4

Working With Data

In Chapter 3 the workflow process was initiated by exploring the defining matters around context and vision. The discussion about curiosity, framing not just the subject matter of interest but also a specific enquiry that you are seeking an answer to, in particular leads your thinking towards this second stage of the process: working with data.

In this chapter I will start by covering some of the most salient aspects of data and statistical literacy. This section will be helpful for those readers without any – or at least with no extensive – prior data experience. For those who have more experience and confidence with this topic, maybe through their previous studies, it might merely offer a reminder of some of the things you will need to focus on when working with data on a visualisation project.

There is a lot of hard work that goes into the activities encapsulated by 'working with data'. I have broken these down into four different groups of action, each creating substantial demands on your time:

- Data acquisition: Gathering the raw material.
- Data examination: Identifying physical properties and meaning.
- Data transformation: Enhancing your data through modification and consolidation.
- Data exploration: Using exploratory analysis and research techniques to learn.

You will find that there are overlapping concerns between this chapter and the nature of Chapter 5, where you will establish your editorial thinking. The present chapter generally focuses more on the mechanics of familiarisation with the characteristics and qualities of your data; the next chapter will build on this to shape what you will actually do with it.

As you might expect, the activities covered in this chapter are associated with the assistance of relevant tools and technology. However, the focus for the book will remain concentrated on identifying *which* tasks you have to undertake and look less at exactly *how* you will undertake these. There will be tool-specific references in the curated collection of resources that are published in the digital companion.

4.1 Data Literacy: Love, Fear and Loathing

I frequently come across people in the field who declare their love for data. I don't love data. For me it would be like claiming 'I love food' when, realistically, that would be misleading. I like sprouts but hate carrots. And don't get me started on mushrooms.

At the very start of the book, I mentioned that data might occasionally prove to be a villain in your quest for developing confidence with data visualisation. If data were an animal it would almost certainly be a cat: it has a capacity to earn and merit love but it demands a lot of attention and always seems to be conspiring against you.

I *love* data that gives me something interesting to do analysis-wise and then, subsequently, also visually. Sometimes that just does not happen.

I *love* data that is neatly structured, clean and complete. This rarely exists. Location data will have inconsistent place-name spellings, there will be dates that have a mixture of US and UK formats, and aggregated data that does not let me get to the underlying components.

You don't need to *love* data but, equally, you shouldn't fear data. You should simply respect it by appreciating that it will potentially need lots of care and attention and a shift in your thinking about its role in the creative process. Just look to develop a rapport with it, embracing its role as the absolutely critical raw material of this process, and learn how to nurture its potential.

For some of you reading this book, you might have interest in data but possibly not much knowledge of the specific activities involving data as you work on a visualisation design solution. An assumed prerequisite for anyone working in data visualisation is an appreciation of data and statistical literacy. However, this is not always the case. One of the biggest causes of failure in data visualisations – especially in relation to the principle I introduced about 'trustworthy design' – comes from a poor understanding of these numerate literacies. This can be overcome, though.

'When I first started learning about visualisation, I naively assumed that datasets arrived at your doorstep ready to roll. Begrudgingly I accepted that before you can plot or graph anything, you have to find the data, understand it, evaluate it, clean it, and perhaps restructure it.' **Marcia Gray, Graphic Designer**

I discussed in the Introduction the different entry points from which people doing data visualisation work come. Typically – but absolutely not universally – those who join from the more creative backgrounds of graphic design and development might not be expected to have developed the same level of data and statistical knowledge than somebody from the more numerate disciplines. If you are part of this creative cohort and can identify with this generalisation, then this chapter will ease you through the learning process (and in doing so hopefully dispel any myth that it is especially complicated).

Conversely, many others may think they do not know enough about data but in reality they already do 'get' it – they just need to learn more about its role in visualisation and possibly realign their understanding of some of the terminology. Therefore, before delving further into this chapter's tasks, there are a few 'defining' matters I need to address to cover the basics in both data and statistical literacy.

Data Assets and Tabulation Types

Firstly, let's consider some of the fundamentals about what a dataset is as well as what shape and form it comes in.

When working on a visualisation I generally find there are two main categories of data 'assets': data that exist in tables, known as datasets; and data that exists as isolated values.

> For the purpose of this book I describe this type of data as being raw because it has not yet been statistically or mathematically manipulated and it has not been modified in any other way from its original state.

Tabulated datasets are what we are mainly interested in at this point. Data as isolated values refers to data that exists as individual facts and statistical figures. These do not necessarily belong in, nor are they normally collected in, a table. They are just potentially useful values that are dispersed around the Web or across reports: individual facts or figures that you might come across during your data gathering or research stages. Later on in your work you might use these to inform calculations (e.g. applying a currency conversion) or to incorporate a fact into a title or caption (e.g. 78% of staff participated in the survey), but they are not your main focus for now.

Tabulated data is unquestionably the most common form of data asset that you will work with, but it too can exist in slightly different shapes and sizes. A primary difference lies between what can be termed *normalised* datasets (Figure 4.1) and *cross-tabulated* datasets (Figure 4.2).

A normalised dataset might loosely be described as looking like lists of data values. In spreadsheet parlance, you would see this as a series of columns and rows of data, while in database parlance it is the arrangement of fields and records. This form of tabulated data is generally the most detailed form of data available for you to work with. The table in Figure 4.1 is an example of normalised data where the columns of variables provide different descriptive values for each movie (or record) held in the table.

CATEGORY	MOVIE TITLE	CRITIC RATING	REVIEW GROUP
Star Wars	Star Wars: Episode IV - A New Hope	8.3	Fresh
Star Wars	Star Wars: Episode V - The Empire Strikes Back	8.7	Fresh
Star Wars	Star Wars: Episode VI - Return of the Jedi	6.8	Fresh
Star Wars	Star Wars: Episode I - The Phantom Menace	5.8	Rotten
Star Wars	Star Wars: Episode II - Attack of the Clones 3D	6.6	Fresh
Star Wars	Star Wars: Episode III - Revenge of the Sith 3D	7.2	Fresh
X-Men	X-Men	7.0	Fresh
X-Men	X2: X-Men United	7.4	Fresh
X-Men	X-Men: The Last Stand	5.9	Rotten
X-Men	X-Men Origins - Wolverine	5.1	Rotten
X-Men	X-Men: First Class	7.4	Fresh
X-Men	The Wolverine	6.3	Fresh
X-Men	X-Men: Days of Future Past	7.6	Fresh
Tolkien	The Lord of the Rings: The Fellowship of the Ring	8.2	Fresh
Tolkien	The Lord of the Rings: The Two Towers	8.5	Fresh
Tolkien	The Lord of the Rings: The Return of the King	8.7	Fresh
Tolkien	The Hobbit: An Unexpected Journey	6.6	Fresh
Tolkien	The Hobbit: The Desolation of Smaug	6.8	Fresh
Tolkien	The Hobbit: The Battle of the Five Armies	6.3	Fresh

Figure 4.1 Example of a Normalised Dataset

Cross-tabulated data is presented in a reconfigured form where, instead of displaying raw data values, the table of cells contain the results of statistical operations (like summed totals, maximums, averages). These values are aggregated calculations formed from the relationship between two variables held in the normalised form of the data. In Figure 4.2, you will see the cross-tabulated result of the normalised table of movie data, now showing a statistical summary for each movie category. The statistic under 'Max Critic Rating' is formed from an aggregating calculation based on the 'Critic Rating' and 'Category' variables seen in Figure 4.1.

Figure 4.2
Example of a Cross-tabulated Dataset

CATEGORY	MOVIES	MAX CRITIC RATING	REVIEW: FRESH	REVIEW: ROTTEN
Star Wars	6	8.7	5	1
Tolkien	6	8.7	6	0
X-Men	7	7.6	5	2
SUMMARY	19	8.7	16	3

Typically, if you receive data in an already cross-tabulated form, you do not have access to the original data. This means you will not be able to 'reverse-engineer' it back into its raw form, which, in turn, means you have reduced the scope of your potential analysis. In contrast, normalised data gives you complete freedom to explore, manipulate and aggregate across multiple dimensions. You may choose to convert the data into 'cross-tabulated' form but that is merely an option that comes with the luxury of having access to the detailed form of your data. In summary, it is always preferable, where possible, to work with normalised data.

Data Types

One of the key parts of the design process concerns understanding the different types of data (sometimes known as *levels of data* or *scales of measurement*). Defining the types of data will have a huge influence on so many aspects of this workflow, such as determining:

- the type of exploratory data analysis you can undertake;
- the editorial thinking you establish;
- the specific chart types you might use;
- the colour choices and layout decisions around composition.

In the simplest sense, data types are distinguished by being either qualitative or quantitative in nature. Beneath this distinction there are several further separations that need to be understood. The most useful taxonomy I have found to describe these different types of data is based on an approach devised by the psychologist researcher Stanley Stevens. He developed the acronym NOIR as a mnemonic device to cover the different types of data you may come to work with, particularly in social research: Nominal, Ordinal, Interval, and Ratio. I have extended this, adding onto the front a 'T' – for Textual – which, admittedly, somewhat undermines the grace of the original acronym but better reflects the experiences of handling data today. It is important to describe, define and compare these different types of data.

Textual (Qualitative)

Textual data is qualitative data and generally exists as unstructured streams of words. Examples of textual data might include:

- 'Any other comments?' data submitted in a survey.
- Descriptive details of a weather forecast for a given city.
- The full title of an academic research project.

- The description of a product on Amazon.
- The URL of an image of Usain Bolt's victory in the 100m at the 2012 Olympics.

In its native form, textual data is likely to offer rich potential but it can prove quite demanding to unlock this. To work with textual data in an analysis and visualisation context will generally require certain natural language processing techniques to derive or extract classifications, sentiments, quantitative properties and relational characteristics.

An example of how you can use textual data is seen in the graphic of CEO swear word usage shown in Figure 4.3. This analysis provides a breakdown of the profanities used by CEOs from a review of recorded conference calls over a period of 10 years. This work shows the two ways of utilising textual data in visualisation. Firstly, you can derive categorical classifications and quantitative measurements to count the use of certain words compared to others and track their usage over time. Secondly, the original form of the textual data can be of direct value for annotation purposes, without the need for any analytical treatment, to include as captions.

