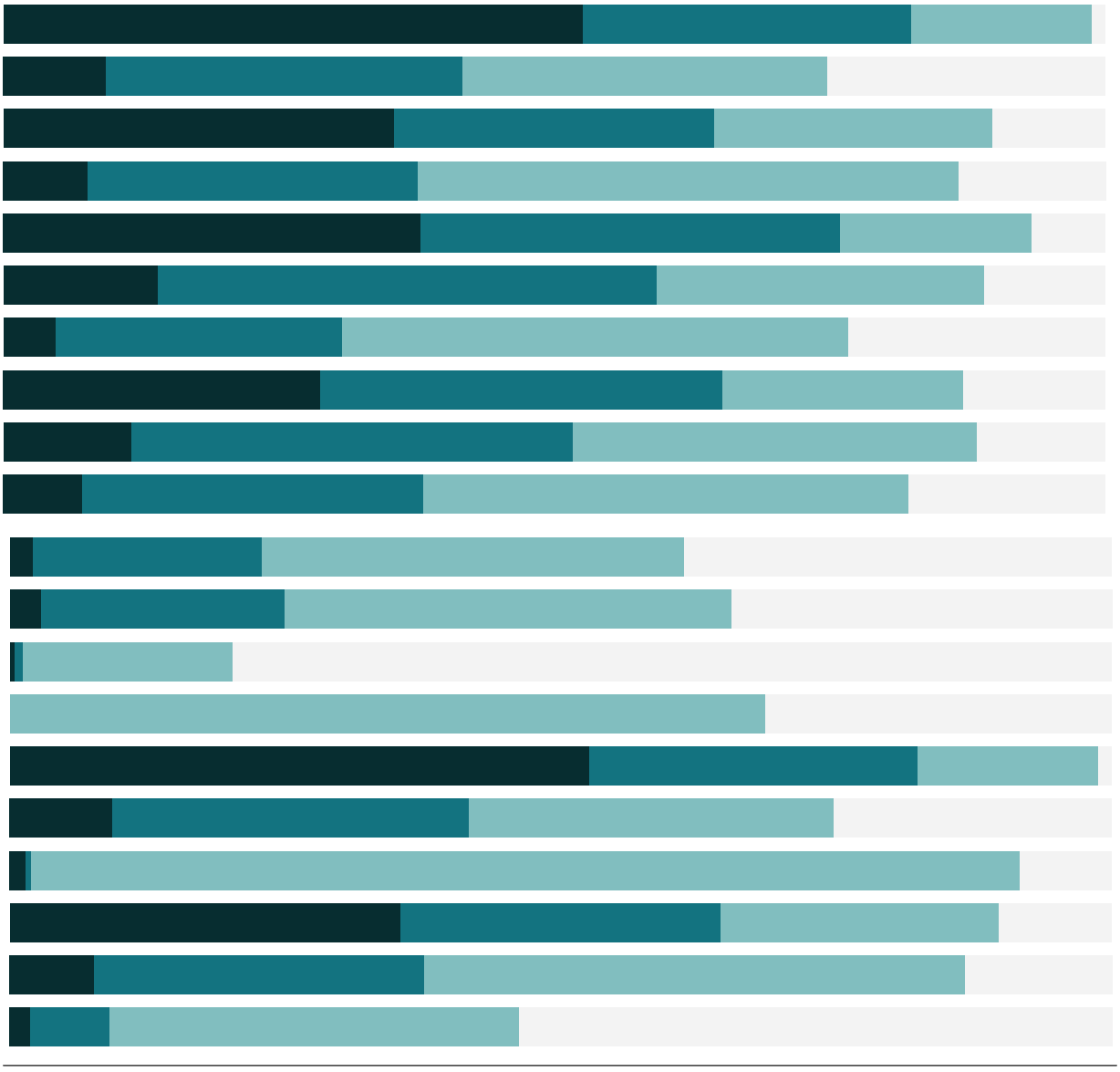

Free Training Transcript: When to Blend and When to Join

Welcome to this video comparing data blends and cross-database joins. You can download the Exercise workbook to follow along in your own copy of Tableau.

