Sampling Data

Why Sample?
When working with data, you may find that you need to create a subset of that information to better analyze or understand it. Sampling data is an easy way to remove empty rows from the top or bottom of a dataset. Sampling is also helpful for paring down data based on a pattern, like selecting the first record of a group, allowing you to carry relevant data through your analysis. Or, sampling can generate dynamic, random samples of data which can be used to create robust solutions.

A dataset contains information on transactions from the month of March. The goal is to remove the header information and find the three highest daily transaction values. These tasks can be accomplished using the sample tool. Drag a sample tool onto the canvas.

Methods of Sampling
The Sample tool's configuration window provides six different methods for sampling data, each of which can be used with "Grouping" functionality. Of note, there is one output anchor, which means that regardless of which method of sampling is selected, rows removed during the sampling process are not output.

Select "Skip 1st N rows". Since the header information appears in the first 4 rows of the dataset, type 4 in the "N" box to specify that the first 4 rows should be skipped.

All 6 available options reference the value N. N can be specified by entering an integer in the N box.

After running the workflow, the first 4 rows have been removed from the dataset. Next, the data should be sorted so dates are in ascending order and transaction values are in descending order. Then, another sample tool is dragged onto the canvas.

This time select the "First N Rows" option.

The N value should be set to 3 to return the 3 highest transaction values.

Use the Group By functionality to get the first 3 rows for each group. Select "Search Date".

Grouping Samples
By selecting a Group By option, Designer will return 3 results for each unique value in the column specified. Without grouping, Designer would only return the first 3 rows in the dataset.

After running the workflow, the 3 highest transactions are displayed for each day. Note that some days had fewer than 3 transactions. In those instances, all values are returned.

Random Sampling

Finally, the sample tool can be used to take a randomized sample of the records. To better illustrate what records are processed, a Record ID tool is added to the workflow before another sample tool is dragged onto the canvas. The only configuration option with an element of random-ness is the “1 in N chance to include each row”. In this configuration, each row is evaluated individually and is assigned a 1 in N chance to be included. The row is included or removed, and the next row is evaluated. A 1 in 2 chance is much more likely to include any particular record than a 1 in 8 chance, but neither guarantees any individual record will be included. When we run the workflow, we note that [42] rows appear in the output. Without changing the configuration, running the workflow again shows [24] rows in the output. Inspecting the record IDs reveals that records included have also changed.