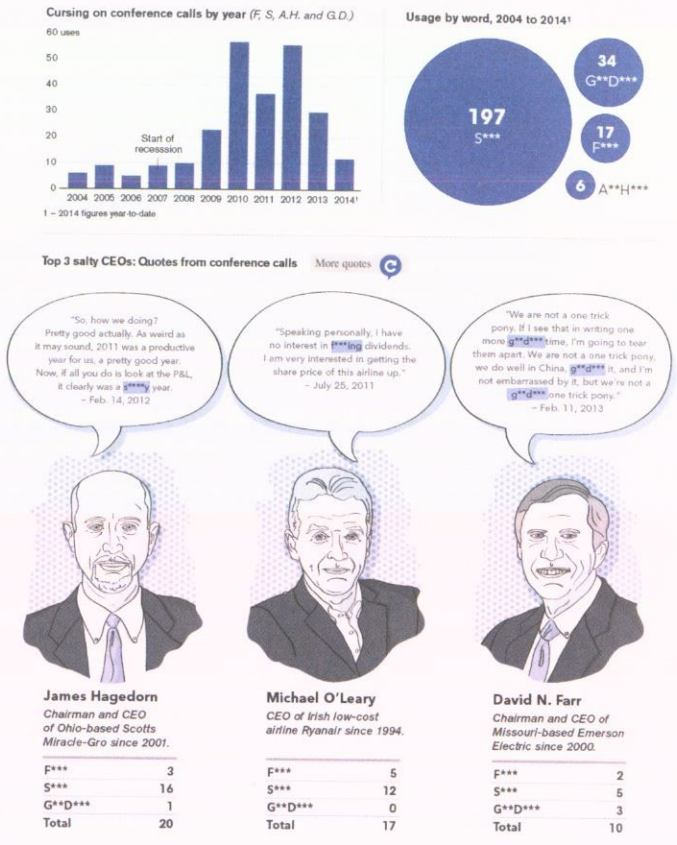


Figure 4.3 Graphic Language: The Curse of the CEO

Working with textual data will always involve a judgement of reward vs effort: how much effort will I need to expend in order to extract usable, valuable content from the text? There are an increasing array of tools and algorithmic techniques to help with this transformational approach but whether you conduct it manually or with some degree of automation it can be quite a significant undertaking. However, the value of the insights you are able to extract may entirely justify the commitment. As ever, your judgment of the aims of your work, the nature of your subject and the interests of your audience will influence your decision.

Nominal (Qualitative)

Nominal data is the next form of qualitative data in the list of distinct data types. This type of data exists in categorical form, offering a means of distinguishing, labelling and organising values. Examples of nominal data might include:

- The 'gender' selected by a survey participant.
- The regional identifier (location name) shown in a weather forecast.
- The university department of an academic member of staff.
- The language of a book on Amazon.
- An athletic event at the Olympics.

Often a dataset will hold multiple nominal variables, maybe offering different organising and naming perspectives, for example the gender, eye colour and hair colour of a class of school kids.

Additionally, there might be a hierarchical relationship existing between two or more nominal variables, representing major and sub-categorical values: for example, a major category holding details of 'Country' and a sub-category holding 'Airport'; or a major category holding details of 'Industry' and a sub-category holding details of 'Company Names'. Recognising this type of relationship will become important when considering the options for which angles of analysis you might decide to focus on and how you may portray them visually using certain chart types.

Nominal data does not necessarily mean text-based data; nominal values can be numeric. For example, a student ID number is a categorical device used uniquely to identify all students. The shirt number of a footballer is a way of helping teammates, spectators and officials to recognise each player. It is important to be aware of occasions when any categorical values are shown as numbers in your data, especially in order to understand that these cannot have (meaningful) arithmetic operations applied to them. You might find logic statements like TRUE or FALSE stated as a 1 and a 0, or data captured about gender may exist as a 1 (male), 2 (female) and 3 (other), but these numeric values should not be considered quantitative values – adding '1' to '2' does not equal '3' (other) for gender.

Ordinal (Qualitative)

Ordinal data is still categorical and qualitative in nature but, instead of there being an arbitrary relationship between the categorical values, there are now characteristics of order. Examples of nominal data might include:

- The response to a survey question: based on a scale of 1 (unhappy) to 5 (very happy).
- The general weather forecast: expressed as Very Hot, Hot, Mild, Cold, Freezing.
- The academic rank of a member of staff.
- The delivery options for an Amazon order: Express, Next Day, Super Saver.
- The medal category for an athletic event: Gold, Silver, Bronze.

Whereas nominal data is a categorical device to help distinguish values, ordinal data is also a means of classifying values, usually in some kind of ranking. The hierarchical order of some ordinal values goes through a single ascending/descending rank from high or good values to low or bad values. Other ordinal values have a natural 'pivot' where the direction changes around a recognisable mid-point, such as the happiness scale which might pivot about 'no

feeling' or weather forecast data that pivots about 'Mild'. Awareness of these different approaches to 'order' will become relevant when you reach the design stages involving the classifying of data through colour scales.

Interval (Quantitative)

Interval data is the less common form of quantitative data, but it is still important to be aware of and to understand its unique characteristics. An interval variable is a quantitative and numeric measurement defined by difference on a scale but *not* by relative scale. This means the difference between two values is meaningful but an arithmetic operation such as multiplication is not.

The most common example is the measure for temperature in a weather forecast, presented in units of Celsius. The absolute difference between 15°C and 20°C is the same difference as between 5°C and 10°C. However, the relative difference between 5°C and 10°C is not the same as the difference between 10°C and 20°C (where in both cases you multiply by two or increase by 100%). This is because a zero value is arbitrary and often means very little or indeed is impossible. A temperature reading of 0°C does not mean there is no temperature, it is a quantitative scale for measuring relative temperature. You cannot have a shoe size or Body Mass Index of zero.

Ratio (Quantitative)

Ratio data is the most common quantitative variable you are likely to come across. It comprises numeric measurements that have properties of difference *and* scale. Examples of nominal data might include:

- The age of a survey participant in years.
- The forecasted amount of rainfall in millimetres.
- The estimated budget for a research grant proposal in GBP (£).
- The number of sales of a book on Amazon.
- The distance of the winning long jump at the 2012 Olympics in metres.

Unlike interval data, for ratio data variables zero means something. The absolute difference in age between a 10 and 20 year old is the same as the difference between a 40 and 50 year old. The relative difference between a 10 and a 20 year old is the same as the difference between a 40 and an 80 year old ('twice as old').

Whereas most of the quantitative measurements you will deal with are based on a linear scale, there are exceptions. Variables about the strength of sound (decibels) and magnitude of earthquakes (Richter) are actually based on a logarithmic scale. An earthquake with a magnitude of 4.0 on the Richter scale is 1000 times stronger based on the amount of energy released than an earthquake of magnitude

If temperature values were measured in kelvin, where there is an absolute zero, this would be considered a ratio scale, not an interval one.

2.0. Some consider these as types of data that are different from ratio variables. Most still define them as ratio variables but separate them as non-linear scaled variables.

Temporal Data

Time-based data is worth mentioning separately because it can be a frustrating type of data to deal with, especially in attempting to define its place within the TNOIR classification. The reason for this is that different components of time can be positioned against almost all data types, depending simply on what form your time data takes:

Textual: 'Four o'clock in the afternoon on Monday, 12 March 2016'

Ordinal: 'PM', 'Afternoon', 'March', 'Q1'

Interval: '12', '12/03/2016', '2016'

Ratio: '16:00'

Note that time-based data is separate in concern to duration data, which, while often formatted in structures such as hh:mm:ss, should be seen as a ratio measure. To work with duration data it is often useful to transform it into single units of time, such as total seconds or minutes.

Discrete vs Continuous

Another important distinction to make about your data, and something that cuts across the TNOIR classification, is whether the data is discrete or continuous. This distinction is influential in how you might analyse it statistically and visually.

The relatively simple explanation is that discrete data is associated with all classifying variables that have no 'in-between' state. This applies to all qualitative data types and any quantitative values for which only a whole is possible. Examples might be:

- Heads or tails for a coin toss.
- Days of the week.
- The size of shoes.
- Numbers of seats in a theatre.

In contrast, continuous variables can hold the value of an in-between state and, in theory, could take on any value between the natural upper and lower limits if it was possible to take measurements in fine degrees of detail, such as:

- Height and weight.
- Temperature.
- Time.

One of the classifications that is hard to nail down involves data that could, on the TNOIR scale, arguably fall under both ordinal and ratio definitions based on its usage. This makes it hard to determine if it should be considered discrete or continuous. An example would be the star system used for rating a movie or the happiness rating. When a star rating value is originally captured, the likelihood is that the input data was discrete in nature. However, for analysis purposes, the statistical operations applied to data that is based on different star ratings could reasonably be treated either as discrete classifications or, feasibly, as continuous numeric values. For both star review ratings or happiness ratings decimal averages could be calculated as a way of formulating average score. (The median and mode would still be discrete.) The suitability of this approach will depend on whether the absolute difference between classifying values can be considered equal.

4.2 Statistical Literacy

If the fear of data is misplaced, I can sympathise with anybody's trepidation towards statistics. For many, statistics can feel complicated to understand and too difficult a prospect to master. Even for those relatively comfortable with stats, it is unquestionably a discipline that can easily become rusty without practice, which can also undermine your confidence. Furthermore, the fear of making mistakes with delicate and rule-based statistical calculations also depresses the confidence levels lower than they need to be.

The problem is that you cannot avoid the need to use *some* statistical techniques if you are going to work with data. It is therefore important to better understand statistics and its role in visualisation, as you must do with data. Perhaps you can make the problem more surmountable by packaging the whole of statistics into smaller, manageable elements that will dispel the perception of overwhelming complexity.

I do believe that it is possible to overstate the range and level of statistical techniques *most* people will need to employ on *most* of their visualisation tasks. The caveats are important as I know there will be people with visualisation experience who are exposed to a tremendous amount of statistical thinking in their work, but it is a relevant point.

