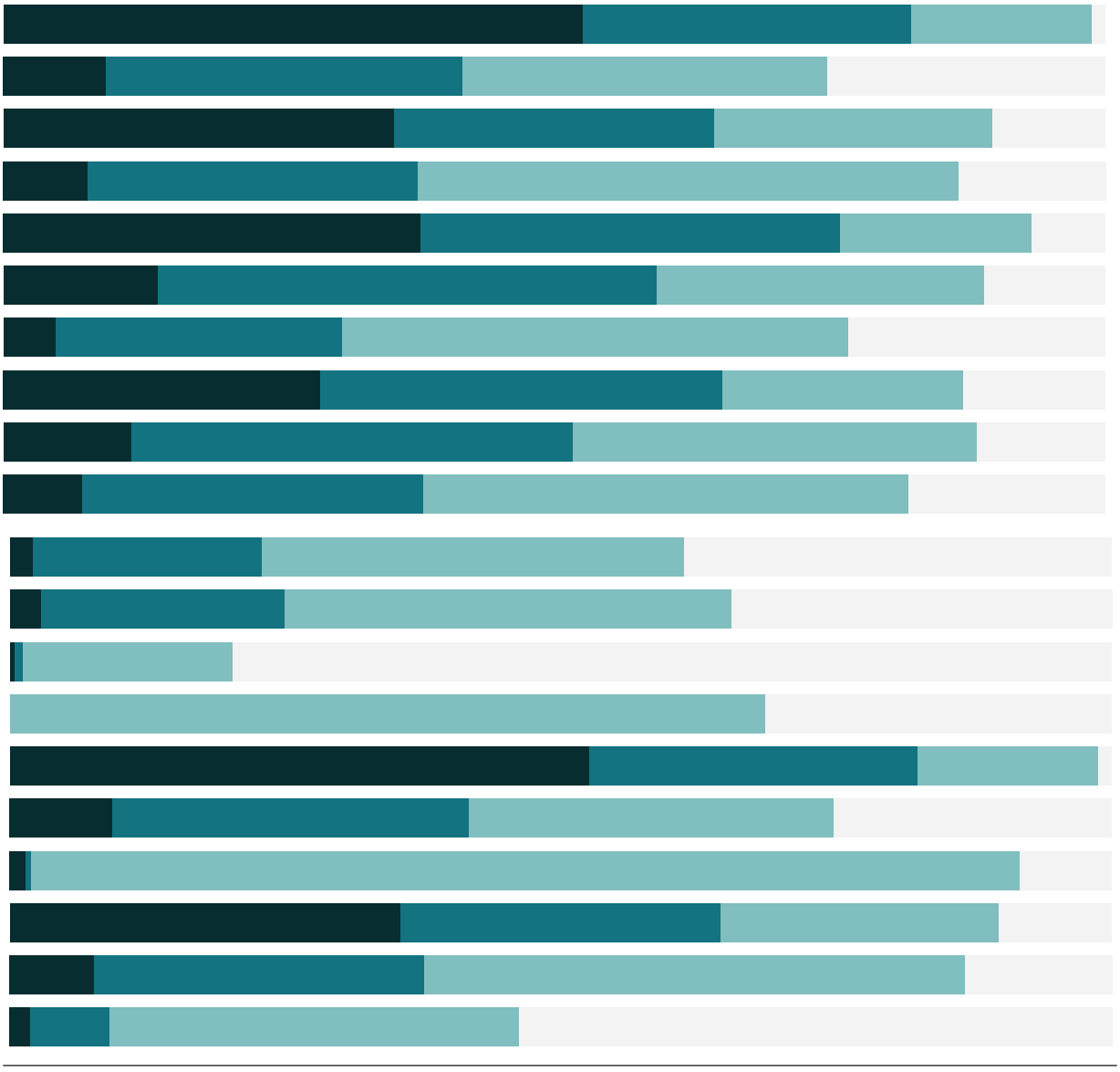

Free Training Transcript: The Profile Pane

Welcome to this video on the Profile Pane in Tableau Prep. You can download the data set and packaged flow file underneath the video to follow along in your own copy of Tableau Prep. We're working with data for bestselling books.

