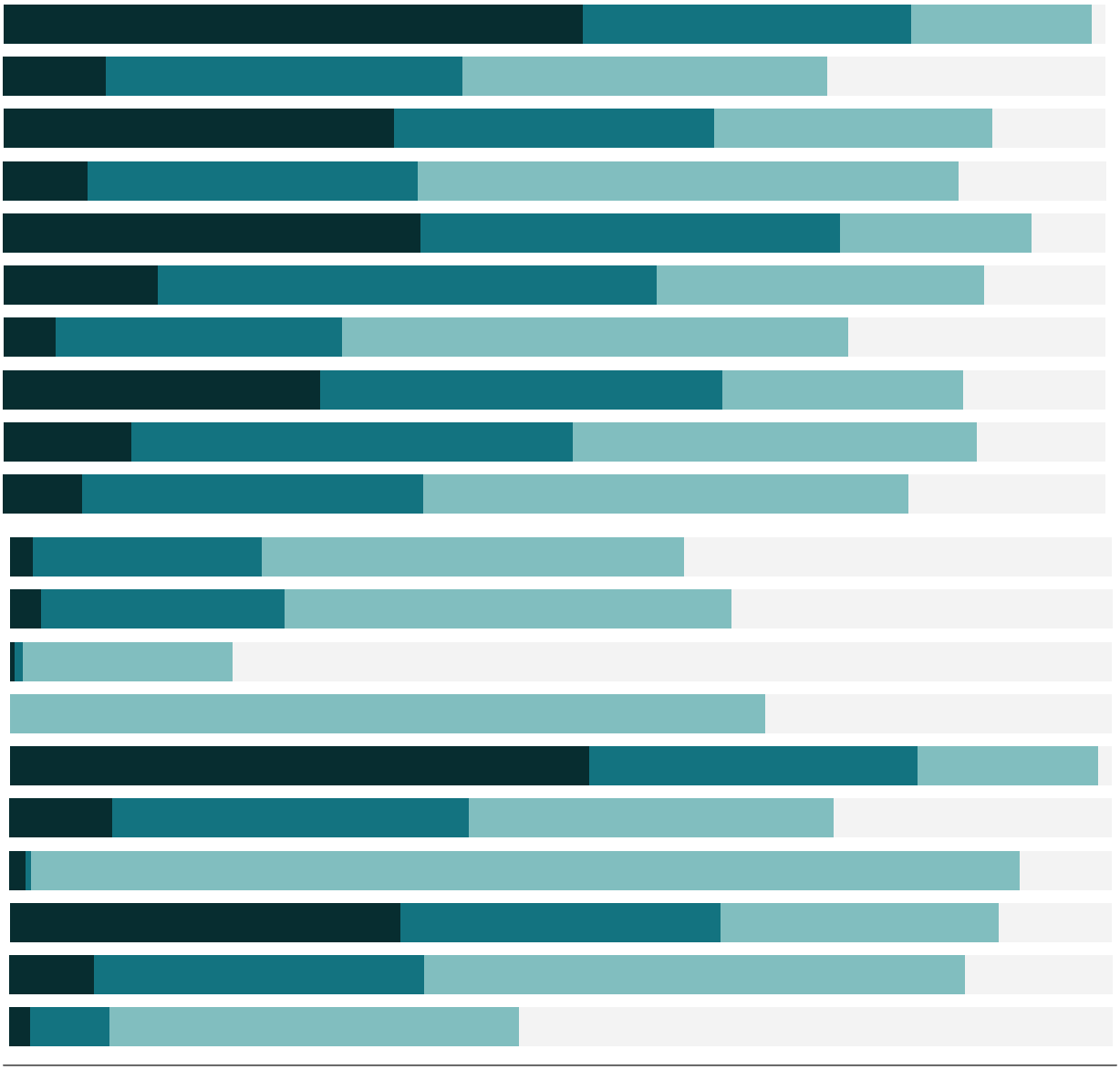

Free Training Transcript: Connecting to PDFs

Welcome to this video on the PDF connector. You can download the pdfs to follow along in your own copy of Tableau.

PDFs can contain valuable data for analysis in Tableau, but getting the data out can be tricky. The PDF connector in Tableau is designed to help pull out data in tables. Because PDFs lack metadata about the data, there may be some post-connection work to be done by you before analysis.

Connecting to PDF data tables

Tableau can read crosstab data tables from PDFs. Ideally, they should look something like this, with columns and rows, with a single line data for each row. Here, the stock data starts on page 2.

Let's connect with Tableau. When we select PDF File from the connect pane, a page-picker appears. We can scan the whole document, a specific page, or a range of pages. We'll look at pages 2-8. One of the lines from the pdf table header is confusing Tableau, but if we turn on the data interpreter, we get our headers as expected.