It all depends, of course. From my experience, however, the majority of data visualisation challenges will generally involve relatively straightforward *univariate* and *multivariate* statistical techniques. Univariate techniques help you to understand the shape, size and range of quantitative values. Multivariate techniques help you to explore the possible relationships between different combinations of variables and variable types. I will describe some of the most relevant statistical operations associated with these techniques later in this chapter, at the point in your thinking where they are most applicable.

As you get more advanced in your work (and your confidence increases) you might have occasion to employ *inference* techniques. These include concepts such as data modelling and the use of regression analysis: attempting to measure the relationships between variables to explore correlations and (the holy grail) causations. Many of you will likely experience visualisation

challenges that require an understanding of probabilities, testing hypotheses and becoming acquainted with terms like confidence intervals. You might use these techniques to assist with forecasting or modelling risk and uncertainty. Above and beyond that, you are moving towards more advanced statistical modelling and algorithm design.

It is somewhat dissatisfactory to allocate only a small part of this text to discussing the role of descriptive and exploratory statistics. However, for the scope of this book, and seeking to achieve a pragmatic balance, the most sensible compromise is just to flag up which statistical activities you might need to consider and where these apply. It can take years to learn about the myriad advanced techniques that exist and it takes experience to know when and how to deploy all the different methods.

There are hundreds of books better placed to offer the depth of detail you truly need to fulfil these activities and there is no real need to reinvent the wheel – and indeed reinvent an inferior wheel. That statistics is just one part of the visualisation challenge, and is in itself such a prolific field, further demonstrates the variety and depth of this subject.

4.3 Data Acquisition

The first step in working with data naturally involves getting it. As I outlined in the contextual discussion about the different types of trigger curiosities, you will only have data in place before now if the opportunity presented by the data was the factor that triggered this work. You will recall this scenario was described as pursuing a curiosity born out of 'potential intrigue'. Otherwise, you will only be in a position to know what data you need after having established your specific or general motivating curiosity. In these situations, once you have sufficiently progressed your thinking around 'formulating your brief', you will need to switch your thinking onto the task of acquiring your data:

- What data do you need and why?
- From where, how, and by whom will the data be acquired?
- When can you obtain it?

What Data Do You Need?

Your primary concern is to ensure you can gather sufficient data about the subject in which you are interested to pursue your identified curiosity. By 'sufficient', I mean you will need to establish some general criteria in your mind for what data you do need and what data you do not need. There is no harm in getting more than you need at this stage but it can result in wasted efforts, waste that you would do well to avoid.

Let's propose you have defined your curiosity to be 'I wonder what a map of McDonald's restaurant openings looks like over time?'. In this scenario you are going to try to find a source of data that will provide you with details of all the McDonald's restaurants that have ever opened. A shopping list of data items would probably include the date of opening, the location details

(as specific as possible) and maybe even a closing date to ensure you can distinguish between still operating and closed-down restaurants.

You will need to conduct some research, a perpetual strand of activity that runs throughout the workflow, as I explained earlier. In this scenario you might need first to research a bit of the history of McDonald's restaurants to discover, for instance, when the first one opened, how many there are, and in which countries they are located. This will establish an initial sense of the timeframe (number of years) and scale (outlets, global spread) of your potential data. You might also discover significant differences between what is considered a restaurant and what is just a franchise positioned in shopping malls or transit hubs. Sensitivities around the qualifying criteria or general counting rules of a subject are important to discover, as they will help significantly to substantiate the integrity and accuracy of your work.

Unless you know or have been told where to find this restaurant data, you will then need to research from where the data might be obtainable. Will this type of information be published on the Web, perhaps on the commercial pages of McDonald's own site? You might have to get in touch with somebody (yes, a human) in the commercial or PR department to access some advice. Perhaps there will be some fast-food enthusiast in some niche corner of the Web who has already gathered and made available data like this?

Suppose you locate a dataset that includes not just McDonald's restaurants but *all* fast-food outlets. This could potentially broaden the scope of your curiosity, enabling broader analysis about the growth of the fast-food industry at large to contextualise MacDonald's contribution to this. Naturally, if you have any stakeholders involved in your project, you might need to discuss with them the merits of this wider perspective.

Another judgement to make concerns the resolution of the data you anticipate needing. This is especially relevant if you are working with big, heavy datasets. You might genuinely want and need all available data. This would be considered full resolution – down to the most detailed grain (e.g. all details about all MacDonald's restaurants, not just totals per city or country). Sometimes, in this initial gathering activity, it may be more practical just to obtain a sample of your data. If this is the case, what will be the criteria used to identify a sufficient sample and how will you select or exclude records? What percentage of your data will be sufficient to be representative of the range and diversity (an important feature we will need to examine next)? Perhaps you only need a statistical, high-level summary (total number of restaurants opened by year)?

The chances are that you will not truly know what data you want or need until you at least get something to start with and learn from there. You might have to revisit or repeat the gathering of your data, so an attitude of 'what I have is good enough to start with' is often sensible.

From Where, How and By Whom Will the Data Be Acquired?

There are several different origins and methods involved in acquiring data, depending on whether it will involve your doing the heavy work to curate the data or if this will be the main responsibility of others.

Curated by You

This group of data-gathering tasks or methods is characterised by your having to do most of the work to bring the data together into a convenient digital form.

Primary data collection: If the data you need does not exist or you need to have full control over its provenance and collection, you will have to consider embarking on gathering 'primary' data. In contrast to secondary data, primary data involves you measuring and collecting the raw data yourself. Typically, this relates to situations where you gather quite small, bespoke datasets about phenomena that are specific to your needs. It might be a research experiment you have designed and launched for participants to submit responses. You may manually record data from other measurement devices, such as your daily weight as measured by your bathroom scales, or the number of times you interacted face-to-face with friends and family. Some people take daily photographs of themselves, their family members or their gardens, in order to stitch these back together eventually to portray stories of change. This data-gathering activity can be expensive in terms of both the time and cost. The benefit however is that you have carefully controlled the collection of the data to optimise its value for your needs.

Manual collection and data foraging: If the data you need does not exist digitally or in a convenient singular location, you will need to *forage* for it. This again might typically relate to situations where you are sourcing relatively small datasets. An example might be researching historical data from archived newspapers that were only published in print form and not available digitally. You might look to pull data from multiple sources to create a single dataset: for example, if you were comparing the attributes of a range of different cars and weighing up which to buy. To achieve this you would probably need to source different parts of the data you need from several different places. Often, data foraging is something you undertake in order to finish off data collected by other means that might have a few missing values. It is sometimes more efficient to find the remaining data items yourself by hand to complete the dataset. This can be somewhat time-consuming depending on the extent of the manual gathering required, but it does provide you with greater assurance over the final condition of the data you have collected.

> Some data acquisition tasks may be repetitive and, should you possess the skills and have access to the necessary resources, there will be scope for exploring ways to automate these. However, you always have to consider the respective effort and ongoing worth of your approach. If you do go to the trouble of authoring an automation routine (of any description) you could end up spending more time on that than you would otherwise collecting by more manual methods. If it is going to be a regular piece of analysis the efficiency gains from your automation will unquestionably prove valuable going forward, but, for any one-off projects, it may not be ultimately worth it

Extracted from pdf files: A special subset of data foraging – or a variation at least – involves those occasions when your data is digital but essentially locked away in a pdf file. For many years now reports containing valuable data have been published on the Web in pdf form. Increasingly, movements like 'open data' are helping to shift the attitudes of organisations towards providing additional, fully accessible digital versions of data. Progress is being made but it will take time before all

industries and government bodies adopt this as a common standard. In the meantime, there are several tools on the market (free and proprietary) that will assist you in extracting tables of data from pdf files and converting these to more usable Excel or CSV formats.

Web scraping (also known as web harvesting): This involves using special tools or programs to extract structured and unstructured items of data published in web pages and convert these into tabulated form for analysis. For example, you may wish to extract several years' worth of test cricket results from a sports website. Depending on the tools used, you can often set routines in motion to extract data across multiple pages of a site based on the connected links that exist within it. This is known as web crawling. Using the same example (let's imagine), you could further your gathering of test cricket data by programmatically fetching data back from the associated links pointing to the team line-ups. An important consideration to bear in mind with any web scraping or crawling activity concerns rules of access and the legalities of extracting the data held on certain sites. Always check – and respect – the terms of use before undertaking this.

Curated by Others

In contrast to the list of methods I have profiled, this next set of data-gathering approaches is characterised by other people having done most of the work to source and compile the data. They will make it available for you to access in different ways without needing the extent of manual efforts often required with the methods presented already. You might occasionally still have to intervene by hand to fine-tune your data, but others would generally have put in the core effort.

Issued to you: On the occasions when you are commissioned by a stakeholder (client, colleague) you will often be provided with the data you need (and probably much more besides), most commonly in a spreadsheet format. The main task for you is therefore less about collection and more about familiarisation with the contents of the data file(s) you are set to work with.

Download from the Web: Earlier I bemoaned the fact that there are still organisations publishing data (through, for example, annual reports) in pdf form. To be fair, increasingly there are facilities being developed that enable interested users to extract data in a more structured form. More sophisticated reporting interfaces may offer users the opportunity to construct detailed queries to extract and download data that is highly customised to their needs.

System report or export: This is related more to an internal context in organisations where there are opportunities to extract data from corporate systems and databases. You might, for example, wish to conduct some analysis about staff costs and so the personnel database may be where you can access the data about the workforce and their salaries.

Third-party services: There is an ever-increasing marketplace for data and many

'Don't underestimate the importance of domain expertise. At the Office for National Statistics (ONS), I was lucky in that I was very often working with the people who created the data – obviously, not everyone will have that luxury. But most credible data producers will now produce something to accompany the data they publish and help users interpret it – make sure you read it, as it will often include key findings as well as notes on reliability and limitations of the data.' **Alan Smith OBE, Data Visualisation Editor, *Financial Times***

commercial services out there now offer extensive sources of curated and customised data that would otherwise be impossible to obtain or very complex to gather. Such requests might include very large, customised extracts from social media platforms like Twitter based on specific keywords and geo-locations.

API: An API (Application Programme Interface) offers the means to create applications that programmatically access streams of data from sites or services, such as accessing a live feed from Transport for London (TfL) to track the current status of trains on the London Underground system.

