Summarizing What You Have Learned

In the previous videos, you saw how to diagnose potential data quality issues using the data table, summary statistics, and basic graphs.

With the components data, you see examples of each of these common problems:

Let's talk about these issues, one at a time.

Incorrect Formatting: The components data table is structured for analysis, in columns and rows, and the only formatting issue was a few incorrect modeling types.

You have changed the modeling types for batch number and part number to nominal, because these are simply labels. You should also change the modeling type for process to nominal, because this is a two-level categorical variable with no real ordering.

Batch size is currently coded as continuous, but there are only four batch sizes. Depending on the types of analyses you run, you might want to change the modeling type for batch size to nominal or ordinal.

Incomplete Data: You didn't have a variable for scrap rate in the data set, so you used a formula to create this variable.

Creating new variables based on existing data is a fairly common step in data preparation. You see some common examples in a future video.

During data analysis, you often find that you need to collect more data. You might need to add observations to your data table, or you might need to add data for other variables.

If you are missing too much information, you might need to create a new data table with this missing information. Or you might be able to add data from another file to your existing data table.

You learn how to combine data tables in upcoming videos.

Missing Data: You are missing a few values for many of your variables, and this might not impact your analyses. But, for the variable temp, you are missing 265 values. You need to investigate why so many values are missing.

There are many strategies for how to handle missing values, and much of this discussion is beyond the scope of this course.

In an upcoming JMP demonstration video, you learn how to create a Missing Data Pattern to explore and understand missing values.

For additional information about how to deal with missing values, see the Read About It for this module. Dirty or Messy Data:

You need to investigate why there are negative values for scrapped parts and determine how to address this issue. You might learn that there is an issue with the reporting system. For example, the negative values might be the result of parts that were incorrectly returned to the wrong batch.

You should also investigate the outlier for speed and determine what caused this reported value. It might be the result of a typo, or there might have been some sort of equipment malfunction. If you can't determine the root cause of this outlier, you might decide to hide and exclude this observation from your analyses. An alternative is to use a missing value code for speed to specify this value as missing. This excludes the value from analyses, but the values for the other variables are still included.

For information about using missing value codes, see the JMP Help at www.jmp.com/help.

Most of the orders were from six customers. If you are going to use customer number in an analysis, you might need to combine some of the smaller customers into an "other" or "miscellaneous" group.

You have 10 levels of supplier, but there is inconsistent naming and abbreviations. There are really only five suppliers. You need to do some data cleanup to fix the supplier names.

In a future JMP demonstration video, you learn how to use Recode in JMP to address these issues with messy categorical data.

Statistical Thinking for Industrial Problem Solving

Close