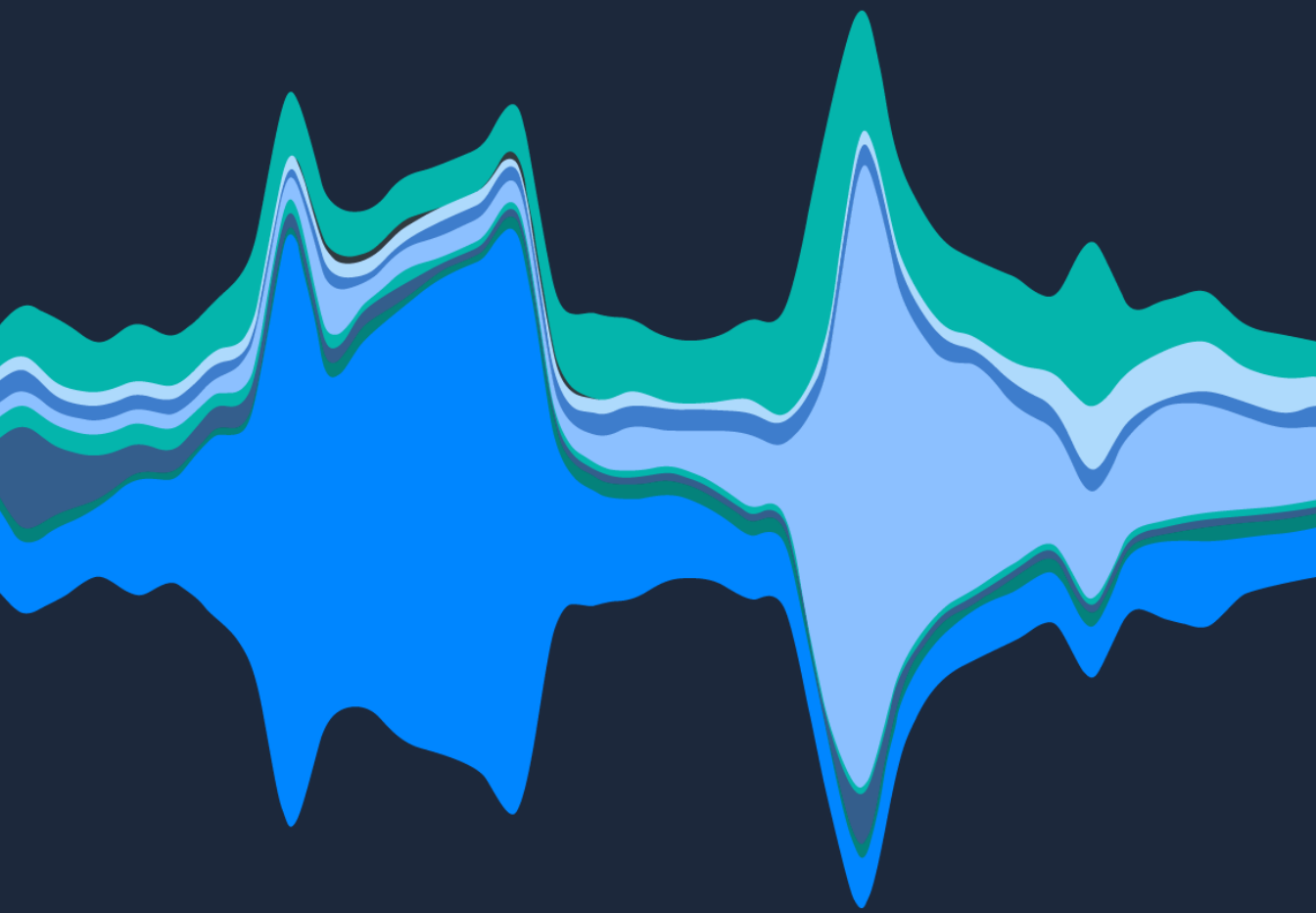
DISPLAYR

RESEARCHSOFTWARE
A DIVISION OF DISPLAYR

Data Visualization



Analysis and reporting software

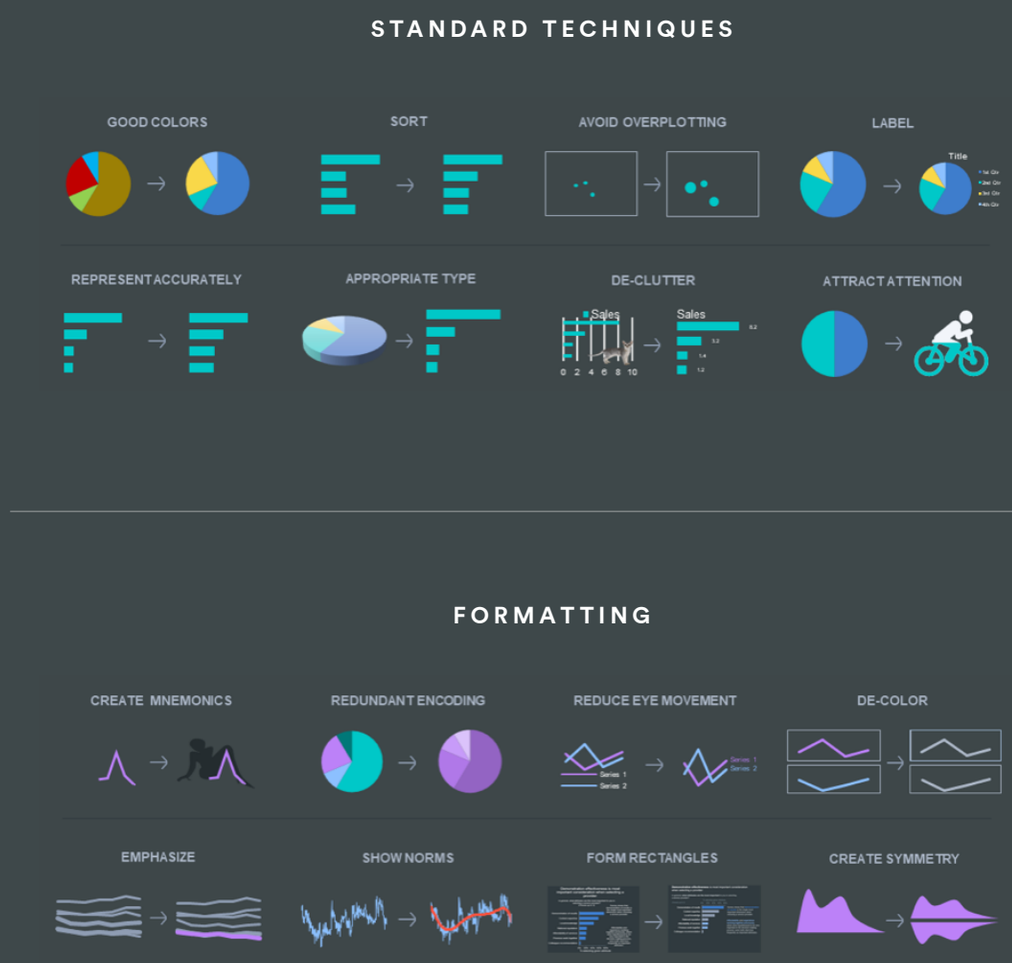
displayr.com

Goal and overview of this book

This book is intended as a resource to help you create visualizations to allow users quickly to discover — and remember — the key stories in data.

This book is designed to help present and make the most of research findings, using foundational (and some maverick) ideas as well as insights into how the target audience/reader/viewer processes and takes in information. The book focuses primarily on market research examples using survey data. A few iconic non-market research examples are included when they are the most effective way of communicating a key idea. Twenty-four techniques for improving visualizations are illustrated with almost one hundred examples of the good, the bad, and the downright ugly.

Table of Contents

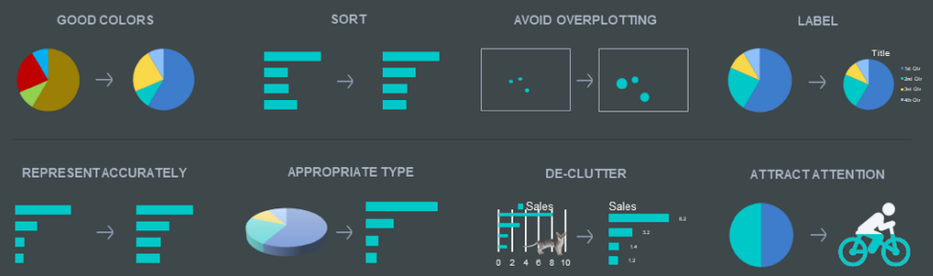


Although many of the examples in this book can be provided as interactive visualizations, the focus will be on examples that can be shown in static documents (e.g., PowerPoint). The constraints of statistical reporting on visualization is best appreciated by an example. The heatmap below would be a better visualization with the numbers removed interactively by the viewer (e.g., by moving a mouse over a region of the heatmap). However, most market research visualizations ultimately end up in a PowerPoint slide. A PowerPoint slide with no numbers is of minimal use, because often the end user needs numbers; merely knowing that darkest blue stands for Aldi and low prices is not enough.

