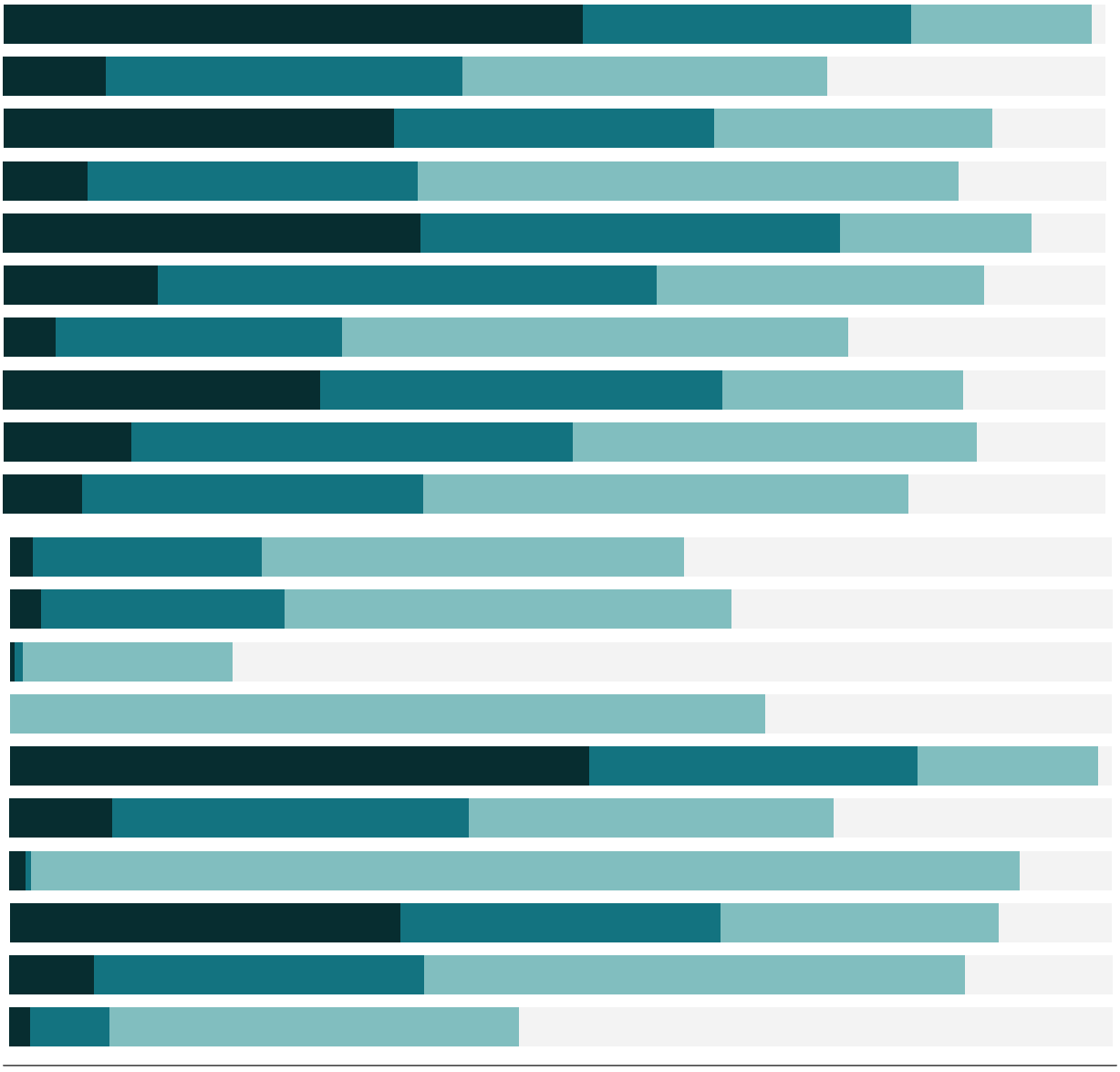

Free Training Transcript: Group and Replace

Welcome to this video on the Group and Replace feature in Tableau Prep. You can download the packaged flow file underneath the video to follow along in your own copy of Tableau Prep.

