Data Visualisation: From Theory to Practice

James Baglin

2020-02-12

Contents

Preface

Acknowledgements

Chapter 1

Design and Integrity

1.1 Summary

Chapter 1 will introduce data visualisation as a design process. The chapter will start with a discussion of the definition of data visualisation and the various reasons why data need to be visualised. This will help introduce the first stage of design, "determining your audience and visualisation objective". The two other stages, "Focusing, justifying and choosing methods" and "Construction and evaluation" will also be introduced. You will examine and apply two handy tools for guiding the critique and preparation of data visualisations for publication. The second part of the chapter introduces data visualisation ethics and data integrity. You will consider the principles of ethical data visualisation and the responsible use of data, including how to identify reliable data sources.

1.1.1 Learning Objectives

The learning objectives of this chapter are as follows:

- Define data visualisation
- List and explain the different types of data and why data types are important to data visualisation

- Identify the various components of a basic data visualisation plot
- Explain the data visualisation design process including the following three stages:
  - Identifying a targeted audience and a data visualisation design objective
  - Focusing, justifying and choosing methods
  - Construction and evaluation
- Apply the Trifecta Check-up to guide the critique of a data visualisation
- Apply the Data Visualisation Check-list to produce publication quality data visualisations
- Discuss ethical principles and data integrity as it relates to the practice of data visualisation
- Locate and identify reliable and reputable sources of data for visualisation

1.1.2 Chapter Video

In the video below, David McCandless (2012) discusses the beauty of data visualisation during a TEDEd talk in 2012 (Only available online).

1.2 Defining Data Visualisation

The classic saying, "seeing is believing", effectively articulates the importance of data visualisation. Whether you are exploring vast datasets; communicating your data analysis in meaningful ways; presenting the story behind your data in order to captivate your audience, data visualisation is the most powerful tool at your disposal. So, what is data visualisation? Kirk (2012) defined data visualisation as "the representation and presentation of data that exploits our visual perception abilities in order to amplify cognition" (p. 17). Let's break this down:

- **Representation**: There isn't much we can discern from raw, unprocessed data. However, if we can represent data in forms that we are familiar with, like geometric objects, we can start to gain insight. Data visualisation represents data in a visual form ready for our brains to process.
- **Presentation**: Careful presentation of data is necessary to ensure that the story behind the data comes to light. There are infinite choices and decisions that need to be made when presenting your visualisation.
- **Visual perception**: Our brain is a very complex and powerful pattern recognition and processing machine. We can exploit our visual processing capabilities to quickly and accurately interpret data. Good data visualisation exploits our visual systems and avoid its pitfalls.

- **Amplify cognition**: Data visualisation should always inform and increase knowledge.

What are the specific advantages of visualising data? Let's explore a few classic examples. Anscombe (1973) published a quartet, reproduced in Figure 1.1, which warns us of the perils of not visualising data. All four scatter plots share the same line of best fit and correlation, however, visually the data all tell very different stories. Some of the datasets have outliers present while one is not linear. Visualisation serves an important perceptual role in exploring your data.

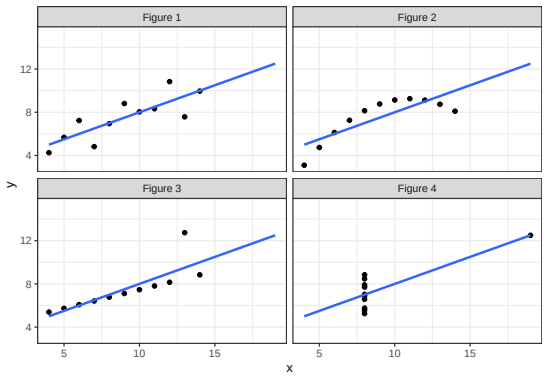


Figure 1.1: Anscombe's quartet (Anscombe, 1973).

Another pioneering example of the power of data visualisation was John Snow's 1854 cholera map shown in Figure 1.2 (Snow, 1854). This visualisation clearly showed authorities that the reported cases of cholera (highlighted in black) were centered around the Broad Street water pump.

Few (2014) lists the following main advantages:

1. Allows you to see the big picture
2. Allows you to easily and rapidly compare values
3. Allows you to see patterns amongst the data
4. Allows you to compare patterns

Additionally, Ware (2013) identified that data visualisation has the advantage of being able to:

1. Comprehend vast volumes of data
2. Facilitate the emergence of properties that might otherwise be hidden
3. Highlight problematic data and assist in quality control
4. Facilitate an understanding of data at all scales, small and large

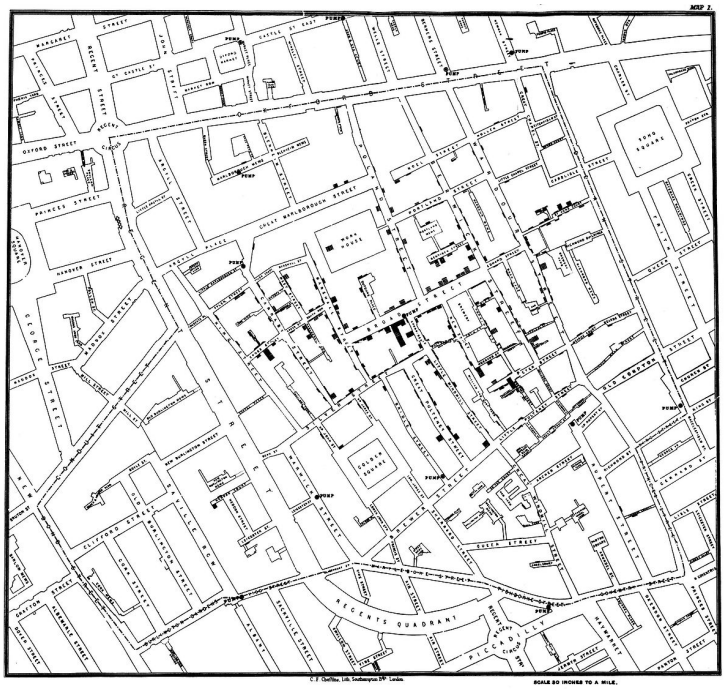


Figure 1.2: John Snow's cholera map (Snow, 1854).

5. Promote hypothesis formation

However, data visualisation is often much more than a window into our data. It is also used to persuade, inform, educate and tell data-based stories because of it's often intuitive and accessible nature. The goal of visualisation is to leave a lasting impression and, thus, Kosara (2016), termed these types of visualisations 'presentation-orientated techniques'. For example, Charles Minard (1869) published a famous visualisation of Napoleon's Russian Campaign in 1869. This visualisation was, and still remains, quite unique (see Figure 1.3). It takes the viewer some time to comprehend as the story is composed of many layers and variables. Despite this, Minard's work is still considered one of the best data visualisations of all time. That's quite the lasing impression.

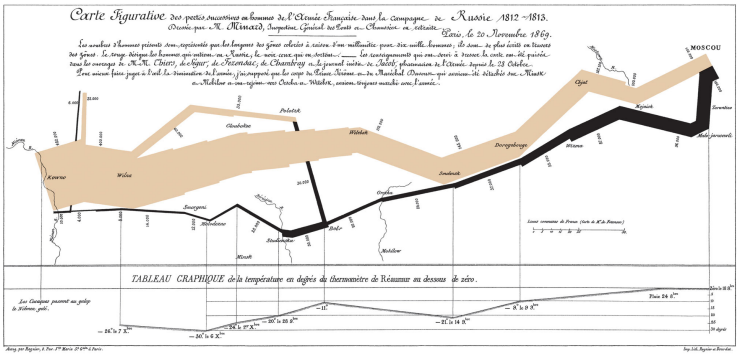


Figure 1.3: Charles Minard's famous visualisation of Napoleon's Russian Campaign (Minard, 1869).

Data visualisation, which sits within the broader area of information visualisation, is concerned with statistical data. Statistical data can mean many different things, but all statistical data have one major characteristic in common - variability. It's this variability that we try to measure, describe and predict that sits at the heart of statistics. As you will learn, data visualisation plays a central role in statistical practice, but it's rarely taught formally. Times are changing and data visualisation has been developing very quickly since the widespread availability of cheap computing power and the internet.

Rarely are we given the opportunity to learn about data visualisation. Often the way we visualise data is not informed by any targeted training. In fact, many of the visualisations we have previously created are, embarrassingly, guided by the default settings available in software like Excel, SAS and SPSS. We use gut instincts to decide if it looks right, instead of making informed decisions based on good design principles and an understanding of human perception.

As a truly interdisciplinary field, data visualisation continues to be heavily influenced by research in visual perception and psychology, statistics, computer

science, art and many other fields. Therefore, becoming an effective data visualisation designer requires a specialised body of multidisciplinary knowledge. This book will help you to design intuitive, accessible and compelling data visualisations that communicate the story behind the data and address practical, real-world problems. However, before we delve into data visualisation design, we need to do a bit of revision and take a quick look at plot anatomy.

1.3 Data Types - A Quick Revision

When visualising data it's important to be able to identify the types of variables present in our data. Types of variables govern the appropriate methods of data visualisation. As such, we will do a quick revision of variable types.

When you measure a variable, qualitative and quantitative variables can take on different scales or levels of measurement. Levels of measurement have a direct bearing on the choice of data visualisation methods you choose. We need to understand the language used to describe different scales. The following short video by Petty (2011) provides a great overview (Only available online).

- **Categorical or Nominal (Qualitative):** Categorical variables are group variables, or categories if you will. There are no meaningful measurement differences such as rankings or intervals between the different categories. Categorical or nominal variables include binary variables (e.g. yes/no, male/female) and multinomial variables (e.g. religious affiliation, hair colour, ethnicity, suburb).
- **Ordinal (Qualitative):** Ordinal data has a rank order by which it can be sorted, but the differences between the ranks are not relative or measurable. Therefore, ordinal data is not strictly quantitative. For example, consider the 1st, 2nd and 3rd place in a race. We know who was faster or slower, but we have no idea by how much. We need to look at the race times.
- **Interval (Quantitative):** An interval variable is similar to an ordinal variable except that the intervals between the values are equally spaced. Interval variables have an *arbitrary zero*-point and therefore no meaningful ratios. For example, think about our calendar year and the Celsius scale; 1000 AD is not half of 2000 AD, and 20 degrees Celsius is not twice as "hot" as 10 degrees Celsius. This is because our calendar and Celsius scale have an arbitrary value for zero. Zero AD and zero degrees Celsius do not imply the presence of zero time or zero heat energy.
- **Ratio (Quantitative):** A ratio variable is similar to an interval variable; however, there is an absolute zero point and ratios are meaningful. An example is time given in seconds, length in centimetres, or heart beats per minute. A value of 0 implies the absence of a variable. We can also make statements like 30 seconds is twice the time of 15 seconds, 10 cm is

half the height of 20 cm, and during exercise a person's resting heart beat almost doubles. Zero heart rate, call 000!

Interval and ratio variables might also be described as being *discrete* (can only take on a particular value, e.g. the number of times a person exercises each week) or *continuous* (a variable can theoretically take on an infinite number of values within a given range, e.g. someone's height).

1.4 Plot Anatomy

Most data visualisations are plotted or graphed, usually on a two-dimensional plane. That makes sense because in recent history, most publications were in print. The dominance of two-dimensional visualisations continues even today, however, we must not forget that technology allows us to move beyond an x and y axis. For example, EarthWindMap visualises global wind patterns using an interactive 3D globe.

However, for now, as we get started, let's take a look at the components of a simple plot, based on a two-dimensional Cartesian plane. This will help us to build some common terminology used throughout the book. Figure 1.4 shows a hypothetical plot visualising the relationship between an x and y variable for two groups, A and B. This is an example of a scatter plot.

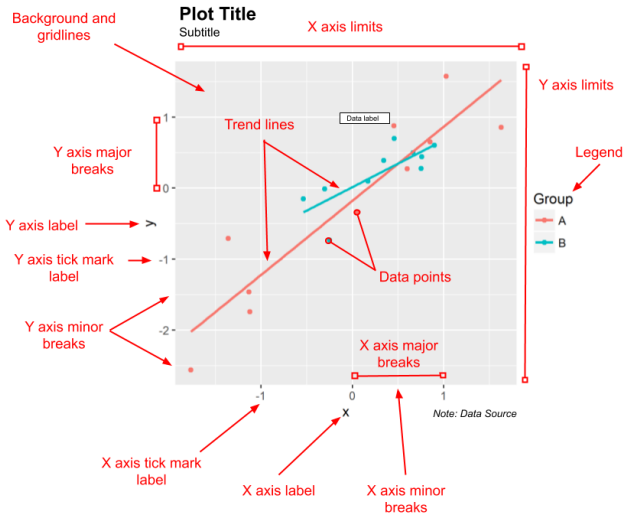


Figure 1.4: Plot anatomy.

Looking at this plot, ensure you can locate and name the following basic components:

- Plot title
- Subtitle
- Background and gridlines
- x and y axis labels
- Major and minor x and y axis breaks and tick marks
- x and y axis limits
- Data points and trend lines
- Data labels
- Legend
- Notes

1.5 A Visual Design Process

The following sections will summarise a visual design process outlined by Kirk (2012). Having a design process ensures you approach a visualisation task in a systematic way that is efficient and effective. However, no matter how much planning and coordination, the design process rarely goes exactly to plan and in a step-by-step fashion. Problems will occur. Remain flexible, learn from your mistakes and don't be afraid to take a step backwards before you move on again.

Also realise that data visualisation is both a science and art. There are many ways to approach a visualisation and many suitable solutions. You will often have to choose between multiple solutions that best serve the purpose of the design. Remain open-minded and don't be afraid to be creative. However, also exercise constraint and don't forget about the important role that data visualisation plays in accurately conveying the story behind the data. The following sections discuss Kirk (2012)'s guiding principles and organises his structured design process into three distinct stages as follows:

- Identifying a targeted audience and a data visualisation design objective
- Focusing, justifying and choosing methods
- Construction and evaluation

1.5.1 Guiding Principles

There are four overarching princples that Kirk (2012) considers to govern the entire design process. These principles are as follows:

- **Strive for form and function:** Form versus function (or style versus substance) relates to the perceived tension between making a data visualisation look good, but also accurately conveying the story behind the data. Data visualisation sits at a cross-roads between art and science. Go too much towards form and you risk losing statistical accuracy. Go too much

towards function and you risk losing the eye-catching impact of a visualisation. The trick is to balance the two, maybe sometimes leaning a little to one side or the other depending on your intended audience. For example James (Cheshire, 2014)'s "Population Lines" visualisation (Figure 1.5) focuses on form for his intended audience who appreciate data visualisation art. Despite the perceived tension between these two goals, it can be argued that good form can help enhance function. Making your data visualisation aesthetically pleasing is always a goal of the design process.

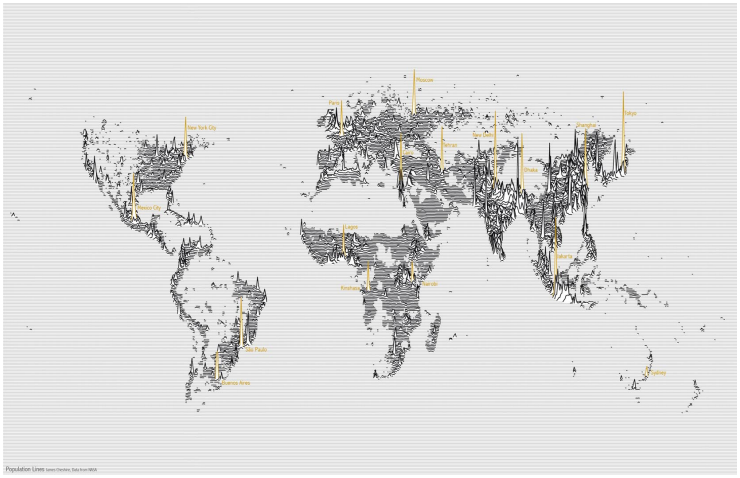


Figure 1.5: Population Lines (Cheshire, 2014).

- **Justify the selection of everything you do:** Just because you know how to create Sankey or network diagrams, doesn't mean you should, especially when a simple bar chart might do the trick. Data visualisation is no different to statistical data analysis. Every stage and decision you make along the way should be recorded, explained and justified. Even more important, it should be reproducible. If you can't explain why you are doing something a certain way, stop and think. Spending extra time thinking about what you are going to do can save you from making a lot of false starts.
- **Creating accessibility through intuitive design:** We should always aim to design intuitive and simple-to-interpret visualisations. Keep it simple, if you can, and tap into the power of people's innate perceptions and prior knowledge. There will be times when there is no way to keep the visualisation simple due to the complexity of the data or story, but under such circumstances, you must still do your very best to make the visualisation accessible.
- **Never deceive the receiver:** Intentionally or unintentionally, data visualisation can deceive the viewer or distort the story behind the data.

You should be conscious of common pitfalls of bad visualisation and vow to never intentionally use these poor design principles.

These principles of data visualisation design effectively summarise the learning objectives of this book. Memorise these principles. The content that follows will help you to abide by these values.

1.5.2 Identifying a targeted audience and a data visualisation design objective

In the first stage, you will need to identify your audience (who you are communicating with) and your design objective or purpose. With these two points clearly articulated, you will be ready to commence the next stage. If you do not complete this stage, you will risk wasting time and designing a visualisation that lacks a clear purpose and misses its mark.

1.5.2.1 Audience

Your target audience is broadly defined as who you are trying to communicate with. Sometimes your audience is as broad as the general population, similar to what a news journalist will target. Sometimes it is an audience of one, for example, during your exploratory data analysis. Regardless, you need to take your audience into account when designing a data visualisation. Consider Figure 1.6 showing Arctic sea ice area published in The New York Times by Watkins (2015). Who do you think the audience is?

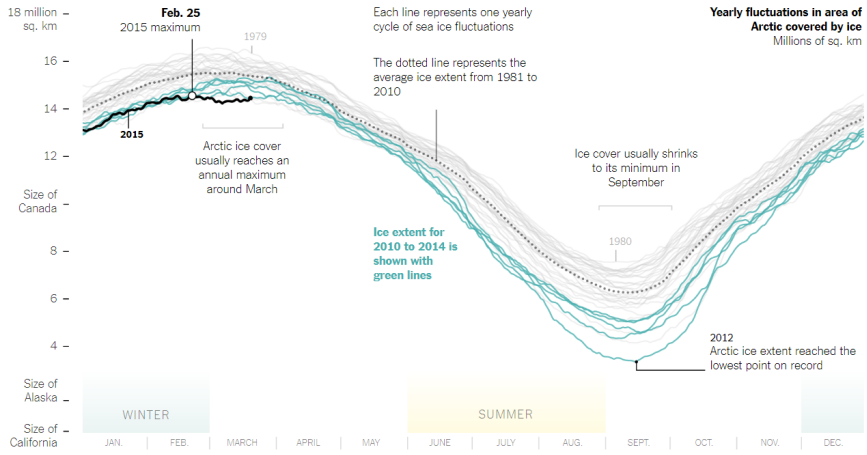


Figure 1.6: Arctic ice reaches a low winter maximum (Watkins, 2015).

The fact that it is published in The New York Times, a prestigious news publication, provides a good idea, but there are clear features in the plot that provide additional clues. Notice the use of annotations that explain key features of the visualisation, aid in interpretation and help to tell a story? The x and y axis are contexualised. We can see months of the year linked to seasons, which helps the viewer to understand arctic seasons. The y axis has reference lines that equate the area of ice to the area of well known states and countries in North America. Colour is used to help the viewer draw their attention to key points. Based on the contextualisation, easy to understand annotations and highlighting, you can assume that this plot was designed for a general audience in North America, mainly the U.S. No surprise there, but this does highlight how the intended audience of a visualisation shapes the design.

Understanding or empathising with your audience is an essential characteristics of design thinking. As Brown (2008) explains, being able to empathise with your audience means you can understand multiple perspectives, which will help you to deliver a design that meets the needs and expectations of your audience. The best way to empathise would be to talk to your audience in order to understand them. However, this is not always possible, so you will often need to make some educated guesses. Here are some things to consider:

- How diverse or broad is the audience? Are they relatively homogeneous (e.g. a group of engineers) or diverse (e.g. the general Australia population)? How do they vary in terms of age, education, and other background factors?
- How big is your audience? Small audiences allow for personalisation.
- How technical is your audience? Can we assume they understand data visualisation? Do they have subject-knowledge expertise? Do they know about statistics?
- Does the audience have any special requirements? Your audience might have colour-blindness, poor vision, cognitive impairments, English as a second language.
- How much time does your audience have? Many people consider themselves time-poor.
- What makes your audience tick? Understanding their interests and motivations can help you to engage them.

There a many more considerations, but you get the point. Take the time to put yourself in your audiences' shoes. This will greatly improve the final design.

1.5.2.2 Objective

All data visualisation must have a clear objective or purpose. Evergreen and Emery (2016) referred to this as the "so what?" question. Why does the visualisation exist? What effect will it have?. If you don't have a good answer to

this question, you need to reconsider your design. The objective of your design is different to the question or practical problem that underlies it. That probably sounds confusing, so let's apply this to the Arctic Sea Ice visualisation.

The question or problem the visualisation answers is "How is the arctic sea ice volume changing across time?" The question is readily answered by the visualisation. The viewer can clearly see that records from 2010 to 2014 are trending down compared to previous years. Question answered, but so what?

If you can answer this "so what?" question, you will understand the objective or the purpose of the visualisation. Fortunately, it was used in a New York Times article discussing climate change, so we can get a very clear idea of the objective. Even without the article you can get a clear idea thanks to the accessible design and informative annotations. Watkins (2015)'s objective was to educate and empathise the issue of climate change to their readers.

There are a multitude of objectives. A visualisation can and often will serve more than one purpose. According to Kirk (2012), broadly speaking, you can split the objective of a visualisation into two mains functions.

- **Function**
  - **Explanatory:** When the function is to explain (i.e. 'presentation-orientated techniques', Kosara (2016)) the visualisation is often carefully constructed around a narrative. Every feature of the visualisation has been carefully crafted to facilitate the telling of a compelling story.
  - **Exploratory:** When the function is to explore, a single story does not dominate and the focus of the visualisation promotes exploration and self-discovery of stories hidden in the data. Exploratory visualisations often make use of interactive features to help immerse the viewer in the data.

You also need to consider tone:

- **Tone:** A visualisation's tone refers to features of the visualisation that are used to trigger an emotive response. Pragmatic or analytical visualisations, or those used for technical purposes, are often characterised by a clear-cut design that favours precision and detail over form. On the other hand, visualisations may use abstraction and manipulate aesthetic qualities (e.g. colour) to convey different emotive tones.

Kirk (2012) provides the following phrases to give you a sense of how you might communicate the objective of your visualisation:

- Persuade; Shape opinion; Inspire; Change behaviour; Shock; Make an impact

- Learn; Increase knowledge; Answer questions; Trigger questions; Enlighten
- Conduct analysis; Monitor; Find patterns; no patterns; lookup
- Familiarise with data; Play with data
- Tell a story; Contextualise data
- Serendipitous discoveries
- Emphasize issues; Grab attention
- Present arguments; Assist decisions
- Experimentation
- Art; Aesthetic pleasure; Creative technique

Use these phrases to help you to articulate a clear objective to your design. Now that you know your audience and have a clear objective, you can move to the next stage.

1.5.3 Focusing, justifying and choosing methods

In the second stage, you will determine how to turn your objective into a deliverable. Kirk (2012) refers to this as editorial focus, or the ability to identify the salient stories behind the data and delivering a visualisation targeted to your audience and objective.

The first major challenge faced during this stage is acquiring and preparing the data for visualisation (see the later sections in this chapter for tips in locating reliable data). Most of your time will often go into this process. During this stage, you will have to source your data (a lot easier these days with the Internet), examine it for completeness and quality, familiarise yourself with the variables and data types and transform or compute additional variables. Data transformation might include tasks such as the following:

- **Parsing:** Splitting variables. For example, extracting the day of week from a date string or breaking a person's name between their first and last.
- **Merging**: Creating a new variable by combining two or more variables. For example, merging a first and last name into a full name variable.
- **Converting:** Transforming variables into new variables. For example, converting age into age bands or converting qualitative variables (Male and Female) into numeric codes (0 and 1).
- **Deriving:** Inferring a new variable from another. For example, calculating someone's age using their date of birth.
- **Calculations/computations:** Creating a new variable by performing a calculation on existing variables. For example, calculating BMI based on height and weight.
- **Removing:** variables that are not needed. For example, removing sensitive identifying information and replacing them with unique codes. Be careful though. You never know what you might end up needing.

You might also be faced with adjusting the resolution or scale of large data. Plotting millions of data points is often unnecessary. You may be faced with adjusting the resolution of your data using one of the following methods:

- **Full:** All data.
- **Filtered:** Exclusion criteria applied.
- **Aggregate:** Data are parceled up into a lower resolution (e.g. per month, year etc).
- **Sample:** Visualisation based on a randomly representative sample of the entire dataset.
- **Summaries:** Data are based on summary statistics only. There are no raw data portrayed.

Once you have completed the data stage, you should be intimately familiar with it. You should be starting to refine your focus in order to draw out the story behind the data that will meet your objective. You might begin to get an idea of the variables that you will use and the method of visualisation that will suit your audience. At this point you should re-clarify and fine tune your focus if needed before moving on.

Data visualisations are used to tell data stories (Chapter 2 will discuss data story-telling in-depth). Just as a good story teller will use well known techniques to convey their intended narrative, so too does a good data visualisation. The main techniques can be organised into the following categories:

- **Comparisons and proportions**
  - Range and distribution
  - Ranking
  - Measurement
  - Context
- **Trends and patterns**
  - Direction
  - Rate of change
  - Fluctuation/variance
  - Significance
  - Intersections
- **Relationships and connections**
  - Exceptions
  - Correlations/Associations
  - Clusters and gaps
  - Hierarchical relationships

You will be learning all about these features through the book. Purpose and editorial focus will help you to use these techniques in a compelling manner.

As you develop your data visualisation, you will be faced with many design choices, all the while, keeping our objective and audience in mind. Having a strong editorial focus and knowledge of data visualisation will help you to make the best choices. You will need to make decisions for how to best represent your data as well as how best to present your visualisation. Presentation concerns the overall look and appearance of the visualisation. This book will focus on many of these issues. As such, we will outline some of the major considerations at this point in the design process:

- **Representation:** This includes choosing an appropriate visualisation method, while taking into account the characteristics of the data, the story to be told and the audience. You also have to think about the degree of precision (form vs. function), and, at the end of the day, settle on a final solution. Keep in mind that there might be many suitable solutions and choosing the right one might come down to personal preference or the requirements of the project.
- **Presentation:** When presenting your data visualisation, there is often a lot of work to be done. You need to think about the appropriate use of colour, interactive features (manipulating parameters, adjusting views, annotated details, animation), annotation (titles, introductions, user guides, labels, captions with narratives, visual annotations, legends and units, data sources and acknowledgements!), and arrangement. Fortunately, Evergreen and Emery (2016) provide a Data Visualisation Checklist for this purpose. This checklist in discussed in a later section.

