Intermediate Machine Learning: 2nd lesson – Missing Values

In this tutorial, you will learn three approaches to dealing with missing values. Then you'll compare the effectiveness of these approaches on a real-world dataset.

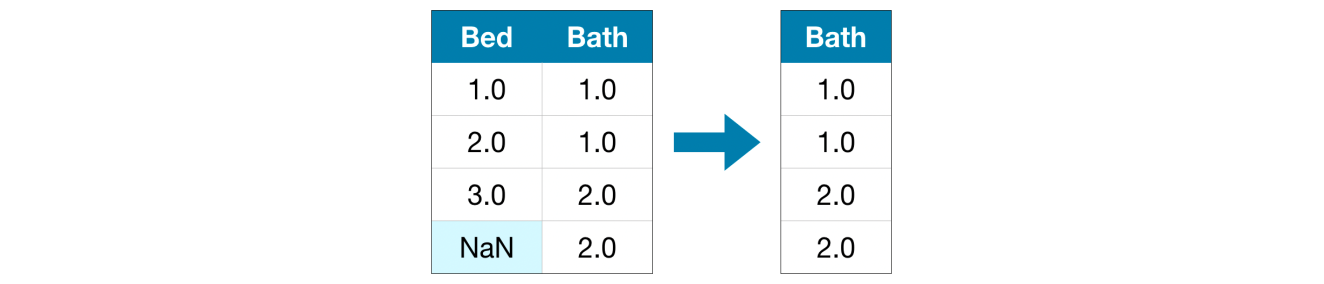
There are many ways data can end up with missing values. For example:

* A double-bedroom house won’t include a value for the size of a third bedroom
* A survey respondent may choose not to share his income

Most machine learning libraries (including scikit-learn) give an error if you try to build a model using data with missing values. So you'll need to choose one of the strategies below.

* A simple option: drop columns with missing values

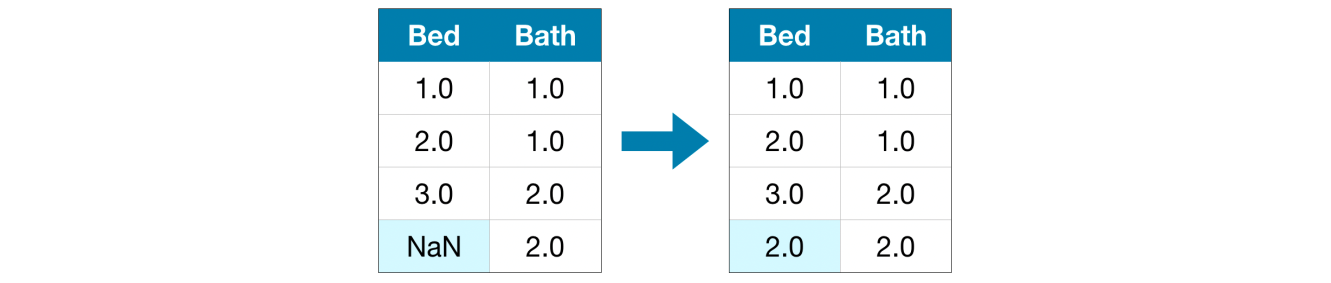
The simplest option is to drop columns with missing values.



Unless most values in the dropped columns are missing, the model loses access to a lot of (potentially useful!) information with this approach. As an extreme example, consider a dataset with 10,000 rows, where one important column is missing a single entry. This approach would drop the column entirely!

* A better option: imputation

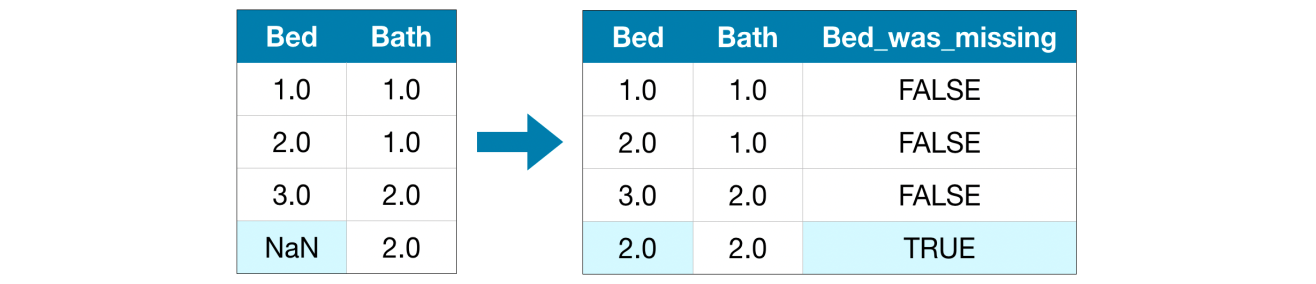
Imputation fills in the missing values with some number. For instance, we can fill in the mean value along each column.



The imputed value won't be exactly right in most cases, but it usually leads to more accurate models than you would get from dropping the column entirely.

* An extension to imputation

Imputation is the standard approach, and it usually works well. However, imputed values may be systematically above or below their actual values (which weren't collected in the dataset). Or rows with missing values may be unique in some other way. In that case, your model would make better predictions by considering which values were originally missing.



In this approach, we impute the missing values, as before. And, additionally, for each column with missing entries in the original dataset, we add a new column that shows the location of the imputed entries. In some cases, this will meaningfully improve results. In other cases, it doesn't help at all.