	Aldi	Costco	Coles	Woolworths	Harris Farm	Foodland	IGA
Low prices	78%	46%	49%	40%	29%	29%	23%
Ease of access	40%	15%	63%	61%	31%	37%	41%
Range	25%	38%	68%	70%	31%	47%	30%
Freshness of produce	36%	25%	60%	64%	54%	41%	35%
Quality	45%	47%	71%	69%	60%	51%	43%

Standard Techniques

The standard techniques discussed in this section are widely known and practiced. If you are experienced in data visualization, skip this section and go to **Formatting.**



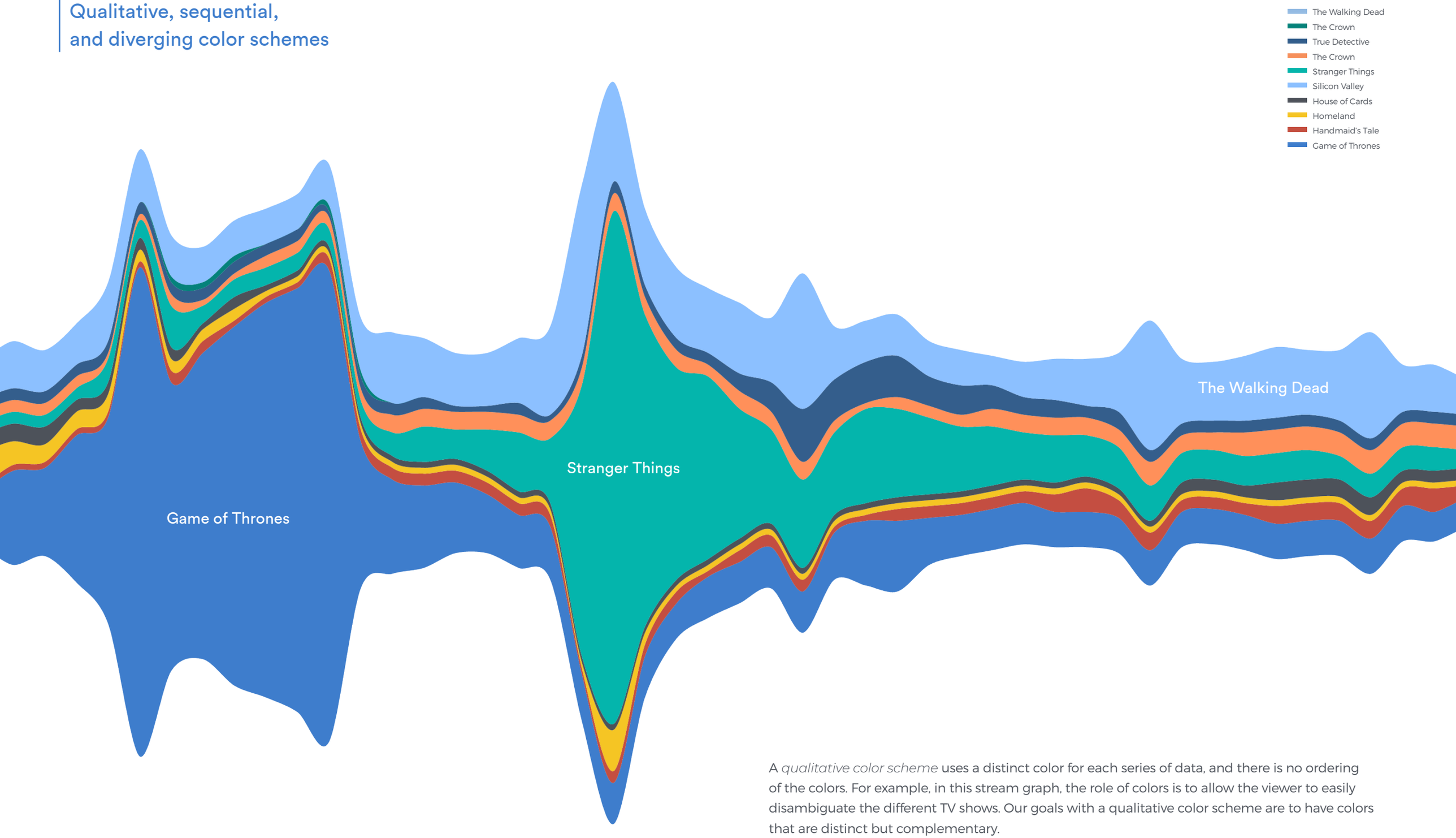
Good colors

Choosing the best colors for a visualization involves:

- Choosing between a qualitative, sequential, or diverging color scheme.
- Choosing complementary colors that look good together and permit easy discrimination by viewers, including those who are color-blind.

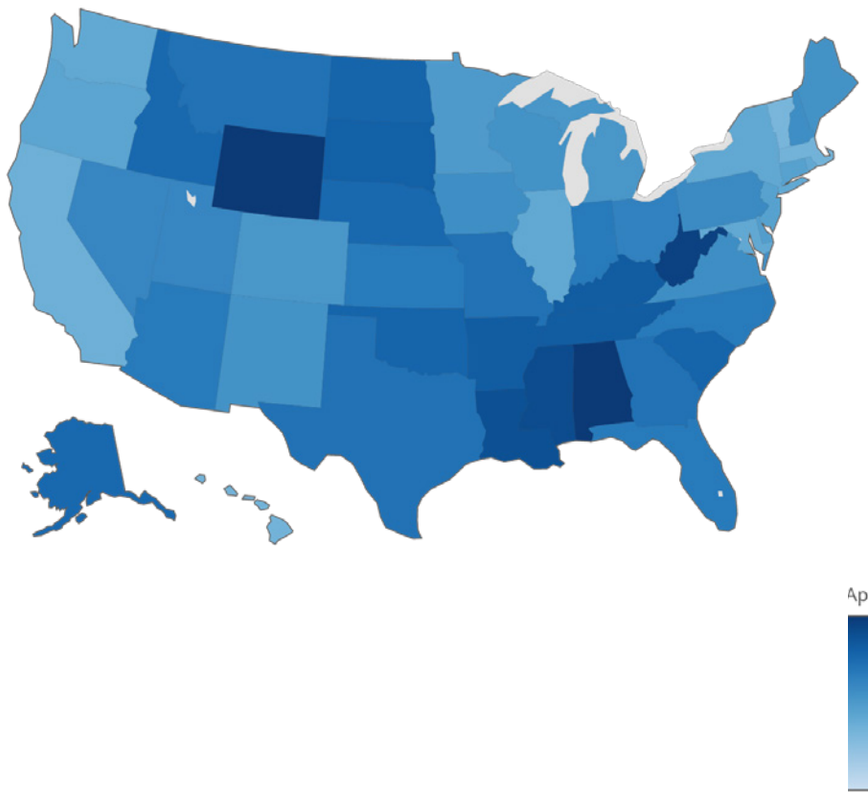
01

Qualitative, sequential, and diverging color schemes

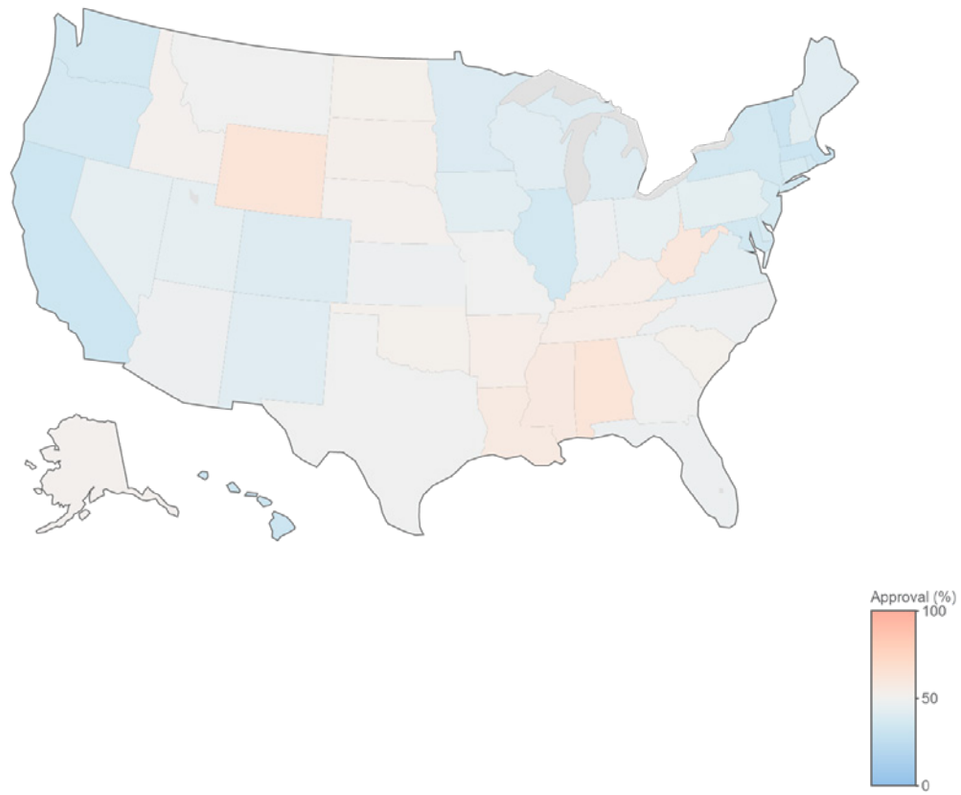


A *qualitative color scheme* uses a distinct color for each series of data, and there is no ordering of the colors. For example, in this stream graph, the role of colors is to allow the viewer to easily disambiguate the different TV shows. Our goals with a qualitative color scheme are to have colors that are distinct but complementary.