There are a plethora of data visualisation methods currently available. For example, The Data Vis Project (Ferdio, 2019) is an online list of the most common information visualisation methods. The list is exhaustive and includes many data and non-data visualisation methods (e.g information visualisation). There is no attempt to critique the limitations of each method, so use with caution. However, in terms of representing the extent of the known methods available to a designer, point proven. Tried and tested, many of these methods are the best place to start. However, sometimes, you might need to invest in an original design. The main focus of this book will be on using the existing methods appropriately for a range of data visualisation tasks. You will learn which methods are suited to your data and story, the strengths and weaknesses of different approaches and how to use open source data visualisation tools to create beautiful visualisations.

1.5.4 Construction and evaluation

The third stage is construction and evaluation. Constructing or building your design is largely done using specialised data visualisation tools. Just as there

are a plethora of data visualisation methods, there are also a plethora of data visualisation packages that can help you construct your designs. This book will focus on using open source tools because they are free and highly powerful. They also force us to be deliberate with our design choices because they require knowledge and time for coding. Many commercial packages are dangerously attractive because they are very easy to use. With very little understanding of data, statistics and data visualisation, users are able to visualise data very quickly using very powerful tools. However, this is often a recipe for disaster as the user has little understanding about effective data visualisation. They are completely reliant on the tool to guide them through the design and good data visualisation practice.

Construction should be a relatively straight-forward process if you have the necessary coding knowledge and a clear design. It will just be a case of sitting down to code the visualisation, fix errors, add features one-by-one and add any finishing touches. Once you have finished construction, you need to start evaluate the accuracy. This includes the following:

- **Data and statistical accuracy:** Double check your calculations and make sure the visualised data make sense in the context of your familiarity with the data.
- **Visualisation accuracy:** Check that your visualisation accurately portrays the data. Do not deceive the viewer.
- **Functional accuracy:** Are all the functions of the visualisation working as intended? This is more relevant to interactive or animated designs.
- **Visual inference:** Does the visualisation design facilitate correct inference from the data? Again, we want to avoid deceiving the viewer, both intentionally and unintentionally.
- **Formatting accuracy:** Check the visualisation for consistency including things like font type/size, colour and terminology.
- **Annotation accuracy:** Proofread all your labels and annotations.

Evergreen and Emery (2016)'s Data Visualisation Checklist, discussed in a following section, is perfect for this pre-publication check.

Evaluation continues after finalising your design. This can include small (clients, focus-groups, mum) or large group feedback (social media, data visualisation blog, comments from a news site). These methods may include the following:

- **Metrics and benchmarks:** Google analytics, Tweets, Facebook likes, Google +1, hits on the visualisation's webpage etc.
- **Client feedback:** Direct feedback from stakeholders.
- **Peer review:** Feedback from other visualisation experts.
- **Unstructured feedback:** Comments or emails from viewers.
- **Invited user assessment:** Use a form or survey to receive structured feedback.

- **Formal case study:** A usually independent, academic evaluation of the visualisation written into a report.

And let's not forget about self-reflection. Ask yourself the following questions:

- Did you accomplish your objective?
- Did it have an impact?
- Did you meet the needs of your audience?
- Did you create something you were satisfied with?
- Were you satisfied with how you justified your choices?
- Did you enjoy the work? Was it rewarding and worth the time?

Evaluation will ensure you continue to learn and develop your data visualisation skills.

Kirk (2012)'s design approach provides a comprehensive and useful overview of the data visualisation design process. Who would have thought that there was so much to developing a data visualisation?

1.6 Trifecta Check-up

Fung (2014), from the excellent blog Junk Charts, provides a very simple and powerful framework to use when quickly evaluating a data visualisation. This framework, named the **Trifecta Check-up**, is useful for both evaluating your own work and the work of others.

Fung's framework helps you to explain why a particular data visualisation works, fails or falls somewhere in between. Being a "tri"-fecta, there are three questions that we use to evaluate the visualisation as shown in Figure 1.7.

- (Q) What is the question?
- (D) What does the data say?
- (V) What does the visual say?

All three questions should result in the same answer. Any discordance between two of these questions results in a poor visualisation. Let's consider an example of a good and not so good data visualisation.

Figure 1.8 shows a data visualisation by Stiles (2016) which considers "How Popular is Your Birthday?". You can see the interactive version here.

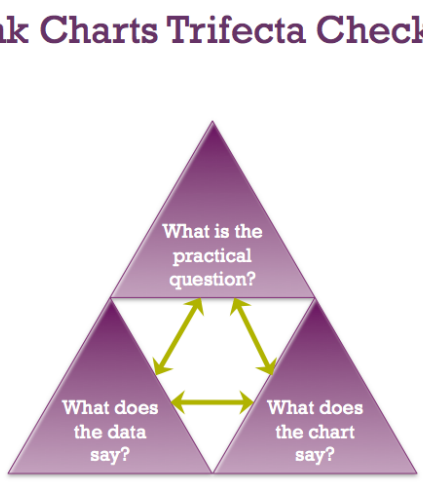


Figure 1.7: The Trifecta Check-up (Fung, 2014).

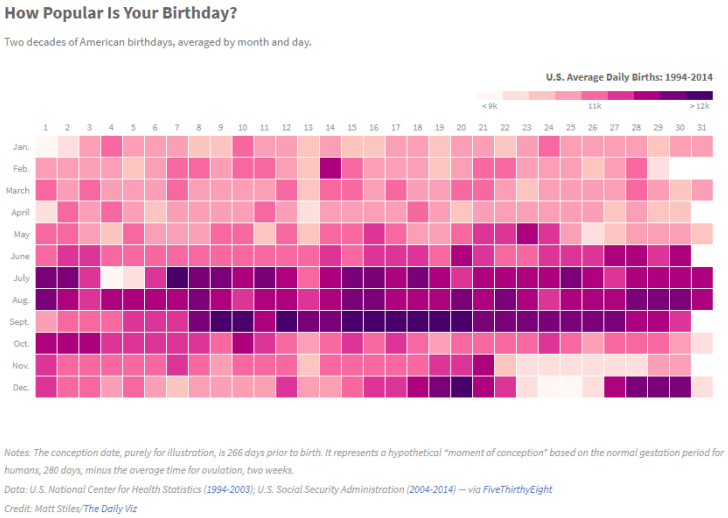


Figure 1.8: How Popular Is Your Birthday? (Stiles, 2016).

1.6.1 (Q) What is the question?

All data visualisations aim to answer a question using data. Without a question, a data visualisation doesn't really have a point. Therefore, the Q in the trifecta sits at the top of the check-up. We use the question to evaluate the other two questions.

The "How Popular is Your Birthday" visualisation answers a clear question, and many would agree (particularly those from the US), an interesting question. The answer to the question is likely to appeal to a wide audience and the data are sufficiently complex to be aided by a data visualisation. The ability to address this question with a data visualisation is a good objective.

1.6.2 (D) What does the data say?

You can have a really good question, but fail to find the right data to answer the question. The D of the Trifecta check-up ask whether the data presented addresses Q. This often requires a designer to make many decisions about what data to use, how to clean, aggregate and transform data ready for visualisation. During the data stage, the decisions made by the designer will determine the success of the visualisation. The viewer must be able to connect the data with the question and be assured of its quality. If the data doesn't connect with the question or questions are raised about the source or quality of the data, the data visualisation may fail the data question of the check-up.

Looking back to the "How Popular is Your Birthday" example, the data visualisation includes annotations that reference the context of the data (U.S. daily birth rates from 1994 - 2014). We can read the source of the data was the U.S. Census data (generally very reliable), and also a note how birth rates were transformed to reflect the average between 1994-2014. The data also directly relates to the question.

1.6.3 (V) What does the visual say?

You can have a question and good data, but unless you can visually communicate the answer using an effective visualisation method, a data visualisation may fail the V of the check-up. Again, there are many ways to visualise the same data, some will be excellent, some will be OK and some that will be plain wrong. The challenge for the designer will be to link an appropriate method with the type of data and the question being addressed.

"How Popular is Your Birthday" does an excellent job of answering the V question. A heat map, with days of the month on the x axis and month on the y axis, presents a familiar, almost calendar-like, grid. A discrete colour scale is used to visualise the magnitude of the average birth-rate for a particular day. While

colour scales lack visual accuracy, the viewer can still glean the high density of births between July and September. These correspond to conception times in cooler months of the year and during the Christmas holiday period. The interactive version of the plot has a hover-over effect where the viewer can read the actual average values. Overall, this data visualisation brings a whole new meaning to the Christmas holiday period in the U.S.

1.6.4 Failing the Trifecta

Figure1.9 shows an example of a not so great data visualisation.

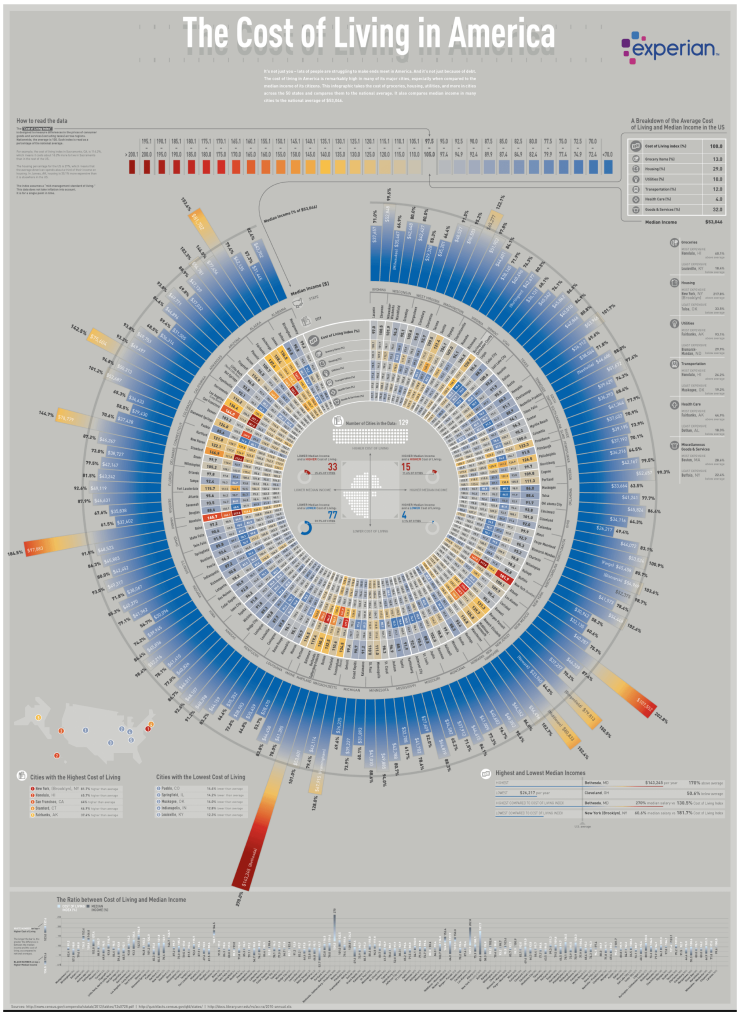


Figure 1.9: The cost of living in America (EDQ.com, 2019).

The previous polar bar chart from EDQ.com (2019) (click here to see full-sized image), entitled "The Cost of Living in America" is an example of where the D and V in the Trifecta check-up needed improvement. The Q is clear and interesting. Visually it looks impressive. The source of the data is the U.S Census, so quality is not a concern. Where this data visualisation starts to fail is in the complexity. There is a lot going on. So when trying to link the question with the data, the visualisation doesn't lead to quick and powerful insight. It takes a lot of time to scan the visualisation and take into account the large amount of information. It could be improved by focusing first on the main story and using sub-plots to dig deeper into a break down of living costs and median income. The aggregation also needs to be reconsidered. The massive polar bar chart breaks the statistics into cities within states. Is this really required to answer the question?

In terms of V, I believe there are a few major issues. First, it is very visually "busy". The eyes are drawn to multiple features. The brain isn't quite sure where to start. Much time is needed to decode the features and put them back together again before interpretations can be made. I can't help but feel that the complexity could be avoided by presenting more aggregated data first and then by drilling down into detail in smaller sub-plots.

Another issue is the polar coordinate system. The visualisation is a simple bar chart, but, perhaps in order to fit on the page, the coordinate system was changed to polar (circle). The issue with this coordinate system is that it makes visual comparisons very difficult because the x and y axis are curved and therefore, the baseline and top of the bars never align to a common scale. The actual values are reported inside the bars, but this defeats the purpose of the visualisation if we rely on the actual values to make comparisons. We could just report a table of values instead.

The bars are also organised in alphabetical order. Does this make sense? Visually it results in a random scattering of bar heights that aesthetically looks pleasing. However, it doesn't aid in rapid comparison or identification of states and cites with low or high incomes/costs of living. Ordering the cities lowest to highest by one of these variables would greatly improve the visualisation. However, I expect that doing so would make the polar bar chart look unusual and less appealing.

It's not all bad. As I said, it's visually impressive and draws the viewer in. While I expect most people to scratch their head for 5-10 minutes taking this in, if it gets people thinking about wages and costs of living across the country, that's a good thing. I just wonder if this could be achieved using a better focused visualisation.

1.6.5 Critiques

According to Kaiser's Trifecta check-up, there are eight possible critques for a data visualisation.

Type	Description
Q	Poor question
D	Poor data
V	Poor visuals
QD	Poor question and problematic data, but visuals are OK.
QV	Data are good, but question and visuals are out of sync
DV	Good question, but issues with data and visuals
QDV	The data visualisation fails everything
Trifecta	Q, D and V are in sync. Good data visualisation

1.7 Publication Ready Data Visualisations

How do we know if a data visualisation is ready for sharing and publication? There are many things to consider outside the obvious like spelling and attributing (referencing) your data. Fortunately, Evergreen and Emery (2014) and later Evergreen and Emery (2016) developed a very helpful data visualisation checklist for this purpose. Using this check-list, you can avoid many of the common issues found in data visualisation and greatly improve the presentation of your plots. The Updated Data Visualisation Checklist can be downloaded and used to rate a visualisation. There is also an online version of the tool. Visualisations scoring between 90 and 100% are said to be well-formatted. The check-list is summarised here, but please refer to the full checklist which is freely available for further detail. The checklist is broken into five main sections: Text, Arrangement, Colour, Lines an Overall

1.7.1 Text

Data visualisations use text sparingly. Just enough to get the job done. It should not detract or draw attention away from the data. Specific considerations include the following:

- Descriptive title that answers a "so what" question. This helps the viewer to take away a key message and contextualise the visualisation. Evergreen and Emery (2016) suggest that the title be top and left justified.
- Subtitles and annotations are used to provide important explanations and to highlight key data points.

- Text size is hierarchical. Important text (e.g. titles) is larger than less important text, such as axis labels.
- Text is horizontal. Vertical text is hard to read. You can re-orientate plots if you are having trouble fitting your labels.
- Data are effectively labelled. Try to avoid legends if you can because reading back and forth is difficult. If you use legends, get them as close as possible to the data (e.g. embedded within the plot assuming it doesn't overlap other elements). Ideally, locate data labels next to data points.
- Use labels sparingly. For example, in a time-series, you don't need to label every date. Avoid numeric labels if the plot already has a y-axis scale.

1.7.2 Arrangement

The arrangement of a data visualisation's elements, including axes, scales, ordering and non-data elements (e.g. graphics) must make the visualisation easy to interpret. Specific considerations include the following:

- Proportions are accurate. For example, a bar representing 50% is twice as big as a bar that is 25%.
- Data are ordered in a meaningful way. For example, in a visualisation of the highest number of page visits for a website, the pages are ordered from highest to lowest, instead of a default ordering such as alphabetically.
- Axis intervals are distanced equally. Breaks in an interval must be carefully noted.
- Graph is two-dimensional. Avoid 3D plots, shadows, bevels etc.
- Avoid decorations such as images and graphics. Icons that support interpretation may be used.

1.7.3 Colour

Colour is used with purpose and carefully selected. Specific considerations include the following:

- Keep colour associations in mind. For example, red is the colour of love and violence.
- Use colours to highlight and draw attention to the key message. Other elements are deemphasised (e.g. less-saturated colours, transparency added).
- Be wary of using colour if your plot has to be printed in black and white.
- Where possible, avoid red-green combinations for those with the most common form of colour blindness.
- Text and data elements are sufficiently contrasted with the background.

1.7.4 Lines

Avoid using excessive lines in your plots because they cause clutter. Specific considerations include the following:

- Grid lines are muted into the background (e.g. grey). They are visible just enough to help interpretation, but do not detract from the data. If numeric value labels are included, grid lines are not needed.
- Remove graph borders.
- Axes use tick marks and axis lines sparingly. Avoid axis lines where possible.
- Avoid secondary or dual axis plots.

1.7.5 Overall

A good visualisation will attract attention, be focused on the data, and have a clear purpose. Specific considerations include the following:

- The data visualisation answers a practical question. Comparison data are used effectively to provide context and meaning.
- The data visualisation type is appropriate for the data and the question being asked.
- The data visualisation has an appropriate level of accuracy. For example, do you really need four decimal places for your numerical variable?
- All data visualisation elements work together to succinctly and effectively convey a takeaway message.

The Data Visualisaiton Checklist is a comprehensive tool to keep in mind when building and finalising your visualisation. Following these guidelines will help you avoid many of the most common issues. However, many of these guidelines require in-depth discussion and extensive knowledge and experience to apply. That will be the objective of the remainder of this book. There is still much, much more to learn.

1.8 Ethical Principles

The Cambridge English Dictionary defines ethics as follows:

> "the study of what is morally right and wrong, or a set of beliefs about what is morally right and wrong"

Many professions codify ethics as set of rules or guiding principles that govern the professional practice and behaviour of its members. For example, the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (National Health and Medical Research Council et al., 2018) provides guidelines for human research performed in Australia. Another examples is the Media Entertainment & Arts Alliance (MEAA) Journalist Code of Ethics (Media Entertainment and Arts Alliance, 2019). Most peak professional bodies include a code of ethics. However, data visualisation lacks a unified set of ethical guidelines most likely because it is practiced widely across many professions and is informed by numerous disciplines (Correll, 2019; Skau, 2012). As such, ethical guidelines in data visualisation need to be informed by other professional codes.

It is important to note that a code of ethics is not the same as the law. Ethical guidelines are an agreed upon set of standards that help us to determine if what we are doing is right or wrong relative to a professional. If we breach a code of ethics, it doesn't mean we have broken the law. For example, a guiding principle of many journalism code of ethics is fairness, or the idea of presenting both sides of a story. If the journalist doesn't fairly present both sides of the story, will they have broken a law? Probably not. Will they be sued? Possibly, but it would depend on the circumstances (e.g. defamation). Another guideline for journalists is not to plagiarise or copy the work of others. Plagiarism is a breach of the Australia Copyright Act 1968 which can lead to criminal charges or civil action. Therefore, plagiarism is both an ethical and legal issue.

Ethical guidelines are not black and white. If we all agreed on what is right or wrong and this never changed or depended on context, ethics would not need to exist. Therefore, ethical codes are often referred to as guidelines. They help point us in the right direction and help us to critically think about the morality of our practice and behaviour. However, they do not always provide a clear answer and sometimes guidelines can be contradictory. For example, journalists are required to attribute the sources of their information, but also protect confidential sources if the need arises. There will always be situations where the ethics are unclear. In these situations, it is always a good idea to discuss the issue with someone more experienced or qualified to advise (many professional or ethical bodies have ethical advisors) or with someone who is considered to have a high level of integrity. This brings us to the next definition.

According to the Cambridge English dictionary, integrity can be defined as follows:

> "the quality of being honest and having strong moral principles that you refuse to change"

Integrity is a personal quality and something that we must strive for in our professional careers and in data visualisation design. Without integrity, our

data visualisations risk losing their credibility, deceiving our viewers, misrepresenting facts, and failing in their objective. Following ethical guidelines is a good way to practice integrity, but integrity is much broader than ethics alone. Integrity applies to the entire design process from choosing data, being transparent about our decisions, producing reproducible work, following guidelines for the responsible use of data and giving and responding to feedback.

Without a dedicated set of ethical principles for data visualisation, we need to draw upon the ethical standards of other related fields and experts (Skau, 2012; Cairo, 2014; Correll, 2019). In the following sections, the major ethical principles that apply to data visualisation will be briefly outlined followed by the key considerations of data integrity.

The principles outlined below were adapted from the following codes of ethics:

- NHMRC National Statement on Ethical Conduct in Human Research (National Health and Medical Research Council et al., 2018)
- MEAA Journalist Code of Ethics (Media Entertainment and Arts Alliance, 2019)
- Engineers Australia Our Code of Ethics (Engineers Australia, 2019)

You might find the inclusion of an Engineering code of ethics a little odd. However, Cairo (2014) rightfully suggested that engineering ethics are relevant to data visualisation because the designer must be technically competent to deliver a solution that balances efficiency and effectiveness.

The following principles are a work in progress. The discussion of each principle is kept relatively brief. The goal of this section is to help you reflect more critically on the data visualisation design process from an ethical perspective. The principles are not mutually exclusive and upholding one principle might raise issues in others. You won't always get it right, or find the perfect outcome to an issue. However, you can commit to taking the time to reflect and always strive to do better. Therefore, the overarching principle of an ethical mindset is integrity or being honest about our shortcomings.

1.8.1 Beneficence

Your data visualisation must serve a valuable purpose by succinctly and accurately representing data in a way that leads to new knowledge and better decision making. Creating visualisation that don't have a clear purpose, use unreliable data, misrepresent the truth, and deceive or confuse the viewer can be said to be maleficent. Maleficent data visualisations can be trivial data visualisations which waste peoples' time or dangerous data visualisation that misrepresents the truth (e.g. fake news). It might be hard to think that a data visualisation can be immoral (Skau, 2012), but the flaws of humanity never disappoint.

Consider the following data visualisation taken from Wikipedia (2009) shown in Figure 1.10. This choropleth map visualisation is based on data from the book *IQ and the Wealth and Poverty of Nations* by Lynn and Vanhanen (2002). Each country includes a colour which represents the proposed average IQ of that nation. The authors conclude from their analysis that intelligence, measured using IQ, is a major determinant of national wealth around the world. You can see obvious indications of this in Figure 1.10, where the continent of Africa, which historically has the lowest wealth as measured by GDP, is dominated by colours indicative of low IQ. However, the work of Lynn and Vanhanen (2002) has drawn heavy criticism. Palairet (2004) pointed out substantial statistical issues with the estimates of country IQ as only 81 out of 185 countries included in the analysis had estimates available. Missing data were imputed by using IQ data from adjacent countries. Furthermore, Ervik (2003) writes that the book fails to establish the reliability and cross-cultural validity of IQ test scores and control for other important factors, not to mention the fact that correlation does not equal causation. For example, richer countries invest more in education which is also highly correlated with IQ.

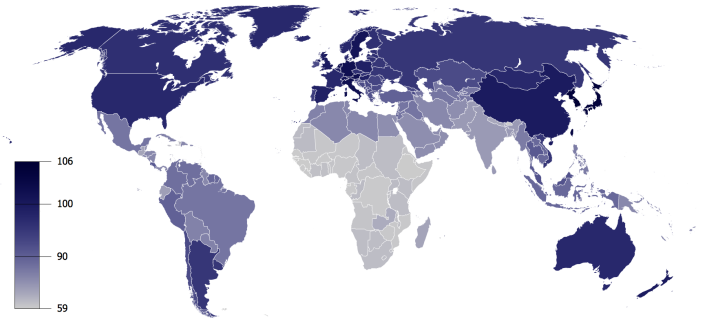


Figure 1.10: IQ by Country (Wikipedia, 2009) based on data from Lynn and Vanhanen (2002).

Figure 1.10 presents data devoid of the critical discourse contributed by the book reviewers. Viewed in isolation it presents questionable data and a potentially dangerous idea. It is not hard to see how predjudice would shape viewer's interpretation. Fortunately, other researchers have developed much better explanations of wealth and economic development. For example, Hausmann et al. (2011) show that economic complexity, or the more diverse and specialised jobs are in a given country, the higher a country's income. Complex economies require citizens with a high level of collective knowledge which might explain variability in the average IQ of nations. Highly complex nations are more likely to invest in knowledge through education and skilled migration which would translate to higher average IQ when compared to nations with less complex economies. Our World in Data (2016) show this relationship in Figure 1.11 based on 2016 data. The higher a country's economic complexity (lower ranks

indicate higher complexity), the higher the GDP per capita.

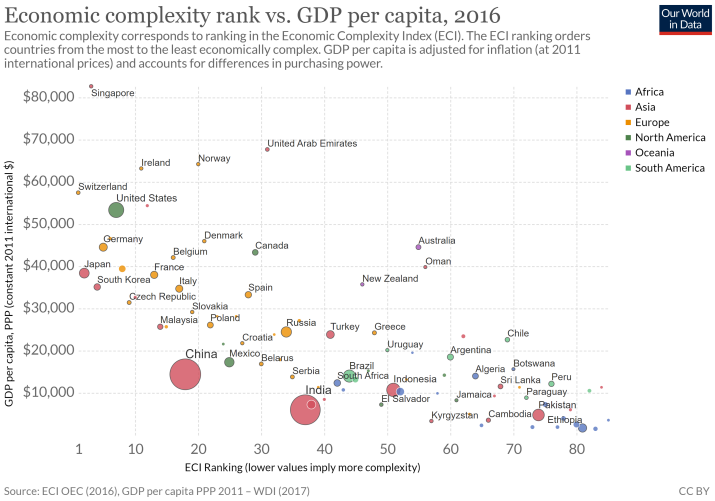


Figure 1.11: Economic complexity rank vs. GDP per capita, 2016 (Our World in Data, 2016).

1.8.2 Transparency

Correll (2019) referred to transparency as making the invisible visible. Many stages of data visualisation are not visible to the viewer. The viewer sees the "end product" and has little insight into data collection, data preprocessing, the numerous designs tested (or ignored) and the technology used to generate the visualisations (Correll, 2019). This means the viewer places a high degree of trust in the designer. In order to earn this trust you should always document and be able to explain and justify these invisible steps. You need to prove you can be trusted. Sharing your data visualisation project code, datasets (assuming you have permission) and correctly attributing external data sources promotes transparency. This will allow others to verify your data and designs, reproduce your work, and produce alternate visualisations. Transparency and reproducibility are cornerstones of scientific research, and data visualisation should be no different. There are limitations to transparency. Depending on your data and topic, you might be restricted in what you can disclose. However, you should still keep this principle in mind because it will help you go back over your work at a later time and allow you to show technical details to others who have permission

1.8.3 Accuracy

Accuracy refers to the the overall validity of our data visualisation. It relates to the quality of the data used, the rigour of data preprocessing and statistical analysis undertaken, and the method and choices used to represent the data. It also extends to the choice of variables visualised and how they relate to the objective of the visualisation. Your designs must be able to be verified and withstand critique. Where limitations are present, you raise the caveats and avoid overstating the findings.