The Context for Group and Replace

A common issue in data cleanliness is the presence of multiple values that should be a single value, such as GB and Great Britain. To deal with this in Tableau Prep, we can leverage the Group and Replace feature in a cleaning step. This feature allows us to group multiple values and replace them with a single value--essentially realiasing.

To begin, we're on a cleaning step in the flow. In the profile pane, we can see the fields in this data set. This is clearly nonsense data to illustrate a feature, not data we should try to analyze.

Manual Selection

The Airbnb listings field displays a common data cleanliness issue--inconsistencies of how a specific piece of information is captured. Here, the room information (bedroom or bathroom) was recorded multiple different ways. There's no easy way to programatically remap these to be consistent, so we'll do a manual group and replace.

First, we'll click on the card for the Airbnb listings field and open the menu. Under Group and Replace, we'll choose Manual Selection. Whatever we select first will become the replacing value--we'll use "bath". With that replacing value selected, the right side of the editor will display the remaining values. Select all the values that should be grouped under the replacing value.

The left side of the editor will display a preview of the new values for the field--with a grouping paperclip icon beside bath, indicating it's a grouped value.

Let's do the same thing for "bed". Select it on the left, and choose the values to be grouped from the right. We can see at the top on the right side, the grouping value's checkmark is greyed out, but the others can be toggled, allowing us to remove them from the group if desired. Since we took something out of the bath group, when we removed it from "bed", it's now an ungrouped value. Let's add it back to "bath". Click the "bath" bar to the left, and add it back to the group.

Out of Domain Values

It's worth noting that if we were to refresh this data and a new value, such as "beds", popped up in this field, it would not be added to any group. If we want to make sure that "beds" is added to the "bed" group--even though it's currently not present in the domain of the data--we can manually add it. Over in the left side of the editor, click the plus icon. We can now type in a net new value--BEDS. It shows with a red dot, indicating it's outside the domain, but otherwise it functions the same as any other value. If we click back into the "bed" group, we can add the new, red-dot BEDS value.

Now if the data is refreshed and "BEDS" shows up in the field, when the flow is run, that value will be grouped appropriately.

Fuzzy Matching Algorithms - Pronunciation

Manual Selection is good for instances where incorrect values are irregular or fairly different from the desired value. However, there are two options that apply algorithms to help group values.

In the Misspellings column, we have four commonly misspelled words and several variations for each word. In each instance, the incorrect versions are at least approximately pronounced the same way as the correct spelling. This is the kind of error a spell checker may catch. We can use the Pronunciation option for Group and Replace, which uses the metaphone 3 algorithm, to automatically address these misspellings.

We'll click on the card and open the menu, then select Group and Replace > Pronunciation. The grouping is automatic this time. We can see on the left of the editor the new groups. Because the correct spelling was the most common value, that is what is chosen as the replace value. If the Replace value was ever not the desired value, we can right click on the new group and select "Edit Value", then modify it to be what we want.

Clicking on a group also opens the right side of the editor and exposes the values in the group, and we can manually add to or remove from the group if desired.

Fuzzy Matching Algorithms - Common Characters

The last Group and Replace option is Common Characters. In the Name Formatting column, we see that some of the names are listed as last-name-comma-first-name rather than the more common first name last name. We don't want to have two values for the same person, so we should group them. Common characters uses the n-gram

fingerprint algorithm to identify words by their unique characters, thus recognizing that Dorian Gray and Grey, Dorian are the same.

As with Pronunciation, the grouping is automatic, and the most commonly occuring value in the group becomes the replace value.

Notes on Algorithms

It's important to remember that algorithms are not perfect. We recommend checking the groups to ensure data is not incorrectly grouped. A group that is automatically created can be modified just like a manual group by clicking on the bar and using the right side of the editor to choose the correct values.

Although only one method of Group and Replace can be performed at a time, if an algorithm misses a group that should exist, we can create a manual group in the editor in addition to the automatic groups.

Edit Value

Finally, if we need to simply address a typo or other small-scale issue, we can always directly edit a value. If, for example, we fix this misspelling of Tableau by right clicking and Editing the Value, it will automatically be grouped with the correct value.

Conclusion

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