Formatting Data - Transcript

Why Would You Use the Select Tool?

Applying changes to columns of data early in the workflow offers distinct advantages in the course of workflow development. First, defining a column's datatype and size ensures that values will be available for downstream calculations or processes that are type dependent. Second, renaming columns whose names are confusing or re-ordering column position reduces the likelihood that values will be used incorrectly or inappropriately in the workflow. Finally, removing columns that are unnecessary for your analysis can greatly optimize a workflow by reducing the number of columns that are processed when a workflow is run.

Workflow Background

This workflow's inputs include information on trees surveyed in New York City. However, before these inputs can be used for data blending and analysis, changes to their current data types, column names and structures must be made. To do so, use a Select tool.

Drag a Select tool onto the Canvas and connect it to the Input Data tool.

Configuring the Select Tool

The Select tool allows you to make changes to a column's metadata such as its name, data type, size and description. Already, you might observe a few features of this input that require modifications. Because this input is sourced from a .CSV file, every column has been assigned a datatype of V_String. Use the Select tool to assign a more appropriate data type to the columns [Planting Date], [Height], and [Status].

Use the Drop Down in the column “Type" to assign a Date data type to the column [Planting Date].

The height of a tree is a numeric value. Use the Drop Down to assign the column [Height (m)] to Int16.

Because the column [Status] contains True and False values to indicate if a tree is alive or dead, categorize it as a Boolean data type (Bool).

If a data type's size is fixed with a defined column size, the value in the column Size will become grayed out. To manually increase or decrease the size of a column, enter a value into the column "Size". Change the size of the column [Tree ID] to seven (7).

Different Column Names

You may decide to rename columns for a couple of reasons. First, columns that are not clearly or intuitively named can cause confusion for workflow developers and data consumers. Second, columns that are not identically named across datasets can cause issues when blending data on the assumption that columns in different sources are named consistently. This input dataset contains two columns whose values correspond to those found in other inputs but are not named the same.
Rename the columns [spc_latin] and [spc_common] to match the columns containing similar data in other inputs. In the column "Rename", enter new column names: [Latin Name] and [Common Name].

Remove Columns

Columns that you do not plan to use in further analysis should be removed from the data stream as early as possible to optimize the processing speed of a workflow. Remove columns from the data stream by simply deselecting a column in the Select tool's configuration window.

The columns [Latitude] and [Longitude] are not needed for the analysis. Deselect the checkboxes next to these columns to remove them from the data stream.

Column Order

Columns can also be rearranged to for visual purposes, ease of use or strategic grouping. Columns at the top of the list in the Select tool's configuration window will appear at the leftmost side, and columns at the bottom of the list will appear on the rightmost.

Unknown Column

Move the column [Planting Date] to the top of the dataset. Click the row containing the column name [Planting Date]. Then, click the up arrow.

You may notice that the last column name, [*Unknown], is not actually a column in the input dataset. However, leaving this column selected is an important consideration in the Select tool's configuration. This unknown column represents the foresight of the Select tool to handle columns that are currently unknown to the tool during configuration time but could appear in the future if an input's structure changed to include a string column like [Leaf Color] or a numeric column to indicate the number of birds present in the tree. To automatically pass this new column through the Select tool, keep [*Unknown] selected. If not, deselect this column.

Review Results

After running the workflow, the changes made in the Select tool are applied to the dataset: columns have been re-ordered, re-named and reclassified appropriately. It's important to remember that changes made in the Select tool apply to entire columns of data, not individual values in rows.

Embedded Select Windows

Other tools, such as the Join tool, in Designer contain what is called an "embedded Select" in their configurations, allowing you to apply many of the same changes to column names, order and datatype in other steps in a workflow. Rather than add another tool to the workflow, take advantage of these opportunities to optimize and organize your data as you build and develop your process.