Statistics teaches us methods of expressing uncertainty using probabilistic frameworks. Uncertainty is often ignored in data visualisation and Correll (2019) and Brodlie et al. (2012) reminds us that as designers we need to do a better job of visually representing uncertainty inherent in measurements, sampling and statistical estimates. This is a challenge because it usually involves adding complexity and risks confusing our viewers, especially in lay audiences (Spiegelhalter et al., 2011). However, there are many examples of data visualisation where including depictions of uncertainty improves viewers' understanding of a situation. A good example is visualising the path of hurricanes. Figure 1.12 shows the projected path of Hurricane Sandy which swept across the the U.S. in 2012 (Huffington Post, 2012). The shaded areas surrounding the hurricane's line presents the uncertainty at the time about it's forecasted path. Visualising this uncertainty was crucial for informing the public so they could take action to stay safe if they were in vicinity of the projected path.

1.8.4 Objectivity

When designing data visualisations we need to be aware how our own personal expectations, biases and experiences can shape decisions and design. For example, using an easy to access dataset might bias data visualisations because the convenience of accessing the data means it leads to an over-representation of data visualisations using the same source. All data has some degree of bias or limitations, so data visualisation can propagate a bias. Objectively, you should match the best data needed to achieve your objective and minimise sources of bias such as convenience. You should also maintain an open, but sceptical mindset. Data visualisation can often uncover unexpected results and you need to be careful not suppress them because they don't fit with our preconceived ideas. We also need to be equally careful to question and validate outcomes that fit with our expectations. Research has found that we are far less critical of facts that fit with our world view which is a phenomenon known as confirmation bias (see Nickerson, 1998). Bias can creep in during all stages of our designs. For example, removing outliers because they ruin the appearance of a plot, removing subgroups of data because it doesn't fit well with the story you want to tell, or failing to explain important context behind the data that will impact the

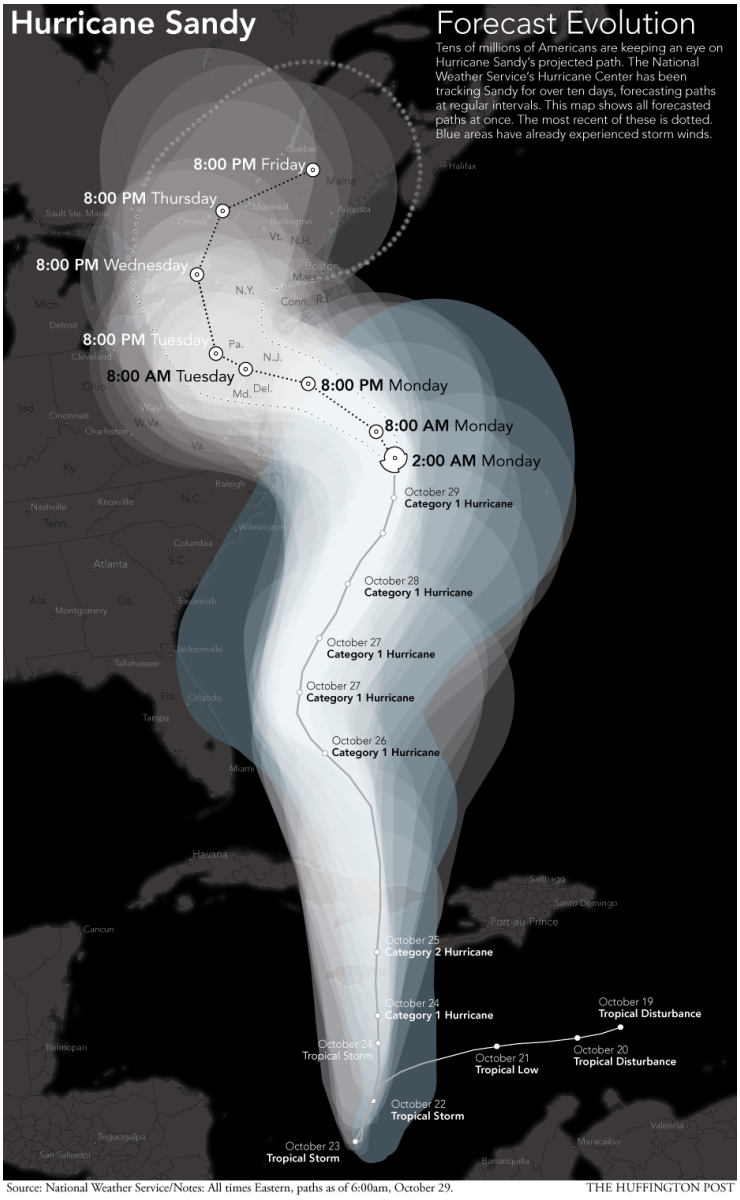
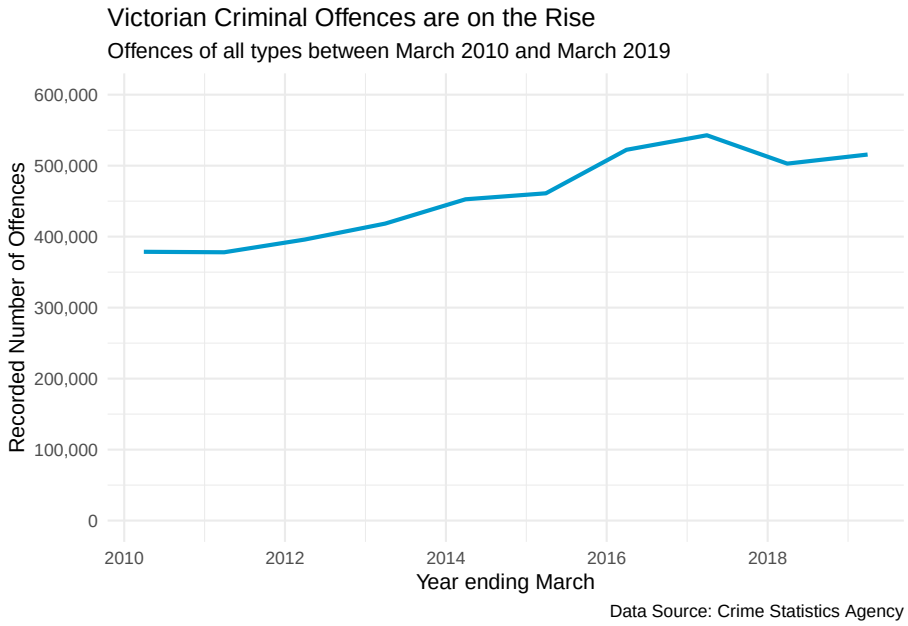
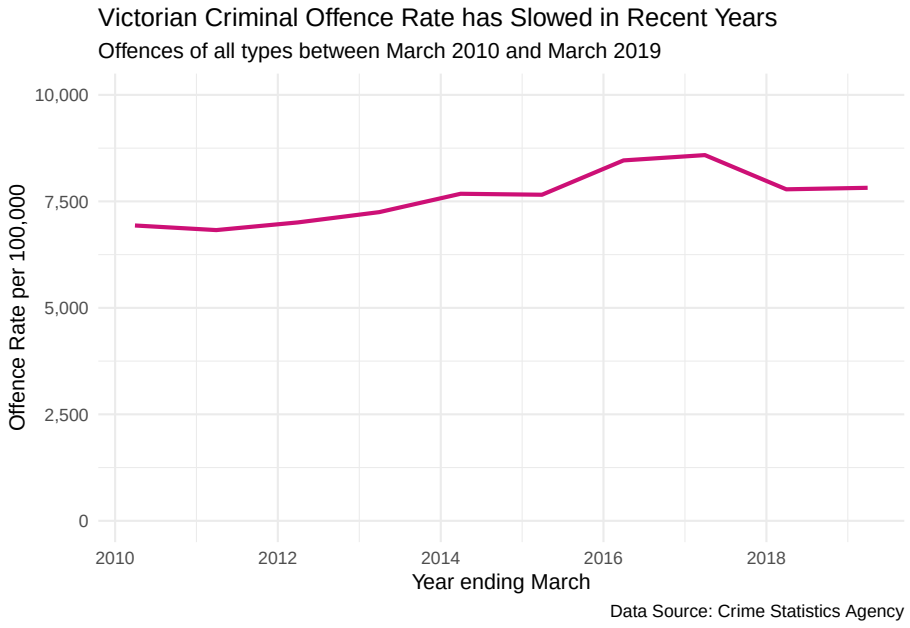


Figure 1.12: Hurricane Sandy's project path (Huffington Post, 2012).

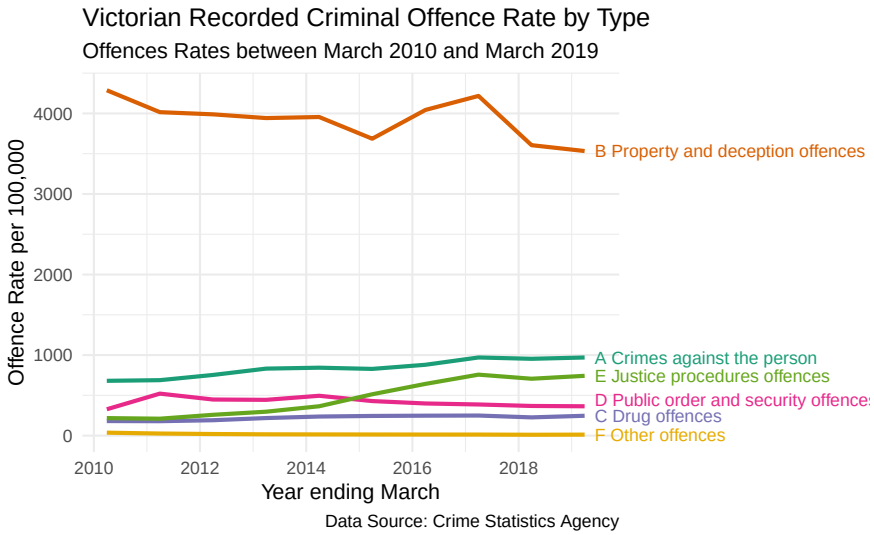
viewer's interpretation. Consider the following data visualisation of criminal offending in Victoria between 2010 and 2019 (Crime Statistics Agency, 2019). Now assume you have just read a news article stating that Victorian crime is out of control, the usual rhetoric you get from politicians. If this fits with your experience and belief, you might not think critically about the following plot. Crime is clearly on the rise.



Now what if I told you that Victoria and Melbourne are undergoing massive population growth and the increase in the number of criminal offences can be partially explained by an increase in population. Instead, we can look at the criminal offence rates per 100,000 Victorian residents. Does this change the story? The change doesn't appear as drastic and since 2017, the crime rate has remained steady. It does present a different, and a more objective side to the story.



Thinking again, is it even useful to look at overall crime? Wouldn't it be more accurate and objective to look at this problem by considering major crime types...



Now this plot paints a far more informed and balanced picture of Victorian crime. We can see what are the most common offence types and which types have changed across time. Better yet, we have controlled for population growth by standardising the offence rate to 100,000 people.

1.8.5 Respect

When practicing data visualisation you need to respect your position of power, the rights of others and the law. We have already looked at examples of how unethical data visualisations can be used to misinform others in order to promote ideological and political agendas. Data visualisation has power because it can present very powerful ideas succinctly and accessibly. For example. studies have shown that the mere presence of a data visualisation can add instant credibility to information being presented about the efficacy of medication (Tal and Wansink, 2016).You must respect that power and do your best not to abuse it, especially when your audience might lack the the knowledge and training to critically interpret a data visualisation. You must also be aware that other people may use your data visualisations in unintended and unethical ways. You must commit to respecting the rights of individuals and the law, especially privacy and copyright. Your designs must avoid bias towards others especially in respect to ethnicity, religion, gender, age, sexual orientation, or disability. This doesn't mean to avoid these topics. In fact data visualisation is a powerful way to draw attention to many issues of discrimination (see the gender pay gap visualisation from Kommenda et al. (2018) in Figure 1.13 for an example). However, when dealing with sensitive topics, we need to be especially careful so as to avoid contributing to the problem.

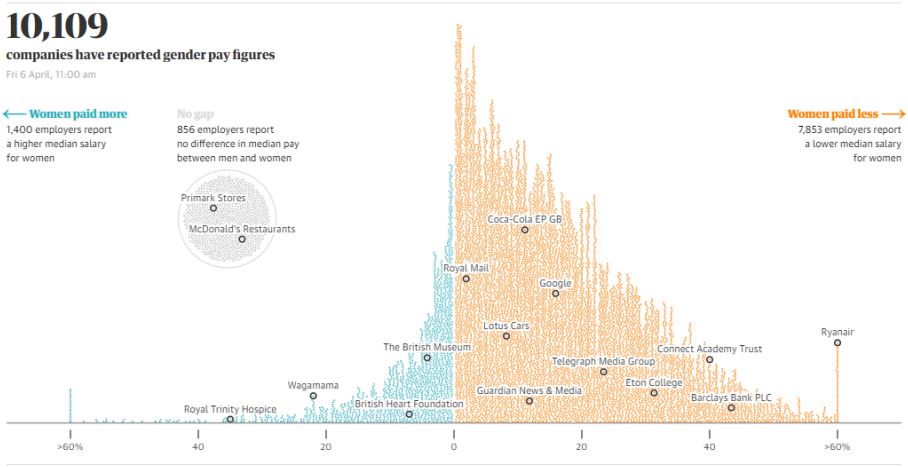


Figure 1.13: Gender pay gap by Kommenda et al. (2018).

You must also respect your clients (who you are designing for), the work of your colleagues and other designers, students (who are still learning) and your audience. When providing feedback or critique, avoid ridicule. Make helpful suggestions and try to understand other peoples' perspective (even if they are wrong). It is OK to disagree, but do so in a dignified and respectful way.

1.8.6 Accountability

You are accountable for your designs. You take credit where credit is due, and you are responsible when you make a mistake or do not achieve your objective. You strive to always improve and continue learning. When you are doing something outside your area of expertise or experience, you take steps to learn the required skills, seek supervision from someone qualified and get feedback from experts.

An outstanding example of accountability was published by Leo (2019) from *The Economist* (see Figure 1.14). The article discusses examples of improving visualisation practice at *The Economist* by looking back at previously published plots, explaining common issues and designing an improved version. The article is a rare introspection into the practice of a data visualisation powerhouse.

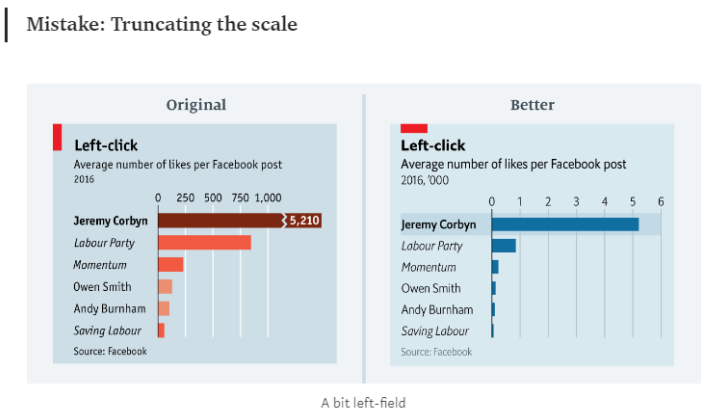


Figure 1.14: The Economist learning from their mistakes (Leo, 2019).

1.9 Data Integrity

Data are powerful and must be used responsibly. Our data visualisations rely on it. As the old saying goes, "Garbage in, garbage out." Scandals in the misuse of social media data also remind us that we need to improve knowledge about the responsible use of data, especially around privacy and consent. A recent example

was the Cambridge Analytica scandal that was credited to assisting the election of U.S. President Donald Trump in 2016 (Cadwalladr and Graham-Harrison, 2018). Cambridge Analytica took Facebook user data from 50 million Americans to help manipulate voters. Facebook was aware of data being harvested in 2015, but failed to notify users or take steps to retrieve the data. In another example, fitness social media app, Strava, published a massive heatmap of its users GPS fitness data in 2018 (Blue, 2018). It wasn't long until major security and privacy concerns were raised. Viewers were able to see GPS data from users in military bases, Government buildings and other sensitive areas. The heatmap was so detailed, it was possible to potentially identify individuals. These two examples highlight illegal and unethical misuse of data that are becoming all too common in today's data-driven world. You need to help put a stop to this lack of integrity by understanding how to use data responsibly. The following sections will discuss the major issues surrounding the use of data and what you need to keep in mind.

1.9.1 Permission

The first thing you need is permission to access and use data for the purpose of visualisation. This might be simple to ascertain, for example, the data come from your workplace and you require it to complete your job. Sometimes it is not clear. For example, you do not automatically have permission to use data published on a website. Check the website's policies or whether the data have a license for reuse. For example, many sites that publish data have a Creative Common's License that will clearly outline how the data can be used and shared. If you cannot find any information on a licence for the published data, contact the site and seek written permission. Some sites ask you to submit requests for permission to access data sources. This allows the data owner to audit and control access. You might be asked questions about your identity, who you represent, what do you intend to do with the data, how you will store it, and who else will have access. Be truthful or you might risk violating a policy or license which can have legal implications. Sites will also sell data which effectively buys a license. It can be a bit of a minefield understanding licenses and permission. The important thing is that you take reasonable steps to verify permission before you start a visualisation.

1.9.2 Security

Once you have your data, you are responsible for security if the license or conditions of use requires it. Again, this might be a simple task. For example, working on company data, using company computers and servers. However, can you copy the data to a portable drive and work on it at home? Maybe, maybe not. Get permission. What if you lose the portable drive or someone

steals it? Are the data encrypted and password protected? Is your password secure? Accessing data remotely using databases is a more secure. However, what if your computer at home is compromised and your data are stolen? What if your computer hardware fails and the data are destroyed? Do you have a back-up? How long will you retain the data after a project is complete? Security is your responsibility.

1.9.3 Consent

Informed consent is a complex ethical issue that relates to an individual's voluntary permission to collect, use or disclose their personal data. Consent is needed prior to the collection of data. Consent must be informed. Informed consent is when the individual providing consent is fully aware of the purpose and risks associated with collection of their data and has the capacity to make an informed decision (National Health and Medical Research Council et al., 2018). For example, obtaining informed consent from minors and people with cognitive impairments often requires consent from a guardian. Consent is still relevant for previously collected data used for a secondary purpose (i.e. a purpose that wasn't explained to the individual when consent was first gained). If it is reasonable to assume that an individual would consent to the secondary use of data and the data are anonymous, consent can sometimes be assumed. The Cambridge Analytica scandal is a relevant case study here. Facebook requires users to sign a user agreement, which outlines how Facebook will collect, store and use user data. Users must agree to this before they can access the site. Facebook will claim that they use this data for operational purposes, for example finding friend connections, making content recommendations and targeted advertising (revenue). Most users consider this a reasonable way to use their data for access to Facebook's powerful social media services. However, Cambridge Analytica harvested that data for a different purpose (political) without the user's knowledge and Facebook sat on the information. Therefore, the use of Facebook data by Cambridge Analytica did not have informed consent, nor was it reasonable to assume that the users would ever consent to the use of their data in this way.

You might not think that informed consent is relevant to you. But think again. What if you were the analyst at Cambridge Analytica that was tasked with mining this data for political gain? Would you have questioned your manager about permission and consent? You might not worry too much because it is the company that is taking the risk. But think again. Unauthoried use of personal data is considered breaking the law in many countries (including ones that you can be extradited to). These laws typically come under privacy legislation such as the Australian Privacy Act (1988).

Serious concerns have also been raised about the incomprehensible nature of privacy policies used by many organisations that routinely collect user information. Litman-Navarro (2019) analysed and visualised the features of 150 major

tech and media company's privacy policies. The policies were often very long, complicated and filled with legal jargon. Litman-Navarro (2019) suggests that many companies might not be sufficiently informing its users and therefore, failing to achieve informed consent. This is irresponsible and unethical.

1.9.4 Privacy and Sensitive Information

Most countries have privacy laws that aim to protect personal and sensitive information about individuals. In Australia, the Privacy Act (1988) defines personal information as follows:

> …information or an opinion, whether true or not, and whether recorded in a material form or not, about an identified individual, or an individual who is reasonably identifiable.

Examples of private information includes names, health records, phone numbers, finance information, and internet usage data etc. Anonymous data are not private. Companies often need to collect private information for the purpose running their business. For example, hospitals need medical histories to ensure their patients receive proper care. With the consent from an individual, this information can be collected, stored, used and disclosed in accordance with the Privacy Act. Companies that come under privacy legislation in Australia have to abide by a set of privacy principles that relate to transparency of data collection and management, the right to anonymity (if practical), use and disclosure of personal information, maintenance of data, data security and the right of an individual to correct information (Office of the Australian Information Commissioner, 2019). Not all private information is equal. The Privacy Act 1988 has even more stringent rules about the use of sensitive information such as health, ethnic origin, political opinions, religious beliefs, sexual orientation, and criminal records.

The current globalised environment allows data to flow effortlessly between national borders, resulting in complications when trying to understand data privacy laws (Svantesson, 2016). To address possible privacy law issues, we have to be aware of how the data has been sourced. 'The right to be forgotten' case involing Google Inc. against the Spanish DPA is a great example. As the Advocate General Jääskinen (Judgment of the Court (Grand Chamber), 2014) explained; Google Inc. main offices are in California, USA and have subsidiaries across Europe. The office in charge of processing data in Europe is based in Ireland and has data centres in Belgium and Finland. Google Inc.'s subsidiary in Spain provides only support for business and advertising services. If a case of data misuse by Google Inc. occurs in Spain, which privacy laws apply? Who is responsible? Confused? You are not alone. You don't need to be an expert in data privacy laws in all countries; however, it is your responsibility to raise possible issues and get appropriate legal advice to evaluate the situation.

Today more than ever, privacy is under threat due to the ease at which personal information can be collected using technology and the internet. When visualising data, ensure you do not violate privacy laws. Ensure your designs do not use data that can be used to identify or potentially identify an individual where you are obliged to protect their privacy.

1.9.5 Data Quality

The quality of a data visualisation can only be as good as the data source. As Cairo (2014) explains:

> "Stories are sometimes built without assessing the quality of their sources or applying proper reporting and analysis methods. This can lead to disastrous results. (p. 26)"

Take the time to locate and identify quality data sources. Here are some tips:

- Data taken from primary sources are more reliable than secondary sources. Primary sources are those that originally collected the data. If another individual or organisation republishes the data, it is a secondary source. Don't be lazy. Track down the original source and confirm the data for yourself.
- Use reliable sources which can be trusted. Reliable sources have the following characteristics:
  - Who collected the data (qualifications and authorisation)
  - Provide clear details on how and when the data were collected including sampling
  - Disclose potential conflicts of interest
  - Use quality control processes
  - Data have been collected in an ethical way
  - A data dictionary has been provided to help users understand the dataset.
- Be wary of sample size.
- Use up to date data or data relevant to your problem.
- Use variables that have known reliability and validity.
- Check missing values. If there appears to be a lot, make sure you understand the reasons before using the data. Some degree of missing data are expected, but a quality data source will provide details.

Getting access to data has never been easier. You are spoilt for choice. However, when searching for data, keep the above guidelines in mind. Not all sources of data are equal. For example, data taken from the Australian Bureau of Statistics, or most other official statistics agencies, is the gold standard of data collection. Data are collected in a routine and systematic way, rigorously processed

and quality checked, and summarised into topic articles with extensive technical detail. Contrast this with data derived from Kaggle. Kaggle is a online community of data scientists where users can post data, usually scraped from websites (therefore by default Kaggle datasets are a secondary source), and compete in prediction challenges. Kaggle is a massive repository of interesting and diverse datasets, however, because the data are contributed by its users, it must be used with caution because much of the advice given above will not be determinable. No doubt, reliable data can be found on Kaggle. You just can't automatically rely on it. Do your homework.

Here is a list of reliable data sources to help you get started.

- Australian Bureau of Statistics
- Australian Bureau of Meteorology
- World Bank Open Data
- Global Health Observatory (GHO) data
- U.S. Government's Open Data
- DataVic
- Data.gov.au
- EU Open Data Portal

Here are a few of the best sites with interesting open data (unknown reliability):

- UCI Machine Learning Repository
- Kaggle Datasets
- FiveThirtyEight
- Google Public Data Search - Google search engine that links to sources of data.
- Registry of Open Data on AWS
- Google Cloud Platform Datasets

There are heaps more, but this will keep you busy for now.

1.9.6 Citing a Data Source

Often a data visualisation will incorporate data, structured or unstructured, from an external source. When this occurs, you should always cite the data source in a note. For example, Figure 1.15 shows a visualisation of vaccine coverage by measles cases worldwide by Vanderslott and Roser (2018). You will notice the inclusion of a source at the bottom of the plot - Global Health Observatory Data Repository (2017).

Citing a data sources used in a data visualisation is important for several reasons:

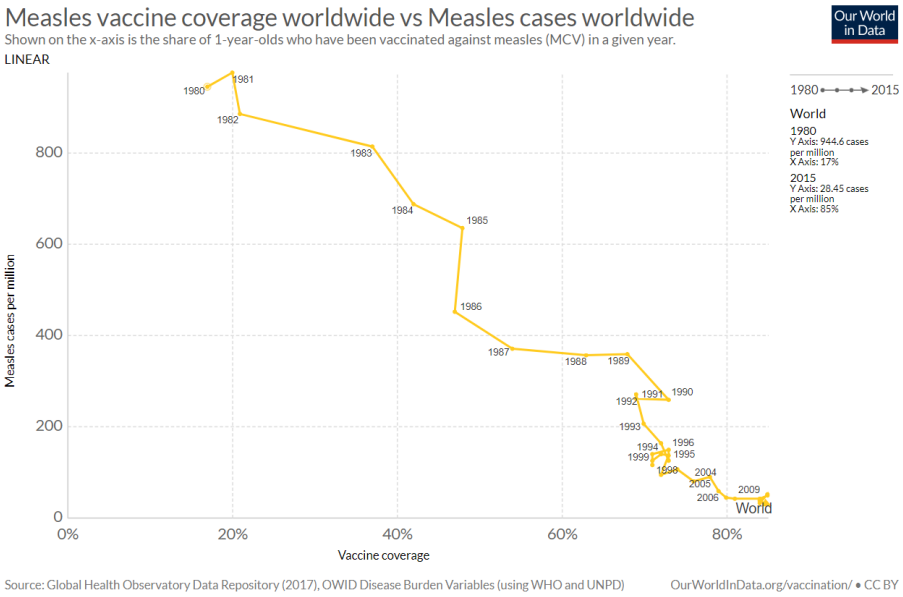


Figure 1.15: Cite your data source in your visualisations (Vanderslott and Roser, 2018).

- Avoids copyright issues because it credits the data to the original source (assuming the data are licensed for reuse)
- Informs the viewer where the data underlying the visualisation was sourced which is important for assessing the quality of the data.
- Allows the viewer to independently source and verify the data and resulting visualisation.

Citations add credibility to your work. Where possible, use a recognised referencing style such as APA, Harvard or Vancouver, to name a few. Often, a company, journal, news site etc will have a preferred style. The goal of any referencing style is to provide a clear, succinct and consistent set of rules to help others locate your sources of information.

