Joining Data - Transcript

Introduction

Nearly every analytics process requires blending data that is found in different data sources to join information together to produce deep and impactful analytic insights. Defining a relationship between data sources is what links data together. Once that relationship is established, a row of data in one input can be associated to a row of data in another input.

Three input datasets on trees surveyed in New York City have been combined to form a single data stream. Trees in good health and at least ten meters tall have been filtered from the dataset. However, before continuing with any further analysis, it would be helpful to clarify the meaning of the values in the column [Borough Code], which contains the digits one (1) through five (5). Luckily, a look-up table in a Text Input tool contains the [Name] and [Population] associated with each borough code. Join these two data streams together to enrich the tree data with additional information on the borough in which the tree is planted.

Join Tool Inputs

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

The Join tool has two input anchors: Left, indicated by the letter "L" and Right, indicated by the letter "R". Connect the "True" output anchor from the Filter tool to the Left input anchor of the Join tool.

Then, connect the output anchor of the Text Input tool to the Join tool's Right input anchor.

Join Tool Configuration

The Join tool can associate rows in one data stream to rows in another by one of two ways: Record Position and Specific Columns. Select a method to learn more.

Joining data based on the position of rows in the Join tool's input datasets assumes that the data in the first row of the Left input should horizontally align with the data in the first row in the Right input. Each row of available data is matched in this way. If you are absolutely sure that data should be associated by position, then this method quickly produces the results you need. However, if this is not the case, your results will be incorrect.

To join data by values in specific columns, you need a column of values in both datasets that match each other. Think of those values like a key that can be used to map one row of data in one input to

another row in another input. The most useful value to use as a key is an identifier that can be used to match data across datasets, like a [Borough Code]. Most importantly, this identifier must be present in both datasets as a column. Using the values in the columns [Borough Code], the row with a code of one (1) in the Left input will match to the row with a code of one (1) in the Right input. Then, the associated values for those records will horizontally align. The resulting data table widens as new columns, like [Borough Name] and [Borough Population], are added to the table.

When using the configuration option to join by specific columns, you will need to specify the columns in the Left and Right inputs to link data together.

Use the Drop Down to select the column [Borough Code] in the Left input.

The column name in the Right input will automatically populate if a matching column name is found. Otherwise, select the corresponding column to use to join data using the Drop Down. More than one column can be used to join data for a more restrictive match.

The Join tool includes what is called an “embedded select” window. This feature provides you with the same functionality available in the standard Select tool: you can remove fields from the output data by deselecting them, you can rename fields, and you can change the order of the outgoing data. This functionality is particularly useful for dealing with field names that the Join tool detects as duplicates. Any duplicate field names are highlighted and given the prefix of “Right_”. Unless that is an appropriate name for your analysis, it’s recommended that the field name be changed to something more representative of the information in that column, or even removed altogether from the outgoing data.

Deselect the duplicate column of data to remove it from the outgoing data stream.

Results

After running the workflow, an error message appears in the Results window: "String Fields can only be Joined to Other String Fields". Only fields of the same data type can be used to join data together. A column classified as a string data type, for example, can only be joined to another field that is also classified as a string data type. In the Left input, the column [Borough Code] is a string datatype, but in the Right input, this same column is classified as a Byte, which is numeric. Recognizing, and fixing, issues with data types and data hygiene can ensure that a Join executes successfully and produces expected results.

The Join tool produces three output data streams, each of which is represented by an output anchor on the Join tool. The output you are probably most interested in is the center Join, which is indicated by the letter J, and represents the data that is now linked together from both data sets. The Join between the Tree and Borough data has expanded the data stream to define the name and population

of the borough that, originally, was just represented by a coded value. Data from the left input that did not join to any data from the right input will fall out of the anchor indicated by the letter L. Similarly, data from the right input that did not join to any data from the left will fall out of the Right output anchor, indicated by the letter R.