Data Cleaning

Better data beats fancier algorithms...

Better Data > Fancier Algorithms

Everybody has to do it, no one really talks about it.

Different types of data will require different types of cleaning.

The systematic approach laid out in this lesson will serve as a good starting point.

Remove Unwanted Observations

Duplicate Observations

Usually come from the data collection process such as combining datasets or scraping data.

Irrelevant Observations

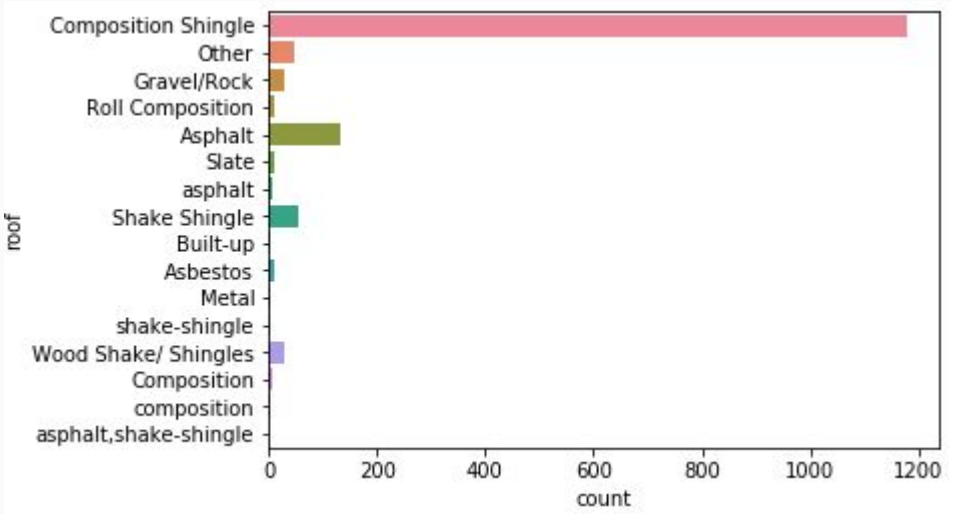
Observations that don’t help with the problem you’re trying to solve.

For example, if building a model for single-family homes only, you wouldn’t need observations for apartments.

Fix Structural Errors

As you can see:

- ‘composition’ is the same as ‘Composition’
- ‘asphalt’ should be ‘Asphalt’
- ‘shake-shingle’ should be ‘Shake Shingle’
- ‘asphalt, shake-shingle’ could probably just be ‘Shake Shingle’ as well



Filter Unwanted Outliers

Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers that decision tree models.

In general, if you have a legitimate reason to remove an outlier, it will help your models performance.

Handle Missing Data

You cannot simply ignore missing values in your dataset.

Dropping vs. Imputing:

- Dropping is suboptimal because when you drop observations, you drop information.
- Imputing is suboptimal because the value was missing and you filled it in, which always leads to a loss in information.