Demo: Exploring Missing Values

In this video, you learn how to explore missing values using the Components data.

First, we use Columns Viewer to see how many values are missing for each of the variables. We select Columns Viewer from the Cols menu, select all of the variables, and click Show Summary.

Several variables are missing values. For example, part number is missing 4 values, and supplier is missing 10 values. Temp is missing 265 values. Only 104 of the batches has a value for temperature.

With Columns Viewer, you can see how many values are missing for a given variable. You might also be interested in seeing the missing values across your variables.

To do this, we use Missing Data Pattern from the Tables menu. We select all of the variables, click Add Columns, and click OK.

A new linked data table, called Missing Data Pattern, is produced. The columns in this table describe the pattern of missing values in the original data table.

The first row represents 99 rows in the Components data table. These rows are not missing values for any of the columns.

Let's look at the second row. This represents five rows in our original table. These rows are missing values for one column. When we look at the Patterns column, the 1 in the last position tells us that these rows are missing values for the last variable. When we scroll through the variables, we see that this last variable is supplier.

There are 255 rows that are missing only the temp. However, temp is also missing with some of the other values. For example, four rows are missing both temp and supplier.

You can run the Cell Plot script that was saved to this table to visualize the Patterns column. This makes it easy to see that the biggest problem is the missing values for temp. The last row of the cell plot corresponds to the last missing data pattern. One row is missing values for many of the variables.

The Treemap script creates a treemap of the missing values. When you hold your mouse pointer on the treemap, you see the values for each block. Approximately 27% of the rows are complete records. That is, they are not missing any values. You see that 69% of the rows are missing just temp. So approximately 96% of the rows in the Components file either have complete data or are just missing temp.

This is good. We have a problem with missing values for temp. But otherwise, missing data might not be a big problem with this data set.

Statistical Thinking for Industrial Problem Solving

Close