1.10 Concluding Thoughts

This chapter defined data visualisation, revised types of variables, introduced plot anatomy, and outlined a data visualisation design process. You were also introduced to two very useful tools, the Trifecta Check-up and the Data Visualisation Checklist. In the final sections, you examined the ethics of data visualisation and data integrity. Would you have thought there was so much

to data visualisation? Are you keen to learn more? If so, you are in luck. You have only begun to scratch the surface.

Chapter 2

Storytelling with Data

2.1 Summary

Using storytelling techniques in data visualisation can help you to better engage your audience and leave a lasting impression. Building upon the data visualisation design process introduced in Chapter 1, this chapter will introduce the notion of telling visual data stories. Examples and case studies will be explored and the scholarly literature discussed in order to define and characterise "narrative visualisation". You will ultimately learn to identify and apply common strategies and techniques used for data visualisation storytelling.

2.1.1 Learning Objectives

The learning objectives of this chapter are as follows:

- Define storytelling and why it is important for data visualisation
- Define data visualisation storytelling
- Identify the explain the following common elements of data visualisation storytelling:
  - Genre
  - Structure
  - Highlighting
  - Transition
  - Ordering
  - Interactivity
  - Messaging
- Apply tools and strategies to help incorporate data visualisation storytelling techniques.

2.1.2 Chapter Video

Watch Hans Rosling, a master data storyteller, in action (Only available online).

2.2 The Power of Storytelling

A story can be defined as a sequence of causally related events that unfold over time at a pace tailored to an audience in order to hold their attention and leave a lasting impression (Kwan-Liu Ma et al., 2012). Storytelling continues to play an important role in human history across all cultures (Chaitin, 2003). Stories are still used today to entertain, educate and pass on important cultural norms, customs and morals. For example, storybooks read to young children are not just about introducing reading and writing. They are a way to entertain, develop imagination, and help children understand the world around them. However, storytelling isn't restricted to books. Before written language, our ancestors would have told oral stories, which continues even today. Many cultures also used dance, music and art. Technology has allowed us to merge many of these modes of storytelling into radio, movies, TV, computer games and other media. From a very young age and throughout life, humans learn through story.

This is no coincidence. Stories allow you to structure information in a way that sends a clear and memorable message (Kosara and Mackinlay, 2013). Human memory is greatly enhanced when information is organised in a meaningful way. Perhaps this explains why we can often recount stories many years on. Stories are also often tailored to their audience in order to maximise their engagement by using characters, settings, events and interactions that resonate with an audience. For example, children books employ colourful illustrations and characters to capture kids' attention, while adult media tends to focus on real images and actors. In the next section, you will consider how storytelling can be extended to data visualisation.

2.3 Storytelling and Data Visualisation

Storytelling and data visualisation can be used together because the characteristics of a story can be applied in almost every communication activity (Dahlstrom, 2014) and, given the power of storytelling, you have good reason to use it. Using storytelling technique for data visualisation is argued to make visualisation more interesting and memorable (Kwan-Liu Ma et al., 2012), or as Kosara (2016) puts it "getting a point across and making it stick" (p. 80). Data visualisation storytelling is different to general storytelling because data and context replace traditional story elements such as a linear sequence of causally related events, settings and characters (Segel and Heer, 2010). Therefore, while a narrative is central to the definition of data stories, it has a slightly different

meaning in data visualisation. Lee et al. (2015) built upon the work of Segel and Heer (2010) by proposing three key elements that characterise a visual data story. These elements are as follows:

- A series of **story pieces** that present data-driven facts
- Story pieces are **visualised** to support each intended message and each visualisation is supported by annotations or narrations that focus on the intended message
- Each story piece is linked and presented in a **meaningful order** that aims to maximise the objective of the story

The last point specifically relates to the importance of order. For example, time is used in traditional stories as a way to understand cause and effect (Kosara and Mackinlay, 2013). Lee et al. (2015) argue that storytelling data visualisations must help guide the viewer and not wholly leave the viewer to their own interpretations. Therefore, data visualisations produced during exploratory data visualisation are not considered examples of visual data stories.

While a relatively new and emerging field, data visualisation storytelling has established itself as a persuasive field that specialises in the intersection of storytelling and data visualisation. Before we describe data visualisation storytelling strategies, let's take a close look at a case study to explore a data visualisation storyteller in action.

2.4 Case Study

In order to understand the strategies employed by a visual data storyteller, the following section will examine an article that appeared in the New York Times written by Litman-Navarro (2019) (see Figure 2.1). You can view the article here. The title of the article was "We read 150 privacy policies. They were an incomprehensible disaster". The author skillfully used data visualisation to highlight the issues of privacy policies used by major tech and media companies. They effectively used data visualisation and storytelling to turn a potentially dry topic into an engaging and memorable article. Before you continue reading, click on the image below to read the article.

There are very deliberate features of this article that highlight effective data visualisation storytelling. The opening statement paints an immediate picture of the problem and objective of the article. The first visualisation is the background of the first page. A splattering of legal writing comprising the contents of several privacy policies reviewed by the author. It is deliberately overwhelming and seems to be printed in the same size font that many companies use to publish (and perhaps hide) their privacy policies. Without even realising it, the viewer is already looking at a visualisation of the raw data.

Figure 2.1: Title and summary (Litman-Navarro, 2019).

As the viewer scrolls down the page, the background disappears, new text appears and the first data visualisation is presented (see Figure 2.2). It is clear the author will take control of the story and guide the viewer through the story piece by piece. This article is similar to a slideshow. This first piece of information looks at how long each privacy policy took to read. A dot plot was used to show the distribution of reading times in minutes. Companies like Facebook, Uber and AirBnB are labelled to put the data in perspective. The user can also hover over individual data points for further labels to appear. The message is clear. Reading many of the privacy policies was a time consuming process.

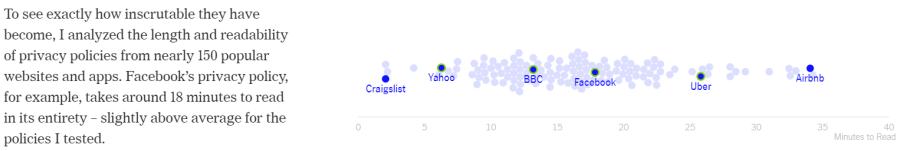


Figure 2.2: Visualising privacy policy reading time (Litman-Navarro, 2019).

Next, Litman-Navarro (2019) measures the complexity of the text using the Lexile test. A scatter plot shows the relationship between a privacy policy's complexity and reading time (see Figure 2.3). This is a positive relationship, but not strong. What was surprising was the large degree of variability. Again, labels are used to identify well known companies.

Figure 2.3: The relationship between privacy policy readability and reading time (Litman-Navarro, 2019).

The same scatter plot is then shaded by educational bands, helping the reader to put the Lexile scores into perspective (see Figure 2.4). By relating Lexile scores to level of education, you can quickly see an issue. Most privacy policies appear to require a college degree or higher to understand. This was a very effective way to help the viewer understand a metric that they were unlikely to be familiar with. This demonstrates an excellent understanding of the audience and the commitment of the storyteller to guide the audience in their interpretation of the data visualisation.

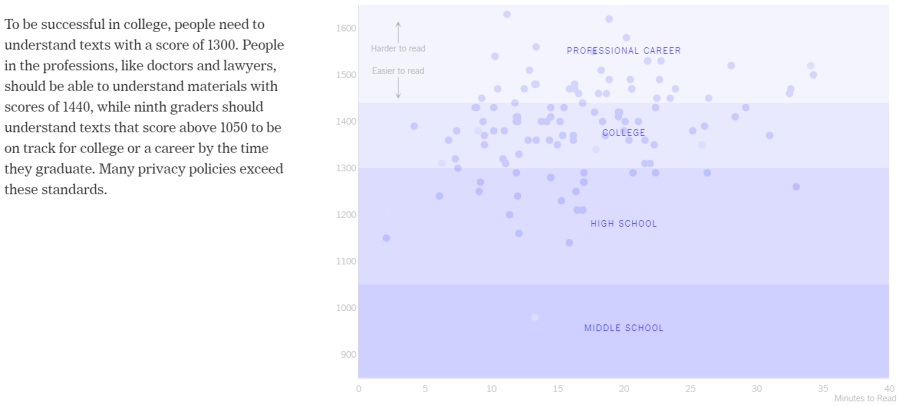


Figure 2.4: Adding context using education bands (Litman-Navarro, 2019).

The next transition reinforces the first comparison by overlaying the reading time and Lexile scores of the first chapters of some well known texts, such as Stephen Hawking's "A Brief History of Time" (see Figure 2.5). It becomes clear

than popular texts are likely be more readable than privacy policies.

For comparison, here are the scores for some classic texts. Only Immanuel Kant's famously difficult "Critique of Pure Reason" registers a more challenging readability score than Facebook's privacy policy. (To calculate their reading time, I measured the first chapter of each text.)

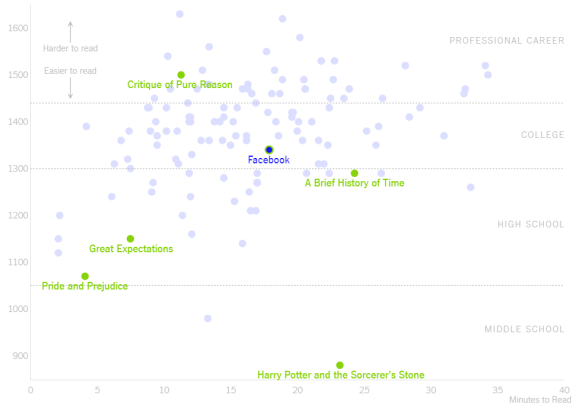


Figure 2.5: Adding additional context using classic texts (Litman-Navarro, 2019).

The major point of the scatter plot is stated in the next transition by highlighting all the privacy policies that require a college degree or higher to understand (see Figure 2.6).

The vast majority of these privacy policies exceed the college reading level. And according to the most recent literacy survey conducted by the National Center for Education Statistics, over half of Americans may struggle to comprehend dense, lengthy texts. That means a significant chunk of the data collection economy is based on consenting to complicated documents that many Americans can't understand.

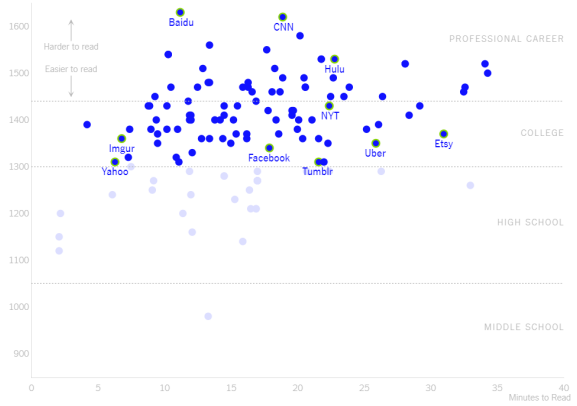


Figure 2.6: Highlighting policies requiring higher education (Litman-Navarro, 2019).

The next two transitions focus on outliers (see Figure 2.7). This is common strategy of storytellers, as outliers often provide interesting comparisons and discussion. The BBC shows that it is possible for privacy statements to be quick and easy to understand. AirBnB, on the other hand, is presented as a cautionary tale of just how bad the situation can get, being both difficult to understand and time consuming to read.

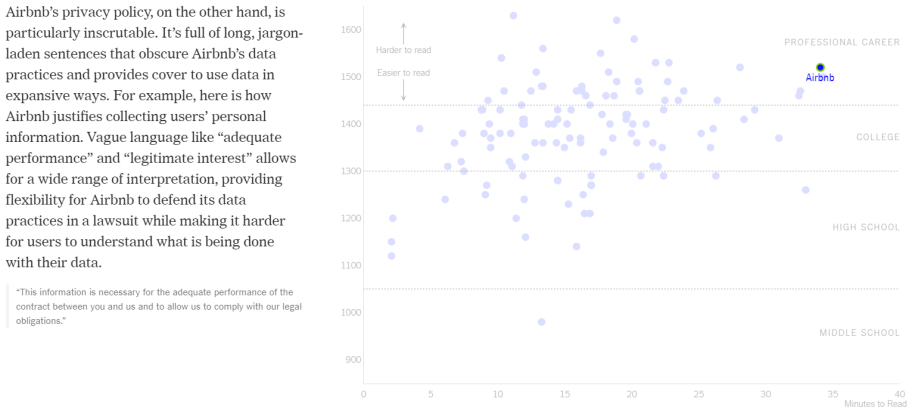


Figure 2.7: Highlighting and discussing outliers (Litman-Navarro, 2019).

Another transition and Google's various versions of their privacy policy are plotted across time by linking data points using a directional line (see Figure 2.8). This highlights how things have gotten worse overtime as the value of data and the complexity of technology has increased. The author also suggests that legislation might play a mixed role in changing the situation. The author draws the viewer's attention to the fact that the readability of Google's privacy policy was improved at the expense of reading time after the General Data Protection Regulation from European Union came into effect in 2018.

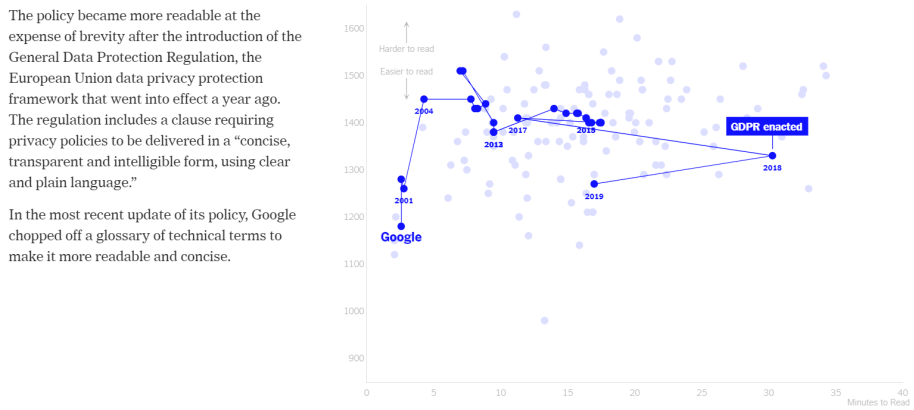


Figure 2.8: Tracking Google's privacy policy across time (Litman-Navarro, 2019).

The data visualisations end at this point and the author sums up in text. Issues regarding the accessibility of privacy policies remain and is likely to continue. Technology users beware.

There were many techniques identified in the literature about data visualisation story-telling used by Litman-Navarro (2019). Consider the following features and think how they helped to tell a story:

- A compelling introduction and graphic (background text) to hook you.
- A clear and concise, piece-by-piece, ordering of information achieved using a slide-like presentation style.
- Labeling of data points to provide context (including hover information)
- Supporting text/annotations and highlighting to guide the viewer's attention and summarise key points
- Explanations of variables (Lexile scores) using everyday references (education level, popular texts)
- Consistent use of a scatter plot that transitioned layers, labels and annotations to make different points in a logical order.
- A slider bar as a visual reference of the viewer's position in the story.
- A detailed discussion of individual data points (outliers and by time) to highlight interesting trends and to delve deeper into the data.
- Transition animations (moving text, data point labels appearing, data point becoming highlighted etc) to draw peoples' attention to changes and position in the story.

The ordering or structure of a story is a defining element a storytelling data visualisation (Lee et al., 2015). Providing a meaningful order helps to optimise the message being sent. A summary of the main points in Litman-Navarro (2019) are as follows:

- Our data are being unknowingly sold because we don't understand the complex privacy policies used by major media and tech companies (Opening statement)
- Privacy policies are time consuming to read (Dot plot)
- Privacy policies are difficult to understand (Scatter plot)
- Most privacy policies require a college level of education (Scatter plot)
- Some privacy policies are better than others (Scatter plot)
- Privacy polices have progressively gotten worse (Scatter plot with connected lines)
- Legislative changes has had mixed results (Scatter plot with connected lines)
- Things are not likely to get better anytime soon, so beware! (Conclusion)

Lee et al. (2015) refers to each of the above points supported using a data visualisations a "story piece". The logical ordering of these stories piece is what makes a data story.

2.5 Well Known Data Visualisation Storytelling Sites

If you enjoyed the work of Litman-Navarro (2019), you will enjoy the following list of well known sites that publish data visualisation stories.

- FiveThrityEight
- Graphic Detail by The Economist
- The Upshot by The New York Times
- The Financial Times - Data Section
- The Guardian - Data Journalism
- Data Journalism Awards

2.6 Storytelling Strategies

Segel and Heer (2010) developed a framework to organise common storytelling strategies used in data visualisation or what they refer to as a narrative visualisations. This framework is a unique and useful starting point for exploring the specific strategies used by narrative designers. The following section will summarise their framework in order to help you incorporate effective storytelling elements into your own work.

2.6.1 Genre

Segel and Heer (2010) suggested that most narrative visualisations fit within a specific **genre**. Figure 2.9 lists the seven genres identified: magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show and film/video/animation. The genre establishes the framing of the visualisation and how each element or story idea will appear or be presented to the viewer. For example, in an annotated chart, the narrative is contained within a single visualisation, while in a slide show, multiple frames are used to present text, visualisations and other supporting content to tell a more in depth story. Determining which genre to use comes down to the story and situation. Sometimes the genres are mixed. Ojo and Heravi (2018) suggest that the most common genre present in award winning data storytelling by a substantial margin was the annotated graph. The case study of Litman-Navarro (2019) is an example of a slide show genre. Even through the page is "scrolled", the transitions present each story idea as a single frame or slide.

Hannen and Burn-Murdoch (2019) (Figure 2.10, click here to view online), published in the Financial Times, is an example of video of a bar chart race looking at the most populous cities across times. Hannen and Burn-Murdoch (2019) skillfully use data visualisation, animation, narration and photography to tell a compelling tale of human history. If only all history lessons could be like this.

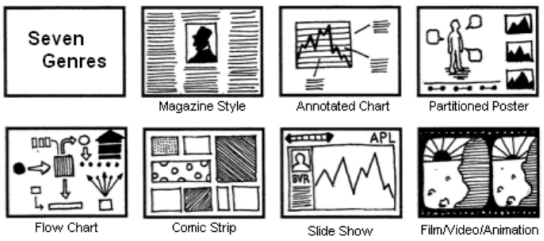


Figure 2.9: Genres of Narrative Visualisation (Segel and Heer, 2010).

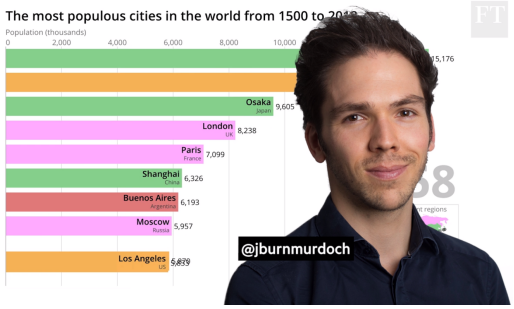


Figure 2.10: Telling a story using a bar chart race (Hannen and Burn-Murdoch, 2019).

2.6.2 Approach

Segel and Heer (2010) explain that narrative visualisations exist on a continuum between **author-driven** and **reader-driven** approaches. Author driven approaches employ a linear ordering of ideas, extensive messaging (annotation, text, highlighting etc) and minimal interactivity on the user's behalf. At the other end of the spectrum, reader-driven approaches have no clear ordering of information, minimal messaging and extensive interactivity (filtering, searching, changing views etc.). Litman-Navarro (2019) and Hannen and Burn-Murdoch (2019) were examples of author-driven approaches. Strict reader driven approaches are not common because a user will risk missing the point, so, many narrative visualisations employ a hybrid-type approach by incorporating both author-driven and reader-driven elements.

For example, Smith et al. (2018) visualise broadband speeds across Britain and conclude that, despite widespread belief to the contrary, some rural areas experience speeds far greater than many urban areas (Figure 2.11, click here to view online). The article's initial approach was author-driven, but they also include a data visualisation app that allows the user to explore different postcodes and compare them to the national average. At this point the article introduces a reader-driven element. Segel and Heer (2010) referred to this as a hybrid approach, which can be more engaging than a purely author-driven method.

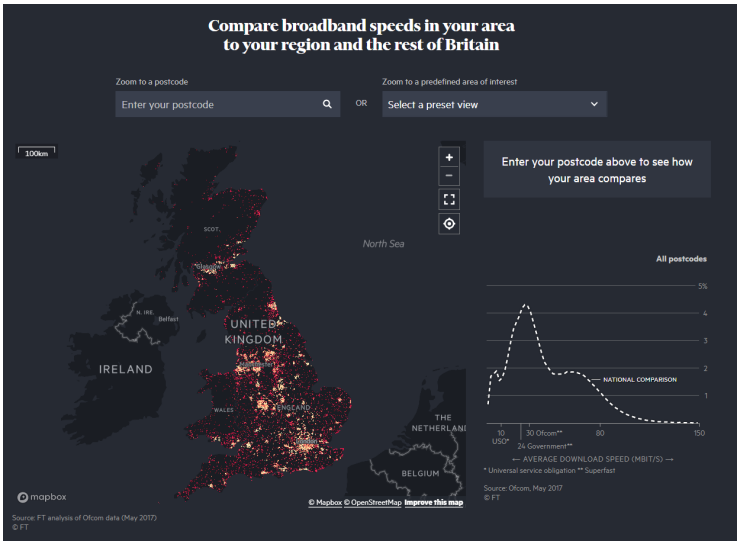


Figure 2.11: Comparing broadband speed in the UK (Smith et al., 2018).

2.6.3 Design Strategies

The design of a narrative visualisation is broken into two major concepts, **narrative structure** and **visual narrative**. Narrative structure refers to design elements that determine how the narrative will unfold. This includes the ordering of elements (linear or user directed), the degree of user interaction and control over the story and the use of messaging (annotations, introductory and summary statements). The visual narrative refers to specific strategies used to elicit a narrative experience. This includes visual elements that direct a user's attention to important points and transitions between frames as well as the user's position within a story. In the following sections, the most important design strategies discussed by Segel and Heer (2010) will be summarised.

2.6.3.1 Annotation

Annotations are used to explicitly summarise and draw viewers' attention to key story elements. These can appear as text overlaid on plots or audio narration. They aim to focus the narrative and ensure the key objective of the story is clear. For example, Holder et al. (2018) used annotations in their map of the 2018 U.S. midterms published in The Guardian (Figure 2.12, click here to view online). The annotations summarised key outcomes of the midterms that helped to support the broader narrative of the article. However, Kosara and Mackinlay (2013) warns that you must be careful to balance text and data. You don't want the text to detract from the visuals.

2.6.3.2 Visual highligthing

Visual highlighting is any method that draws attention to key data observations, statistics, outliers or trends present in the data. There are many ways we can highlight or draw peoples' attention. For example, Watkins (2015) uses colour to draw attention to the shrinking arctic sea ice area across time (Figure 2.13). The most common strategy is using colour. However, size, boldness, connections and animations are also relevant. When using highlighting, other data elements remain visualised, but are de-emphasized, for example, by adding transparency, de-saturating colours or using neutral colours like grey. Highlighting ensures that viewers draw their to the right elements in the the right order. In later chapters you will learn about human perception and the rules that govern our visual attention.

2.6.3.3 Matching Content

Matching content refers to the consistent use of visual elements to show the relationships between different sections of a data visualisation. This can refer to the consistent use of colour to refer to the same categories or the consistent

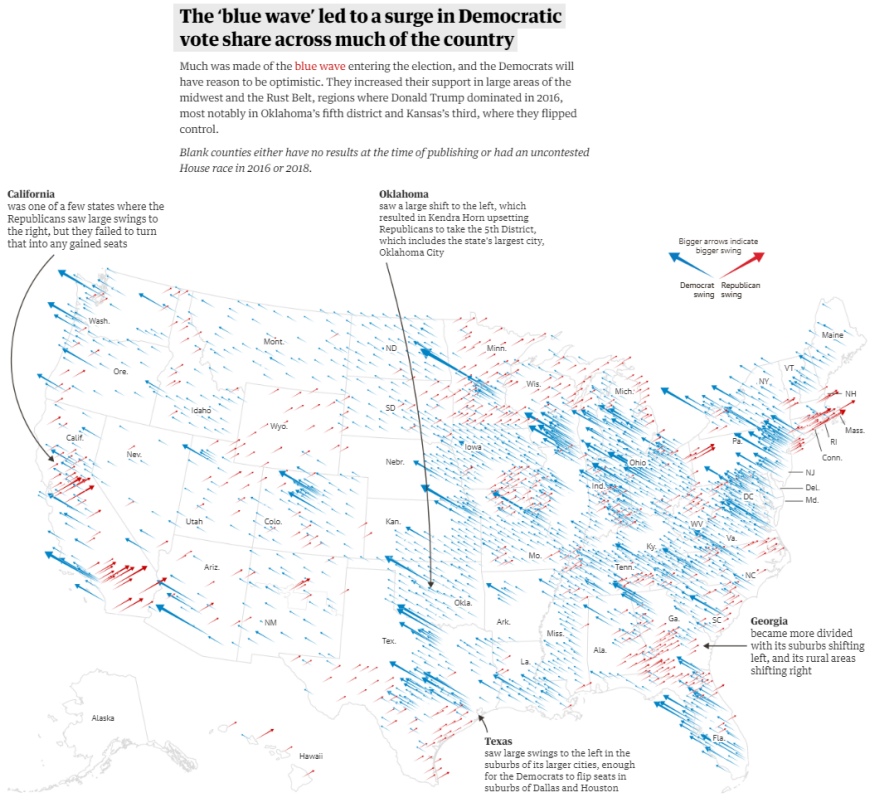


Figure 2.12: Annotations ensure the main story cannot be missed (Holder et al., 2018).

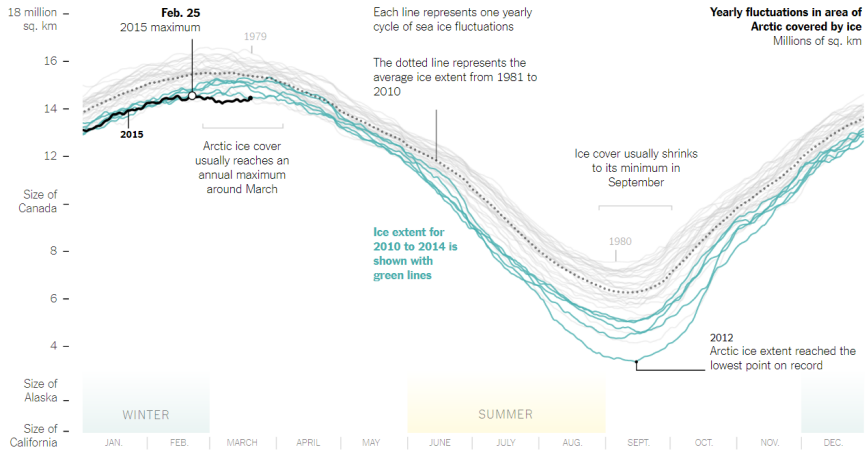


Figure 2.13: Highlighting draws attention to key information (Watkins, 2015).

ordering of categories. Matching content creates efficiency and predictability as the viewer can make assumptions about the layout of related elements.

