Sorting Data - Transcript

Sorting Data

Sorting values in a particular order may provide more meaningful insights for your analysis, especially when it comes to understanding how values compare to others in your data. Typically, you will sort three different types of values: strings, numbers, and dates. These values can be sorted in two ways: in ascending order or descending order.

For columns classified as string data types, values are sorted alphabetically. When sorted in ascending order, values starting with letters at the beginning of the alphabet will appear at the top of the dataset, and words beginning with letters at the end of the alphabet will be at the bottom of the dataset. Sorting the values in descending order reverses these results. For numeric data, sorting in ascending order will place the smallest values at the top of the column and the largest value at the bottom; descending order will be the reverse, from largest to smallest.

When Dates are sorted in ascending order, the oldest date will appear at the top of the column with the newest date at the bottom.

Configuring the Sort Tool

Data on trees planted in New York City in 2016, 2017 and 2018 has been prepared and blended over the course of the analysis. Now, you want to sort the data to better visualize and understand the values this dataset contains by Borough Name, and Planting Date, as well as a tree's Common Name.

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

The Sort Tool's configuration requires two pieces of information: the column whose values you want to sort, and the order in which to sort them.

Use the dropdown menus in the Sort tool's configuration window to sort the column "Borough Name" in Ascending Order.

After running the workflow, the values in the column "Borough Name" are sorted from Bronx to Staten Island. Even a basic sort provides a more organized view of the trees that have been planted in each borough.

Sorting on Multiple Columns

Drill down into more specific trends in your data by sorting multiple columns at a time, creating levels of ordering.

Add an additional level of sorting by ordering the values in the column "Planting Date" in descending order.

Now, for each borough, the most recently planted trees are at the top of the dataset, while the oldest trees are at the bottom of the dataset. When duplicate values are found in the columns used for sorting, the output will preserve the order in which the data appeared in the input dataset.

Sorting Strings

The Sort tool offers the ability to sort columns classified as string data types in dictionary order for a wide range of global alphabets. Choosing this option will sort columns of text using rules that account for the language characters in the source data, and also produces different results from the default sorting behavior of string values.

Default Behavior

Explore the default sorting behavior of string values by sorting the column "Common Name" in Ascending Order.

By default, string values will be sorted based on the ASCII value of the characters a value contains. This means that strings containing punctuation, such as a single quote, will appear at the top of the dataset, followed by values starting with uppercase letters, and then lowercase letters.

Enabling Dictionary Order

Now, configure the Sort tool to "Use Dictionary Order" in English from the United States.

After running the workflow with Dictionary Order enabled, the output has changed. Now, values appear how they would be ordered in a dictionary, no matter their case.