Sequential color schemes order the colors in a meaningful way. Typically, they are created from two colors, with gradations between them. In the choropleth below, darker blues denote higher levels of approval for President Trump in March 2018, and gradations are between light blue and dark blue.



A *diverging color scheme* is created by combining two sequential color schemes. In the example below, gray is used to represent people with an approval of 50%, with redder colors for higher approval and bluer for lower approval. In this example, the color scheme also has a qualitative component, with red signifying majority approval and being the color of the Republican Party.



Choosing qualitative colors

Some colors work better together than others. Numerous color theories have been developed to assist in choosing colors that work well together. Fortunately, there are lots of great online tools that can be used to help choose complementary colors, such as https://coolors.co and https://color.adobe.com. These tools even allow you to select colors complementary to any that you are required to use (e.g., your brand's colors).

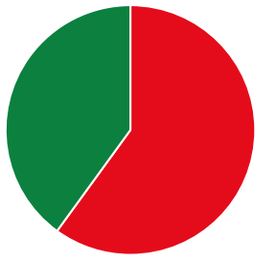
Other things to do include:

- Avoid using pure, bright, or strong colors if they will appear as large swathes in a visualization. Such colors are best used in small areas.
- Use light grays and other muted colors as background.[1]
- Choose colors that work for those with color-blindness.
- Avoid placing bright colors mixed with white next to each other.[2]
- Use colors that are in some way associated with the data. If representing Coca-Cola, try to use a red; if Pepsi, a blue; trees as green, etc.

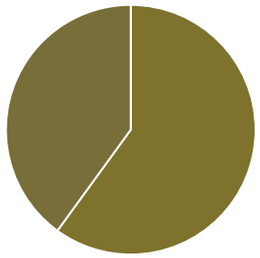
[1] Edward R. Tufte (1990), Envisaging information, Graphics Press LLC.
[2] Eduard Imhof (2007), Cartographic Relief Presentation, Esri Press Classics.

Around 4% of people have some degree of color-blindness. Most are men. There are various types of color-blindness which all have some effect on color choice, but the one with the biggest impact is *protanopia*. If you can see green and red clearly, you can easily tell them apart in the visualization on the left.

If however you suffer from protanopia, they both appear as the nearly indistinguishable green-brown shown on the right.

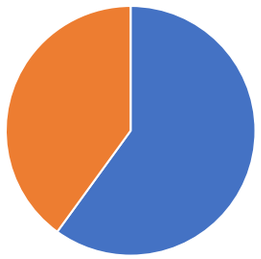


Green and red

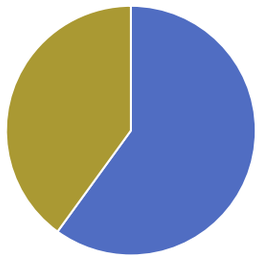


With protanopia

The first step in addressing color-blindness in data visualization is to use colors that are more readily seen by people with color-blindness. The rigorous way to do this is to review color palettes using specialty tools like the webpage *projects.susielu.com/viz-palette*. A simpler hack is to use blues instead of greens and oranges instead of reds: the blue is still largely visible, and the lightness of the orange turns into a lighter green-brown color. The second technique to deal with color-blindness is to use redundant coding, as discussed in Chapter 10, Redundant encoding..

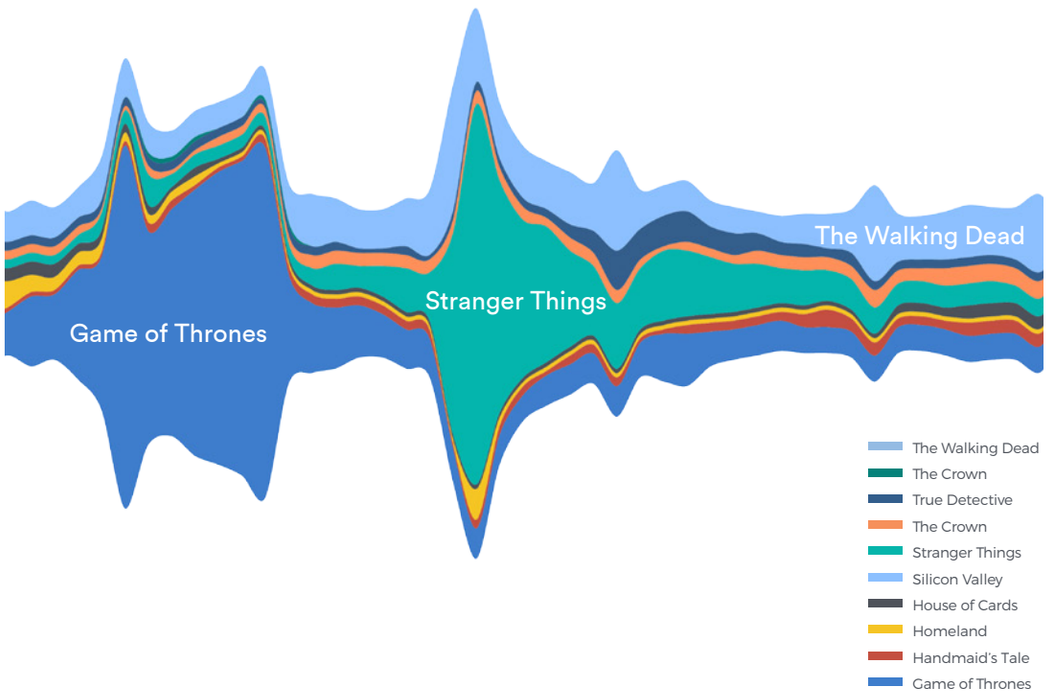


Orange and blue



With protanopia

Most of these tools are designed to allow you to choose five or fewer colors. If you need to have more qualitative colors, the most straightforward solution is to use the color palette(s) that come with your visualization software.



If you wish to create a new color palette with a large number of colors but do not have access to a designer, you can:

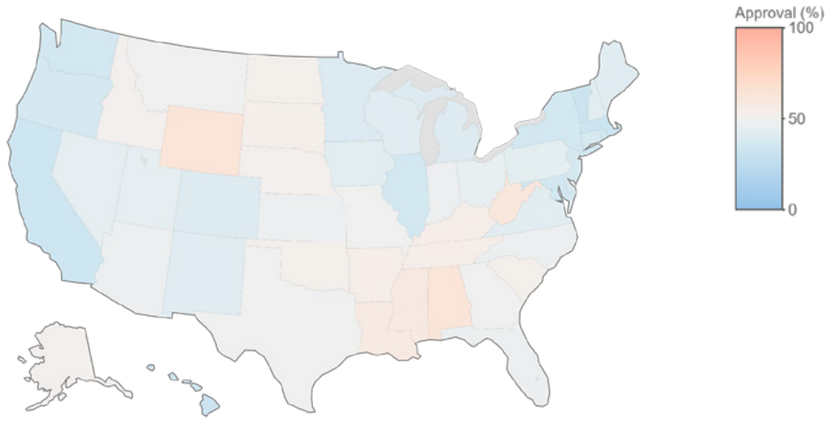
- Use one of the tools to choose an initial small palette. Ideally, try to choose one with many colors on the same side of the color wheel (this instruction will make sense when you are in one of the apps).
- Choosing complementary colors from the other side of the palette.
- Choosing lighter or darker versions of the same color.
- Adding in various shades of gray.

Choosing sequential and diverging color schemes

Choosing sequential and diverging color schemes is a little more challenging than choosing qualitative color schemes. The goal is to choose colors that have a natural gradation, where the degree of gradation is defined by perception as opposed to technical measures such as the percentage of blue or transparency. A widely used resource for choosing sequential and diverging color schemes is *http://colorbrewer2.org*.