When Can the Data Be Acquired?

The issue of *when* data is ready and available for acquisition is a delicate one. If you are conducting analysis of some survey results, naturally you will not have the full dataset of responses to work with until the survey is closed. However, you could reasonably begin some of your analysis work early by using an initial sample of what had been submitted so far. Ideally you will always work with data that is as complete as possible, but on occasions it may be advantageous to take the opportunity to get an early sense of the nature of the submitted responses in order to begin preparing your final analysis routines. Working on any dataset that may not yet be complete is a risk. You do not want to progress too far ahead with your visualisation workflow if there is the real prospect that any further data that emerges could offer new insights or even trigger different, more interesting curiosities.

4.4 Data Examination

After acquiring your data your next step is to thoroughly examine it. As I have remarked, your data is your key raw material from which the eventual visualisation output will be formed. Before you choose what meal to cook, you need to know what ingredients you have and what you need to do to prepare them.

It may be that, in the act of acquiring the data, you have already achieved a certain degree of familiarity about its status, characteristics and qualities, especially if you curated the data yourself. However, there is a definite need to go much further than you have likely achieved before now. To do this you need to conduct an examination of the physical properties and the meaning of your data.

As you progress through the stages of this workflow, your data will likely change considerably: you will bring more of it in, you will remove some of it, and you will refine it to suit your needs. All these modifications will alter the physical makeup of your data so you will need to keep revisiting this step to preserve your critical familiarity.

Data Properties

The first part of familiarising yourself with with your data is to undertake an examination of its physical properties. Specifically you need to ascertain its type, size and condition. This task is quite mechanical in many ways because you are in effect just 'looking' at the data, establishing its surface characteristics through visual and/or statistical observations.

What To Look For?

The *type* and *size* of your data involve assessing the characteristics and amount of data you have to work with. As you examine the data you also need to determine its *condition*: how good is its quality and is it fit for purpose?

Data types: Firstly, you need to identify what data types you have. In gathering this data in the first place you might already have a solid appreciation about what you have before you, but doing this thoroughly helps to establish the attention to detail you will need to demonstrate throughout this stage. Here you will need to refer to the definitions from earlier in the chapter about the different types of data (TNOIR). Specifically you are looking to define each column or field of data based on whether it is qualitative (text, nominal, ordinal) or quantitative (interval, ratio) and whether it is discrete or continuous in nature.

Size: Within each column or field you next need to know what range of values exist and what are the specific attributes/formats of the values held. For example, if you have a quantitative variable (interval or ratio), what is the lowest and the highest value? In what number format is it presented (i.e. how many decimal points or comma formatted)? If it is a categorical variable (nominal or ordinal), how many different values are held? If you have textual data, what is the maximum character length or word count?

Condition: This is the best moment to identify any data quality and completeness issues. Naturally, unidentified and unresolved issues around data quality will come to bite hard later, undermining the scope and, crucially, trust in the accuracy of your work. You will address these issues next in the 'transformation' step, but for now the focus is on identifying any problems. Things to look out for may include the following:

- Missing values, records or variables – Are empty cells assumed as being of no value (zero/nothing) or no measurement (n/a, null)? This is a subtle but important difference.
- Erroneous values – Typos and any value that clearly looks out of place (such as a gender value in the age column).
- Inconsistencies – Capitalisation, units of measurement, value formatting.
- Duplicate records.
- Out of date – Values that might have expired in accuracy, like someone's age or any statistic that would be reasonably expected to have subsequently changed.
- Uncommon system characters or line breaks.
- Leading or trailing spaces – the invisible evil!
- Date issues around format (dd/mm/yy or mm/dd/yy) and basis (systems like Excel's base dates on daily counts since 1 January 1900, but not all do that).

How to Approach This?

I explained in the earlier 'Data literacy' section the difference in asset types (data that exists in tables and data that exists as isolated values) and also the difference in form (normalised data or cross-tabulated). Depending on the asset and form of data, your examination of data types may involve slightly different approaches, but the general task is the same. Performing this examination process will vary, though, based on the tools you are using. The simplest approach, relevant to most, is to

describe the task as you would undertake it using Excel, given that this continues to be the common tool most people use or have the skills to use. Also, it is likely that most visualisation tasks you undertake will involve data of a size that can be comfortably handled in Excel.

As you go through this task, it is good practice to note down a detailed overview of what data you have, perhaps in the form of a table of data descriptions. This is not as technical a duty as would be associated with the creation of a data dictionary but its role and value are similar, offering a convenient means to capture all the descriptive properties of your various data assets.

'Data inspires me. I always open the data in its native format and look at the raw data just to get the lay of the land. It's much like looking at a map to begin a journey.' **Kim Rees, Co-founder, Periscopic**

Inspect and scan: Your first task is just to scan your table of data visually. Navigate around it using the mouse/trackpad, use the arrow keys to move up or down and left or right, and just look at all the data. Gain a sense of its overall dimension. How many columns and how many rows does it occupy? How big a prospect might working with this be?

Data operations: Inspecting your data more closely might require the use of interrogation features such as sorting columns and doing basic filters. This can be a quick and simple way to acquaint yourself with the type of data and range of values.

Going further, once again depending on the technology (and assuming you have normalised data to start with), you might apply a cross-tabulation or pivot table to create aggregated, summary views of different angles and combinations of your data. This can be a useful approach to also check out the unique range of values that exist under different categories as well as helping to establish how sub-categories may relate other categories hierarchically. This type of inspection will be furthered in the next step of the 'working with data' process when you will undertake deeper visual interrogations of the type, size and condition of your data.

If you have multiple tables, you will need to repeat this approach for each one as well as determine how they are related collectively and on what basis. It could be that just considering one table as the standard template, representative of each instance, is sufficient: for example, if each subsequent table is just a different monthly view of the same activity.

For so-called 'Big Data' (see the glossary definition earlier), it is less likely that you can conduct this examination work through relatively quick, visual observations using Excel. Instead it will need tools based around statistical language that will *describe* for you what is there rather than let you *look* at what is there.

Statistical methods: The role of statistics in this examination stage generally involves relatively basic quantitative analysis methods to help describe and understand the characteristics of each data variable. The common term applied to this type of statistical approach is univariate, because it involves just looking at one variable at a time (the best opportunity to perform the analysis of multiple variables comes later). Here are some different types of statistical analyses you might find useful at this stage. These are not the only methods you will ever need to use, but will likely prove to be among the most common:

- *Frequency counts*: applied to categorical values to understand the frequency of different instances.
- *Frequency distribution*: applied to quantitative values to learn about the type and shape of the distribution of values.

- Measurements of *central tendency* describe the summary attributes of a group of quantitative values, including:
  - the mean (the average value);
  - the median (the middle value if all quantities were arranged from smallest to largest);
  - the mode (the most common value).
- Measurements of *spread* are used to describe the dispersion of values above and below the mean:
  - Maximum, minimum and range: the highest and lowest and magnitude of spread of values.
  - Percentiles: the value below which *x*% of values fall (e.g. the 20th percentile is the value below which 20% of all quantitative values fall).
  - Standard deviation: a calculated measure used to determine how spread out a series of quantitative values are.

Data Meaning

Irrespective of whether you or others have curated the data, you need to be discerning about how much trust you place in it, at least to begin with. As discussed in the 'trustworthy design' principle, there are provenance issues, inaccuracies and biases that will affect its status on the journey from being created to being acquired. These are matters you need to be concerned with in order to resolve or at least compensate for potential shortcomings.

Knowing more about the physical properties of your data does not yet achieve full familiarity with its content nor give you sufficient acquaintance with its qualities. You will have examined the data in a largely mechanical and probably quite detached way from the underlying subject matter. You now need to think a little deeper about its meaning, specifically what it does – and does not – truly represent.

'A visualization is always a model (authored), never a mould (replica), of the real. That's a huge responsibility.' **Paolo Ciuccarelli, Scientific Director of DensityDesign Research Lab at Politecnico di Milano**

What Phenomenon?

Determining the meaning of your data requires that you recognise this is more than just a bunch of numbers and text values held in the cells of a table. Ask yourself, 'What is it about? What activity, entity, instance or phenomenon does it represent?'.

One of the most valuable pieces of advice I have seen regarding this task came from Kim Rees, co-founder of Periscopic. Kim describes the process of taking one single row of data and using that as an entry point to learn carefully about what each value means individually and then collectively. Breaking down the separation between values created by the table's cells, and then sticking the pieces back together, helps you appreciate the parts and the whole far better.

'Absorb the data. Read it, re-read it, read it backwards and understand the lyrical and human-centred contribution.' **Kate McLean, Smellscape Mapper and Senior Lecturer Graphic Design**

You saw the various macro- and micro-level views applied to the context of the Texas Department for Criminal Justice executed offenders information in the previous chapter. The underlying meaning of this data – its phenomenon – was offenders who had been judged guilty of committing heinous crimes and had faced the ultimate consequence. The availability of textual data describing the offenders' last statements and details of their crimes heightened the emotive potential of this data. It was heavy stuff. However, it was still just a collection of values detailing dates, names, locations, categories. All datasets, whether on executed offenders or the locations of MacDonald's restaurants, share the same properties as outlined by the TNOIR data-type mnemonic. What distinguishes them is what these values mean.

What you are developing here is a more semantic appreciation of your data to substantiate the physical definitions. You are then taking that collective appreciation of what your data stands for to influence how you might decide to amplify or suppress the influence of this semantic meaning. This builds on the discussion in the last chapter about the tonal dimension, specifically the difference between figurative and non-figurative portrayals.

A bar chart (Figure 4.4) comprising two bars, one of height 43 and the other of height 1, arguably does not quite encapsulate the emotive significance of Barack Obama becoming the first black US president, succeeding the 43 white presidents who served before him. Perhaps a more potent approach may be to present a chronological display of 44 photographs of each president in order to visually contrast Mr Obama's headshot in the final image in the sequence with the previous 43. Essentially, the value of 43 is almost irrelevant in its detail – it could be 25 or 55 – it is about there being 'many' of the same thing followed by the 'one' that is'different'. That's what creates the impact. (What will image number 45 bring? A further striking 'difference' or a return to the standard mould?)