The Profile Pane

When we're on a cleaning step, indicated in the flow by the bar icon, the pane below the flow is the Profile Pane. (For information on cleaning data, check out the video on the Cleaning Step.)

The Profile Pane helps us explore our data and understand its contents--it's a powerful way of interacting with our data. For discrete data, each grey bar we see represents a value in the field itself. The length of the bar represents the number of records with that value, and the visual scrollbar provides an overview of the distribution of the data. For example, we can see that most titles show up only once in the data, but Ready Player One appears twice. Similarly, most authors appear once, but we can sort and bring the authors with the most records to the top.

Highlighting

If we click on a bar such as this author with the most records, it's highlighted in blue. Across all the other cards, the values associated with this author are also highlighted. Using the visual scrollbar in the Title column, we can find the titles of her books. We can see that her books are in the 10-20 dollar range, they're various ranks, and all occur in the Early & Middle list.

Similarly, if we click on the nulls for the "Weeks on List" field, we see null values here correspond to the nulls in the "Previous Rank" field and the entirety of the Early & Middle and Young Adult lists. It appears those bestseller lists don't provide this information. This highlighting makes it easy to examine the structure of our data and see how the distributions and values of various fields are related.

Distributions of Data

Discrete data shows as grey bars with each value in the field represented. Continuous data shows as blue bars in a histogram, representing ranges of the data. Let's look at the Price field. We see the most common prices are 10-20 dollars, with 53 rows of data in this

bar, and only one row in the 50-60 dollar range.

If we want to see the actual prices themselves, we can open the menu and change the view state from summary (that is, histogram) to detail (showing each value). Now we get a visual scrollbar on the side that shows the more detailed distribution, and we can see three peaks--cheaper books, the slightly pricier range of likely trade paperbacks, and a longer tail with peaks in the upper twenties. Going back to the summary view, if we multiselect the Hardcover lists, we can see sure enough, those higher prices are for hardcovers.

The default view for continuous data is the binned, summary view. This is very useful for outlier detection. For example, in the "Weeks on List" field, it's easy to visually determine there are several records that have been on the bestseller lists far longer than others. We could dig in to see if those are errors of data recording or if they're just abnormally popular books. If we were in the detail view, it'd be much harder to see the gap between the rest of the records and those values.

Using the Profile Pane to Identify Errors

Let's click into a more complex flow. This time, we're looking at all 4 weeks worth of data for February. We can see that there are a lot of nulls for ISBN, Author, and Title. This is unexpected.

If we click on one of these null bars, we see they're from one specific week. Let's go back to that input. If we click on the plus and select "Insert Step", we can bring up the profile pane for just this week's worth of data. There are no nulls in the Information field, and if we insert another step to compare, all the delimiters look the same as other weeks. There must be something else going on.

Ahhh, wait. It looks like this field is called "Info" for one week, and "Information" for the other. Let's go into the Union and sure enough, we can use these colored bars to identify that the Info/Information columns are mismatched. We'll drag Information onto Info to merge them, and now if we go back to the combined cleaning step, we see those nulls are gone for title and author. (For more information on Unions, check out the video on the Union Step).

However, there are still some odd bars showing in the profile pane. There are some null prices, and some blanks for ISBN. If we click on either of those bars, we see it's just Trade Paperback Fiction from the week of the 21st. Let's go back and look at that data.

Insert a step, and click “Trade Paperback Fiction”. In the data grid, we can see just those values (almost like a temporary filter.) And it looks like there’s a comma instead of a pipe between the price and ISBN. Let’s fix this with a quick calculation. REGEXP_ REPLACE([Info], “, ISBN:”, “ | ISBN:”)

Because we made this change upstream in the flow, before the union and cleaning step, when we go back to that later clean step, the change has carried through and we have the appropriate data in place. The profile pane shows some good looking data!

Conclusion

Thank you for watching this video on The Profile Pane in Tableau Prep. We invite you to continue with the free training videos to learn more about using Tableau.