When choosing sequential color schemes and diverging color schemes, key decisions are:

- What color to start with and which to end with. Ideally these are colors that have some meaning in the domain of interest.
- What color to have in the middle (if using a diverging color scheme). For example, light gray was used in the earlier example; sometimes white is appropriate and other times black, depending on the background.
- Whether to use a stepped color scheme (e.g., only five unique colors) or as many colors as there are unique data values.
- Whether to shade based on order, or with a mapping from the actual values of the data to the colors. For example, if the largest number is 100 and the second largest is 5, you can represent 5 as a marginally lighter blue, or by a blue that is close to 5% of the perceptual strength of the blue represented by 100.
- What value to assign to the start and end colors. For example, in the *choropleth* below, the pure blue was set to 0% and the coral (pink) to 100%, which allows the viewer to see that there are no states with extreme values. Alternatively, the pure blue could be set to the lowest observed value and the pure choral to the highest value, which would make the gradations between the states clearer.



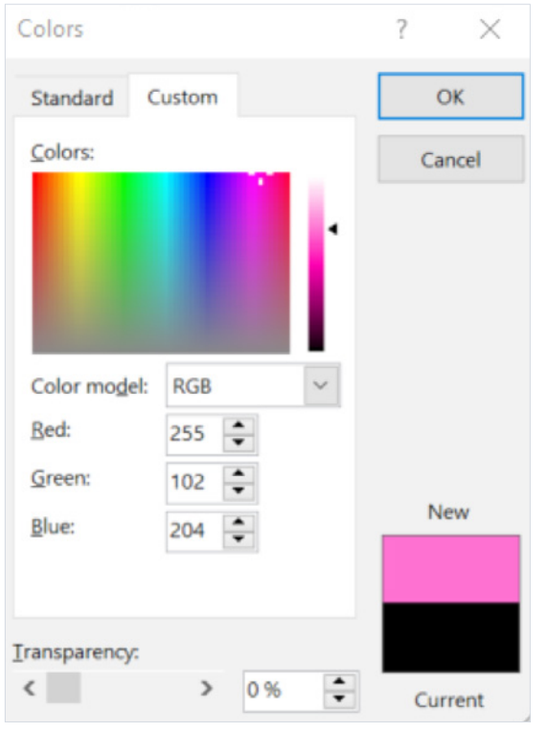
Color naming systems

There are many ways of representing colors in software; this section describes the most common ones. The key thing to appreciate is that if the colors are named in one system, it is usually possible to convert to another naming.

RED GREEN BLUE (RGB)

Most software programs represent colors using the RGB system, where each color is represented as mixtures of the colors red, green, and blue. For example, in PowerPoint, the **Custom** tab for color shows that the pink is made up of 255 Red, 102 Green, and 204 Blue (see screenshot at right).
These numbers are measured on a 256-point scale, from 0 to 255. Some software convert these to proportions, where, for example, a value of 0.5 is equivalent to 128.

In addition to the RGB specification, there is also often a **Transparency** setting, which is provided as a percentage. Alternatively, it can be specified as an alpha value, which is 100% minus the transparency.



HEXADECIMAL (“HEX”)

When colors are represented in computer code they are usually represented as strings (i.e., text). For example, the pink above is
represented as `#ff66cc` or as `#ff66ccff`. This is just another way of writing the RGB scheme, using hexadecimals (base 16 math rather than the normal base 10). That is, in hexadecimal, `ff` means 255, `66` means 102, `cc` means 204 and the last two characters, which are not always required, show the alpha value, which in this case is 100% (i.e., 0% transparency).

WORD REPRESENTATIONS (E.G., “BLUE”)

Sometimes software can understand word representations of colors, such as `red` and `blue`.

HSL AND HSV

The HSL and HSV models are alternatives to RGB, designed with the goal of representing colors in a way that is more intuitive when making design choices, as they separate the notion of the qualitative color (be it red, blue, or green) from notions of saturation and lightness (i.e., extent of mixing with white and black).

CMYK

CMYK (stands for cyan, magenta, yellow, and black) is the color model used in physical printing (e.g., laser printers).

PMS

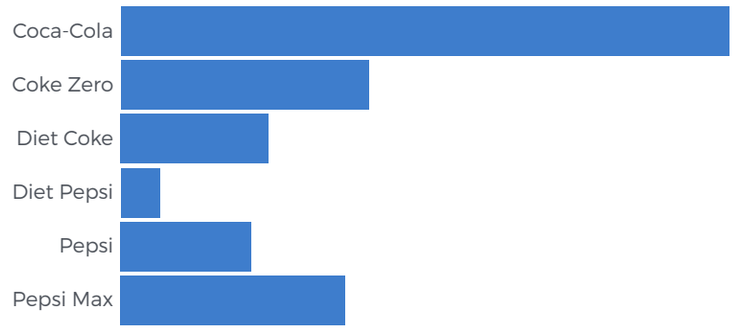
Used by professional designers, the *Pantone Matching System* is a commercial color system in which many pre-made colors are identified and printed in books and on cards for easy reference and precise comparison.

Sort

Sorting the data in a visualization — typically from highest to lowest — improves almost all visualizations.

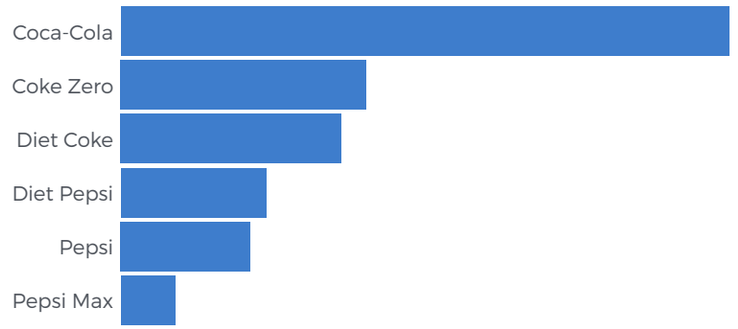
02

Many visualizations show the data sorted alphabetically by the category labels, as shown below. Where there is a natural order for the visualization (e.g. age categories), this is typically the best presentation. However, in almost all other situations, it is better to sort the data from highest to lowest.



Sorting is useful for several reasons. Sometimes the viewer is interested in the ranking of the items, so sorting saves them time. Sorting also reduces the potential error when drawing conclusions about order. In the visualization above, it seems that Coke Zero is bigger than Pepsi Max, and that Diet Coke is bigger than Pepsi. In the visualization below this conclusion is unmissable. With more data, this benefit increases. Sorting helps the user by providing redundant encoding, a topic revisited in Chapter 10, Redundant encoding.

The principle of sorting is also applicable when there are multiple series, in which case the data can be sorted either by one of those series, or by the difference between series (examples of this are shown in Chapter 20, Force contrasts).

