Print

# Common Data Quality Issues

Before you analyze any data, you need to make sure the data are prepared for analysis.

In this video, you learn about the most common data quality problems, along with issues you might need to address during data preparation.

Common data quality issues include incorrect formatting, incomplete data, missing data, and

dirty or messy data.

Let's talk about each of these issues. Later, you learn how to diagnose and address some of these problems.

We start with incorrectly formatted data. This occurs when data are in the wrong form or format for analysis. This can apply to the data table as a whole, or to the formatting of individual variables in the data table.

When the data table is incorrectly formatted, the data might be stored in separate columns but an analysis requires the data to be stacked in one column.

Within a data table, the individual variables might have the incorrect modeling type. Because the modeling type drives the analysis, having an incorrect modeling type might lead to the wrong analysis.

Another issue relates to dates and times. Columns containing date or time measurements need to be formatted as date or time variables. This enables you to perform calculations, such as the elapsed time between events.

Incomplete data can relate to not having data on important variables, or not having enough data to perform an analysis.

If you are missing critical variables, you might not be able to identify the root cause of the problem.

Not having enough data can impact the conclusions you draw when you conduct formal analyses or create statistical models.

For example, you can have a high degree of uncertainty when you estimate a parameter, or you might not be able to detect a difference when you conduct a hypothesis test. You learn more about methods for making sure you collect enough data in the Decision Making with Data module.

One of the most common problems is missing data (or missing values). When you have a missing value, this means that the value for a variable is not available for an observation.

There are different types of missing data. Data can be missing completely at random, missing at random, or missing not at random.

When data are missing completely at random, there is no structure or pattern to the values that are missing, and there is no identifiable reason for the missing values. The pattern of missing values, either within a variable or across variables in a data table, is completely random.

When data are missing at random, there might be a reason the values are missing, but the missingness generally does not impact your study or your conclusions. For example, suppose that you are studying heights and weights of men and women. If women, in general, are less likely to

report their weight than men, weight might be missing at random for women. The distribution of weights for women, at least theoretically, is the same with or without the missing values. You simply have fewer weight values for women.

In both of these cases, analyses might not be seriously impacted by the missingness unless a large number of observations have missing values.

However, you should be concerned about values that are missing not at random. When this happens, the fact that a value is missing is related to the reason it is missing. For example, this happens when a question on a survey is skipped on purpose, or when sensitive information is omitted. Consider the weight example. If women who weigh more are less likely to report their weights, weight might be missing not at random. The observed distribution of weights for women is different from the true distribution. This can lead to biased analyses or incorrect conclusions.

The last issue is dirty or messy data.

When your data are messy, your data values might be incorrect, inaccurate, have typographical errors or typos, or be dated or obsolete. Or there might be other issues with the data.

For categorical variables, the data might have inconsistent capitalization, abbreviations, and spacing. Or there might be an overwhelming number of categories, some of which have few values.

Continuous data might have unusual values, or values that aren't physically possible. For example, you might have negative values when the variable can only take positive values.

There are other types of messy data, which occur at the observation and variable level.

For example, records might be duplicated, meaning that the same observations are included more than once.

Or variables might be redundant to one another. That is, you might have two or more variables that contain essentially the same information about the observations.

Now that you're aware of potential data quality issues, let's see how to diagnose some of these issues. In the upcoming videos, you learn how to scan the data table for trouble, and how to use numeric and graphical summaries to identify possible issues.

---





Close