

Demo: Using Recode

In this video, you learn how to use Recode to clean supplier names in the file Components.jmp.

First, let's take a look at the supplier names using the Distribution platform from the Analyze menu. You can see that there are 10 levels (or categories) of supplier, and that 10 observations are missing values.

However, you can see that there are really only five unique suppliers. There are issues with spelling and capitalization, and you can see that the supplier Cox is entered two different ways.

To clean up the supplier names, we use Recode.

To do this, we right-click the column head for supplier and select Recode.

The current values for supplier are summarized under Old Values. The 10 missing values are at the top. Anderson is listed three times. One value has a leading space, and the others have slightly different spelling.

Let's look at the supplier Hersh. This is also listed three times. There are a couple of issues with capitalization and spelling.

To clean up some of the most common issues with categorical data, we'll use options under the red triangle.

For example, we'll select Convert to Titlecase to address the inconsistent capitalization, and we'll select Trim Whitespace to remove leading spaces.

All of the changes are tracked in the Changes box.

To address some of the other issues, we'll take a shortcut and group the values. For example, when we select all of the labels for Anderson and click the Group button, all of these labels are grouped together.

We'll also group the labels for Cox and Hersh. Now there are six values: the five suppliers, plus the missing values.

In many analyses, observations with missing values are not included. One technique for including missing data for categorical variables in your analysis is to recode the missing values.

Let's do this. We'll recode these values with the label Missing.

Now, we need to determine how to save the recoded values. The default is to save the values to a New Column. Another option is to save the recoded values, with a formula for recoding, to a Formula Column. The last option, which isn't recommended, is to overwrite your current values by saving the recoded values in the same column.

We'll select Formula Column and click Recode. A new column, supplier 2, has been saved to the data table with a formula. The formula provides a trail of the changes that you made. You can also update the formula if you need to make additional changes.

One final step is to graph the original values and the recoded values to make sure the recoding was correctly applied. This looks good!

Statistical Thinking for Industrial Problem Solving

Copyright © 2020 SAS Institute Inc., Cary, NC, USA. All rights reserved.

Close