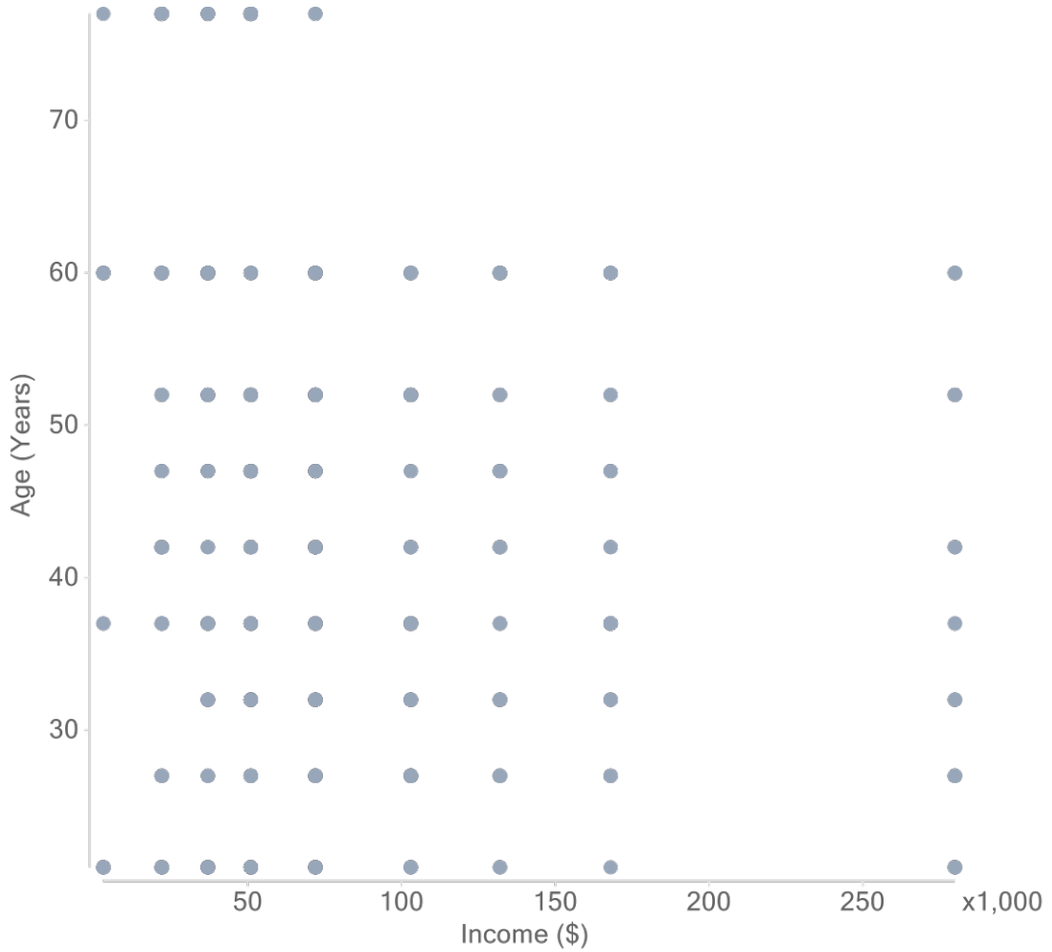


Avoid overplotting

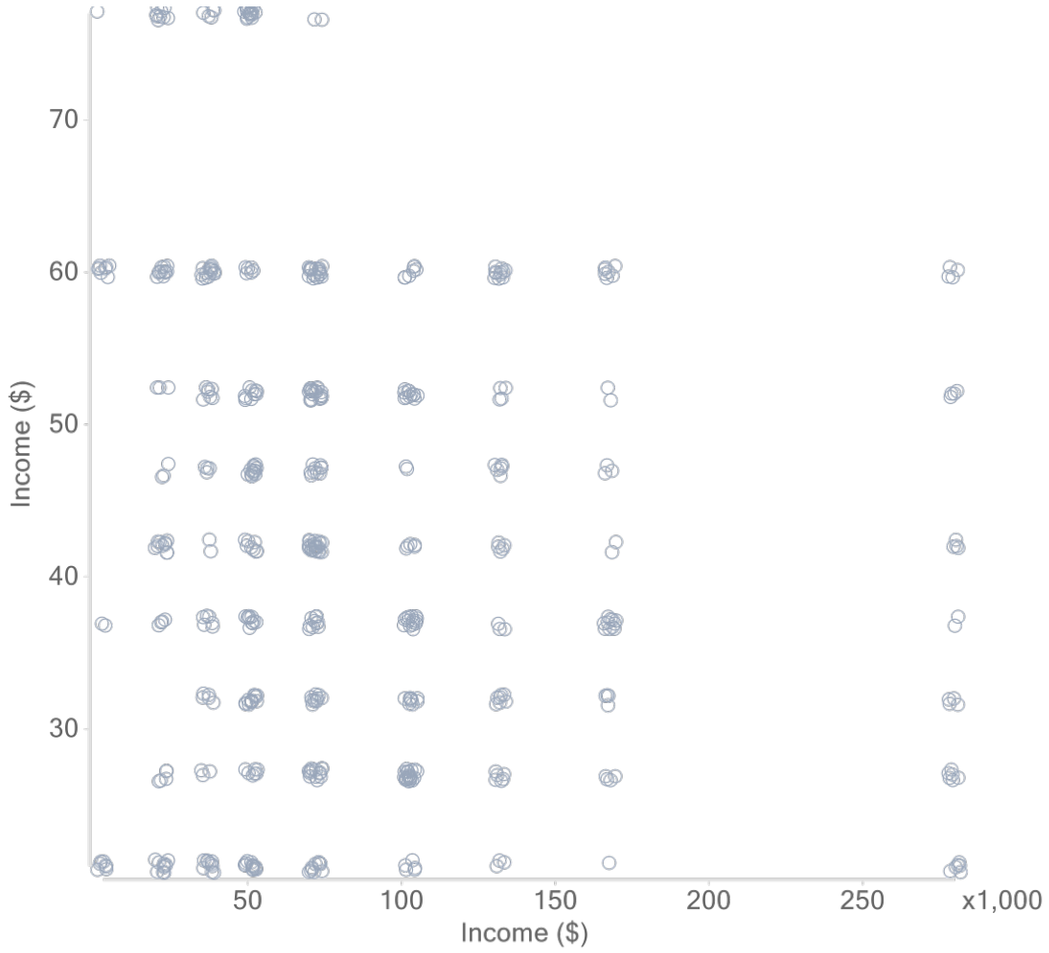
Overplotting is where data or labels in a visualization overlap, making it impossible to interpret the visualization correctly.

03

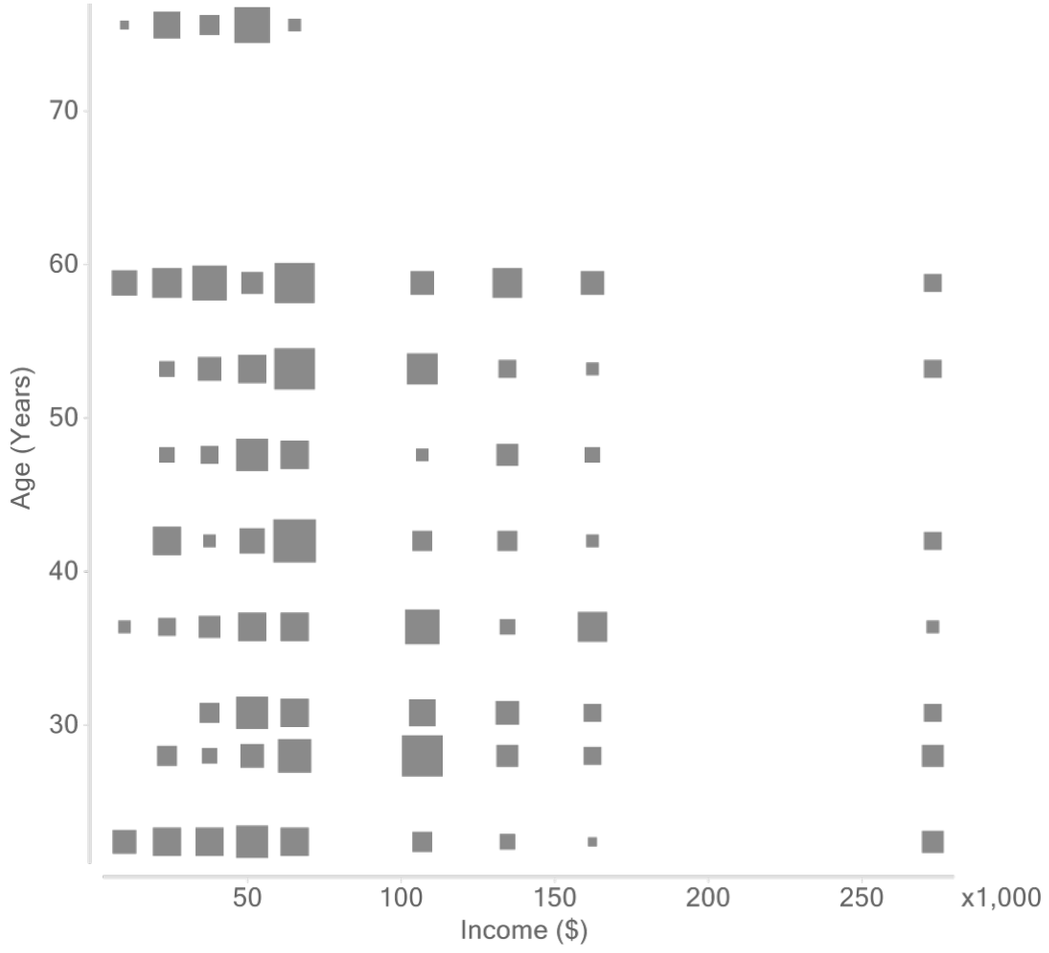
The *scatterplot* below shows age by income. This visualization exhibits the telltale sign of *overplotting*, which is that the data appears in neat rows and columns. There is no way to determine from this visualization if, say, there is only one person aged 60 with income of $50,000 or more.