2.6.3.4 Progress Bars

Story books have pages which are a visual and numerical way to measure progress through a story and a way to find your way back to specific information. Movies and videos have a linear play bar which reports time, progress and provides a way to navigate the video. Narrative visualisations that incorporate a slide show-like structure use progress bars for a similar purpose. They help to situate the viewer in terms of the story progress and a way to navigate the story.

2.6.3.5 Consistent Visual Platform

Litman-Navarro (2019) reuse of the scatter plot in the privacy policy article is an example of using a consistent visual platform. Contents within the platform are changed, while the general layout remains consistent. Again, this creates efficiency and predictability. Once the user understands the platform, they can focus on the story as it unfolds, reading annotations, drawing their attention to highlighted elements and interacting when prompted.

2.6.3.6 Multi-messaging

Multi-messaging refers to the use of a combination of text, annotations or graphics that work together to enrich the narrative. Narrative visualisations can be well distinguished from exploratory data visualisations in this respect. Narrative visualisations are often accompanied by both text and annotations to support the story and reveal further detail.

2.6.3.7 Details on Demand

Extensive annotations and text can overwhelm a viewer. Interactive data visualisations can overcome this by providing details on demand. For example, Litman-Navarro (2019) make company labels appear when a user hovers over a point (see Figure 2.2). Showing all the labels on the same plot would create a visual mess of overlapping elements. Instead, hovering over the data point will quickly reveal the company name. Another example was the *new.com.au* interactive map visualising the results of the Australian 2017 Same Sex Marriage Vote reported in Reynolds (2017). The viewer could click on an electorate to reveal a detail break-down of participation rates by gender and age distribution (Figure 2.14, click here to view online).

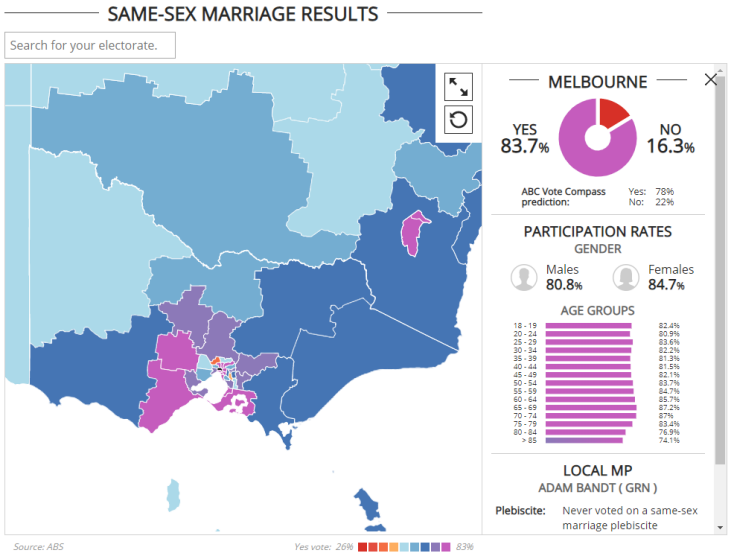


Figure 2.14: Clicking on an electorate reveals detailed analysis (Reynolds, 2017).

2.6.3.8 Timeline slider

Timeline sliders are a common interactive element used in data visualisation that encourages a viewer to transition the back and forth. The act of sliding and transitioning the visualisation allows the viewer to control and experience a trend unfold before their eyes. The U.S. Geological Survey (USGS, 2015) created an app to explore U.S. water use across the time (see Figure 2.15, click here to view online. The viewer transitions water use categories across time to see how different states compare. Sliding through the visualisation, the viewer can see a story of peak water in the 1980's and a subsequent decline. The ability to readily control the transition allows the viewer to isolate specific states to drill deeper into the history of water use.

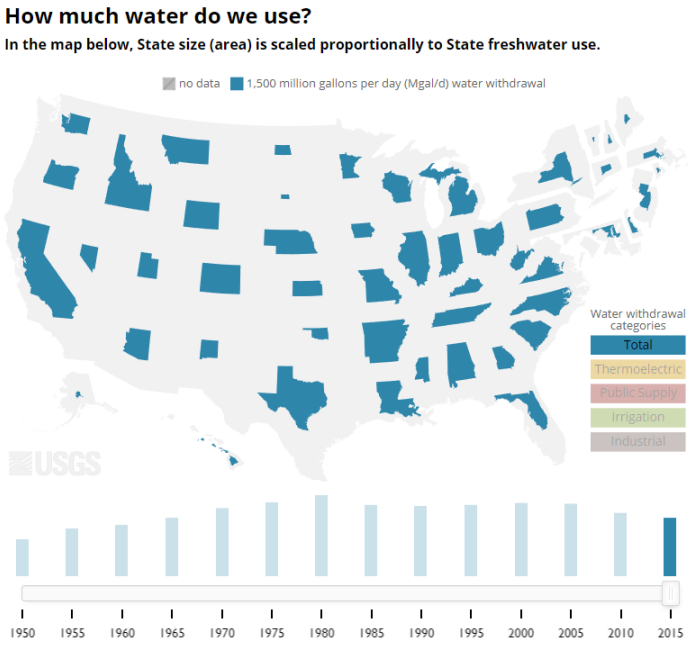


Figure 2.15: A slider controls the transition of time (USGS, 2015).

2.6.3.9 Tacit tutorial

Many narrative visualisation include interactive features designed to engage the viewer in a deeper and more memorable experience. This often requires the user to interact with controls and data visualisations that they might not be familiar with or be immediately aware how to use in the intended way. Designers can help the viewer by using **tacit tutorials** that imply how to understand or engage with a visualisation. The tutorials are tacit because an explicit tutorial

will detract and possible disengage viewers. People don't like reading instructions. Therefore, the challenge is to hide the instructions within the story. A simple, yet effective example of this can be seen in the work of Hanrahan et al. (2017) from the ABC News who published a narrative visualisation of data taken from the 2016 Australian Census (click here to view online). The story explains the Census using a frequency based approach by standardising the Australian population as 100 people. Therefore, each dot in the visuals represents approximately 240,000 Australians. To explain this point, Hanrahan et al. (2017) include a frame that presents a visualisation and provides a brief explanation of how to interpret the dots (Figure 2.16). This very simple explanation is embedded in the narrative and sets the consistent visual platform for the remainder of the story.

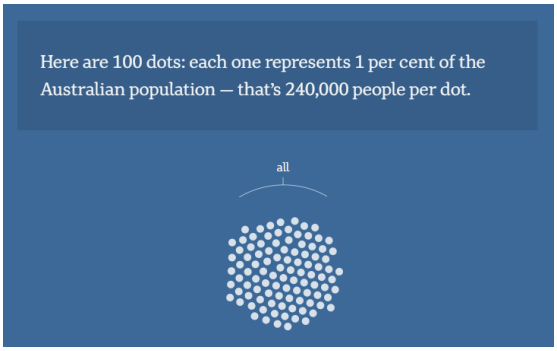


Figure 2.16: A tacit tutorial helps the viewer to understand the visuals (Hanrahan et al., 2017).

2.6.3.10 Semantically Consistent

Semantic consistency refers to the consistent use of visual aesthetics to encode information. A clear example of this is the consistent use of colour and sized used to represent a variable across related visualisations. This aims to create predictability and efficiency for the viewer as the representation of variables remains consistent. For example, The New York Times article, *How Much Hotter Is Your Hometown Than When You Were Born?* (Figure 2.17, click here to view online) uses a consistent colour scale to represent temperature throughout the article (Popovich et al., 2019).

2.6.3.11 Markers of Interactivity

Interactive features used in narrative visualisation can be easily missed. Designers need to be sure that the viewer's attention is draw to features and points in the narrative where the viewer can interact. This might include a tacit tutorial

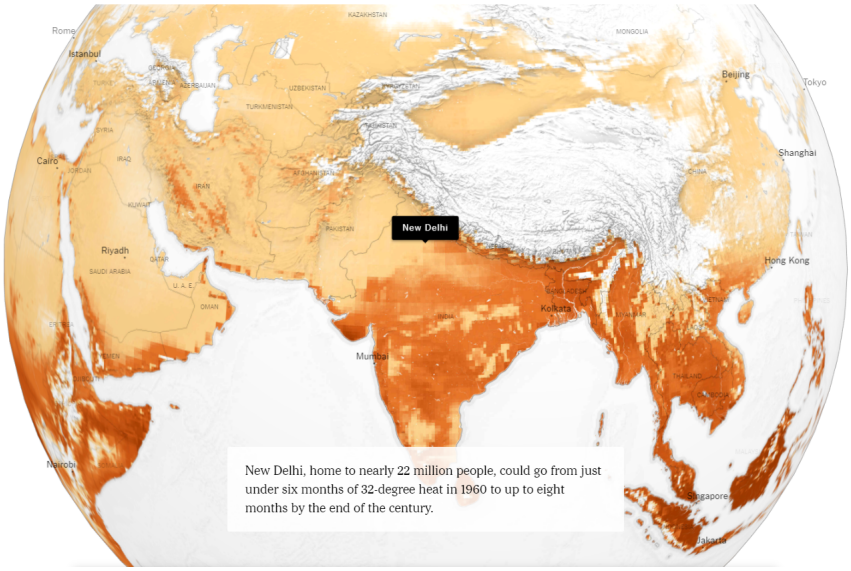


Figure 2.17: Using colour consistently to encode information creates efficiency (Popovich et al., 2019).

or any method including highlighting and animation to distinguish interactive elements to the viewer. The previous example by Popovich et al. (2019) presents the user with input boxes to type in their home city and age. The page will stop the viewer from scrolling and progressing the story until they fill in the boxes. This ensures the interactive element of the story cannot be missed.

2.6.3.12 Animated Transitions

Without transitions, changes in narrative visualisations can go unnoticed. Animated transitions act as a implicit visual cue to ensure the viewer is aware that the narrative has progressed or important visual elements have changed. Thinking back to many of the examples in this chapter, can you remember the many instances of animated transition? If not, work back through a few of the examples. The transitions are hard to miss.

You now have a great overview of the main strategies employed by narrative visualisation designers. The strategies you choose will always depend on the objective of the story, the nature of the data and visuals chosen. Most narrative visualisations use a combination of strategies, but never should the aim be to use as many as possible. Mindlessly adding as many strategies as possible will be a recipe for disaster. Make sure each element is justified and supports the overall story.

2.7 Storytelling Structure

The key to a good data visualisation story is an underlying structure or ordering of information. This structure should be implicit to the viewer. However, when designing your narrative, explicitly outlining the structure is important. Mannon (2018) adapted Freytag's Pyramid, to generalise storytelling to data stories, which provides a useful starting point for data visualisation storytelling (see Figure 2.18). Mannon (2018)'s ideas are visualised in the following graphic and briefly explained in the following sections.

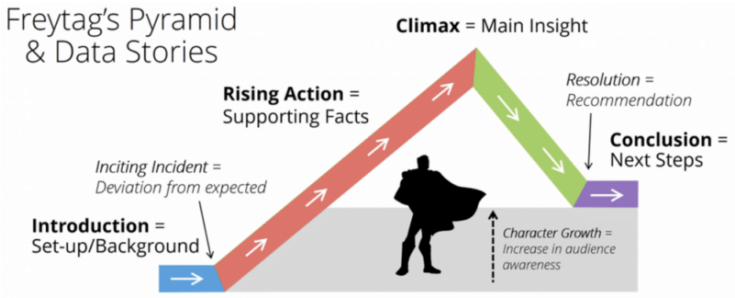


Figure 2.18: Good stories are well structured (Mannon, 2018).

2.7.1 Set-up

The first step is to hook your audience. This is the point where your audience will decide whether to continue viewing the data visualisation story. Think carefully about your audience and how you might capture their attention. There are many ways to achieve this, such as a powerful anecdote, surprising statistic, thought provoking question, images, or a provocative statement. The set-up will also need to provide some context and background, but the main focus should be on "hooking the audience". At the end of the set-up, the audience is clear on the purpose or objective of the visualisation story. Once you have the audiences' attention, you are ready to give them the facts.

2.7.2 Supporting Facts

Use a series of ideas supported by data visualisations that will guide the viewer in understanding the key points of your narrative. These ideas must all tie into the overall objective and main insight of the story. Stick to a small number of important ideas. Do not bombard the viewer with too much information or detail. Less is often more. Interactive elements can be used to allow the viewer to explore the detail if necessary. Think of each idea as a piece of evidence

or fact. You need just enough to achieve your objective. Too few, and your audience won't be convinced. Too much, and your audience will be left behind.

2.7.3 Main Insight

Your supporting ideas will culminate in the overall insight or message that you want the audience to remember. Even if they forget some of the supporting facts, they must reach this point convinced of the veracity of the main insight.

2.7.4 The Solution

A compelling story doesn't stop at the main insight. There needs to be a resolution, or what is commonly referred to as a "call to action". This is where the objective of your story is reiterated. You must explain the implications of your insight. A compelling story has a lasting effect on your audience. They might be better informed, they might be motivated to change something, they might start to worry about an issue they were unaware of, or they might have been persuaded to change their mind. State this final conclusion explicitly. Don't make your audience read between the lines. You have failed if your audience gets to the end of the story and thinks "so what?".

Now let's think back to the privacy policy case study by Litman-Navarro (2019) and determine if this structure is present.

Set-up

The background image of the page contains the text of several privacy policies from major tech and media companies. It is small and difficult to read. The first paragraph introduces the background of the article and paints a depressing picture of the issue that the article will address.

Supporting Facts

An ordered series of key ideas, supported by data visualisations, used to support the main insight.

- Privacy policies are time consuming to read
- Privacy policies are difficult to understand
- Most privacy policies require a college level of education
- Some privacy policies are better than others
- Privacy polices have progressively gotten worse
- Legislative changes has had mixed results

Main Insight

After the presentation of the supporting facts, the conclusion is obvious. Privacy policies are a massive issue in our digital lives.

The Solution

The story concludes by discussing what a good privacy policy should do, and warns viewers to assume that someone is always watching.

It is a pretty good fit. The structure outlined by Mannon (2018) won't work in all situations, nor should you try. However, Mannon (2018) does remind us of one thing. Structure is key to good storytelling.

2.8 Concluding Thoughts

This chapter introduced storytelling and generalised storytelling to the design of data visualisation. The chapter explored many examples of narrative visualisations and discussed data visualisation storytelling strategies using the framework developed by Segel and Heer (2010). In the final section, Mannon (2018) showed how Freytag's Pyramid story structure can be applied to narrative visualisations. This chapter was an extension of the data visualisation design process introduced in Chapter 1. Both chapters emphasise the point that data visualisation requires deliberate design. It is not something that can be replaced by a software package. Data visualisation software doesn't understand your audience, objective, dataset, ethics, integrity or narrative. Designers do the hardwork. Designer have to think before they create. Chapters 1 and 2 have given you some powerful thinking tools to help you along the way.

Chapter 3

Visual Perception and Colour

3.1 Summary

Data visualisation designers need to understand the laws and limitations of human visual perception. Humans do not directly perceive reality. Vision is our brain's reconstruction of reality put together using neural impulses triggered by light in the photoreceptors of our eyes. Our brains follow well known heuristics in it's construction of vision. Understanding and exploiting these heuristics will help you to understand and design effective data visualisations. Humans can also see in colour. Colour is one of the most powerful visual properties of human vision as it encodes vast amounts of information present in the environment. Therefore, colour is one of the most versatile tools used by data visualisation designers. This chapter will discuss some of the important rules about using colour effectively and responsibly.

3.1.1 Learning Objectives

The learning objectives for this chapter are as follows:

- Discuss how visual illusions provide insight into visual perception and its limitations.
- Outline the three stages of Ware's visual information processing model and its implications for designing data visualisations.
- Explain the concept of preattentive processing and identify common features that are preattentively processed.

- Define the Gestalt laws of proximity, similarity, connectedness, continuity, symmetry, closure, figure-ground and common fate and explain how they inform data visualisation design.
- Define and differentiate between change and inattention blindness and explain their implications on data visualisation design.
- Identify common data visualisation features and the types of variables they are used to represent.
- Rank the accuracy of different data visualisation features used to represent quantitative variables for comparative purposes.
- Define colour as perceived by the human visual perception system, and the RGB (red, green, blue) and HSV (hue, saturation and value) colour models.
- Explain the hexadecimal (hex) colour code system
- Define colour blindness, identify the most common types and apply colour blind friendly colour schemes to your data visualisations.
- Identify common natural and cultural colour associations and be sensitive to these associations when designing data visualisations.
- Be aware of the most common colour rules and considerations related to data visualisation and apply good colour sense to design accurate and impactful visualisations.

3.2 Visual Complexity

The following video by Lotto (2009) discusses the surprising ways in which optical illusions help us to understand human vision (Only available online).

Visual illusions demonstrate the sensitive nature of our powerful visual perception system. Our eyes and brain do not operate like a video recorder. What we perceive is our brain's interpretation of light entering our eye and triggering electrical impulses from the cones and rods in our eyes. The brain uses its enormous power to turn these signals into what we perceive as sight. Vision is a construct of our brain.

However, because of the complexity involved in converting light to vision, our brains use some short-cuts or assumptions to ensure that we can perceive vision as accurately and quickly as possible. Most of the time, these assumptions hold and our brain's construction of reality is good enough for us to survive. Visual illusions, on the other hand, mess with these assumptions or exploit limitations in our visual processing system with surprising results. Check out the following slideshow of some famous visual illusions (Only available online).

Visual illusions remind us of the limitations of our visual perception system. We must be conscious of this fact and respect our perceptual boundaries. By understanding a little about our visual perception system we can ensure we design intuitive data visualisation that do not deceive.

3.3 Our Visual Information Processing System

In order to design good data visualisations you need to be aware of some of the important theory underlying visual perception. The goal of data visualisation "is to amplify cognition" (Kirk, 2012), but to do so, we need to be able to get a signal (data visualisation) through the human sensory system (eye, retina, optic nerve) and processed correctly by the brain. You might believe this process is simple and instinctual, but I assure you there is a lot going on and there is a lot that can go wrong. Visual perception is an enormous area of research, so, in this chapter we will focus on the big ideas and take home messages related to data visualisation. For the authoritative reference on this area please see Ware (2013).

We will first take a look at an overview of how visual information is processed by the brain. Ware (2013) proposes a simplified three stage model to help explain our eye's and brain's complex visual information processing system (see Figure 3.1). To start the process, our eyes scan a scene, such as a data visualisation. Our eyes convert light into electrical signals that travel along the optic nerve and into the visual cortex of the brain. Visual processing enters stage 1.

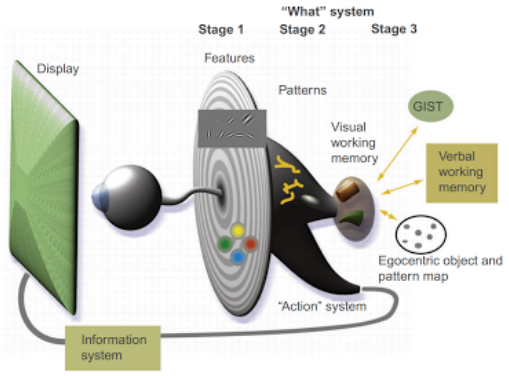


Figure 3.1: Ware's (2013) three stage model of visual information processing, p. 20.

Stage 1: Parallel Processing to Extract Low-Level Properties of the Visual Scene

Once the signal from our neurons in the eye reach the visual cortex, networks of billions of neurons work in parallel to extract low level properties of the visual field, such as colour, texture, orientation and movement. This happens automatically and unconsciously. The results of this process are held temporarily in iconic memory (<1000 ms), which is just enough time for our conscious processes to divert attention if required and move to the next stage. During stage 1, information that is optimised for recognition by the vast neuronal networks

in the visual cortex will be readily detected and processed. This will enable efficient interpretation of our data visualisation.

Stage 2: Pattern Perception

During the pattern perception stage, information obtained from stage 1 is further processed in order to identify patterns. Our visual field is broken into regions and our brains analyse contours, regions of colour, texture and motion. Stage 2 starts to involve our attention as we choose to make visual queries of the display. This stage is slower and less automatic than stage 1 but is still very rapid. Patterns perceived in the visual display can be held for a few seconds in our memory. Our brains will also transition between our object perception pathways and our action pathways. Data visualisation mainly concerns the object recognition pathway or the "what" system. The action pathway relates to our body's reaction to the environment. For example, the action of catching a ball rapidly approaching you.

Stage 3: Visual Working Memory

This is the fully conscious stage of visual information processing. Our attention has been drawn to a visual task, for example, interpreting a data visualisation that has caught our attention in an online news article. Our identification of the data visualisation based on the broad layout of the graphic is referred to as "gist". We are all familiar with visual gist. Think about how quickly we can identify the scene of a TV show as you rapidly flick through channels. Almost instinctively, we can identify the broad spatial layout of the scene as something like a beach, park, bush, studio, house etc. This suggests another important implication for data visualisation. Using visualisations that have "gist" will lead to more rapid interpretation. For example, most people are familiar with common data visualisations such as bar charts and scatter plots. Using these familiar data visualisation methods will allow the viewer to get the "gist" of the data visualisation very quickly.

Once we get the gist of the visualisation, we commence a series of visual queries driven by stages 1 and 2. We employ visual search strategies to bring the information together. Our working memory allows us to keep a few pieces of the visualisation in mind at any one time, however, we can also exploit our long-term memory stores to help fill in the gaps and activate contextual cues.

With this in mind, Ware (2013) compiled the following list of the costs versus benefits considerations for data visualisation (p. 24 - 25):

- Where two or more tools can perform the same task, choose the one that allows for the most valuable work to be done per unit of time.
- Consider adopting novel design solutions only when the estimated pay-off is substantially greater than the cost of learning to use them.
- Unless the benefit of novelty outweighs the cost of inconsistency, adopt tools that are consistent with other commonly used tools.

- Effort spent on developing tools should be in proportion to the profits they are expected to generate. This means that small-market custom solutions should be developed only for high value cognitive work.

Essentially, however you choose to visualise your data, the benefit (knowledge and insight) must outweigh the cost in your time (creating the visualisation) and your audience's time (in processing and interpretation). Always try to keep it "perceptually" simple, or "kips" for short.

3.4 Important Visual Laws

A good data visualisation will allow the viewer to quickly find the important patterns that tell the story behind the data. There are well known rules and laws of human visual processing that allow us to do this job in the most efficient way possible. We are all, innately, but unconsciously familiar with these laws. Through hard-wiring and learning from our environment, our brains are highly tuned to process visual information in certain ways. We will start from the beginning with the most important law related to data visualisation - preattentive processing.

3.4.1 Preattentive Processing

What first draws your attention in Figure 3.2?

Figure 3.2: Delicious.

The delicious cherries! Why? Why not the leaves or branches of the tree or the blue sky?

Consider another example. What features in Figure 3.3 “pop-out” to you?

Figure 3.3: Paradise.

The curvature of the beach, the rows of coloured beach chairs, the vibrant blue of the ocean? Why did these things so readily draw our attention? These are examples of preattentive processing.

Let’s consider another, more sterile, example. Count the number of 3s in the following sequence of numbers.

45929078059772098775972655665110049836645
27107462144654207079014738109743897010971
43907097349266847858715819048630901889074
25747072354745666142018774072849875310665

It takes a while, doesn’t it! You have to visually scan each digit. Now, try again.

45929078059772098775972655665110049836645
27107462144654207079014738109743897010971
43907097349266847858715819048630901889074
25747072354745666142018774072849875310665

Much quicker, I bet. Now you only had to scan the colour red to count all the 3s. Colour is said to be preattentively processed. What exactly does this mean? Researchers discovered long ago that certain visual features appear to stand-out. Preattentive features were so quickly processed by the brain, that

researchers originally believed such features were processed prior to conscious attention. This turned out to be not quite true, as attention is still necessary, but the name stuck. Ware (2013) defines preattentive processing as the degree to which a visual object is made available for our attention. Camouflage can be thought of the opposite of preattentive processing. Camouflage seeks to conceal objects by reducing the degree to which an object draws attention. Preattentive processing is a very powerful idea and governs much of "why" we design data visualisations using particular methods.

Colour, and many other features, are known to be preattentively processed. Figure 3.4 provides some concrete examples that demonstrate how other visual properties are preattentivelty processed. For contrast, the last two boxes (juncture and parallelism) are examples of features that are not preattentive.

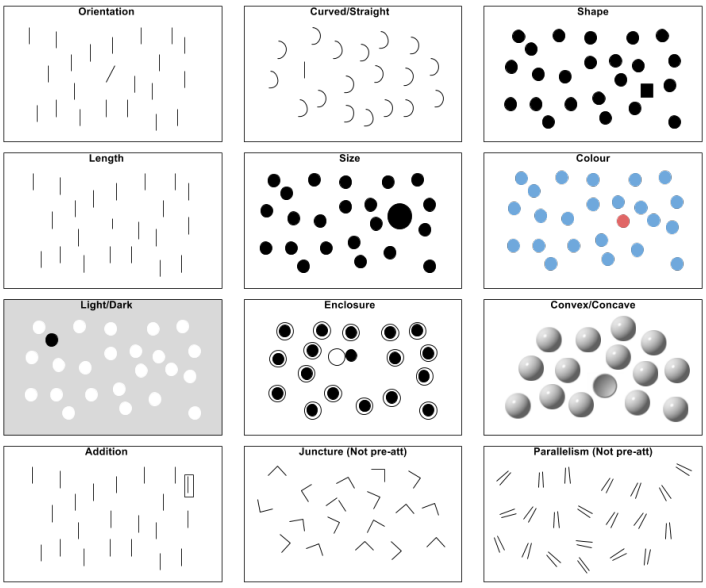


Figure 3.4: Examples of preattentively processed features adapted from Ware (2013).

Using preattentive features helps us to design data visualisations that allow efficient processing by the perceiver. For example, we use colour and shape to rapidly distinguish groups, and size, length and colour intensity to allow rapid comparisons. Many common data visualisation methods are based on preattentive principles. When designing your own, heed preattentive theory. Use it to draw the viewer's attention to the most important elements of the visualisation, while ensuring that other features of the visualisation don't detract attention from the story.

3.4.2 Gestalt Laws

The Gestalt (German translation meaning "pattern") school of psychology, founded in 1912, sought to understand the ways in which humans recognise visual patterns. It turns out that our brains are incredible at this task. Let's put ourselves to the test. Can you see the dog in Figure 3.5?

Figure 3.5: Our brains fill in the gaps (Boyer and Sarkar, 2000).

To someone who has never seen a dog, this image would appear to be a collection of black blobs. However, our brains use what are known as Gestalt laws to "fill in" the perceptual gaps and allow us to see the dog. Beware though, because we are highly tuned to see patterns in our environment, we are also known to get it wrong, sometimes very wrong! Our psychological predisposition to see patterns, where no patterns exist, is referred to as pareidolia. A prime example was the face on Mars (Figure 3.6).

