Data wrangling

A code-free introduction to wrestling your bits and bytes.

Guy Maskall Follow

Jul 15 · 4 min read ★

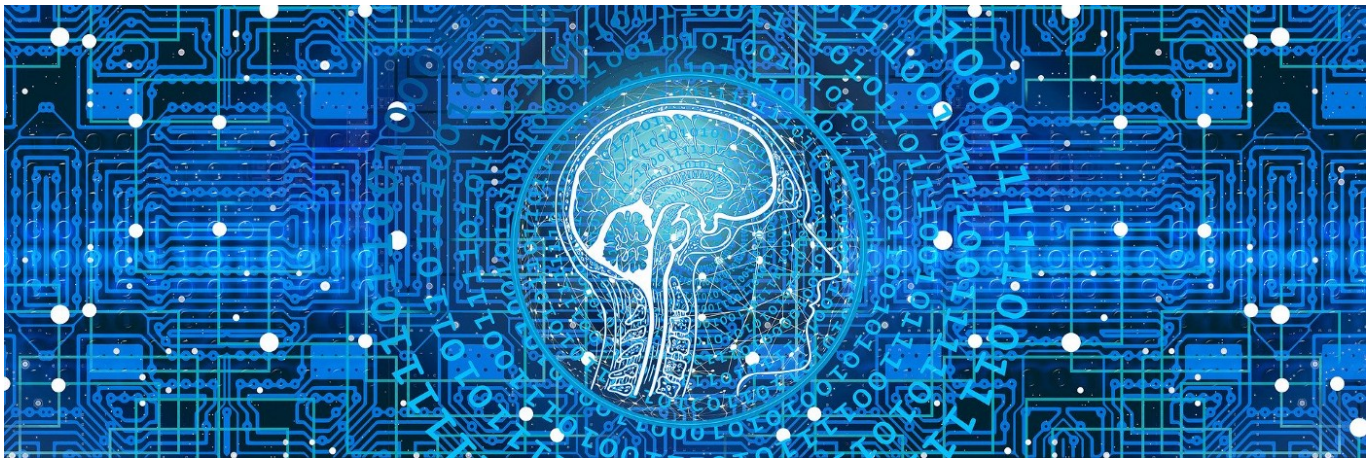

Image by Gerd Altmann from Pixabay

Data acquisition

First of all, you have to get your data! This can involve extricating the data you want from a larger dataset. It can involve merging two or more datasets.

Data shape

A dog isn't just for Christmas, it's for life, right? Well, data wrangling isn't just something you do once in a data science project and never revisit. It's an ongoing activity. Some aspects of an analysis are suited to data in a "long" format, whilst others are more suited to data in a "wide" format. You'll understand these terms more deeply in due course, but for now, we can give an example. Say you have a longitudinal study of a group of people where you've measured their weight over time whilst they were on a dietary regime. Some people were on regime A and some on regime B. Your data might be organized like this:

Subject ID | Date | Regime | Weight

One analysis you were interested in might be to plot each subject's weight change over time. In this case, the above "long" format is suitable and you can immediately plot weight as a function of date. But what if you wanted to calculate the weight change for each subject? You'd like to reformat the above to

Subject ID | Regime | Weight at first date | Weight at last date

and you might want to swap between such formats as your analysis progresses. The take away point here is that no one data format will likely be right throughout all stages of a project. But we highlight data wrangling as a first step because there are some crucial data wrangling steps to perform before you progress onto anything else.

Data quality

Another important aspect of the initial data wrangling is assessing data quality, and improving upon it if possible. A common mistake here is to rush in with some technique to impute (guess) any missing values. This is really bad practice. Imagine there were answers to a question "Have you ever taken illegal drugs?" and there were a number of blanks where people had declined to answer. Most people would likely, truthfully, answer "No". Does it seem reasonable to fill the missing values with "no"? The fact that a subject declined to answer the question would likely be a piece of information you'd want to retain.

If you were working on building a machine learning model to predict in advance when a river would flood and there were missing values in river level sensor data, it would be worth exploring the data before blindly replacing missing values with the average river level. Such analysis may well reveal that the missing values were preceded by a rapid rise in river level up to some extreme value — the value at which the monitoring station itself flooded and ceased operation! Replacing those missing values with the (much lower) average level would not be a great idea!

Yes, some data quality issues can be fixed immediately. Let's say people could enter their sex in freeform 'f', 'F', 'Female', 'female', 'Fem', 'Woman'… Your subsequent exploration would be greatly aided by cleaning these up to a consistent standard first.

An assessment of data quality can also help refine the actual question you're tackling. Let's revisit the dietary regime example above. One data quality check might be to calculate the days between the first and last weight measurement for each subject. What if 80% of the samples came up with a month, and 20% for two months, but you knew the trial actually had lasted for three? Your data quality seems insufficient to answer the question "Did either regime ultimately work better than the other?" but you could adapt the question to answer a still meaningful "Did either regime work better than the other in the first month?"

Summary

The acquisition of data clearly tends to be preloaded at the start of an analysis. But that's not to say you won't decide some extra information would be really helpful and include it later on. The appropriate shape for your data will change throughout a project, so expect wrangling of the data shape to be an ongoing activity. Indeed, grouping and aggregating is a form of data wrangling and affects the shape of the data. And data quality is something important to assess at the start; is your project a non-starter? There are issues that can, and should, be fixed right at the outset, but many should just be flagged, to be addressed in more detail later on. Remember, the fact that something is missing can be important information in itself! Being critical of your data quality is something you should do constantly throughout your work, not just once at the start.

About this article

This is the third article of a linked series written to provide a straightforward introduction to getting started with the data science process. You can find the introduction here, the previous article here, and the next article in the series here.

Data Science