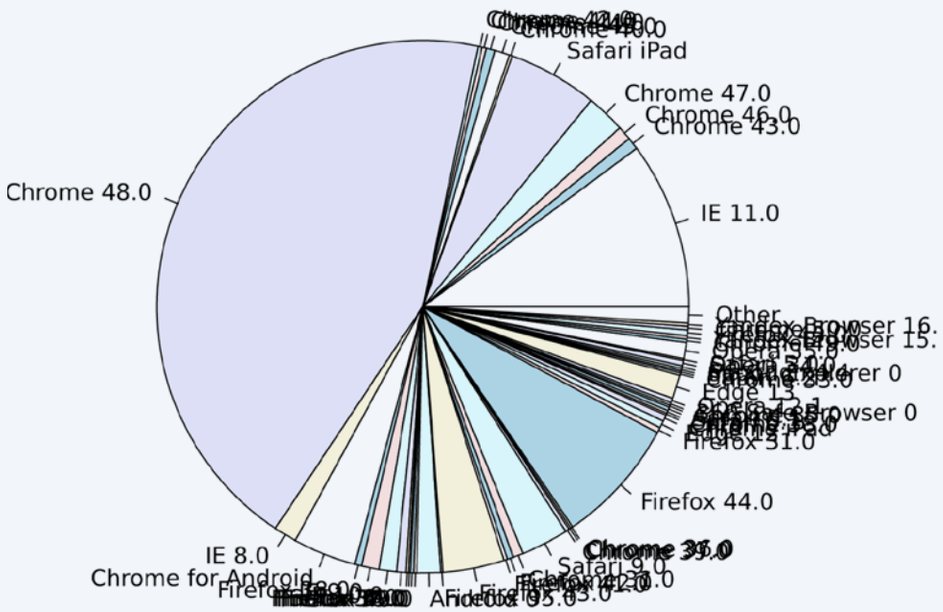
The simplest solution to overplotting is to replace the gray dots with partially transparent dots or circles, and then add small random numbers to the data. This is known as *jittering*. See the example below.



An alternative option is to use *tiles*, where the area of the tiles is proportional to the data.

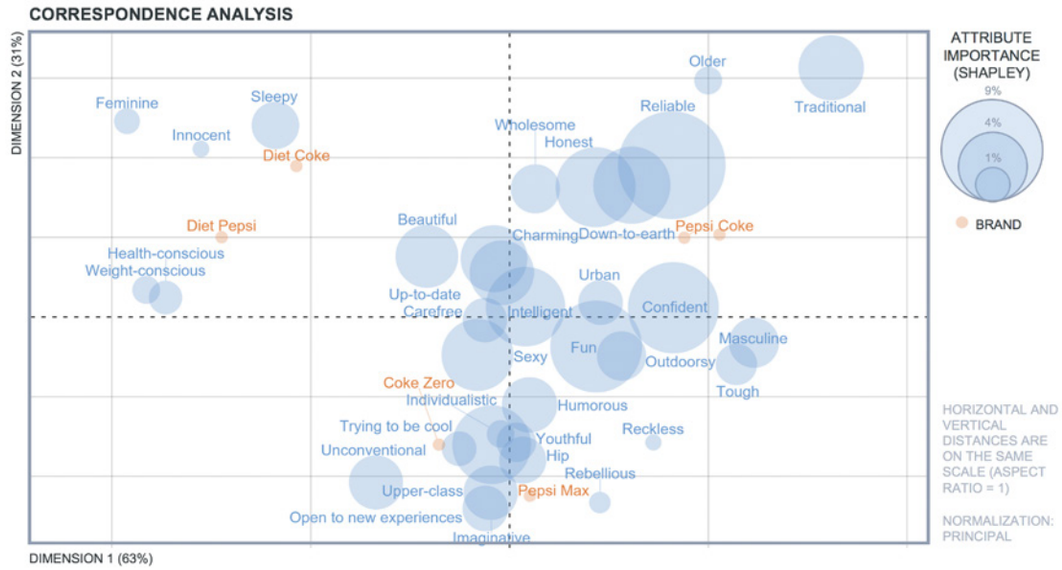


Overplotting also occurs with labels. The overplotting in the pie chart below makes it difficult to discern much (an improved version is shown in Chapter 10, Redundant encoding).



A variety of solutions to overplotting have been developed. The two traditional solutions are to shorten descriptions (e.g., substitute one or two letters for a longer description, with a key on the side of the visualization), and to use legends (e.g., show the colors on the pie chart with a legend explaining their meaning). Neither of these solutions are desirable, as they make the visualization harder to read (see Chapter 11, Reduce eye movement).

The *bubble chart* below presents an example of a better solution to the problem of overplotting. The bubbles are permitted to overlap, with transparency making this overlap intelligible. The labels have been automatically positioned by the software, so they don't overlap and are still close to the relevant bubbles. Where it is not possible to position the label in an optimal position, lines are drawn connecting the labels to their bubbles. Another modern strategy is to use hover effects, where few or even no labels are shown but appear when a mouse pointer is moved over the visualization.

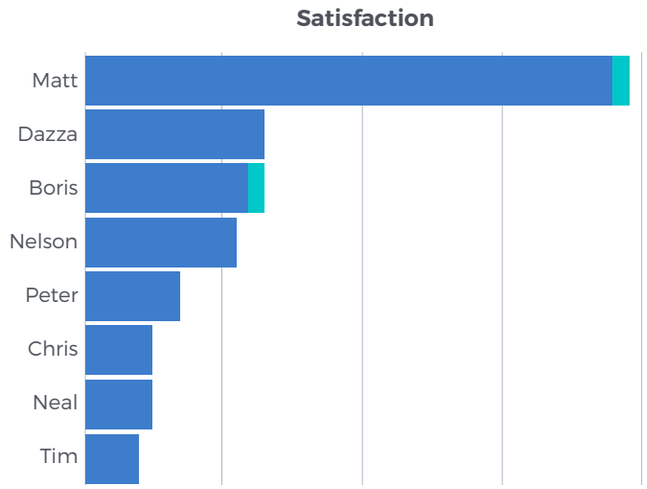


Label

A common mistake when presenting data in a visualization is to provide insufficient information, leaving the viewer unable to understand what the data means. This is fixed by adding labels.

04

In the *stacked bar chart* below, the numbers are not defined, so beyond establishing that Matt seems to be winning in satisfaction, it is unclear what the data means. An improved version is shown in the next chapter.



Often context can provide an adequate explanation to compensate for a paucity of labels. In this book, we have used minimal labeling because we wanted to emphasize visual aspects of visualization; but in other contexts we tend to label much more, because a visualization that cannot be read is of little value to anyone.

A simple checklist for ensuring that the visualization is labeled appropriately is:

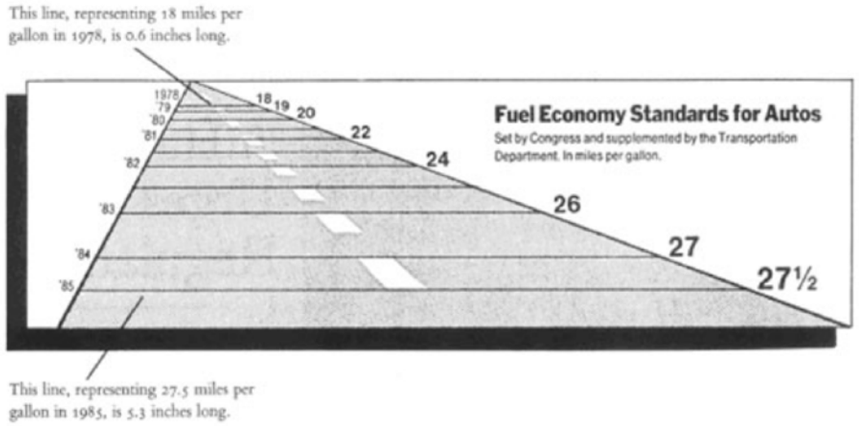
- Label axes.
- Show the units (e.g., kg, liters, $, USD).
- Show where the data comes from.
- Have a title that summarizes the data.

Represent accurately

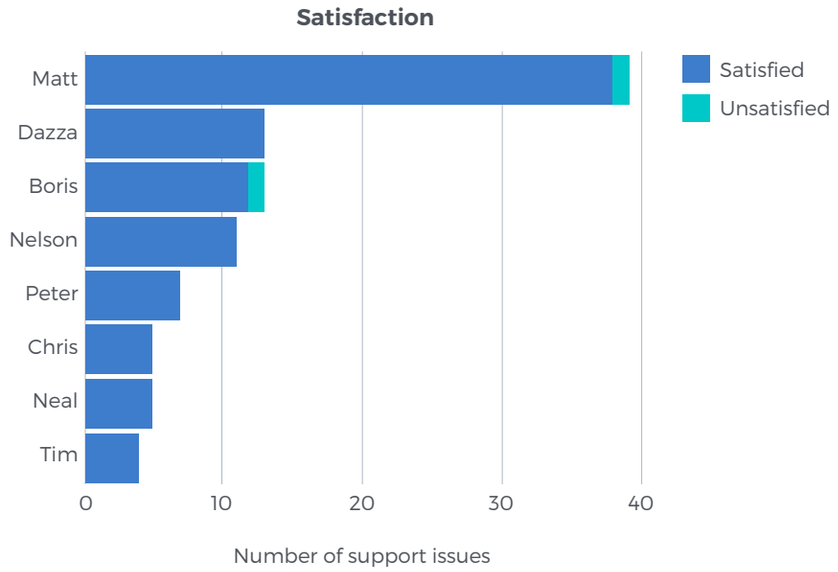
It is possible inadvertently to create visualizations that misrepresent data. A diligent viewer can usually discern the real story, but the imperative is to create visualizations that can be accurately read by even the lazy viewer.

