Filtering Data - Transcript

Filtering Data

Filtering rows from your data can speed up your analysis by quickly isolating rows of data based on conditions that you specify. Rows of data are then evaluated and separated into two output data streams: one which contains the rows that meet the specified conditions and are "True", or another that contains the rows that do not and are "False".

Drag a Filter tool from the Favorites tool palette and drop it onto the Canvas.

Configuring the Filter Tool

The Filter tool's configuration supports two types of filters: basic and custom. Basic filters support simple queries to evaluate one condition in a single column of data. Custom filters, on the other hand, can test more complex queries with more than one condition across multiple columns of data.

Creating Basic filters

The basic filter is constructed of three parts: a column in which to test a condition, an operator, and the condition to test.

Create a filter to determine the rows where trees are in good health.

Use the Drop Down to select the column that contains information on the health of a tree: "Health".

Operators

The types of operators that are available in the Drop Down depend on the data type assigned to the column that is being queried. Numeric columns will display a list of operators to test conditions such as "greater than" or "less than" while DateTime datatypes can test dates before or after a fixed date, like December 1st 2018 or a dynamic date, like "Yesterday". Because the column "Health" is categorized as a string data type, the types of operators that can be applied to this column range from testing for a specific value, to alphabet position, to substrings, or even the presence of null and empty cells. Select "Equals" from the dropdown menu of operators.

Condition

Test the condition that a tree's health is "Good". In the text box, manually type "Good".

Run the Workflow

After running the workflow, data is split into two streams: one containing the rows that are True with respect to the Filter tool's query and another that contains the rows that are False. Click the "T" and "F" output anchors on the Filter tool to view the results.

Over 160,000 rows of data passed through the True anchor of the Filter tool, meaning that the value in the column "Health" is equal to Good.

The remaining rows of the original input were evalulated as "False", meaning that the value in the column "Health" was not equal to Good.

Customizing Filters

Rows can be filtered based on multiple conditions, such as trees that are in good health and are at least 10 (ten) meters tall To create a more complex query, use the Custom Filter's expression editor in the Filter tool's configuration.

Select the Radio Button to enable the Custom Filter's expression editor.

Any basic filters created in the Filter tool are also replicated in the Custom filter, so the first condition, that a tree is in good health, is already present in the Expression Editor. In the text box below, enter the next condition that a tree's Height is greater than or equal to ten meters: [Height] >= 10.

In the Filter tool's expression editor, click the "Variables" button and select "Height" from the menu.

Complete the expression by typing equals 10 (= 10).

Writing a Multi-Conditional Filter

Now, two conditions are specified as conditions to test. However, the Filter tool's configuration requires more than just a list of conditions; the Filter tool must know how these conditions should be tested in relation to each other. Adding a Boolean Operator such as "And" or an "OR" between the statements defines the query to evaluate. Using an "AND" requires that both conditions must be true in the same row of data: a tree must be healthy and tall. An "Or" requires that only one of the conditions be true.

Identify the rows in which trees are both healthy and tall. In the text box below, enter the correct operator to complete the expression.

If an OR is used, results can include trees that are healthy, but short, as well as trees that are tall but not in good health. This result is not correct for identifying the trees that are both healthy and tall.

With the two conditions in place, over 80,000 trees in New York City were identifed as being in good health and at least ten (10) meters tall. Trees that do not meet this criteria may be candidates for a separate type of analysis, or left out of downstream workflow development.