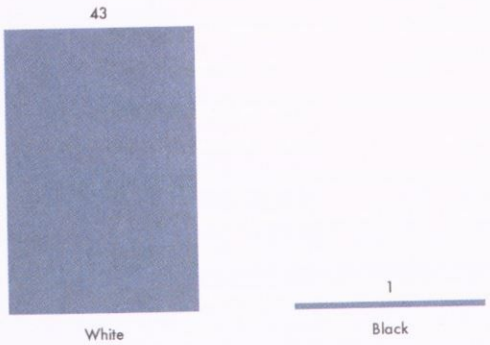


Figure 4.4 US Presidents by Ethnicity (1789 to 2015)

Learning about the underlying phenomena of your data helps you feel its spirit more strongly than just looking at the rather agnostic physical properties. It also helps you in knowing what potential sits inside the data – the qualities it possesses – so you are then equipped the best understanding of how you might want to portray it. Likewise it prepares you for the level of responsibility and potential sensitivity you will face in curating a visual representation of *this* subject matter. As you saw with the case study of the 'Florida Gun Crimes' graphic, some subjects are inherently more emotive than others, so we have to demonstrate a certain amount of courage and conviction in deciding how to undertake such challenges.

'Find loveliness in the unlovely. That is my guiding principle. Often, topics are disturbing or difficult; inherently ugly. But if they are illustrated elegantly there is a special sort of beauty in the truthful communication of something. Secondly, Kirk Goldsberry stresses that data visualization should ultimately be true to a phenomenon, rather than a technique or the format of data. This has had a huge impact on how I think about the creative process and its results.' **John Nelson, Cartographer**

Completeness

Another aspect of examining the meaning of data is to determine how representative it is. I have touched on data quality already, but inaccuracies in conclusions about what data is saying have arguably a greater impact on trust and are more damaging than any individual missing elements of data.

The questions you need to ask of your data are: does it represent genuine observations about a given phenomenon or is it influenced by the collection method? Does your data reflect the entirety of a particular phenomenon, a recognised sample, or maybe even an obstructed view caused by hidden limitations in the availability of data about that phenomenon?

Reflecting on the published executed offenders data, there would be a certain confidence that it is representative of the total population of executions but with a specific caveat: it is all the executed offenders under the jurisdiction of the Texas Department of Criminal Justice since 1982. It is not the whole of the executions conducted across the entire USA nor is it representative of all the executions that have taken place throughout the history of Texas. Any conclusions drawn from this data must be boxed within those parameters.

The matter of judging completeness can be less about the number of records and more a question of the integrity of the data content. This executed offenders dataset would appear to be a trusted and reliable record of each offender but would there/could there be an incentive for the curators of this data not to capture, for example, the last statements as they were explicitly expressed? Could they have possibly been in any way sanitised or edited, for example? These are the types of questions you need to pose. This is not aimless cynicism, it is about seeking assurances of quality and condition so you can be confident about what you can legitimately present and conclude from it (as well as what you should not).

Consider a different scenario. If you are looking to assess the political mood of a nation during a televised election debate, you might consider analysing Twitter data by looking at the sentiments for and against the candidates involved. Although this would offer an accessible source of rich data, it would not provide an entirely reliable view of the national mood. It could only offer algorithmically determined insights (i.e. through the process of determining the sentiment from natural language) of the people who have a Twitter account, are watching the debate and have chosen to tweet about it during a given timeframe.

Now, just because you might not have access to a 'whole' population of political opinion data does not mean it is not legitimate to work on a sample. Sometimes samples are astutely reflective of the population. And in truth, if samples were not viable then most of the world's analyses would need to cease immediately.

> 'This is one of the first questions we should ask about any dataset: what is missing? What can we learn from the gaps?' **Jer Thorp, Founder of The Office for Creative Research**

A final point is to encourage you to probe any absence of data. Sometimes you might choose to switch the focus away from the data you have got towards the data you have not got. If the data you have is literally as much as you can acquire but you know the subject should have more data about it, then perhaps shine a light on the gaps, making that your story. Maybe you will unearth a discovery

about the lack of intent or will to make the data available, which in itself may be a fascinating discovery. As transparency increases, those who are not stand out the most.

Any identified lack of completeness or full representativeness is not an obstacle to progress, it just means you need to tread carefully with regard to how you might represent and present any work that emerges from it. It is about caution not cessation.

Influence on Process

This extensive examination work gives you an initial – but thorough – appreciation of the potential of your data, the things it will offer and the things it will not. Of course this potential is as yet unrealised. Furthering this examination will be the focus of the next activity, as you look to employ more visual techniques to help unearth the as-yet-hidden qualities of understanding locked away in the data. For now, this examination work takes your analytical and creative thinking forward another step.

Purpose map 'tone': Through deeper acquaintance with your data, you will have been able to further consider the suitability of the potential tone of your work. By learning more about the inherent characteristics of the subject, this might help to confirm or redefine your intentions for adopting a utilitarian (reading) or sensation-based (feeling) tone.

Editorial angles: The main benefit of exploring the data types is to arrive at an understanding of what you have and have not got to work with. More specifically, it guides your thinking towards what possible angles of analysis may be viable and relevant, and which can be eliminated as not. For example, if you do not have any location or spatial data, this rules out the immediate possibility of being able to map your data. This is not something you could pursue with the current scope of your dataset. If you do have time-based data then the prospect of conducting analysis that might show changes over time is viable. You will learn more about this idea of editorial 'angle' in the next chapter but let me state now it is one of the most important components of visualisation thinking.

Physical properties influence scale: Data is your raw material, your ideas are not. I stated towards the end of Chapter 3 that you should embrace the instinctive manifestations of ideas and seek influence and inspiration from other sources. However, with the shape and size of your data

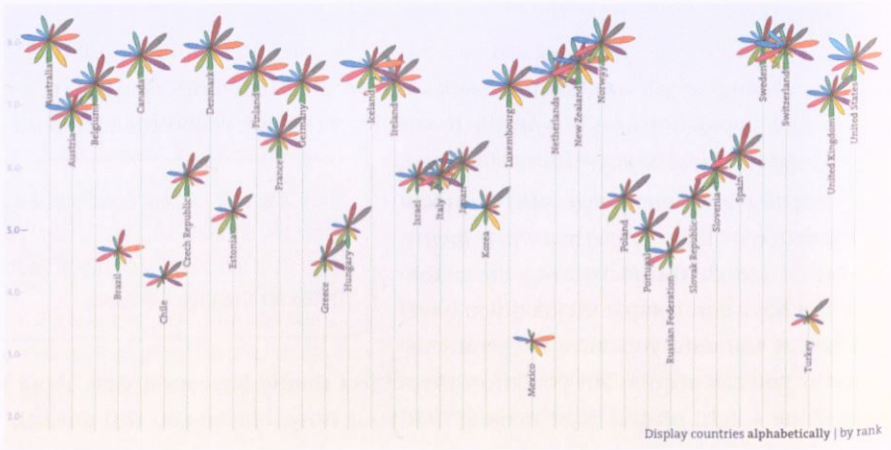


Figure 4.5 OECD Better Life Index

having such an impact on any eventual designs, you must respect the need to be led by your data's physical properties and not just your ideas.

In particular, the range of values in your data will shape things significantly. The shape of data in the 'Better Life Index' project you saw earlier is a good example. Figure 4.5 presents an analysis of the quality of life across the 36 OECD member states. Each country is a flower comprising 11 petals with each representing a different quality of life indicator (the larger the petal, the better the measured quality of life).

'My design approach requires that I immerse myself deeply in the problem domain and available data very early in the project, to get a feel for the unique characteristics of the data, its "texture" and the affordances it brings. It is very important that the results from these explorations, which I also discuss in detail with my clients, can influence the basic concept and main direction of the project. To put it in Hans Rosling's words, you need to "let the data set change your mind set".' **Moritz Stefaner, Truth & Beauty Operator**

Consider this. Would this design concept still be viable if there were 20 indicators? Or just 3? How about if the analysis was for 150 countries? The connection between data range and chart design involves a discerning judgement about 'fit'. You need to identify carefully the underlying shape of the data to be displayed and what tolerances this might test in the shape of the possible design concepts used.

Another relevant concern involves the challenge of elegantly handling quantitative measures that have hugely varied value ranges and contain (legitimate) outliers. Accommodating all the values into a single display can have a hugely distorting impact on the space it occupies. For example, note the exceptional size of the shape for *Avatar* in Figure 4.6, from the 'Spotlight on profitability' graphic you saw earlier. It is the one movie included that bursts through the ceiling, far beyond the otherwise entirely suitable 1000 million maximum scale value. As a single outlier, in this case, it was treated with a rather unique approach. As you can see, its striking shape conveniently trespasses onto the space offered by the two empty rows above. The result emphasises this value's exceptional quality. You might seldom have the luxury of this type of effective resolution, so the key point to stress is always be acutely aware of the existence of 'Avatars' in your data.

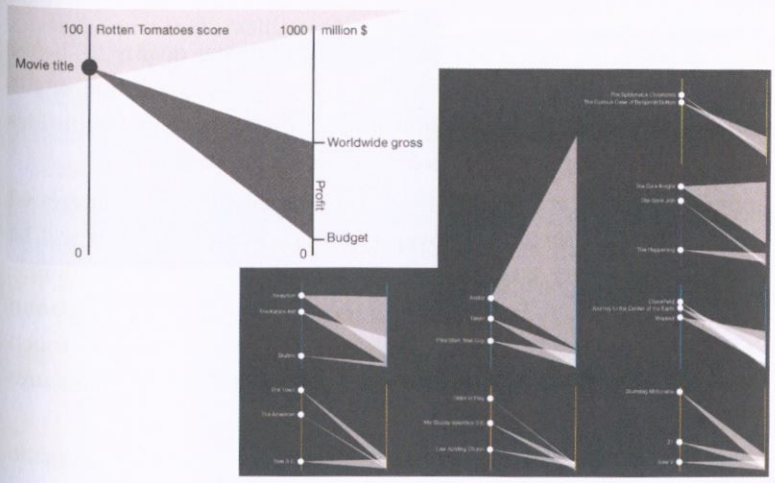


Figure 4.6
Spotlight on Profitability

4.5 Data Transformation

Having undertaken an examination of your data you will have a good idea about what needs to be done to ensure it is entirely fit for purpose. The next activity is to work on transforming the data so it is in optimum condition for your needs.