05

The visualization directly below is highly misleading. The data to the right of the road shows that fuel economy is 27½/18-1 = 53% better in 1985 than 1978, but matching line is almost eight times longer.[3] An instinctive reading of such a visualization will always be wrong. The only hope of a clear reading is if the viewer ignores the visual.

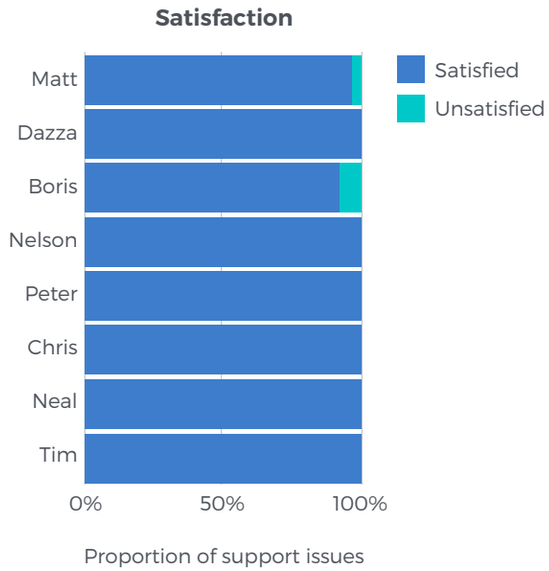


A more prosaic example is shown below. A glance at this *stacked bar chart* would lead to the conclusion that Matt is leading in terms of satisfaction. A more careful reading registers that Matt has the second lowest satisfaction rate: the longer bar means he addresses more customer support issues.

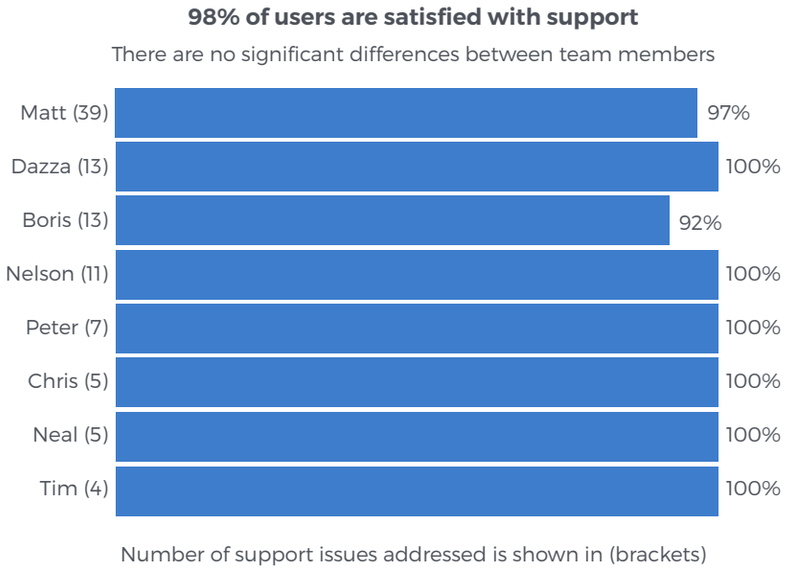


[3] Edward R Tufte (1983), The Visual Displayr of Quantitative Information, Graphics Press, p. 5

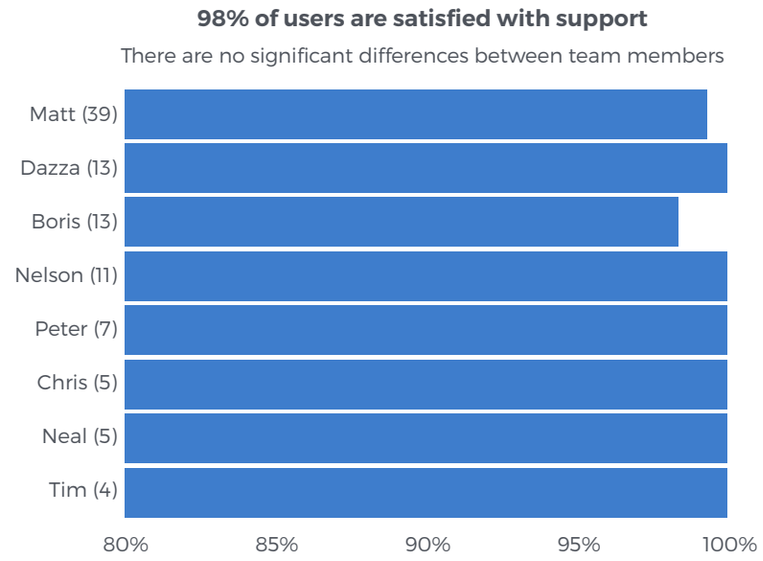
The standard solution to this problem is to express the data as percentages, as shown below. Now we can see that Boris has the lowest satisfaction level, and Matt the second lowest.



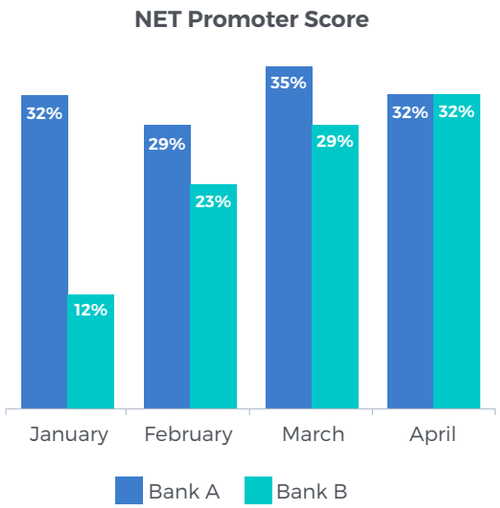
The revised visualization is better but still misleading. One problem is that it fails to reveal how many support issues each person dealt with. This is important: the more support issues somebody resolves, the higher the chances that at least one user will report being unsatisfied. This is addressed below both in the row labels and the inclusion of more informative titles and a footer. A second problem with the chart above is that unsatisfied data is redundant in an unhelpful way. This is also rectified in the visualization below, which shows just the proportion of satisfied users.



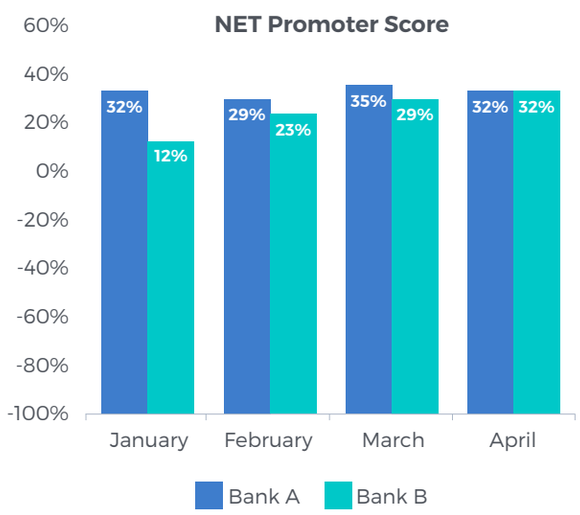
Another type of misrepresentation is shown in the chart below. The bar for Boris implies that he is much worse than the data shows. The problem here is that the horizontal axis intersects the vertical axis at 80%, rather than the more appropriate 0%.



Unfortunately, problems with market research data are usually not this clear. In the example above, the data is on a *ratio scale*, which is to say that it is meaningful to compare the ratio of one number to another (e.g., 97% is about 97%/92%-1 = 5.4% higher than 92%). However, most market research visualizations do not have ratio-scale data, which makes the size of bars more arbitrary.