Briefly, cross-database joins integrate two data sources at the row level to create a single data source. Blending takes separate query results from each data source and aggregates them in the view. For more information on Cross-database Joins and Data Blending specifically, please watch their individual videos.

One-to-many relationships

To begin, we'll set up a blend with our Product 2016 and Sales 2016 data sources from the cross-database joins video. Let's do a quick analysis of our categories sales by market. From the Product data source, bring Category to Rows. Then we'll switch to the Sales data source. Click on the link by Product ID to establish the blend relationship and bring Sales to Columns. This looks great, we have our blend, and we can get a sense of our overall sales per category. But what if we bring Market to color? The majority of our bars are now grey, which is an asterisk. Why?

We blended on product ID, which has a one-to-many relationship with Market. The bars in pale teal show product IDs that are only sold in the USCA market, but everything in grey is sold in multiple markets--represented in aggregate with an asterisk. Because of the way a blend is performed, aggregating the results from each data source separately and displaying them together in the view, we can't easily visualize data when there's that one-to-many relationship.

This is an instance when a cross-database join is a better answer than blending. This data source is joined at the row level on Product ID. Let's recreate that first view: Category to rows, Sales to columns, Market to color. We now have the behavior we wanted, breaking up our category sales by market. When we joined the data, we simply supplemented the product information to the sales information, leaving the original number of records from Sales unchanged.

Joins inflating the number of rows

So is a cross-database join always the best option? Let's switch to another pair of data sets, this time Office City and Coffee Chain from the Data Blending video. These are two retail chains owned by the same parent company. Office City sells in all 50 states in the US, and Coffee Chain sells in 20 states. Let's say we want to see our combined sales from both organizations across all 50 states.

Here we have a joined data source, joined on the common field, State. We'll bring out State from the Office City connection within this data source, since it has all 50 states. And if we right click and edit the calculation "Added Sales" we see that we've added our Office City sales and Coffee Chain sales, with a Zero Null function so for states where Coffee Chain has no sales, the null is handled as a zero. And we'll bring Added Sales to Columns. Wow. These are BOOMING businesses, if we have hundreds of millions in sales for some states. I happen to know this can't be right—let's dig deeper.

This dashboard shows the Office City sales for 4 states, and the Coffee Chain sales for the same four. Our combined sales should be what we see in the equations in the center – simply adding the values. The blended (yellow) combined sales gives the right answer, but the joined data source (purple) has hugely overblown values.

If we look at the number of records on Office City, Coffee Chain, and the joined data source, we see this is clearly inflated. Because we joined on State, Tableau made a row for each combination of State with unique information: Product and Date from Coffee Chain, and Row ID from Office City. This cross product meant we were counting our sales values multiple times.

In contrast, blending provided the correct values because it returned the sales value for one data source for State, the sales value for the other data source for that state, and we were able to just add them together. This is the desired behavior in this instance.

As you can see, what field we join or blend on has a HUGE impact on the analysis. Choosing the correct field, and understanding how that blend or join is being created, is very important to knowing which to use when.

Benefits of each

When performing a cross-database join to create an integrated data source, there are many benefits. Because the join is row-level, it doesn't run into the one-to-many limitations faced by data blending—that is, no more asterisks. The single data source can be extracted, saved, published, and easily shared with others, unlike a data blend. It's just like any other data source. However, if the granularity of joining field isn't correct, performing a join can artificially inflate the size of your dataset and provide misleading information. A data blend sends separate queries to each data source and aggregates them to the desired level of granularity in the view, which can be exactly what the analysis needs.

Conclusion

Thank you for watching this video on combining data sources with blends or cross-database joins. We invite you to continue with the Free Training videos to learn more about using Tableau.