8.3.1 Data-Cleaning Strategies

Wikipedia doesn't have strict standards on how movie data is presented, so it needs a lot of work to clean up the data and make it usable. Like most web-scraped data, it's in the flexible JSON format to store all kinds of data, but Britta needs to organize it in a structured format before she can send it to SQL—and she's asked you to assist with this task. (You do have experience in this, after all.) First, explore your options for cleaning the dataset.

The transform step is largely spent on data cleaning. There are other transformations that aren't strictly data cleaning, but for the most part, the transformation step is used to clean up your data.

Every messy dataset presents its own unique challenges. There's no one right way to clean data, but we can still have a rough game plan to follow.

Bad data comes in three states:

- Beyond repair
- Badly damaged
- Wrong form

The state of the data largely determines which strategy you should use to clean it.

Data beyond repair could be data that has been overwritten or has suffered severe data corruption during storage or transfer (such as power loss during writing, voltage spikes, or hard-drive failures). The worst-case example would be having data with every value missing. All the information is lost and unrecoverable. For data beyond repair, all we can do is delete it and move on.

Data that is badly damaged may have good data that we can recover, but it will take time and effort to repair the damaged data. This can be garbled data, with a lot of missing values, from inconsistent sources, or existing in

multiple columns. Consider trade-offs to pick the best solution (even if the "best" solution isn't perfect, but rather the "best-available" solution). To repair badly damaged data, try these strategies:

- Filling in missing data by
  - substituting data from another source,
  - interpolating between existing data points, or
  - extrapolating from existing data
- Standardizing units of measure (e.g., monetary values stored in multiple currencies)
- Consolidating data from multiple columns

Finally, **data in the wrong form** should usually be fixed—that is, the data is good but can't be used in its current form. "Good" data in the wrong form can be data that is too granular or detailed, numeric data stored as strings, or data that needs to be split into multiple columns (e.g., address data). To remedy good data in the wrong form, try these strategies:

- Reshape the data
- Convert data types
- Parse text data to the correct format
- Split columns

These options are all available to us, but knowing when to perform which strategy can feel overwhelming. There is no simple checklist or flowchart we can use to guide us, and ultimately, that's a good thing. In data cleaning, we have to constantly ask ourselves what we might have missed, and following a rigid plan means we won't be asking ourselves those important questions. Data cleaning requires a lot of improvising.

IMPORTANT

> **It's important to document your data cleaning assumptions as well as**

> **decisions and their motivations.** Later decisions depend on earlier decisions made, which can be too much to remember. Any assumptions that were part of an earlier decision can, if forgotten, ruin later steps.
>
> Transforming a messy dataset into a clean dataset is an iterative process. As you clean one part of the data, you may reveal something messy in another part of the data. Sometimes that means unwinding a lot of work that you've already done and having to redo it with a slight change. Documenting *why* a particular step is necessary will show you how to redo it without introducing more errors.

We're not completely lost—we do have a strategy. We're not going to try and clean all the data at once. Instead, we're going to focus on one problem at a time using an iterative process.