At this juncture, the linearity of a book becomes rather unsatisfactory. Transforming your data is something that will take place before, during and after both the examination and (upcoming) exploration steps. It will also continue beyond the boundaries of this stage of the workflow. For example, the need to transform data may only emerge once you begin your 'editorial thinking', as covered by the next chapter (indeed you will likely find yourself bouncing forwards and backwards between these sections of the book on a regular basis). As you get into the design stage you will constantly stumble upon additional reasons to tweak the shape and size of your data assets. The main point here is that your needs will evolve. This moment in the workflow is not going to be the only or final occasion when you look to refine your data.

Two important notes to share upfront at this stage. Firstly, in accordance with the desire for trustworthy design, any treatments you apply to your data need to be recorded and potentially shared with your audience. You must be able to reveal the thinking behind any significant assumptions, calculations and modifications you have made to your data.

Secondly, I must emphasise the critical value of keeping backups. Before you undertake any transformation, make a copy of your dataset. After each major iteration remember to save a milestone version for backup purposes. Additionally, when making changes, it is useful to preserve original (unaltered) data items nearby for easy rollback should you need them. For example, suppose you are cleaning up a column of messy data to do with 'Gender' that has a variety of inconsistent values (such as "M", "Male", "male", "FEMALE", "F", "Female"). Normally I would keep the original data, duplicate the column, and then tidy up this second column of values. I have then gained access to both original and modified versions. If you are going to do any transformation work that might involve a significant investment of time and (manual) effort, having an opportunity to refer to a previous state is always useful in my experience.

There are four different types of potential activity involved in transforming your data: cleaning, converting, creating and consolidating.

Transform to clean: I spoke about the importance of data quality (better quality in, better quality out, etc.) in the examination section when looking at the physical condition of the data. There's no need to revisit the list of potential observations you might need to consider looking out for but this is the point where you will need to begin to address these.

There is no single or best approach for how to conduct this task. Some issues can be addressed through a straightforward 'find and replace' (or remove) operation. Some treatments will be possible using simple functions to convert data into new states, such as using logic formulae that state 'if this, do this, otherwise do that'. For example, if the value in the 'Gender' column is "M" make it "Male", if the value is "MALE" make it "Male" etc. Other tasks might be much more intricate, requiring manual intervention, often in combination with inspection features like 'sort' or 'filter', to find, isolate and then modify problem values.

Part of cleaning up your data involves the elimination of junk. Going back to the earlier scenario about gathering data about McDonald's restaurants, you probably would not need the name of the

restaurant manager, details of the opening times or the contact telephone number. It is down to your judgement at the time of gathering the data to decide whether these extra items of detail – if they were as easily acquirable as the other items of data that you really *did* need – may potentially provide value for your analysis later in the process. My tactic is usually to gather as much data as I can and then reject/trim later; *later* has arrived and now is the time to consider what to remove. Any fields or rows of data that you know serve no ongoing value will take up space and attention, so get rid of these. You will need to separate the wheat from the chaff to help reduce your problem.

Transform to convert: Often you will seek to create new data values out of existing ones. In the illustration in Figure 4.7, it might be useful to extract the constituent parts of a 'Release Date' field in order to group, analyse and use the data in different ways. You might use the 'Month' and 'Year' fields to aggregate your analysis at these respective levels in order to explore within-year and across-year seasonality. You could also create a 'Full Release Date' formatted version of the date to offer a more presentable form of the release date value possibly for labeling purposes.

Release Date	Month	Year	Full Release Date
21/06/13	6	2013	Friday, 21 June 2013
26/08/13	8	2013	Monday, 26 August 2013
21/12/12	12	2012	Friday, 21 December 2012
22/12/10	12	2010	Wednesday, 22 December 2010
11/06/09	6	2009	Thursday, 11 June 2009
30/11/10	11	2010	Tuesday, 30 November 2010
29/04/07	4	2007	Sunday, 29 April 2007
14/12/07	12	2007	Friday, 14 December 2007
07/11/06	11	2006	Tuesday, 7 November 2006
20/10/06	10	2006	Friday, 20 October 2006
22/12/06	12	2006	Friday, 22 December 2006
12/11/04	11	2004	Friday, 12 November 2004
20/06/03	6	2003	Friday, 20 June 2003
01/01/98	1	1998	Thursday, 1 January 1998
01/01/03	1	2003	Wednesday, 1 January 2003
29/08/03	8	2003	Friday, 29 August 2003

Figure 4.7 Example of Converted Data Transformation

Extracting or deriving new forms of data will be necessary when it comes to handling qualitative 'textual' data. As stated in the 'Data literacy' section, if you have textual data you will generally always need to transform this into various categorical or quantitative forms, unless its role is simply to provide value as an annotation (such as a quoted caption or label). Some would argue that qualitative visualisation involves special methods for the representation of data. I would disagree. I believe the unique challenge of working with textual data lies with the task of transforming the data: visually representing the extracted and derived properties from textual data involves the same suite of representation options (i.e. chart types) that would be useful for portraying analysis of any other data types.

Here is a breakdown of some of the conversions, calculations and extractions you could apply to textual data. Some of these tasks can be quite straightforward (e.g. Using the LEN function in Excel to determine the number of characters) while others are more technical and will require more sophisticated tools or programmes dedicated to handling textual data.

Categorical conversions:

- Identify keywords or summary themes from text and convert these into categorical classifications.
- Identify and flag up instances of certain cases existing or otherwise (e.g. X is mentioned in this passage).

- Identify and flag up the existence of certain relationships (e.g. A and B were both mentioned in the same passage, C was always mentioned before D).
- Use natural language-processing techniques to determine sentiments, to identify specific word types (nouns, verbs, adjectives) or sentence structures (around clauses and punctuation marks).
- With URLs, isolate and extract the different components of website address and sub-folder locations

Quantitative conversions:

- Calculate the frequency of certain words being used.
- Analyse the attributes of text, such as total word count, physical length, potential reading duration.
- Count the number of sentences or paragraphs, derived from the frequency of different punctuation marks.
- Position the temporal location of certain words/phrases in relation to other words/phrases or compared to the whole (e.g. X was mentioned at 1m51s).
- Position the spatial location of certain words/phrases in relation to other words/phrases or compared to the whole.

A further challenge that falls under this 'converting' heading will sometimes emerge when you are working with data supplied by others in spreadsheets. This concerns the obstacles created when trying to analyse a data that has been formatted visually, perhaps in readiness for printing. If you receive data in this form you will need to unpack and reconstruct it into the normalised form described earlier, comprising all records and fields included in a single table.

Any merged cells need unmerging or removing. You might have a heading that is common to a series of columns. If you see this, unmerge it and replicate the same heading across each of the relevant columns (perhaps appending an index number to each header to maintain some differentiation). Cells that have visual formatting like background shading or font attributes (bold, coloured) to indicate a value or status are useful when observing and reading the data, but for analysis operations these properties are largely invisible. You will need to create new values in actual data form that are not visual (creating categorical values, say, or status flags like 'yes' or 'no') to recreate the meaning of the formats. The data provided to you – or that you create – via a spreadsheet does not need to be elegant in appearance, it needs to be functional.

Transform to create: This task is something I refer to as the hidden cleverness, where you are doing background thinking to form new calculations, values, groupings and any other mathematical or manual treatments that really expand the variety of data available.

A simple example might involve the need to create some percentage calculations in a new field, based on related quantities elsewhere within your existing data. Perhaps you have pairs of 'start date' and 'end date' values and you need to calculate the duration in days for all your records. You might use logic formula to assist in creating a new variable that summarises another – maybe something like (in language terms) IF Age < 18 THEN status = "Child", ELSE status = "Adult". Alternatively, you might want to create a calculation that standardised some quantities' need to source base population figures for all the relevant locations in your data in order to convert some

quantities into 'per capita' values. This would be particularly necessary if you anticipate wanting to map the data as this will ensure you are facilitating legitimate comparisons.

Transform to consolidate: This involves bringing in additional data to help expand (more variables) or append (more records) to enhance the editorial and representation potential of your project.

An example of a need to expand your data would be if you had details about locations only at country level but you wanted to be able to group and aggregate your analysis at continent level. You could gather a dataset that holds values showing the relationships between country and continent and then add a new variable to your dataset against which you would perform a simple lookup operation to fill in the associated continent values.

Consolidating by appending data might occur if you had previously acquired a dataset that now had more or newer data (specifically, additional records) available to bring it up to date. For instance, you might have started some analysis on music record sales up to a certain point in time, but once you'd actually started working on the task another week had elapsed and more data had become available.

Additionally, you may start to think about sourcing other media assets to enhance your presentation options, beyond just gathering extra data. You might anticipate the potential value for gathering photos (headshots of the people in your data), icons/symbols (country flags), links to articles (URLs), or videos (clips of goals scored). All of these would contribute to broadening the scope of your annotation options. Even though there is a while yet until we reach that particular layer of design thinking, it is useful to start contemplating this as early possible in case the collection of these additional assets requires significant time and effort. It might also reveal any obstacles around having to obtain permissions for usage or sufficiently high quality media. If you know you are going to have to do something, don't leave it too late – reduce the possibility of such stresses by acting early.

4.6 Data Exploration

The examination task was about forming a deep acquaintance with the physical properties and meaning of your data. You now need to interrogate that data further – and differently – to find out what potential insights and qualities of understanding it could provide.

Undertaking data exploration will involve the use of statistical and visual techniques to move beyond *looking* at data and begin to start *seeing* it. You will be directly pursuing your initially defined curiosity, to determine if answers exist and whether they are suitably enlightening in nature. Often you will not know for sure

'After the data exploration phase you may come to the conclusion that the data does not support the goal of the project. The thing is: data is leading in a data visualization project – you cannot make up some data just to comply with your initial ideas. So, you need to have some kind of an open mind and "listen to what the data has to say", and learn what its potential is for a visualisation. Sometimes this means that a project has to stop if there is too much of a mismatch between the goal of the project and the available data. In other cases this may mean that the goal needs to be adjusted and the project can continue.' **Jan Willem Tulp, Data Experience Designer**

whether what you initially thought was interesting is exactly that. This activity will confirm, refine or reject your core curiosity and perhaps, if you are fortunate, present discoveries that will encourage other interesting avenues of enquiry.

