Summary

2 minutes

We've covered a lot. Let's recap on some of the key messages.

What are data?

Data fall into several conceptual categories. The most common are:

- continuous data (numbers),
- categorical data that have no order,
- ordinal data, which can be treated as numbers or ordered categories.

Data are stored on a computer as distinctive types, and we usually try to match the datatype to the kind of data that it is. For example, continuous data are best stored as floating-point numbers, because these allow fractions to be stored. By contrast, categorical data often arrive as strings (text) and must be converted to one-hot vectors for the computer to understand them properly.

What makes a good dataset?

We learned that a dataset it helpful if it:

- contains relevant information,
- is complete,
- is a good representation of the population (real-world).

If data aren't complete, we can take steps to make sure that incomplete data doesn't cause big issues. When doing so, we need to be careful not to introduce new issues, such as making data no longer representative.

Thinking about data

We showed how visualizing data can help to get an understanding of what might be useful in a model. Using different types of graphs, colors, and so on, can be fun and make complex information much more intuitive.

We learned that understanding our data lets us make better decisions about our models. In the final exercise, we improved our model by exploring how many cabins were on the ship and taking the time to consider why this information was helpful. Yet, overall found the could be improved through simplification into nine Deck labels.

Module complete:

Unlock achievement