There are heaps of other hilarious examples to be found on the internet. As long as you don't actually believe that Jesus is appearing in your food, enjoy this human tendency. There are eight common Gestalt laws that you need to know about. These laws are introduced in the following sections. We will also consider how these laws apply to data visualisation.

3.4.2.1 Proximity

Objects close or clustering together are perceptually grouped (Figure 3.7). Data visualisations use proximity to highlight relationships between categories and

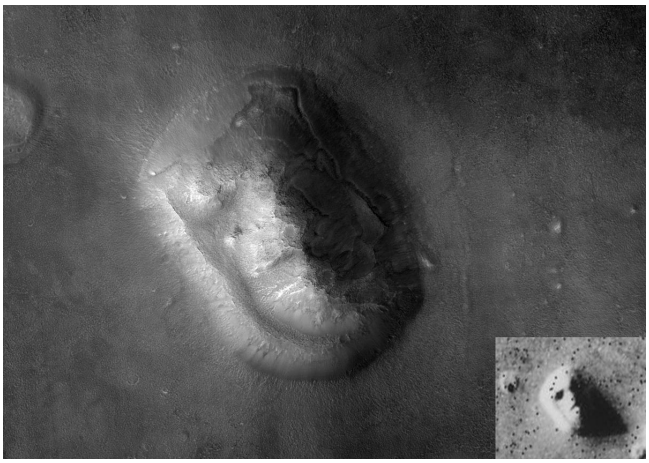


Figure 3.6: Face on Mars (NASA, 2007).

trends in the data. This also means that spacing can be used to visualise no relationship.

Network data visualisations such as *Transport Clusters* by Grandjean (2016) (Figure 3.8) use proximity to show clusters of nodes that share a relationship. The visualisation shows the increased connectedness of airports within continents which form distinct clusters, but also how airports connect between continents.

3.4.2.2 Similarity

Objects of similar characteristics (e.g. size, shape, colour) are grouped (Figure 3.9). Visualisation implication: Use colour, size, shape and other attributes to group related objects or to differentiate between categories.

The choropleth map of the 2019 Australian Federal Election results by the Guardian demonstrates this law (Figure 3.10). Colour is used to represent political parties. Electorates sharing the same colours are grouped together so the viewer can quickly see which seats belong to Labour, the Coalition and other parties.

3.4.2.3 Connectedness

Connectedness is more powerful than proximity, colour, size or shape. Objects connected by lines demonstrate relationships between objects (Figure 3.11). Data visualisations, such as times series plots and network diagrams, use connections between data points to represent temporal changes and highlight relationships.

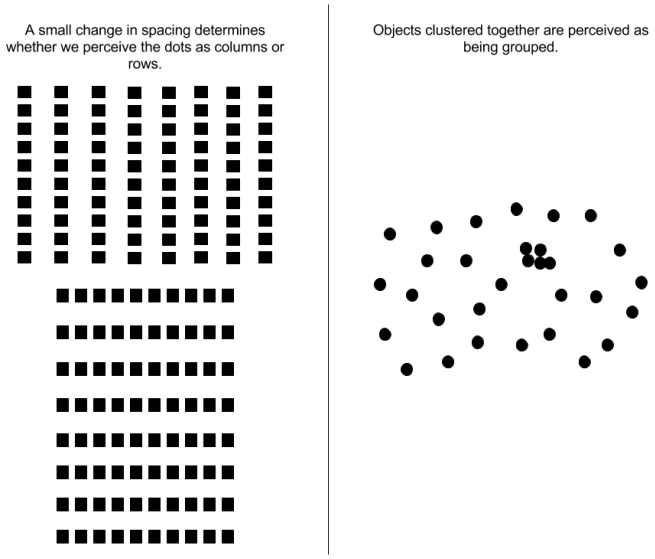


Figure 3.7: Proximity examples.

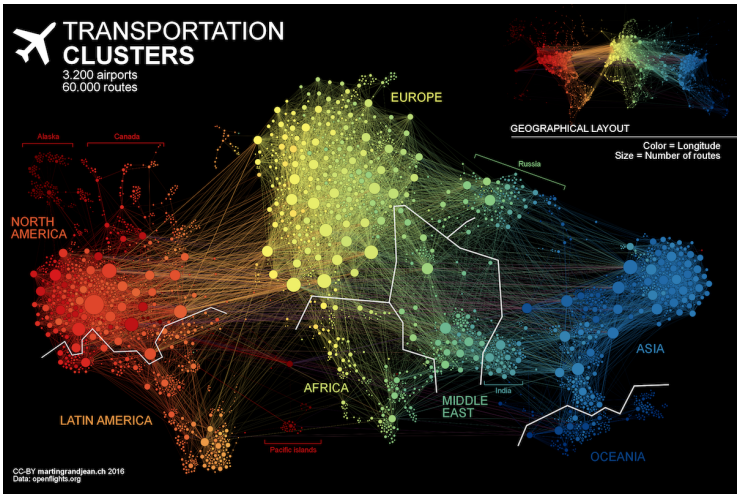


Figure 3.8: Network data visualisations use clustering to visualise relationships (Grandjean, 2016).

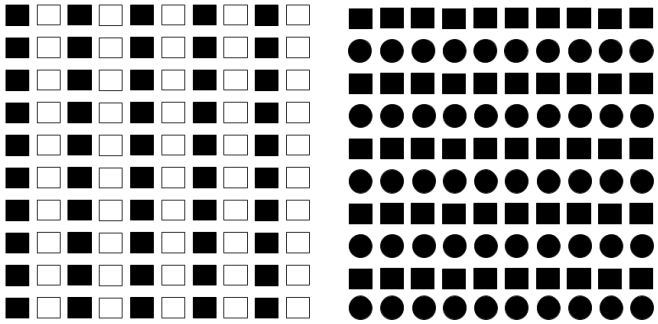


Figure 3.9: Similarity examples.

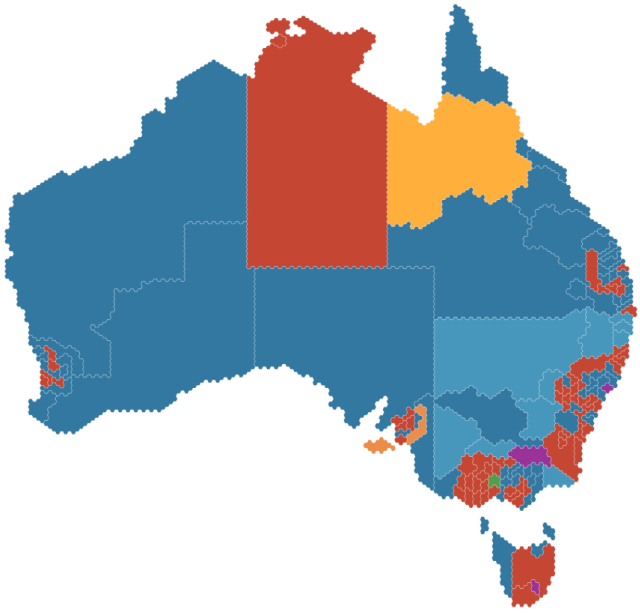


Figure 3.10: The 2019 Australian Federal Election results (The Guardian, 2019).

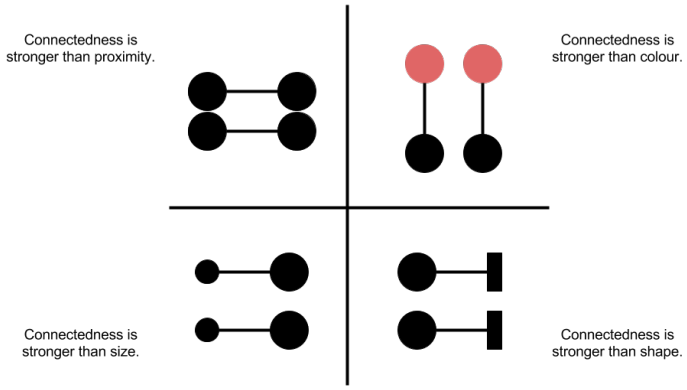


Figure 3.11: Connectedness examples.

Time series plots are the best example of data visualisations that demonstrate connectedness. For example, the *Median values of capital city houses* by Kusher (2018) from CoreLogic shows the significant down turn in the property market experienced in Australia commencing in 2018 (Figure 3.12). The lines connect time series data from each of the capital cities. The connectedness of the lines provides a strong sense of temporal relationship between data points for each city.

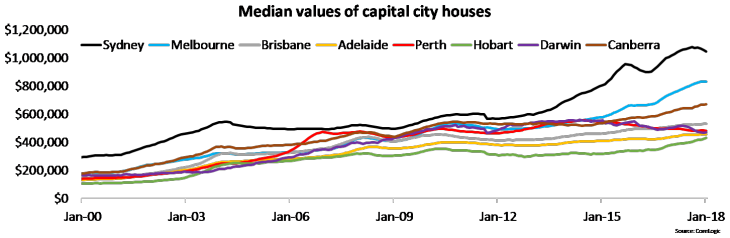


Figure 3.12: Mind the housing value gap (Kusher, 2018).

3.4.2.4 Continuity

This law predicts that we are inclined to perceive objects from elements that are smooth and continuous, versus irregular and jagged (Figure 3.13). Smooth lines are easier to perceive the connection between data points and identify trends.

We should also be careful with how we arrange foreground and background objects that overlap to ensure objects are perceived correctly.

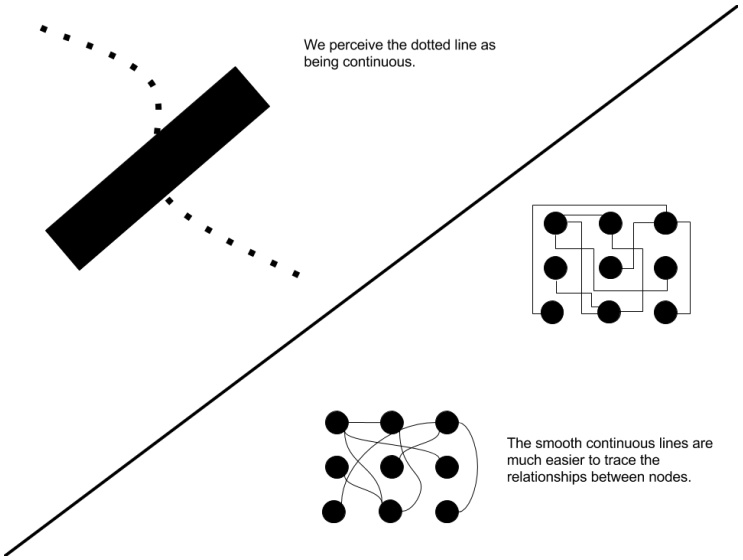


Figure 3.13: Continuity examples.

The smooth continuous lines of a Sankey Diagram demonstrate this law. The Data Team (2015) from the Economist compared the flow of asylum seekers in Europe in 2014 and 2012 (Figure 3.14). The smooth lines that transition between the nodes are easily tracked and are perceptually grouped.

3.4.2.5 Symmetry

We tend to group symmetrical objects together (Figure 3.15). Aligning data visualisations, for example side-by-side, promotes comparisons and allows differences to be readily perceived.

The world population pyramid by Roser (2019) from *Our World in Data* uses symmetry to compare the age distribution of males and females across time and projected into the future (Figure 3.16). By aligning the population distributions as a mirror image, you can see some subtle differences based on the projections into 2100. The projections suggest males will outnumber females, but with females living longer.

3.4.2.6 Closure

Closure refers to our tendency to "fill in the gaps" when we see incomplete patterns that resemble familiar shapes and objects (Figure 3.17). This means we

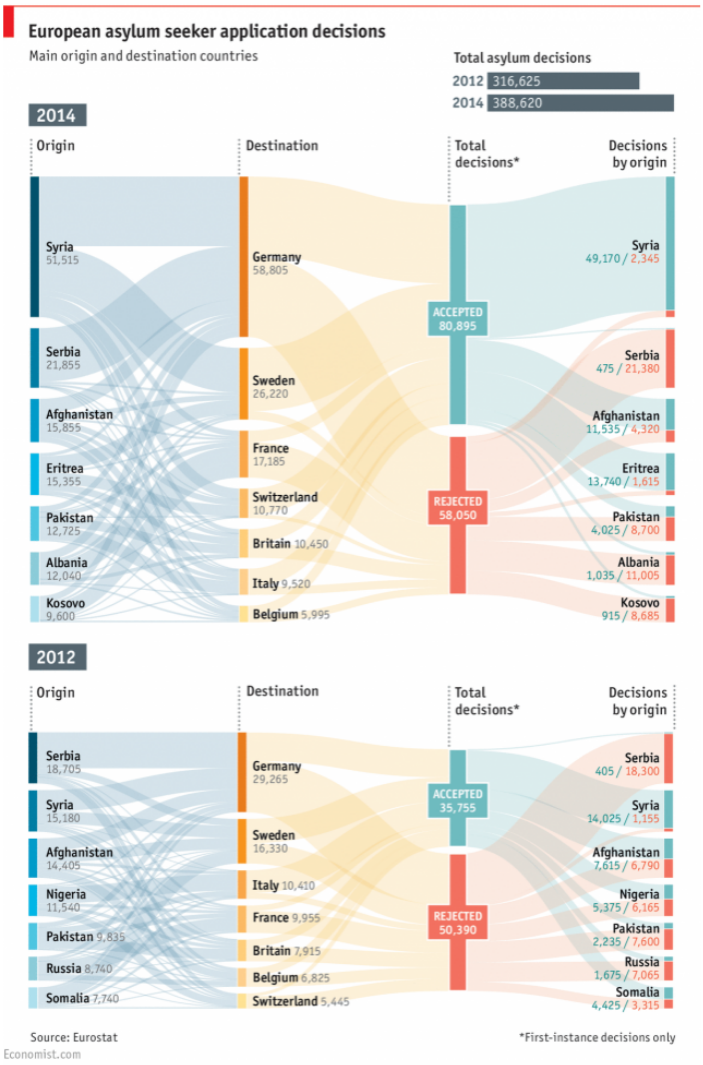


Figure 3.14: The smooth continuous lines of a Sankey Diagram demonstrate continuity (The Data Team, 2015).

We tend to see three pairs of different brackets, as opposed to six brackets.

[ ] { } ( )

Figure 3.15: Symmetry examples.

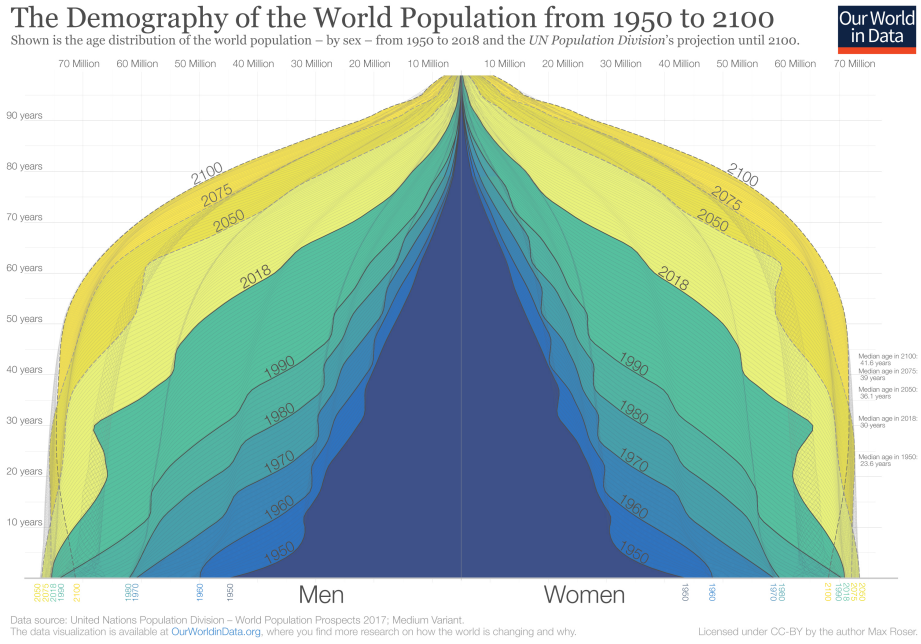


Figure 3.16: Population pyramids use symmetry (Roser, 2019).

should group related objects around closed shapes and ensure that overlapping objects are "closed" correctly by the brain. Contrasting overlapping objects using shapes and colour can help ensure the correct closure is achieved.

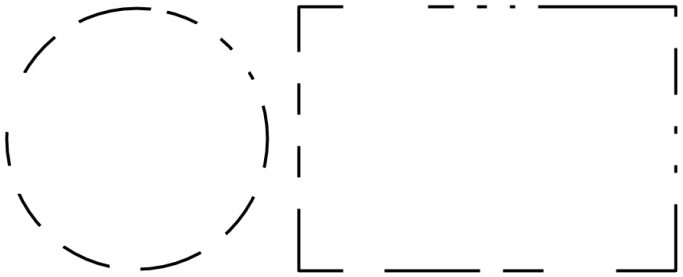


Figure 3.17: Closure examples.

Closure comes in handy when we have overlapping elements. For example, when you have two overlapping points, our brains fill in the shape and perceive two points. Ridgeline plots like Smith (2018) also show this law in action. Here you perceive the full density distribution even through it is obscured by other distributions (Figure 3.18). The transparency also helps.

3.4.2.7 Figure Ground Principle

The figure ground effect tells us that smaller objects within a figure are interpreted as the foreground, while larger objects make up the background (Figure 3.19). Data visualisations often plot data objects to backgrounds. This ensures that the objects within the border of the background are perceived to be representations of the data.

This law explains why non-data elements, such as grid lines, plot background and axis labels are faded into the background. You want the data to draw the audiences' attention. Zev (2016) shows want happens when you over-emphasise background plot colours and grid lines (Figure 3.20). Instead of the data drawing attention, you first must look past the grid lines which compete for the foreground. Think of the data as the prisoner and the bars of the prison cell as the grid lines. The data appear to be locked-up.

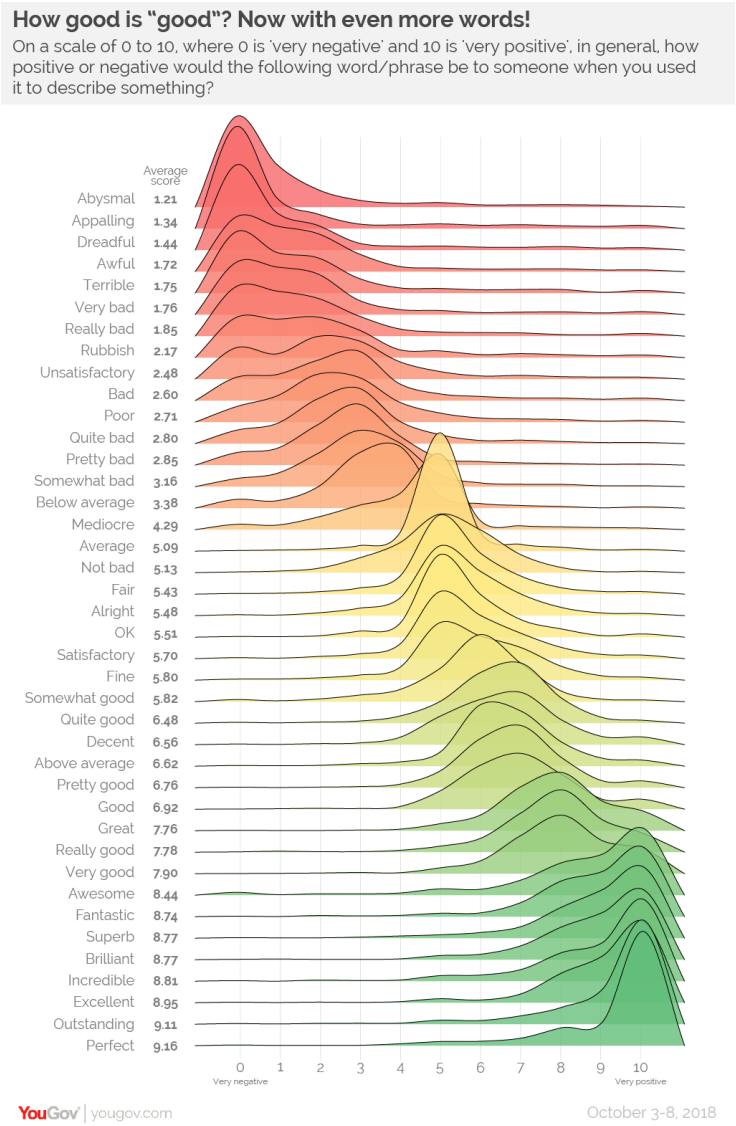


Figure 3.18: Closure helps reduce confusion with overlapping elements (Smith, 2018).

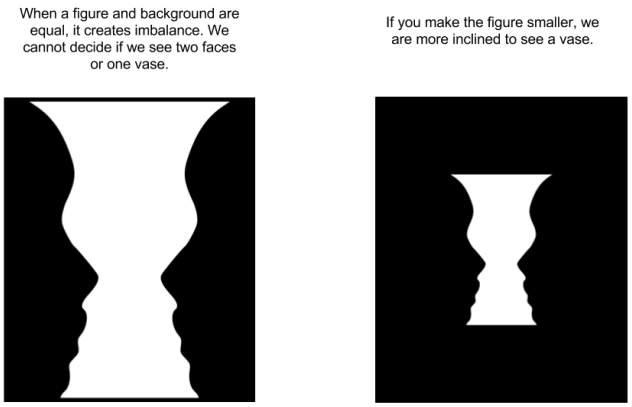


Figure 3.19: Figure group principle examples.

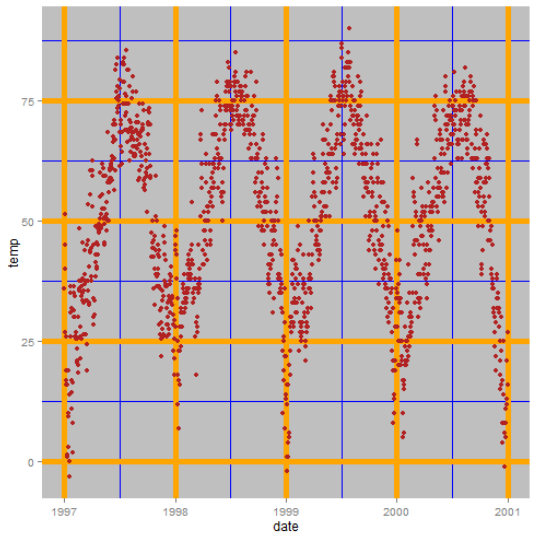


Figure 3.20: Bold grid lines compete with the data for visual attention (Zev, 2016).

3.4.2.8 Common Fate

Objects perceived to be moving in the same direction are grouped together and share a common path (Figure 3.21). Animated and interactive data visualisations use this principle to show relationships between categories and highlight changes across time or based on user input.

Figure 3.21: Common fate examples (Only available online).

Han Rosling's (BBC, 2010) animated bubble charts show this law in action (Figure 3.22, https://youtu.be/jbkSRLYSojo). Countries that move together are perceived to be related. Rosling uses this law to explain the underlying factors that drive these trends.

Figure 3.22: Hans Rosling's animated bubble charts (BBC, 2010)(Only available online).

3.4.3 Change and Inattention Blindness

Watch the following video for a great overview and example of change and inattention blindness (Only available online - https://youtu.be/VkrrVozZR2c).

Change blindness is caused by the finite capacity of our short-term memory. We simply cannot process and store all the information that we receive from our environment. Instead, we must limit our attention and memory to the important things that are occurring and sometimes this means we will miss subtle, sometimes abrupt, changes in the environment due to breaks in attention or interruptions to line of sight. This has important implications for data visualisations that include dynamic or animated features. It's important to ensure that changes are easy to pay attention to and are not hidden by distractions. For interactive visualisations that change multiple parameters, each change to the data visualisation should use visual perception laws to ensure the changes "pop out" and draw attention.

Inattention blindness is also caused by the limitations of our short-term memory. Attentional limitations imposed by a finite short-term memory means that we can only focus our attention on a limited number of objects in the environment. The more attention required, the more likely that other obvious, but irrelevant, objects in the environment will go unnoticed. This phenomenon has been extensively researched and replicated in real-world experiments. Watch the following movie and count how many times the participants wearing white pass the ball (Only available online - https://youtu.be/vJG698U2Mvo). Once you have the answer, visit this website here to see if you are correct. If you haven't seen this video before, you might be very surprised with the answer.

Inattention blindness shows us that human beings have a limited ability to hold information in short-term memory. If we over-burden the viewer with unnecessary visual complexity or too many irrelevant/redundant objects, we run the risk of the viewer missing important details. To avoid inattention blindness ruining our data visualisations, Kirk's process (Chapter 1) cautions us to exercise editorial focus and narrow down on the salient features of the data.

3.5 Visual Variables

Based on what we have learnt about visual perception, we can define a common set of visual objects/features/aesthetics that can be mapped to different types of variables. Figure 3.23 lists the most common examples. There are many more examples out there, but most are variations of the ones listed below. Note that the quantitative scales refer to all interval and ratio variable measures on discrete or continuous scales. Some features can be used to represent multiple data types. As such, the following list represents the most common type that you are likely to see paired with the different features. All demonstrate good visual design principles based on human visual information processing theory.

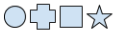
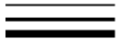
Feature/Object	Example	Notes	Nominal	Ordinal	Quantitative
Position		Powerful and accurate			✓
Length		Powerful and accurate			✓
Size		Less accurate, but still useful			✓
Angle/Slope		Less accurate, but still useful			✓
Shape		Can only use a few different shapes.	✓		
Colour - Hue		Can only use a few different hues.	✓		
Colour - Sequential		Can only use a few different ordered levels.		✓	
Colour - Diverging		Can only use a few different ordered levels.		✓	
Colour - Continuous		Less accurate than other quantitative encodings			✓
Texture		Bygone feature, but may still have applications.	✓		
Line - Type		Can only use a very limited number of types.	✓		
Line - Weight/Boldness		Not very accurate, but still useful.			✓

Figure 3.23: Visual variables.

3.6 Visual Comparison Accuracy

Different methods of comparing quantitative data in visualisations will impact the time it takes for the viewer to accurately complete comparisons. This is based on our visual perception system and our brain's ability to perceive certain patterns more readily than others. Cleveland and McGill (1985) proposed a hierarchy of accurate data visualisation methods commonly used to represent comparisons. Figure 3.24 summarises this hierarchy.