To frame this process, it is worth introducing something that will be covered in Chapter 5, where you will consider some of the parallels between visualisation and photography. Before committing to take a photograph you must first develop an appreciation of all the possible viewpoints that are available to you. Only then can you determine which of these is best. The notion of 'best' will be defined in the next chapter, but for now you need to think about identifying all the possible viewpoints in your data – to recognise the knowns and the unknowns.

Widening the Viewpoint: Knowns and Unknowns

At a news briefing in February 2002, the US Secretary of Defense, Donald Rumsfeld, delivered his infamous 'known knowns' statement:

> Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones.

There was much commentary about the apparent lack of elegance in the language used and criticism of the muddled meaning. I disagree with this analysis. I thought it was probably the most efficient way he could have articulated what he was explaining, at least in written or verbal form. The essence of Rumsfeld's statement was to distinguish *awareness* of what is knowable about a subject (what knowledge exists) from the status of *acquiring* this knowledge. There is a lot of value to be gained from using this structure (Figure 4.8) to shape your approach to thinking about data exploration.

The *known knowns* are aspects of knowledge about your subject and about the qualities present in your data that you are aware of – you are aware that you know these things. The nature of these known knowns might mean you have confidence that the origin curiosity was relevant and the available insights that emerged in response are suitably interesting. You cannot afford to be complacent, though. You will need to challenge yourself to check that these curiosities are still legitimate and relevant. To support this, you should continue to look and learn about the subject through research, topping up your awareness of the most potentially relevant dynamics of the subject, and continue to interrogate your data accordingly.

Additionally, you should not just concentrate on this potentially quite narrow viewpoint. As I mentioned earlier, it is important to give yourself as broad a view as possible across your subject and its data to optimise your decisions about what other interesting enquiries might be available. This is where you need to consider the other quadrants in this diagram.

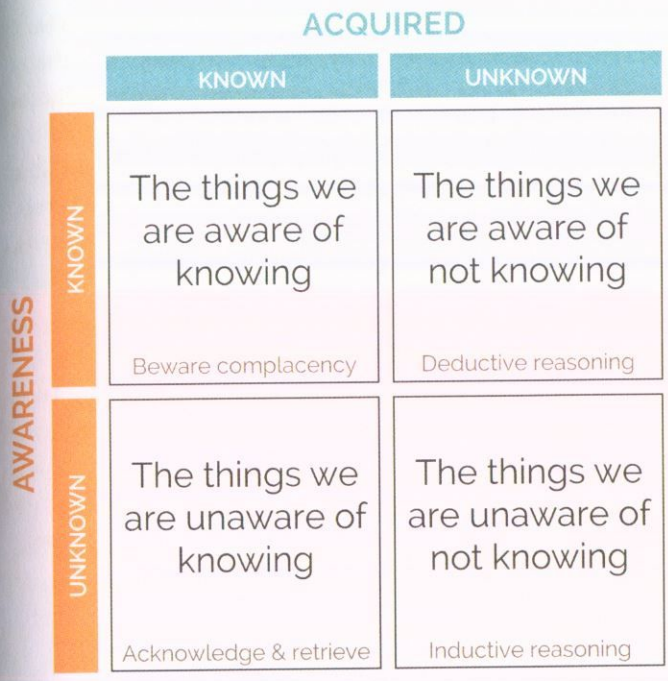


Figure 4.8
Making Sense of the *Known Knowns*

On occasion, though I would argue rarely, there may be *unknown knowns*, things you did not realise you knew or perhaps did not wish to acknowledge that you knew about a subject. This may relate to previous understandings that have been forgotten, consciously ignored or buried. Regardless, you need to acknowledge these.

For the knowledge that has yet to be acquired – the *known unknowns* and the even more elusive *unknown unknowns* – tactics are needed to help plug these gaps as far, as deep and as wide as possible. You cannot possibly achieve mastery of all the domains you work with. Instead, you need to have the capacity and be in position to turn as many unknowns as possible into knowns, and in doing so optimise your understanding of a subject. Only then will you be capable of appreciating the full array of viewpoints the data offers.

To make the best decisions you first need to be aware of all the options. This activity is about broadening your awareness of the potentially interesting things you could show – and could say – about your data. The resulting luxury of choice is something you will deal with in the next stage.

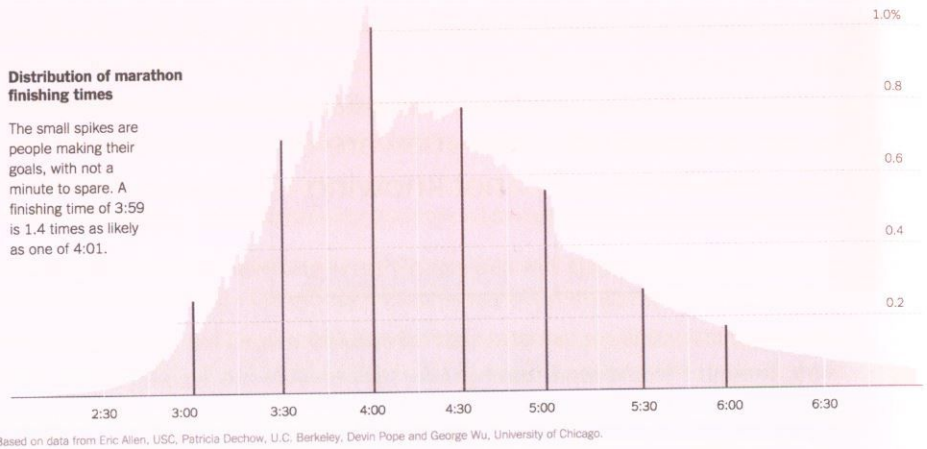
Exploratory Data Analysis

As I have stated, the aim throughout this book is to create a visualisation that will facilitate understanding for others. That is the end goal. At this stage of the workflow the deficit in understanding lies with you. The task of addressing the *unknowns* you have about a subject, as well as substantiating what *knowns* already exist, involves the use of exploratory data analysis (EDA). This integrates statistical methods with visual analysis to offer a way of extracting deeper understanding and widening the view to unlock as much of the potential as possible from within your data.

The chart in Figure 4.9 is a great demonstration of the value in combining statistical and visual techniques to understand your data better. It shows the results of nearly every major and many minor (full) marathon from around the world. On the surface, the distribution of finishing times reveals the common bell shape found in plots about many natural phenomenon, such as the height measurements of a large group of people. However, when you zoom in closer the data reveals some really interesting threshold patterns for finishing times on or just before the three-, four- and five-hour marks. You can see that the influence of runners setting themselves targets, often rounded to the hourly milestones, genuinely appeared to affect the results achieved.

Figure 4.9 What Good Marathons and Bad Investments Have in Common

Arbitrary goals, like round numbers, can be motivating – just ask 9 million marathoners.



Based on data from Eric Allen, USC, Patricia Dechow, U.C. Berkeley, Devin Pope and George Wu, University of Chicago.

Although statistical analysis of this data would have revealed many interesting facts, these unique patterns were only realistically discoverable through studying the visual display of the data. This is the essence of EDA but there is no instruction manual for it. As John Tukey, the father of EDA, described: 'Exploratory data analysis is an attitude, a flexibility, and a reliance on display, not a bundle of techniques'. There is no single path to undertaking this activity effectively; it requires a number of different technical, practical and conceptual capabilities.

Instinct of the analyst: This is the primary matter. The attitude and flexibility that Turkey describes are about recognising the importance of the analyst's traits. Effective EDA is not about the tool. There are many vendors out there pitching their devices as the magic option where we just have to 'point and click' to uncover a deep discovery. Technology inevitably plays a key role in facilitating this endeavour but the value of a good analyst cannot be underestimated: it is arguably more influential than the differentiating characteristics between one tool and the next. In the absence of a defined procedure for conducting EDA, an analyst needs to possess the capacity to recognise and pursue the scent of enquiry. A good analyst will have that special blend of natural inquisitiveness and the sense to know what approaches (statistical or visual) to employ and when. Furthermore, when these traits collide with a strong subject knowledge this means better judgments are made about which findings from the analysis are meaningful and which are not.

Reasoning: Efficiency is a particularly important aspect of this exploration stage. The act of interrogating data, waiting for it to volunteer its secrets, can take a lot of time and energy. Even with smaller datasets you can find yourself tempted into trying out myriad combinations of analyses, driven by the desire to find *the killer* insight in the shadows.

'At the beginning, there's a process of "interviewing" the data – first evaluating their source and means of collection/aggregation/computation, and then trying to get a sense of what they say – and how well they say it via quick sketches in Excel with pivot tables and charts. Do the data, in various slices, say anything interesting? If I'm coming into this with certain assumptions, do the data confirm them, or refute them?' **Alyson Hurt, News Graphics Editor, NPR**

Reasoning is an attempt to help reduce the size of the prospect. You cannot afford to try everything. There are so many statistical methods and, as you will see, so many visual means for seeing views of data that you simply cannot expect to have the capacity to try to unleash the full exploratory artillery. EDA is about being smart, recognising that you need to be discerning about your tactics.

In academia there are two distinctions in approaches to reasoning – deductive and inductive – that I feel are usefully applied in this discussion:

- *Deductive* reasoning is targeted: You have a specific curiosity or hypothesis, framed by subject knowledge, and you are going to interrogate the data in order to determine whether there is any evidence of relevance or interest in the concluding finding. I consider this adopting a detective's mindset (Sherlock Holmes).
- *Inductive* reasoning is much more open in nature: You will 'play around' with the data, based on your sense or instinct about what might be of interest, and wait and see what emerges. In some ways this is like prospecting, hoping for that moment of serendipity when you unearth gold.

In this exploration process you ideally need to accommodate both approaches. The deductive process will focus on exploring further targeted curiosities, the inductive process will give you a fighting chance of finding more of those slippery 'unknowns', often almost by accident. It is important to give yourself room to embark on these somewhat less structured exploratory journeys.

