differences in *Box Plot*. You can also use the original raw time-series data, merge the group/cluster labels based on all sampleid & dataset pairs and visualize selected groups in the *Series Slicer*.





Try some other methods like *Hierarchical Clustering* instead of *MDS*.

Lesson 7: Other data types

We have now familiarized ourselves with all four new widgets from the P&G add-on. However, the power of this data mining toolset comes from the multitude of possible combinations of different widgets and settings as well as using them in different ways on different data and combining them.

Try to perform similar analyses on IFF, DRF, or Instron data.