Visual Comparison Accuracy
(Cleveland and McGill, 1985)

Rank (Most - Least)	Feature
1	Position - common scale
2	Position - same scale , but unaligned
3	Length
4	Angle
5	Slope
6	Area
7	Volume
8	Density
9	Colour saturation
10	Colour hue

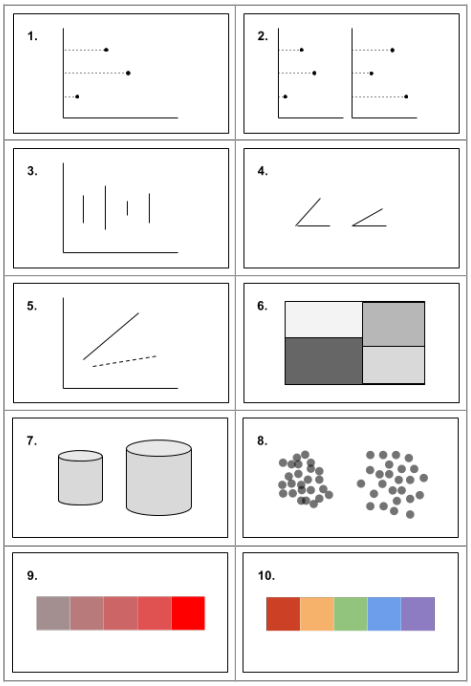


Figure 3.24: Visual comparison accuracy adapted from Cleveland and McGill (1985).

Cleveland and McGill (1985) informs us that position, length, angle and slope are the most accurate representations for comparing quantitative variables. We should carefully use area, volume and density, and only use colour saturation and hue where absolutely necessary. This is not to say that colour is ineffective. The contrary is true. However, colour is better suited to grouping related objects as opposed to reflecting a quantitative variable used for comparison. Judging the degree of difference between saturation and hue is a lot less accurate than position, length and the like. We learn more about colour in the next section.

3.7 Colour

In the following video, Colm Kelleher explains the difference between the physical concept of colour and colour perceptions (Only available online - https://youtu.be/UZ5UGnU7oOI & https://youtu.be/l8_fZPHasdo)

MacDonald (1999) wrote, *"Color is one of the most effective visual attributes for coding information in displays and is capable, when used correctly, of achieving powerful and memorable effects."* (p. 26). Colour is used extensively in data visualisation to make elements "pop out" (preattentive processing), improve aesthetics, trigger certain emotional associations, draw connections between related elements and reflect quantitative values. Colour is a powerful data visualisation tool, but it must be used with care. In this chapter we will take a close look at the topic of colour and what we need to know to use colour effectively in practice.

Colour perception is our visual perception system's response to the visible spectrum of light. Visible light is electromagnetic radiation emitted in wavelengths between 400 - 700 nanometer (nm)(Figure 3.25). Light outside this range is not visible to the human eye and include the infrared range (700 nm to 1 mm) and the ultraviolet range (10 - 400 nm).

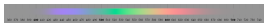


Figure 3.25: The visible light spectrum (Spigget, 2010).

Within the visible light range there are three primary colours that eye's photoreceptors (cones) can respond to, hence humans are said to be trichromatic. These colours are blue, green and red (don't confuse these with the primary colours used in art - blue, yellow, red). The relative activation of the three types of cones within the eye determine the colour we perceive, assuming a person is not colour blind (Figure 3.26).

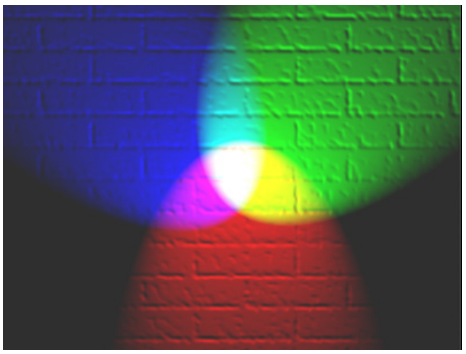


Figure 3.26: The primary colours of human vision are blue, green and red (En:User:Bb3cxv, 2007).

The three primary colours interact with each other to form a three-dimensional colour space (Figure 3.27).

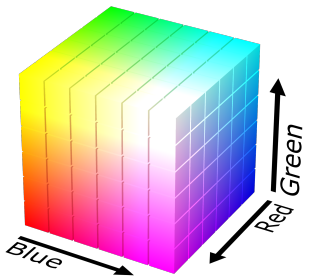


Figure 3.27: Three colours create a 3d colour space (SharkD, 2008).

It's important to note that colour is not discrete. Therefore, defining colour isn't as simple as listing all the colours you know. While our brains automatically transcode colour from electrical impulses sent from the eye, computers, which run data visualisation software, need colour models that can form a basic language for defining colour rendered on computer screens and used in print. In the next section we will take a look at the RGB and HSV colour models.

3.8 Colour Models

Colour models aim to specify colour in a standard way (Silva et al., 2011). This is very important for computing and web technologies which rely heavily on colour. The most efficient way to do this, is to create a colour model that can be used to specific a colour and reproduce that colour on another system using small pieces of code. There are numerous colour models (e.g. RGB, CMYK, HSV and HSL), however, in this section we will pay close attention to one of the most intuitive.

The HSV model specifies colour using three main parameters: hue, saturation and value, or HSV (Figure 3.28). **Hue** refers to what we perceive to be different colour types, e.g. red, green, blue, orange etc. **Saturation** refers to colour purity, and value to the brightness of the colour. To remember the difference between saturation and value, consider the following analogy. When you buy a new bright red (hue) shirt, the colour is very pure. However, after you wash the shirt and dry it on the clothesline, the colour slowly fades. Basically, purity reduces because the colour pigments are slowly being washed out and destroyed by the sun. Value refers to brightness. You are wearing your bright red shirt outside on a sunny day. It captures everyone's attention because of the brightness of the sun. The sun starts to go down and night approaches. There is less

light, so your shirt appears less bright. Eventually, your shirt appears almost black because it is dark outside. Most colour models also include an **alpha** parameter which codes for transparency. While not technically a colour parameter, alpha is a very important scale in data visualisation for dealing with overlapping elements.

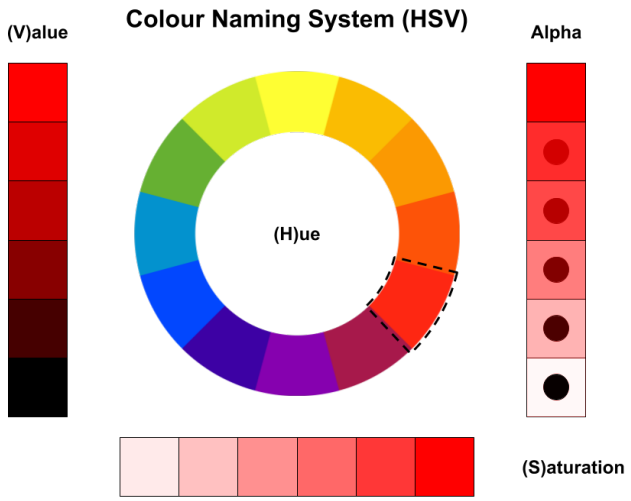


Figure 3.28: The HSV colour model.

Many data visualisation packages are based on the common RGB (red, green and blue) colour model. Therefore, to use HSV, you may need to convert between colour models. This is easy to do using online tools discussed below.

To make coding colours more efficient in web programming, a hexadecimal colour coding system was developed. The hexadecimal system is a base 16 number system. Each colour (red, green and blue) is represented as one of 16 characters using the following format: `#rrggbb`. The 16 characters include:

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F

Therefore… you can select different hues as follows:

Black = `#000000`

White = `#FFFFFF`

Red = `#FF0000`

Green = `#00FF00`

Blue = `#0000FF`

You can also express the hexadecimal systems as a number between 0 - 255 (i.e. 162), to represent an RGB colour. For example, black would be

`rgb(0,0,0)`, white `rgb(255,255,255)` or red `rgb(255,0,0)`. However, the hex codes are more common in web technologies. Working with hex codes can be difficult, so you should explore colour picking tools that will allow you to visually scan the colour space and output the relevant hex code (Figure 3.29).

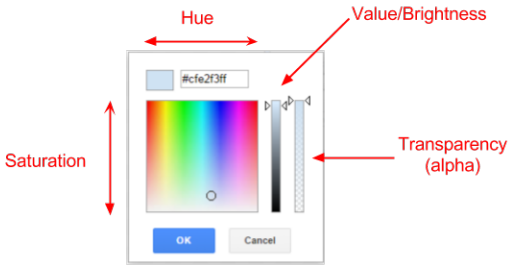


Figure 3.29: Colour pickers make choosing colours easy.

There are also excellent websites like http://colorizer.org/ that allow you to pick colours, check colour codes and convert between different colour models (HSV, HSL, RGB and CMYK). Figure 3.30 summarises the hex based colour coding system used in many data visualisation packages.

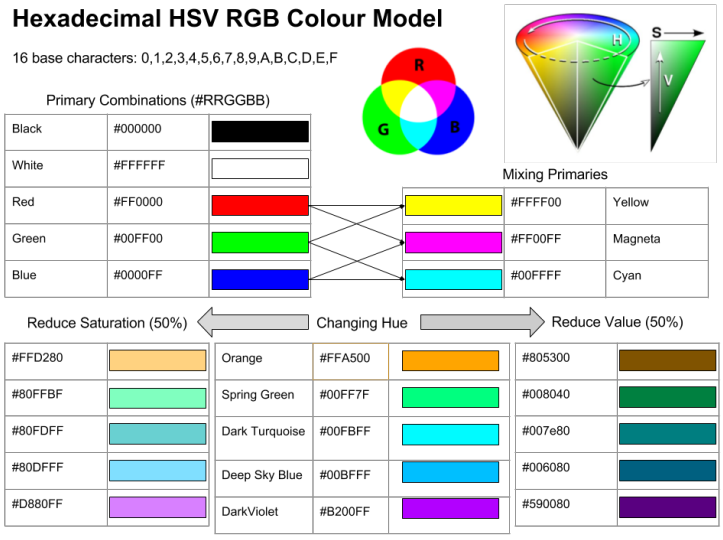


Figure 3.30: The hex colour coding system.

3.9 Colour Scales and Data Types

Colour is a very versatile visual property used in data visualisation. One of the most common ways it is used is to represent a variable. There are a range of different colour scales that can be used to represent different types of variables. Figure 3.31 provides a useful overview. The infographic splits the colour scales by the types of variables - nominal, ordinal, interval and ratio. Nominal and ordinal colour scales are said to be **discete** because each level represents a single colour. Ordinal colour scales incorporate a sequence of discrete colours that represent the ordering present in the variable. **Continuous** colour scales, on the other hand, can take on any colour value within a range specified by the scale, which makes them ideal for interval and ratio variables.

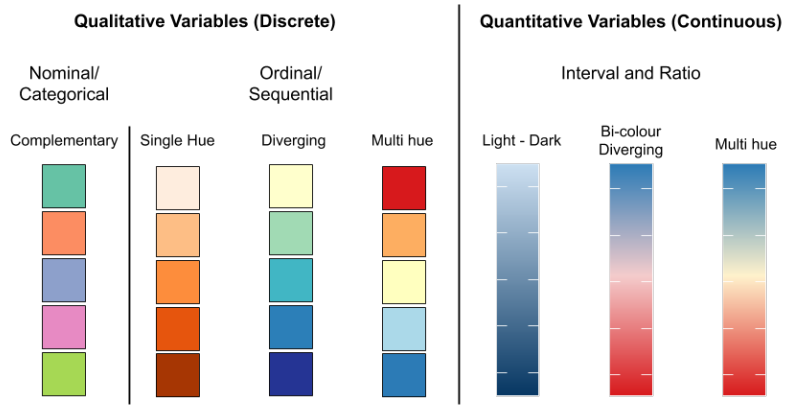


Figure 3.31: Colour can be used to create many different types of scales.

Changing colour scales have sometimes have drastic effects on the appearance of a visualisastion. Find the right colour scale can be tricky. Fortunately, there are tools available to assist.

3.10 ColorBrewer

Choosing colour scales is not an easy task. The first problem you are faced with when choosing colours is that are easy to distinguish. This might be easy for a few colours, but the problem gets harder as the number of colours required grows. How do you know the colours are readily distinguishable? How do you know your colours will appear consistently on other devices? Are your colours colour blind safe? Fortunately, there are tools that can help such as the ColorBrewer scales (Harrower and Brewer, 2003; Brewer and Harrower, 2019).

Cynthia Brewer developed these scales to promote best practice in map representation, being a geographer herself. However, her work has much broader applications to data visualisation. The ColorBrewer 2.0 web tool allows a user to select colour themes based on sequential, diverging or qualitative data, single or multi-hue schemes and colourblind, print- and photocopy-safe palettes. The tool also allows you to preview your colour scheme in order to check suitability and accuracy and allows you to export the scheme as a vector of colour hex codes for easy implementation into applications (Figure 3.32).

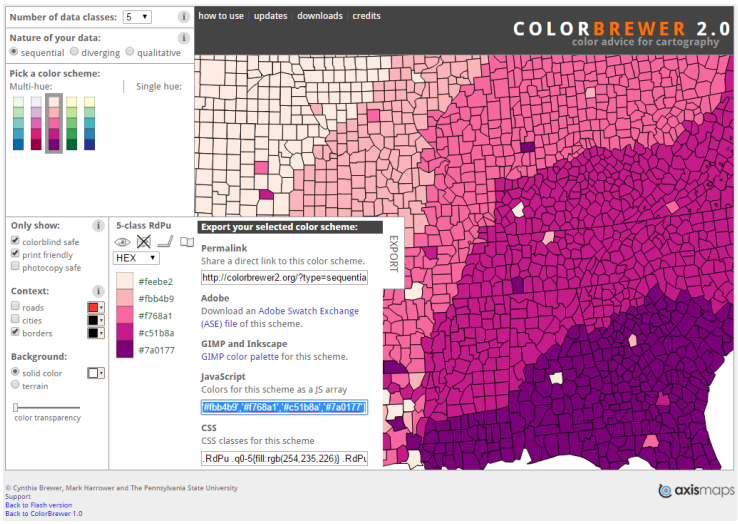


Figure 3.32: ColorBrewer web tool (Brewer and Harrower, 2019).

This tools provides some interesting insight into the limitations of colour scales. The maximum number of colour levels/classes is 12 (Figure 3.33). However, this is only possible for qualitative variables. Looking at the map, I think you would agree that 12 colours is hard work. There is a lot of looking back and forth between the map and the colour legend.

Things are even more interesting if you want to choose a colourblind safe palette. The maximum number of colours considered colourblind safe in the qualitative scales is only four (Figure 3.34). However, colour blind scales for sequential and diverging scales can support up to 11 colours. This is a nice segue to the the next section on colour blindness.

3.11 Colour Blindness

You need to do your best to ensure your data visualisations can be readily interpreted by people with colour blindness. Colour blindness is caused by missing or dysfunctional cones in the eye's retina. Approximately 8% of males and 0.4%

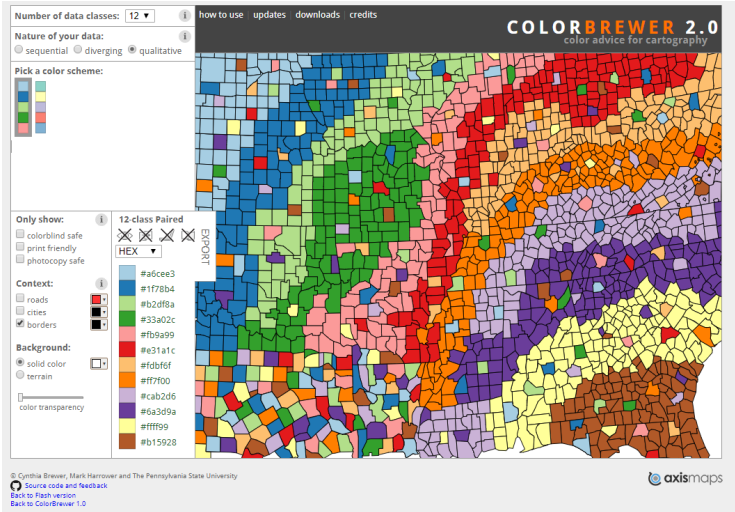


Figure 3.33: Too many colours… (Brewer and Harrower, 2019).

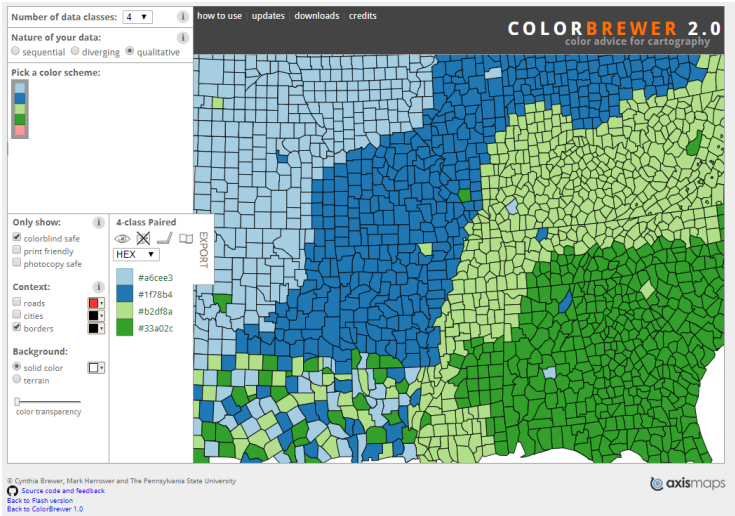


Figure 3.34: Few colour blind safe colour palettes are available (Brewer and Harrower, 2019).

of females have some form of colour blindness. There are many variations based on the type and number of dysfunctional or missing cones in the retina. The red-green colour blindness types are the most common. These types make it difficult to distinguish red from green. Figure 3.35 will help you to understand the different classifications of colour blindness.

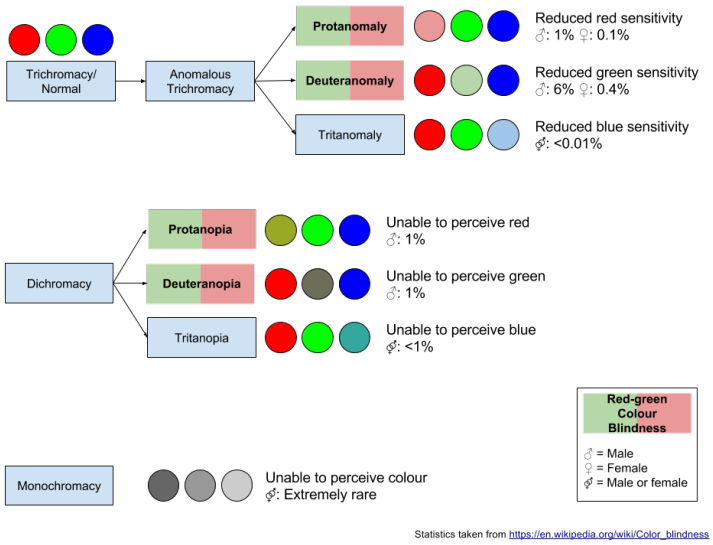


Figure 3.35: Colour blindness.

Figure 3.36 shows how colour blindness changes the perception of colour. Of course, I am assuming you're not colour blind! The images above were created using the Coblis Colour Blindness Simulator (Wickline, 2001). This is a very useful tool.

Colour blind simulators can help us to determine how our data visualisation will appear to people with colour blindness. You won't be able to cater to all forms of colour blindness. Red-green colour blindness is the most common, so checking your visualisation using a red-green colour blindness simulator is recommended.

3.12 Colour Associations

Colours have many natural and cultural associations. We need to keep these in mind when choosing colours as we might need to avoid particular associations or use these associations to trigger the right emotional response. Table 3.1 was reproduced from MacDonald (1999).

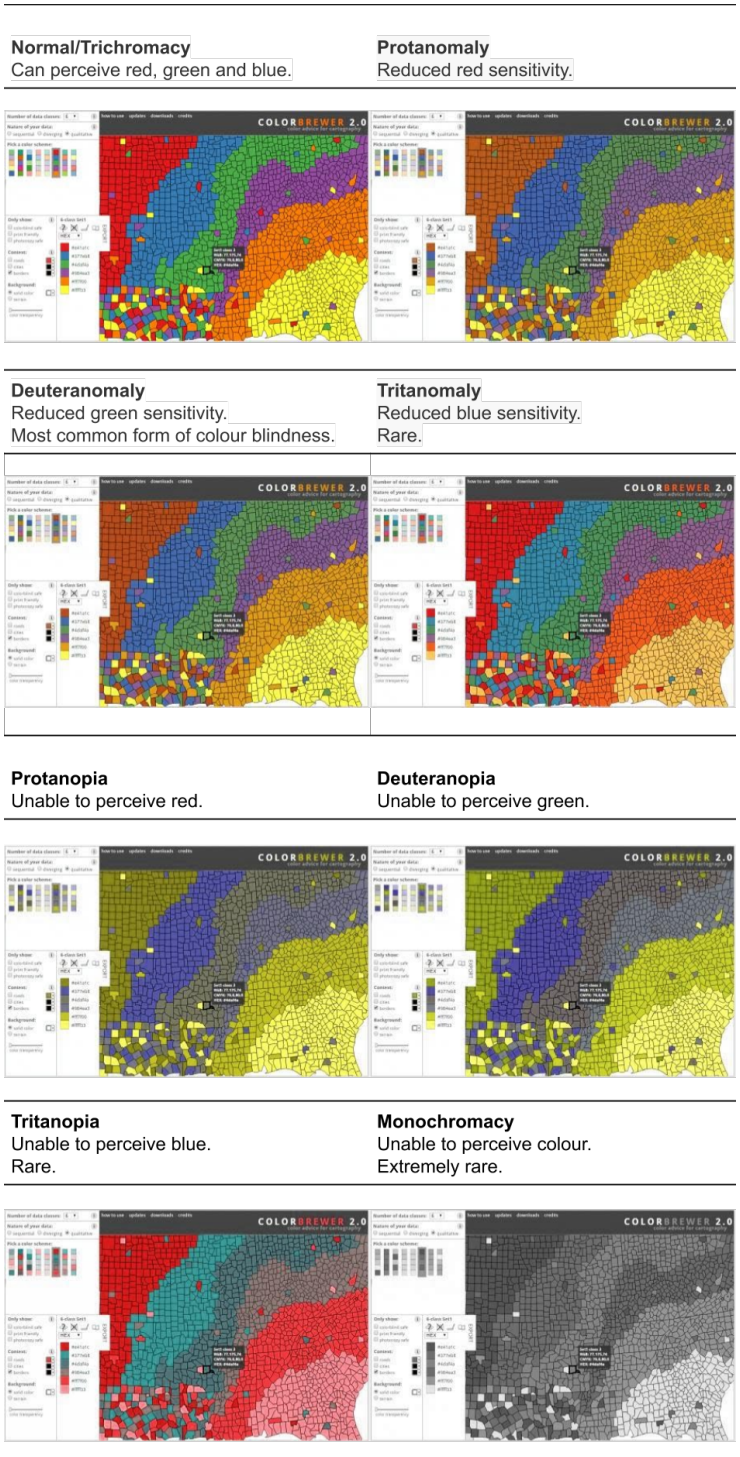


Figure 3.36: Different types of colour blindness and their simulated effect.

Table 3.1: Colour Associations adapted from MacDonald (1999).

Colour	Positive	Negative
Red	Passion, strength, energy, heat, love	Blood, war, fire, danger, anger, aggression
Green	Nature, spring, fertility, safety, environment	Inexperience, decay, envy, misfortune
Yellow	Sun, summer, gold, harvest, optimism	Cowardice, treason, hazard, illness, folly
Blue	Sky, sea, stability, peace, unity, depth	Depression, obscenity, conservatism, passivity
White	Snow, purity, peace, cleanliness, innocence	Cold, clinical, surrender, sterility, death, banality
Gray	Intelligence, dignity, restraint, maturity	Shadow, concrete, drabness, boredom
Black	Coal, power, formality, depth, solidarity, style	Fear, void, night, secrecy, evil, anonymity

3.13 Responsible Use of Colour

Colour is a big topic. We have outlined the basics, but you really need to know how to use colour responsibly in data visualisation. The following sections will outline the most important considerations that all designers must understand. The ideas are all explained using examples. This advice has been adapted from Few (2008), Lujin Wang et al. (2008), MacDonald (1999), Silva et al. (2011) and Ware (2013). If you stick to these rules and you will rarely go wrong.

3.13.1 Use colour with purpose

The first rule is simple. Use colour with purpose (Few, 2008). Figure 3.37 A. uses colour unnecessarily. It does not improve the effectiveness of the data visualisation in communicating insight. Figure 3.37 B. shows that less is more. Colour is used to differentiate the "All employees" categories from the age bands. This signals to the viewers that the "All employee" category is different. No other colours are necessary.

3.13.2 Use colour to differentiate important features

Colour is a powerful way to help the audience differentiate features, groups, or values in a data visualisation (Few, 2008). Figure 3.38 A. uses a different colour for each category on the x-axis. The x-axis is one variable and the categories within the variable are already defined on the x-axis. The use of colour is redundant. Figure 3.38 B., on the other hand, uses colour to differentiate the age categories from a different statistical summary, "All employees".

3.13.3 Ensure colour constancy

Our perception of an object's colour can change based on the colours of the objects surrounding it. Consider the squares in Figure 3.39. The two squares

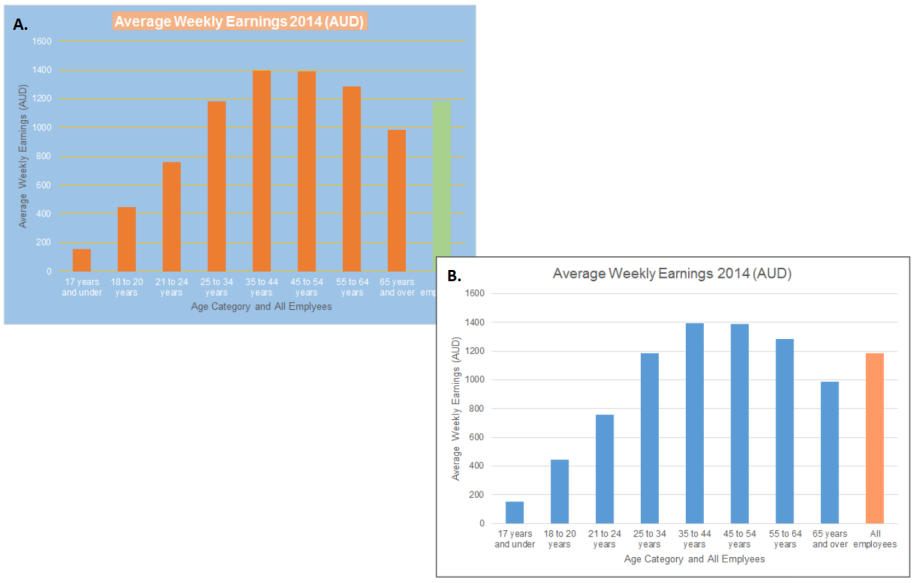


Figure 3.37: A. Use colour sparingly. Less is more. B. Use as few colours as possible to achieve what you intend to communicate.

Figure 3.38: A. Colour is used in a redundant manner. B. Colour is used to differentiate between two different statistical summaries - average earning by age category vs. all employees.

in the image appear to have a different level of saturation.

Figure 3.39: Two squares, different colours?

Now consider what happens when we remove the background colours and use a constant white background (Figure 3.40).

Figure 3.40: Same colour…

The colours are exactly the same. Not convinced? Consider Figure 3.41. In this image, the squares appear to be the same colour and saturation, however, if you trace the objects' colours side-by-side, they are clearly different.