I often think about EDA in the context of a comparison with the challenge of a 'Where's Wally?' visual puzzle. The process of finding Wally feels somewhat unscientific. Sometimes you let your eyes race around the scene like a dog who has just been let out of the car and is torpedoing across a field. However, after the initial burst of randomness, perhaps subconsciously, you then go through a more considered process of visual analysis. Elimination takes place by working around different parts of the scene and sequentially declaring 'Wally-free' zones. This aids your focus and strategy for where to look next. As you then move across each mini-scene you are pattern matching, looking out for the giveaway characteristics of the boy wearing glasses, a red-and-white-striped hat and jumper, and blue trousers.

The objective of this task is clear and singular in definition. The challenge of EDA is rarely that clean. There is a source curiosity to follow, for sure, and you might find evidence of Wally somewhere in the data. However, unlike the 'Where's Wally?' challenge, in EDA you have the chance also to find other things that might change the definition of what qualifies

as an interesting insight. In unearthing other discoveries you might determine that you no longer care about Wally; finding him no longer represents the main enquiry.

Inevitably you are faced with a trade-off between spare capacity in time and attention and your own internal satisfaction that you have explored as many different angles of enquiry as possible.

Chart types: This is about seeing the data from all feasible angles. The power of the visual means that we can easily rely on our pattern-matching and sense-making capabilities – in harmony with contextual subject knowledge – to make observations about data that appear to have relevance.

'I kick it over into a rough picture as soon as possible. When I can see something then I am able to ask better questions of it – then the what-about-this iterations begin. I try to look at the same data in as many different dimensions as possible. For example, if I have a spreadsheet of bird sighting locations and times, first I like to see where they happen, previewing it in some mapping software. I'll also look for patterns in the timing of the phenomenon, usually using a pivot table in a spreadsheet. The real magic happens when a pattern reveals itself only when seen in both dimensions at the same time.' **John Nelson, Cartographer, on the value of visually exploring his data**

The data representation gallery that you will encounter in Chapter 6 presents nearly 50 different chart types, offering a broad repertoire of options for portraying data. The focus of the collection is on chart types that could be used to communicate to others. However, within this gallery there are also many chart types that help with pursuing EDA. In each chart profile, indications are given for those chart types that be particularly useful to support your exploratory activity. As a rough estimate, I would say about half of these can prove to be great allies in this stage of discovery.

The visual methods used in EDA do not just involve charting, they also involve selective charting – smart charting, 'smarting' if you like? (No, Andy, nobody likes that). Every chart type presented in the gallery includes helpful descriptions that will give you an idea of their role and also what observations – and potential interpretations – they might facilitate. It is important to know now that the chart types are organised across five main families (categorical, hierarchical, relational, temporal, and spatial) depending on the primary focus of your analysis. The focus of your analysis will, in turn, depend on the types of data you have and what you are trying to see.

Research: I have raised this already but make no apology for doing so again so soon. How you conduct research and how much you can do will naturally depend on your circumstances, but it is always important to exploit as many different approaches to learning about the domain and the data you are working with. As you will recall, the middle stage of forming understanding – *interpreting* – is about viewers translating what they have perceived from a display into meaning. They can only do this with domain knowledge. Similarly, when it comes to conducting exploratory analysis using visual methods, you might be able to *perceive* the charts you make, but without possessing or acquiring sufficient domain knowledge you will not know if what you are seeing is meaningful. Sometimes the consequence of this exploratory data analysis will only mean you have become better acquainted with specific questions and more defined curiosities about a subject even if you possibly do not yet have any answers.

The approach to research is largely common sense: you explore the places (books, websites) and consult the people (experts, colleagues) that will collectively give you the best chance of getting accurate answers to the questions you have. Good communication skills, therefore, are vital – it

is not just about talking to others, it is about listening. If you are in a dialogue with experts you will have to find an approach that allows you to understand potentially complicated matters and also cut through to the most salient matters of interest.

Statistical methods: Although the value of the univariate statistical techniques profiled earlier still applies here, what you are often looking to undertake in EDA is multivariate analysis. This concerns testing out the potential existence of a correlation between quantitative variables as well as determining the possible causation variables – the holy grail of data analysis.

Typically, I find statistical analysis plays more of a supporting role during much of the exploration activity rather than a leading role. Visual techniques will serve up tangible observations about whether data relationships and quantities seem relevant, but to substantiate this you will need to conduct statistical tests of significance.

One of the main exceptions is when dealing with large datasets. Here the first approach might be more statistical in nature due to the amount of data obstructing rapid visual approaches. Going further, algorithmic approaches – using techniques like machine learning – might help to scale the task of statistically exploring large dimensions of data – and the endless permutations they offer. What these approaches gain in productivity they clearly lose in human quality. The significance of this should not be underestimated. It may be possible to take a blended approach where you might utilise machine learning techniques to act as an initial *battering ram* to help reduce the problem, identifying the major dimensions within the data that might hold certain key statistical attributes and then conducting further exploration 'by hand and by eye'.

Nothings: What if you have found nothing? You have hit a dead end, discovering no significant relationships and finding nothing interesting about the shape or distribution of your data. What do you do? In these situations you need to change your mindset: *nothing* is usually *something*. Dead ends and discovering blind alleys are good news because they help you develop focus by eliminating different dimensions of possible analysis. If you have traits of nothingness in your data or analysis – gaps, nulls, zeroes and no insights – this could prove to be the insight. As described earlier, make the gaps the focus of your story.

> 'My main advice is not to be disheartened. Sometimes the data don't show what you thought they would, or they aren't available in a usable or comparable form. But [in my world] sometimes that research still turns up threads a reporter could pursue and turn into a really interesting story – there just might not be a viz in it. Or maybe there's no story at all. And that's all okay. At minimum, you've still hopefully learned something new in the process about a topic, or a data source (person or database), or a "gotcha" in a particular dataset – lessons that can be applied to another project down the line.' **Alyson Hurt, News Graphics Editor, NPR**

There is *always* something interesting in your data. If a value has not changed over time, maybe it was supposed to – that is an insight. If everything is the same size, that is the story. If there is no significance in the quantities, categories or spatial relationships, make those your insights. You will only know that these findings are relevant by truly understanding the context of the subject matter. This is why you must make as much effort as possible to convert your *unknowns* into *knowns*.

Not always needed: It is important to couch this discussion about exploration in pragmatic reality. Not *all* visualisation challenges will involve *much* EDA. Your subject and your data might be immediately understandable and you may have a sufficiently broad viewpoint of your subject

(plenty of known knowns already in place). Further EDA activity may have diminishing value. Additionally, if you are faced with small tables of data this simply will not warrant multivariate investigation. You certainly need to be ready and equipped with the capacity to undertake this type of exploration activity when it is needed, but the key point here is to judge when.

Summary: Working with Data

This chapter first introduced key foundations for the requisite data literacy involved in visualisation, specifically the importance of the distinction between normalised and cross-tabulated datasets as well as the different types of data (using the TNOIR mnemonic):

- Textual (qualitative): e.g. 'Any other comments?' data submitted in a survey.
- Nominal (qualitative): e.g. The 'gender' selected by a survey participant.
- Ordinal (qualitative): e.g. The response to a survey question, based on a scale of 1 (unhappy) to 5 (very happy).
- Interval (quantitative): e.g. The shoe size of a survey participant.
- Ratio (quantitative): e.g. The age of a survey participant in years.

You then walked through the four steps involved in working with data:

Acquisition Different sources and methods for getting your data.

- Curated by you: primary data collection, manual collection and data foraging, extracted from pdf, web scraping (also known as web harvesting).
- Curated by others: issued to you, downloaded from the Web, system report or export, third-party services, APIs.

Examination Developing an intimate appreciation of the characteristics of this critical raw material:

- Physical properties: type, size, and condition.
- Meaning: phenomenon, completeness.

Transformation Getting your data into shape, ready for its role in your exploratory analysis and visualisation design:

- Clean: resolve any data quality issues.
- Create: consider new calculations and conversions.
- Consolidate: what other data (to expand or append) or other assets could be sought to enhance your project?

Exploration Using visual and statistical techniques to *see* the data's qualities: what insights does it reveal to you as you deepen your familiarity with it?

Tips and Tactics

- Perfect data (complete, accurate, up to date, truly representative) is an almost impossible standard to reach (given the presence of time constraints) so your decision will be when is good enough, good enough: when do diminishing returns start to materialise?

- Do not underestimate the demands on your time; working with data will always be consuming of your attention and effort:
  - Ensure you have built plenty of time into your handling of this data stage.
  - Be patient and persevere.
  - Be disciplined: it is easy to get swallowed up in the potential hunt for discovering things from your data, attempting to explore every possible permutation.
- If your data does not already have a unique identifier it is often worth creating one to track your data preparation process. This is especially helpful if you need to preserve or revert to a very specific ordering of your data (e.g. if the rows have been carefully arranged in order to undertake cross-row calculations like cumulative or sub-totals).
- Clerical tasks like file management are important: maintain backups of each major iteration of data, employ good file organisation of your data and other assets, and maintain logical naming conventions.
- Data management practices around data security and privacy will be important in the more sensitive/confidential cases.
- Keep notes about where you have sourced data, what you have done with it, any assumptions or counting rules you have applied, ideas you might have for transforming or consolidating, issues/problems, things you do not understand.
- To learn about your data, its meaning and the subject matter to which it relates, you should build in time to undertake research in order to equip yourself suitably with domain knowledge.
- Anticipate and have contingency plans for the worst-case scenarios for data, such as the scarcity of data availability, null values, odd distributions, erroneous values, long values, bad formatting, data loss.
- Communicate. If you do not know anything about your data, ask: do not assume or stay ignorant. And then listen: always pay attention to key information.
- Attention to detail is of paramount importance at this stage, so get into good habits early and do not cut corners.
- Maintain an open mind and do not get frustrated. You can only work with what you have. If it is not showing what you expected or hoped for, you cannot force it to say something that is simply not there.
- Exploratory Data Analysis is *not* about design elegance. Do not waste time making your analysis 'pretty', it only needs to inform you.