05 Data preparation

Most of us have at some time been abroad in countries where we don't speak the language. With this basic method of communication closed off to us, it can be very difficult to get our meaning across. Even if we do have some knowledge of the language, the gaps in our vocabulary and grammar will often cause us (and the listener) some frustration.

Language, then, is a fundamental necessity if we want to understand and communicate with another individual. And preparing data is all about establishing a common language between human and machine.

In this chapter, we will learn why data should never be analysed without first having been prepared, the step-by-step process in which data can be prepared and the best methods that I have learned for managing problems in datasets.

Encouraging data to talk

As practitioners, unless we are very lucky, data will often come to us 'dirty'. It is often collected by people who do not standardize their records, managed by people who might tamper with their datasets' column and row names to suit their own projects, and stored in non-optimal locations that can cause damage to the data. With so many different people working on a single dataset and using different methods for adding data, the resulting datasets in many organizations are unsurprisingly riddled with errors and gaps. And we cannot expect a machine to know where the errors lie or how to fix inconsistencies in information.

So it is our job to prepare the data in such a way that it can be comprehended – and correctly analysed – by a machine.

With great power comes great responsibility

Data preparation (also known as 'data wrangling') is a complex component of the entire process and, as it comprises a number of tasks that can only be completed manually, this stage normally takes the most amount of time.[1] The reason behind this close attention to data preparation is that, if the raw data is not first structured properly in the dataset, then the later stages of the process will either not work at all or, even worse, will give us inaccurate predictions and/or incorrect results. This can spell disaster for you and your company, and at the very worst end of the scale, neglect of this stage can result in firings and, in the case of freelance work, even lawsuits.

It is not my intention to scare you – I simply mean to show you how essential it is to prepare our data. Surprisingly, despite its importance, this is where I have found a serious gap in the educational materials on data science, which largely focus on the later stages of the process: analysing and visualizing. These books and courses use datasets that have already been prepared. But while this approach may be fine if you're just getting to grips with the discipline, paying no attention to preparation means that you are effectively only learning the cosmetic ways to work with data.

Working only with datasets from educational courses will merely show you data that has been cleaned up to fit the case example. But in the real world, data is often dirty, messy and corrupt, and without knowing the causes and symptoms of dirty data, we cannot adequately complete our project. If you do not prepare your data, when you get out into the real world on your first project, and your algorithm inevitably returns 'missing data' errors, or 'text qualifier' errors, or 'division by zero', your project will grind to a halt.

How do we know, then, that our data has been sufficiently prepared? Quite simply, once we have ensured that it is suitable for our data analysis stage. It must:

- be in the right format;
- be free from errors; and
- have all gaps and anomalies accounted for.

A common phrase that data scientists use is 'garbage in, garbage out', which means if you put unclean data into an algorithm, you will only get nonsensical results, making your analyses useless. It is true that a number of practitioners struggle with this step, but that is only because they don't have a framework to follow. This usually leads to an unstructured,

undocumented approach and means they have to reinvent the wheel every time they prepare their data; an ineffective and time-consuming approach in the long run.

So, let's get started with the process of preparation.

CASE STUDY Ubisoft – making the case for data preparation

Ulf Morys is Finance Director at the German branch of Ubisoft, a game design, development and distribution company that has created popular game franchises from *Assassin's Creed* to *Far Cry*. Ulf oversees a distribution subsidiary, which distributes Ubisoft's video games in the area of Germany, Switzerland and Austria (GSA) and is also responsible for the financial components of the company's operations in central Europe.

Ubisoft had historically had its data used solely by the production team for in-game analytics and monetization. Until Ulf changed matters, finance did not number among its strategic areas for data science.[2] But ignoring the improvements that data science makes can be a costly oversight, and having prior experience of leveraging data to make important business decisions (in his previous job, Ulf saved his company $40 million in a merger thanks to his attention to the data) meant that he knew that a deliberate strategy for using company data was essential.

He says, crucially:

> Preparing data doesn't add more data, it just improves the way you look at it. It's like that scene in the film *The Wizard of Oz* when Dorothy opens the door of her house to the kingdom of Oz and the black and white world of Kansas is replaced with Technicolor. There's nothing fundamentally different to the way the story functions on a technical level, and yet everything has changed. It has been cleaned up.
>
> (SuperDataScience, 2016)

To better understand how Ubisoft in particular benefited from data preparation, Ulf turned to Ubisoft's production team, who had been collecting data from its thousands of online gamers for years, from the length of time the game was played and the length of time that it took to complete individual levels, to what the player did in the game and where they failed on the game map. Ulf found that they were using this past data to assess the likelihood that customers would purchase in-game products through a 'freemium model'.[3] Having the data to hand

not only helped Ubisoft to find out the purchasing patterns of its core customers but also to identify the behaviours that could be applied to future players.

Talking to his team about the grand steps the production team were making thanks to data science turned their heads to the idea. During a financial strategy meeting, Ulf's team mapped out the sources of all Ubisoft's available data and what was missing from the picture, which offered something tangible from which colleagues could bounce off their ideas. 'Very simply,' says Ulf, 'if you don't know something is there, you can't ask questions of it' (SuperDataScience, 2016).

The gaps showed what key data they needed to gather from their customers (store dimensions, space provided for selling video games, attitudes of typical buyers to video games and consumer feelings about Ubisoft's output) before they could make meaningful analyses. Ulf says:

> It was evident why we weren't taking a more systematic approach to our customers: *we didn't have the data*. When I went to our sales department, I was told which of our customers were good based on their own insider knowledge. But that knowledge hadn't been collected systematically. Getting the data – the hard facts – was absolutely necessary.
>
> (SuperDataScience, 2016)

Gathering that information for 2,000 stores enabled Ulf to prepare statistically relevant data that would finally be suitable for analysis. This helped Ubisoft to better target its customers in a way that had not been possible before.

Preparing your data for a journey

In order to get our raw data (information that has not been prepared) to respond to analysis, we first have to prepare it. The method for doing so comprises just three stages:

1. **Extract** the data from its sources;
2. **Transform** the data into a comprehensible language for access in a relational database;
3. **Load** the data into the end source.

This process is known as ETL, and it will help us to gather our data into a suitable format in our end source – known as the 'warehouse' – which can be accessed and analysed in the later stages of the Data Science Process. A data warehouse stores otherwise disparate data in a single system. Oftentimes, it will comprise relational databases.