Few (2008) summarises this rules as follows: *"If you want different objects of the same colour in a table or graph to look the same, make sure the background colour that surrounds them is consistent."*

Figure 3.41: Our perception of colour depends on the other colours surrounding an object.

3.13.4 Ensure adequate contrast

Data visualisation is first about the data. You want the data to "pop-out" and draw people's attention. You also want non-data elements (e.g. titles, labels, axis labels etc.) to be easily read. Therefore, data visualisation must use background colours that provide sufficient contrast (Few, 2008). Figure 3.42 provides examples of good and bad contrast.

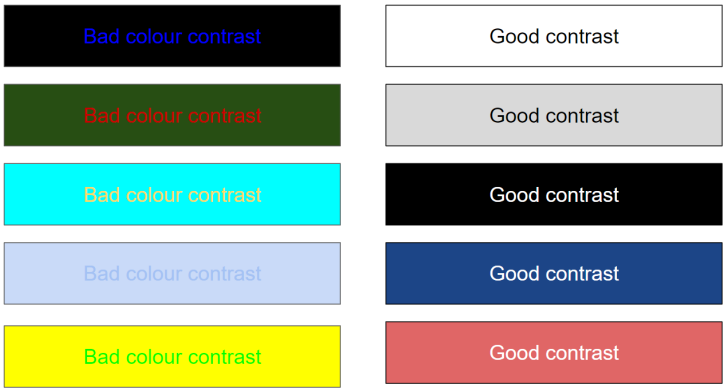


Figure 3.42: Examples of good and bad contrast.

3.13.5 Define objects with equiluminous colour using thin borders

When you have overlapping coloured elements, sometimes the elements can be difficult to discern, especially, if the elements have the same level of brightness (equiluminous). Ware (2013) recommends using a thin border with a higher level of luminosity (e.g. white) to improve object definition. This principle is demonstrated in Figure 3.43.

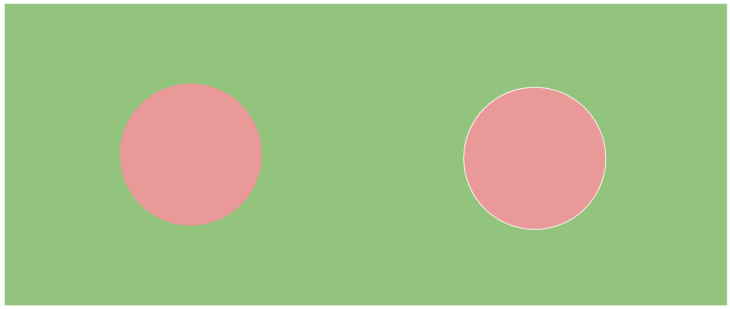


Figure 3.43: The circle with the white border is seen more clearly.

3.13.6 Avoid highly saturated colours

Stare at the dot in the coloured image of Figure 3.44 for 30 seconds. Then look at the dot in the right image for another 10 seconds. What do you notice?

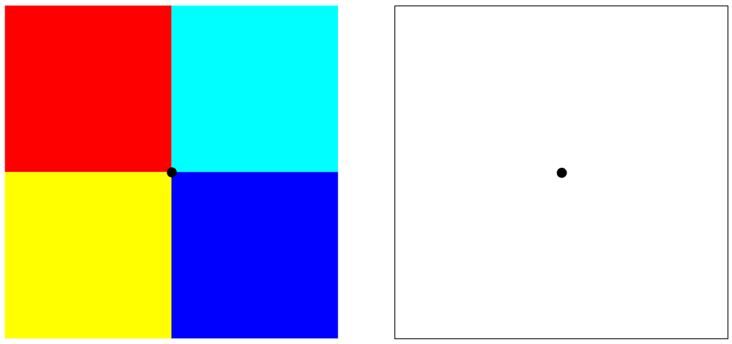


Figure 3.44: After image experiment.

The opposite colours are superimposed on the white background. This is an example of an after image. The reason the colours are inverted (red > cyan, cyan > red, yellow > blue, and blue > yellow) relates to retinal fatigue. The cones in the retina that respond to a given colour become fatigued, for example

the red cones. When you look at the white background, the three cone types (red, green, blue) are equally stimulated, but because the red cone type is fatigued, the other two cone types, green and blue, respond equally producing a cyan coloured after image. The effect is weakened if the colours are weakened. Run the experiment again using Figure 3.45

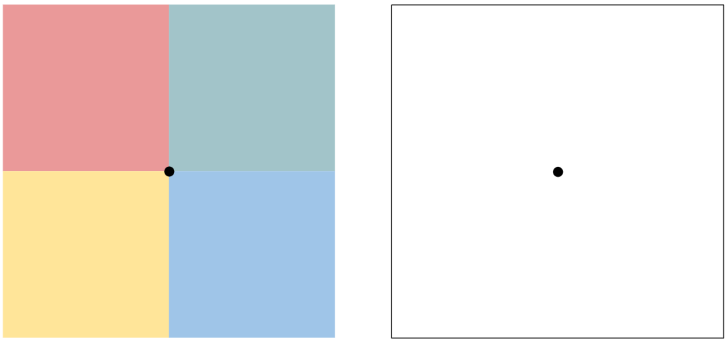


Figure 3.45: The after image is less pronounced and more transient.

So what is the lesson here? MacDonald (1999) explains as follows: *"Areas of strong colour and high contrast can produce after images when the viewer looks away from the screen, resulting in visual stress from prolonged viewing."* (p. 22).

3.13.7 Reserve bright colours to highlight important information

Colour is one of the most versatile tools available to a designer. It is incredibly powerful at drawing viewers' attention. Therefore, when we use it, it must serve an important purpose. Few (2008) recommends using soft or natural colours for most elements and reserving bright or dark colours to draw attention to the most crucial information. Using bright and artificial colours can also produce unsightly visualisations (see Figure 3.46).

3.13.8 Saturated colours can be used for small data points

There are situations when bright colours can be helpful. Small data points are often hard to see (Figure 3.47 A.). Increasing the size of the points as well as increasing their brightness can make them easier to see (Few, 2008) (Figure 3.47 B.).

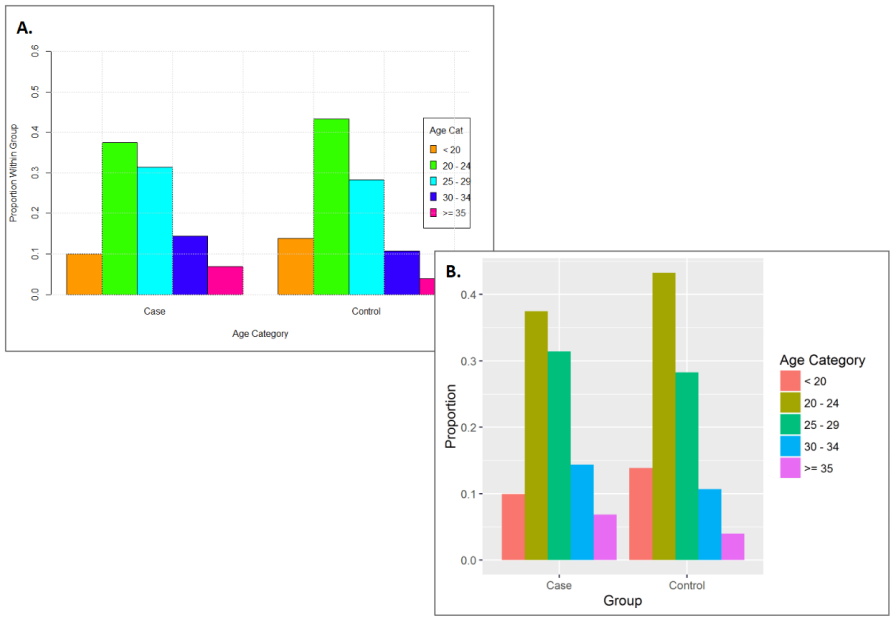


Figure 3.46: A. Bright colours belong in Vegas. B. Reduce saturation and use more natural hues in most situations.

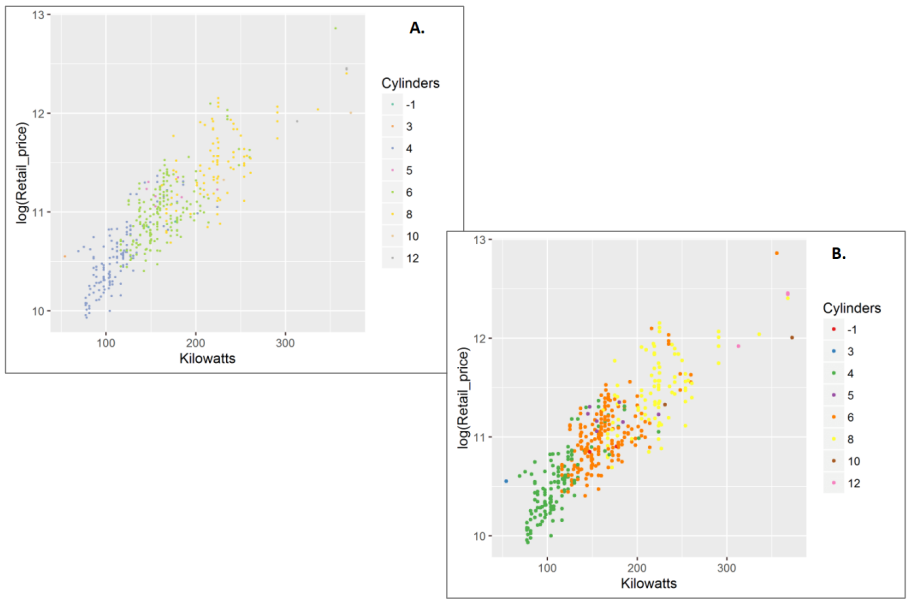


Figure 3.47: A. Small data points are hard to see. B. Increase size and colour brightness to make them easier to see.

3.13.9 Use colour scales to encode important information

Colour can be used to create many scales; nominal, ordinal, interval and ratio When using colour to represent a variable, ensure the colour scale matches the variable type (Few, 2008). For example, a diamond's cut, fair, good, very good, premium and ideal is a sequential or ordinal variable. Figure 3.48 A. uses a complementary or nominal colour scale to encode cut. This is a missed opportunity. Figure 3.48 B. corrects this and uses a single hue, sequential colour scale.

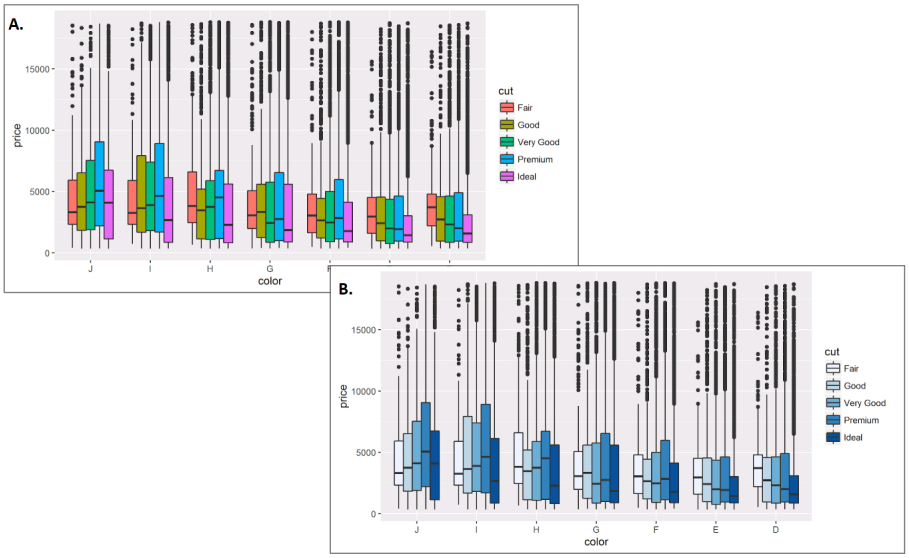


Figure 3.48: A. Cut is an ordinal variable, but the colour scale is nominal. B. The colour scale for cut is now ordinal.

3.13.10 Non-data elements should not compete with the data

Titles, labels, grid lines, and plot backgrounds are all examples of non-data elements. These features should not detract from the data or key elements that communicate insight. They should be visible only enough to perform their supportive role (Few, 2008). For example, the bold gridlines and colourful background in Figure 3.49 A. enhance the accuracy of reading data points from the plot and constrasting data points, respectively. However, they complete for attention with the data. A grey background and white grid lines are still noticable and supportive, but ensure your attention first focuses on the trend in the data (Figure 3.49 B.)

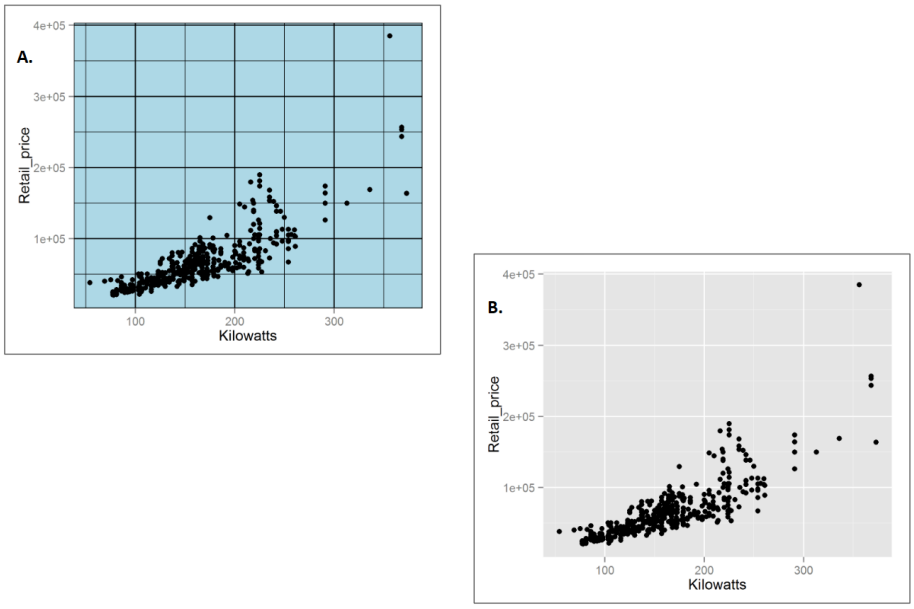


Figure 3.49: A. The bold grid lines and colourful plot background detract from the data. B. A grey background and white gridlines ensure the data speak for themselves.

3.13.11 Try to avoid colour scales that use red and green.

Colour scales that use red and green should be avoided where possible due to the most common forms of red-green colour blindness (Few, 2008). For example, the continuous red-green colour scale used in Figure 3.50 A., appears very different to a person with Dueteranopia (unable to perceive green - Figure 3.50 B., simulated).

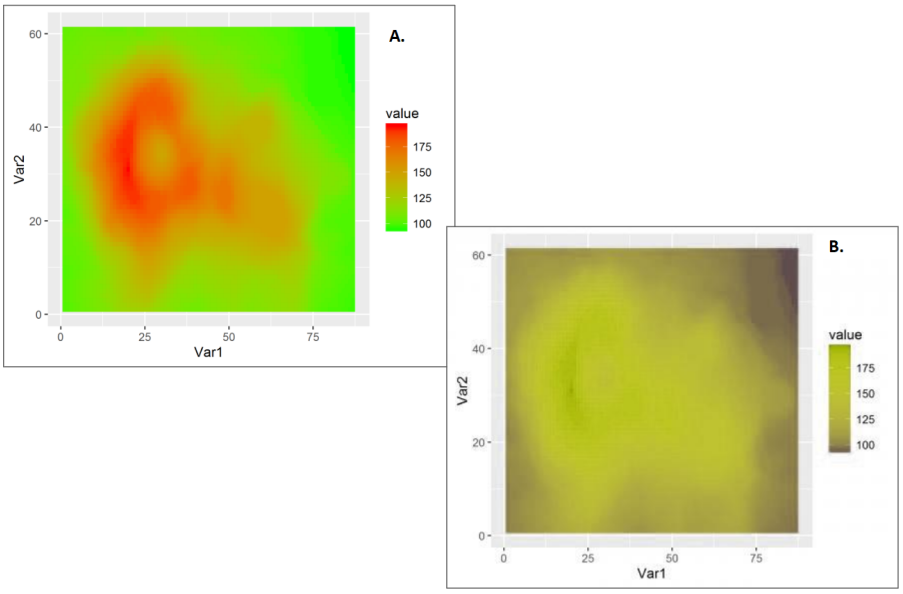


Figure 3.50: A. A red-green continuous diverging colour scale. B. The same plot as A. but the simulated appearance for someone with Dueteranopia.

3.13.12 Avoid visual effects.

Special effects belong in Hollywood movies. Avoid using them in visualisations (Few, 2008). For example, the shadows in Figure 3.51 A. serve no purpose except for a cheap visual thrill. The 3D effect in Figure 3.51 B. is worse. It introduces a depth distortion. Because the "17 year and under" category appears further away, the height of the bar will be underestimated.

3.14 Concluding Thoughts

This chapter introduced and discussed the important relationship between data visualisation and human visual information processing. Vision is a construct of

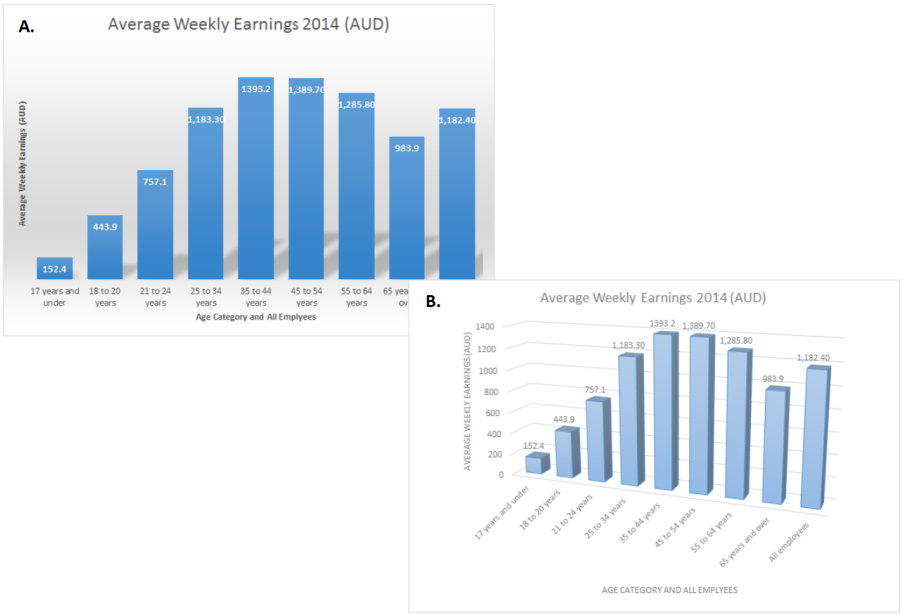


Figure 3.51: A. The shadows serve no purpose. B. 3D effects can distort data visualisations.

the brain enabled by rapid processing and powerful heuristics. However, optical illusions remind us that our visual system has its limitations. As a data visualisation designer, you must respect the power and limits of human vision. Preattentive processing demonstrates that certain visual features are more readily seen and data visualisation takes full advantage of this ability. Gestalt laws help designers to use human pattern recognition laws to relate meaning to features in data. Human vision relies heavily on colour, and therefore, there is little surprise that colour is a designer's most versatile tool. Colour can be used to draw attention, differentiate groups, highlight important features and represent statistical quantities. This chapter introduced the physical and perceptual concepts of colour, computer colour models and, most importantly, the rules of responsible colour use for data visualisation. The theoretical and practical knowledge covered in the chapter has laid a solid foundation that informs many of the design decisions to come.

Bibliography

(1988). Privacy Act.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17.

BBC (2010). Hans Rosling's 200 countries, 200 years, 4 minutes - The joy of stats - BBC Four.

Blue, V. (2018). Strava's fitness heatmaps are a 'potential catastrophe'.

Boyer, K. L. and Sarkar, S. (2000). *Perceptual organization for artificial vision systems*. Springer, New York, NY, 1st edition.

Brewer, C. A. and Harrower, M. (2019). ColourBrewer 2.0 web tool.

Brodlie, K., Allendes Osorio, R., and Lopes, A. (2012). A review of uncertainty in data visualization. In Dill, J., Earnshaw, R., Kasik, D., Vince, J., and Wong, P., editors, *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109. Springer London, London.

Brown, T. (2008). Design thinking. *Harvard Business Review*, June:1–10.

Cadwalladr, C. and Graham-Harrison, E. (2018). Revealed: 50 million facebook profiles harvested for Cambridge Analytica in major data breach.

Cairo, A. (2014). Ethical infographics: In data visualisation journalism meets engineering. *Investigative Reporters & Editors Journal*, (Spring):24–27.

Chaitin, J. (2003). Narratives and story-telling.

Cheshire, J. (2014). Population lines.

Cleveland, W. S. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833.

Correll, M. (2019). Ethical dimensions of visualization research. In Brewster, S., Fitzpatrick, G., Cox, A., and Kostakos, V., editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 188, Glasgow, Scotland.

Crime Statistics Agency (2019). Recorded offences.

Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111(Supplement 4):13614–13620.

EDQ.com (2019). The cost of living n America.

Engineers Australia (2019). Our code of ethics.

En:User:Bb3cxv (2007). File:RGB illumination.jpg.

Ervik, A. O. (2003). Book Review: IQ and the wealth of nations. *The Economic Journal*, 113(488):F406–F408.

Evergreen, S. and Emery, A. K. (2014). Introducing the data visualization checklist.

Evergreen, S. and Emery, A. K. (2016). Updated data visualization checklist.

Ferdio (2019). Data vis project.

Few, S. (2008). Practical rules for using color in charts.

Few, S. (2014). Why do we visualise quantitative data?

Fung, K. (2014). Junk Charts Trifecta Checkup: The definitive guide.

Grandjean, M. (2016). Connected world: Untangling the air traffic network.

Hannen, T. and Burn-Murdoch, J. (2019). Bar chart race: the most populous cities through time.

Hanrahan, C., Elvery, S., and Byrd, J. (2017). Census 2016: This is Australia as 100 people.

Harrower, M. and Brewer, C. A. (2003). ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37.

Hausmann, R., Hidalgo, C. E., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., and A., Y. M. (2011). *The atlas of economic complexity*. MIT Press, Cambride, MA.

Holder, J., Levett, C., Levitt, D., and Andringa, P. (2018). Blue wave or blue ripple? A visual guide to the Democrats' gains in the midterms.

Huffington Post (2012). Hurricane Sandy graphics show storm's changing path.

Judgment of the Court (Grand Chamber) (2014). Google Spain SL & Google Inc. v AEPD & González Case C-131/12.

Kirk, A. (2012). *Data visualization: a successful design process*. Packt Publishing Ltd, Birmingham, UK.

Kommenda, N., Barr, C., and Holder, J. (2018). Gender pay gap: what we learned and how to fix it.

Kosara, R. (2016). Presentation-oriented visualization techniques. *IEEE Computer Graphics and Applications*, 36(1):80–85.

Kosara, R. and Mackinlay, J. (2013). Storytelling: The next step for visualization. *Computer*, 46(5):44–50.

Kusher, C. (2018). Mind the housing value gap.

Kwan-Liu Ma, Liao, I., Frazier, J., Hauser, H., and Kostis, H.-N. (2012). Scientific storytelling using visualization. *IEEE Computer Graphics and Applications*, 32(1):12–19.

Lee, B., Riche, N. H., Isenberg, P., and Carpendale, S. (2015). More than telling a story: Tansforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5):84–90.

Leo, S. (2019). Mistakes, we've drawn a few.

Litman-Navarro, K. (2019). We read 150 privacy policies. They were an incomprehensible disaster.

Lotto, B. (2009). Optical illusions show how we see.

Lujin Wang, Giesen, J., McDonnell, K., Zolliker, P., and Mueller, K. (2008). Color design for illustrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1739–1754.

Lynn, R. and Vanhanen, T. (2002). *IQ and the wealth of nations*. Praeger Publishers, Westport, CT.

MacDonald, L. W. (1999). Using color effectively in computer graphics. *IEEE Computer Graphics and Applications*, 19(4):20–35.

Mannon, N. (2018). Persuasive storytelling with data visualization.

McCandless, D. (2012). The beauty of data visualization.

Media Entertainment and Arts Alliance (2019). MEAA journalist code of ethics.

Minard, C. (1869). File:Minard.png.

NASA (2007). File:Face on Mars with Inset.jpg.

National Health and Medical Research Council, Australian Research Council, and Universities Australia (2018). *National Statement on Ethical Conduct in Human Research 2007 (Updated 2018)*. National Health and Medical Research Council, Commonwealth of Australia, Canberra.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Office of the Australian Information Commissioner (2019). Australian Privacy Principles.

Ojo, A. and Heravi, B. (2018). Patterns in Award Winning Data Storytelling. *Digital Journalism*, 6(6):693–718.

Our World in Data (2016). Economic complexity rank vs. GDP per capita, 2016.

Palairet, M. R. (2004). Book review: IQ and the wealth of nations. *Heredity*, 92(4):361–362.

Petty, N. (2011). Types of data: nominal, ordinal, interval/ratio - Statistics Help.

Popovich, N., Migliozzi, B., Taylor, R., Williams, J., and Watkins, D. (2019). How much hotter is your hometown than when you were born?

Reynolds, E. (2017). Why NSW had most marginal Yes vote of any state.

Roser, M. (2019). The global population pyramid: How global demography has changed and what we can expect for the 21st century.

Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148.

SharkD (2008). File:RGB color solid cube.png.

Silva, S., Sousa Santos, B., and Madeira, J. (2011). Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333.

Skau, D. (2012). A code of ethics for data visualization professionals.

Smith, A., Fildes, N., Blood, D., Harlow, M., Nevitt, C., and Rininsland, Æ. (2018). Broadband speed map reveals Britain's new digital divide.

Smith, M. (2018). How good is "good"?

Snow, J. (1854). File:Snow-cholera-map-1.jpg.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400.

Spigget (2010). File:Rendered Spectrum.png.

Stiles, M. (2016). How common is your birthday? This visualization might surprise you.

Svantesson, D. (2016). Enforcing privacy across different jurisdictions. In De, D. and Hert, P., editors, *Enforcing privacy. Law, governance and technology series*, pages 195–222. Cham, Switzerland.

Tal, A. and Wansink, B. (2016). Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science*, 25(1):117–125.

The Data Team (2015). Seeking safety.

The Guardian (2019). Australian election 2019 live results.

USGS (2015). U.S. Water Use from 1950-2015: How much water do we use?

Vanderslott, S. and Roser, M. (2018). Vaccination.

Ware, C. (2013). *Information visualization: Perception for design*. Morgan Kaufmann, Waltham, MA, 3rd edition.

Watkins, D. (2015). Arctic ice reaches a low winter maximum.

Wickline, M. (2001). Coblis — Color blindness simulator.

Wikipedia (2009). File:IQ by Country.png.

Zev, R. (2016). Beautiful plotting in R: A ggplot2 cheatsheet.