Union

Each page comes in as a separate table, but since they're structured the same way with the same column headings, it's easy to bring them back together with a union. With the first table already on the canvas, we'll drag any other desired pages' tables to the union drop area underneath the first. We'll get a new column for table name, so we always know where the data came from. If we scroll down, we can see that the unioned data from page 3 is aligning perfectly with page 2.

Cleaning up imperfect tables

This PDF was structured really nicely for import, but not all PDFs will be. As a general rule, Tableau will connect best to PDFs with a tabular structure like this, with a single row of data per line, no hierarchies or nested headers, and no sub tables.

But PDFs don't have to be perfect to work. Let's connect to another file, one that's not as easy as the first. The table we want is on page 14. As a note: Tableau looks at absolute

page numbers, which may not correspond to the pagination in the document. The "rescan PDF" option, under the data connection dropdown, lets us re-pick what pages to look at. There's only one table on the page, but there are 3 options to the left. Tableau detected 3 possible ways to pull in that table. If we bring out each one at a time, we can see what they contain.

Note: when unioning tables across pages like we did before, be sure to union the correct table version from each page, not multiple versions of the same page.

We can bring out table 1 and use the data interpreter: It looks like table 1 is everything, but for some reason 1995-1997 are being read as a single column. If we wanted to use this version of the data, we could clean this up by doing a custom split on the field and splitting all columns on spaces, then rename the split fields to the appropriate years.

But let's see what tables 2 and 3 give us. Table 2 looks like the bottom of the original table. And table 3 looks like the top. I like the column delineation better with tables 2 and 3, so let's work with those. First, I'll union them, dragging out table 2 under table 3: there's a mismatch between "inflows" and F1. If we select both and choose "merge mismatched fields", we now have the column as expected. We can rename this "Water Source".

Dealing with null values

There are several rows of nulls, either because a sub header, like "Change in storage" was read as a data row, or because a single row of data was broken up into multiple rows, like with "discharge from hydroelectricity generation", where the header was read as 3 different rows.

To get rid of these nulls, we'll add a data source filter, from up in the right corner, click Add. We'll add a filter, and we would choose any column that has these nulls, I'll use F10, and click OK. Next, select Null, then change the filter to exclude. Then click Ok, and OK again. Now those rows are gone, and we're left with just our data.

But I'll undo that, because we can also see there are several types of water source that are actually totals. Let's filter those out and the nulls all at once. Click Filter > add and we'll select Water Source. Check anything that's a header or a total from the original PDF, or that corresponds to a null row for the values: Abstraction for, Change in Storage, Discharge from, Hydroelectricity appears twice, once with data and once with nulls, so

I'll leave it for now, Outflows, and the three totals at the bottom. Then click Exclude, and ok.

Fixing Headers and Pivoting

Now we just have the rows with data- and one row of nulls for hydroelectric, but we can clean that up in a moment. We don't have any headers, but we can cross-reference the values with the original PDF to see what should be what. And the columns should just be the years from 1995 to 2010. I'll speed this up.

We can now pivot our data so we have a column for Year and a column for Million Cubic Meters. We'll hide the table name column. Change the data type for Year to date. Change the data type for million cubic meters to number (whole). And now that this is the final column, we'll filter any nulls in this column: Filters > Add, Million cubic meters, Select null, Check exclude, and ok.

Realiasing members of a field

There's still some issues with the water source names. Bring up the menu for that field, and click Aliases. OK. Let's clean these up. Just like we could rename columns (or fields), here we can re-alias members of a given field. Right click a value and select Edit Alias. Discharge from hydroelectricity generation. I'll speed up the others. Groundwater. Abstraction for hydroelectricity.

Recreating Groupings and Hierarchies

Finally, I noticed there was a structure in the original PDF table—there are categories of water sources. We can build this structure in the data pane, so we'll click to sheet 1. First off, Million Cubic Meters is really a measure, so we can drag it down there.

We can recreate that here with groups. Abstraction for hydroelectricity, Discharge from hydroelectricity generation, and Evapotranspiration, To sea and net abstraction can be grouped.

As can Groundwater, Ice, Lakes and reservoirs, Snow, Soil moisture.

Let's right click on this new field and edit the group. First, we'll click Precipitation and click group, to make it its own group. And now we can rename them: Inflows, Change in Storage, Outflows. And we'll name the grouped field categories. If we drag water sources

on top of this new categories field, we can create a hierarchy, enabling drill down in the view.

Tips on working with PDFs

The precise steps to clean up any given PDF post-connection can vary, but hopefully this video showed you some tools you can use to clean up the data so it's ready for analysis.

As a reminder, Tableau will have trouble with PDFs containing: Sub tables, Hierarchies in headers, "single rows" that are actually multiple rows of content that should be interpreted as a single row.

Finally, note that colors and shading can change how data is interpreted because of how PDFs must be parsed to cells and tables of data.

Conclusion

Thank you for watching this PDF connector training video. We invite you to continue with the Free Training videos to learn more about using Tableau.