Summarizing Data

Summarizing Your Data

When working with large datasets, it's important to be able to quickly and succinctly get an overview of your data. Whether you need to group information, sum values in a column, or some other summary process, the Summarize tool makes it easy to find what you need.

Because summary processes are applied to columns of data, ensure your data is vertically oriented. This may require you to transform (or transpose) data into an appropriate format.

A dataset contains information on trees in the 5 boroughs of New York City, including the height in meters, Latin name, common name, and a borough code. Use the Summarize tool to find out what different types of trees are in New York and how many there are of each type. Drag a summarize tool onto the canvas and connect it to the Input Data Tool.

To identify the types of trees in NY, start by selecting the "Common Name" column and use the action dropdown to "Group By".

Note that the action appears in our actions window. Now use the "Count" action on the "Common Name" column to see how many of each type of tree are present.

After running the workflow, the types and count of each tree type appear in the Results Window. When scrolling through results, note that there appear to be duplicate values. The Summarize tool is case sensitive, so it is important to ensure the data is consistent before summarizing. After inserting a data cleansing tool upstream and re-running the workflow, the expected results appear. When using the Summarize tool, only the columns that have had summary processes applied are included in the output.

The Summarize tool can also be used to find the tallest and shortest trees in each borough. Drag another summarize tool onto the canvas. This time, the column "Borough code" is selected and the "Group By" action applied. Next, the "Max" action is applied to the "Height" column. Now the "Min" action is applied to the Height column.

After running the workflow, each borough is listed with the height of its tallest and shortest trees. Since all boroughs have a tree less than 2 meters tall, this finding isn't meaningful, it will be replaced by the types of tree in each borough.

Changing an Action

If the new action was going to summarize data in the Height column, the "min" action could be changed by utilizing the dropdown. This dropdown will show all available actions based on the column selected.

Deleting an Action

Since the data containing tree types is not in the height column, this action will need to be deleted. Select the action to be deleted and click the delete button.

Datatypes and Actions

Now select the “Common Name” column and apply the “Concatenate” action which can be found in the string section. Note that the actions dropdown includes sections which represent datatypes, including Spatial. Some actions are only supported for particular datatypes, which means you may need to change your datatype upstream if you want to use a particular action. In this instance, “Common Name” is a String datatype which supports the “Concatenate” action. With the “Common Name” column selected, note that most “Numeric” actions are greyed out.

Action Properties

The concatenate action is selected from the string section. This action “adds” multiple string values together in one cell. This action includes an “Action Properties” window, which is unique to certain String and Finance actions. In this instance, the "action properties" is where the separator used to distinguish unique values is specified. The default is a comma, but a space after each comma makes sense in this instance.

After running the workflow, the list of boroughs contains the height of the tallest tree and a listing of the types of trees in that borough. Before sharing the findings, it makes sense to put the types of trees before the height and rename the columns.

Renaming Outputs

The concatenate action is selected and moved up using the arrows. To rename outputs, double click the “output field name” cell and type a more appropriate name. After running the workflow, the list is ready for sharing.