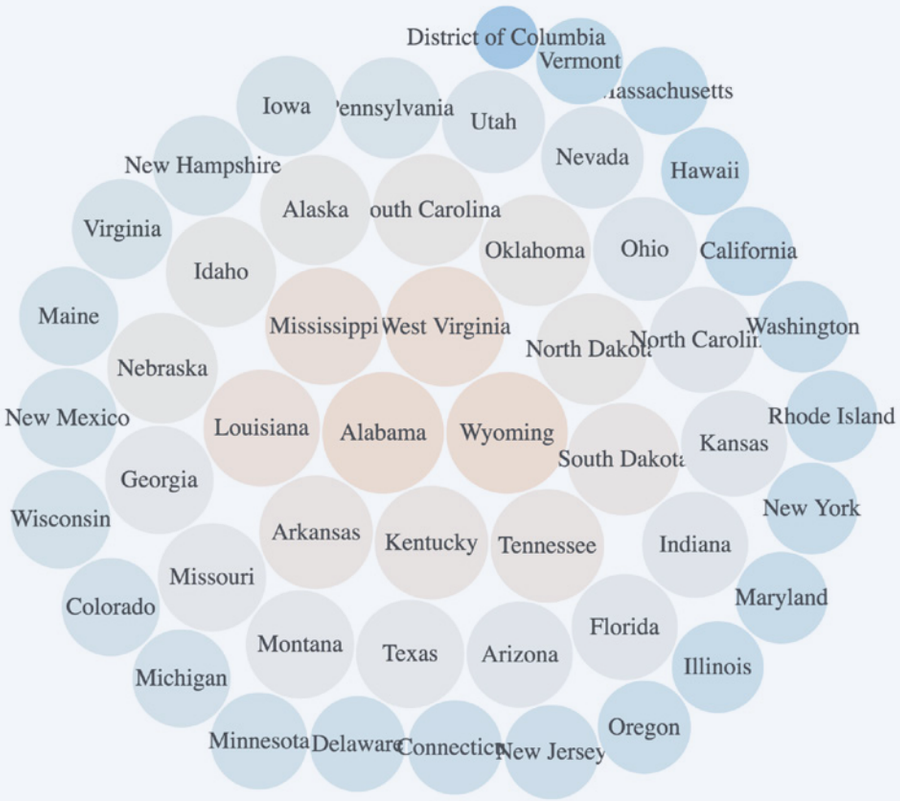
Consider the chart above, which shows net promoter scores (NPS) for two banks over four months. At first glance this seems a sensible visualization, with column heights proportional to the values. However, the NPS is measured on a scale of -100% to 100%, so arguably a more accurate representation is the one below.



In this case the increased accuracy seems unhelpful. There are many other similar situations in market research (e.g., when showing average importance scales, it is rarely ideal to set the axes to start at the lowest possible average and finish at the highest).

The area size issue (in particular, with circles)

The *circle packing (or bubble cloud)* visualization shown below uses the sizes of the circles to communicate the different values. This raises a problematic issue: how does one define the size of a circle? To the mathematically-minded, it is obvious that size should mean area. But most people's ability to read size from the area of an image accurately is poor.[4] Consequently, it is not uncommon to find visualizations where area is represented by the height of the circle. To the math whiz, this is obviously an error. However, the mere existence of such errors demonstrates that the more mathematically-oriented view may itself be incorrect: if many people equate size with height, encoding the values into the area will lead to more errors. The key point here is that when using visual elements where their area could be used by the viewer to make inferences, it is advisable to provide redundant encodings, as area on its own is not sufficient. Area is perhaps best interpreted as a way of showing relative orderings of values, as opposed to being thought to be able to precisely communicating actual values.



[4] Jacques Bertin's (1967), Sémiologie Graphique. Les diagrammes, les réseaux, les cartes. Translation 1983. Semiology of Graphics by William J. Berg

Appropriate type

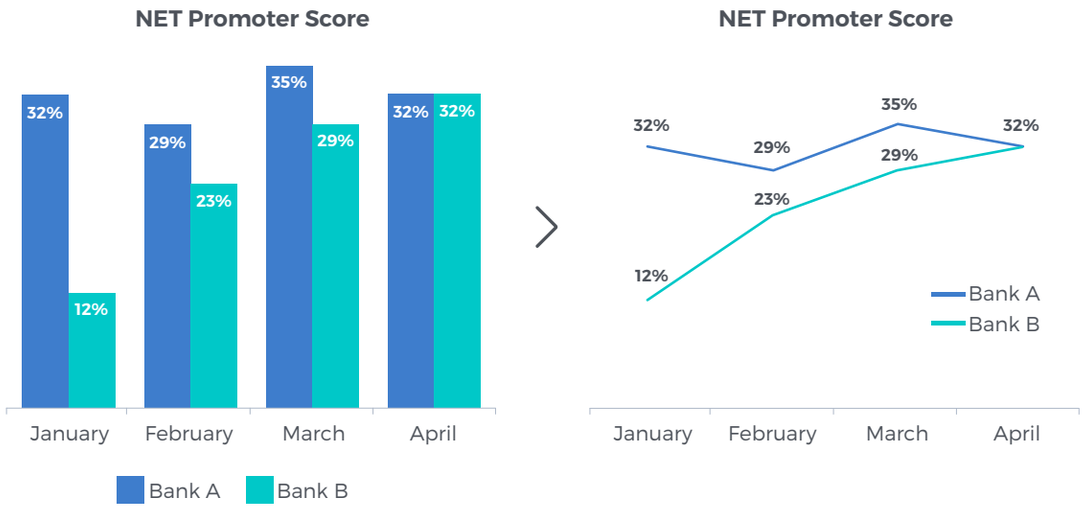
Some visualizations are intrinsically better for showing certain types of data than others.

This chapter covers a few very basic rules of thumb regarding how to select an appropriate visualization. Much of the rest of this book explores this theme in more detail.

06

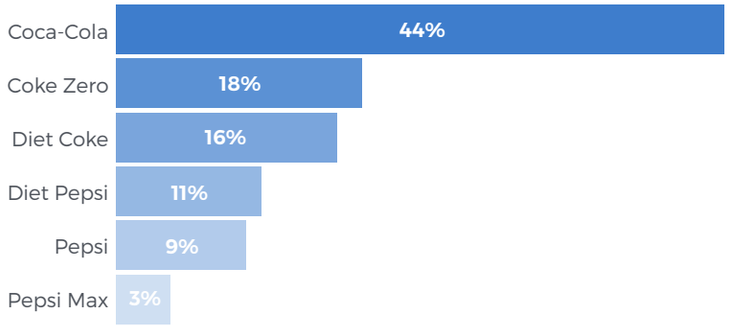
Use line charts to show and compare trends

Perhaps the only widely agreed-upon rule is that line charts are usually best for comparisons of trend data, making the visualization below on the right preferable to the one on the left.



Bar and column charts are good for comparisons

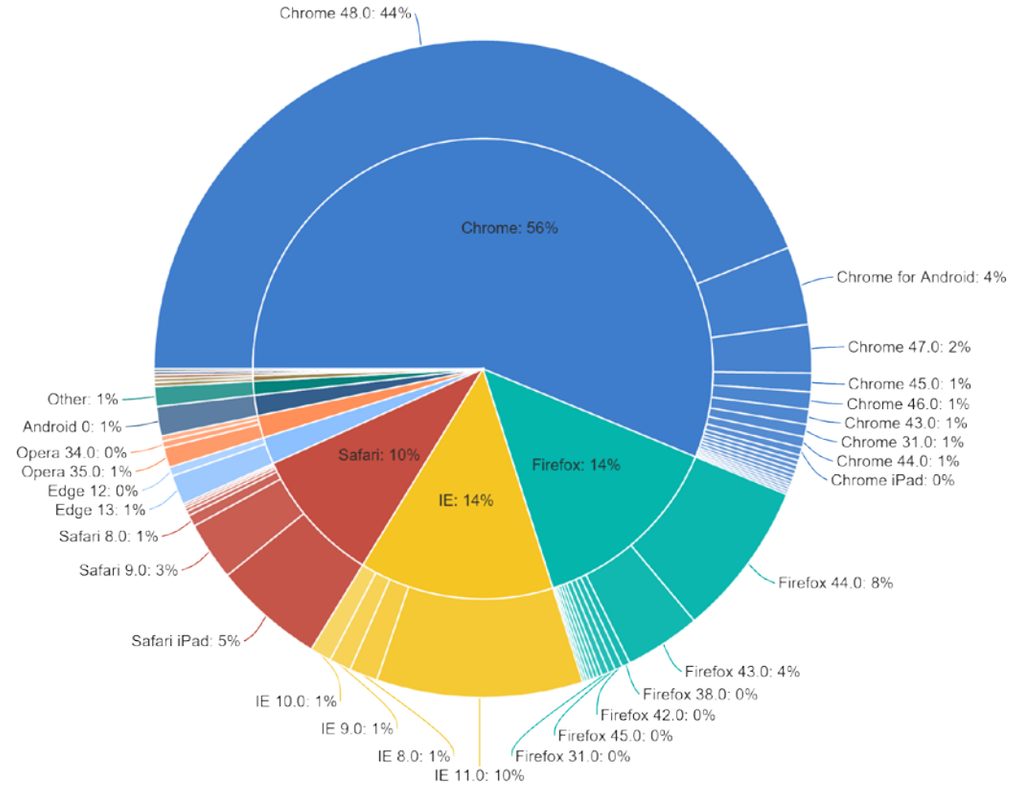
Bar charts and *column charts* are ideally suited for comparing values, where the data has a ratio scale (i.e., where there is a meaningful 0 value for the bar to commence). Column charts are perhaps a bit more accurately read,[5] but in practice bar charts tend to be more generally useful as they make it relatively easy to create charts with long labels, without the need to wrap.



[5] Jacques Bertin's (1967), Sémiologie Graphique. Les diagrammes, les réseaux, les cartes, Translation 1983. Semiology of Graphics by William J. Berg

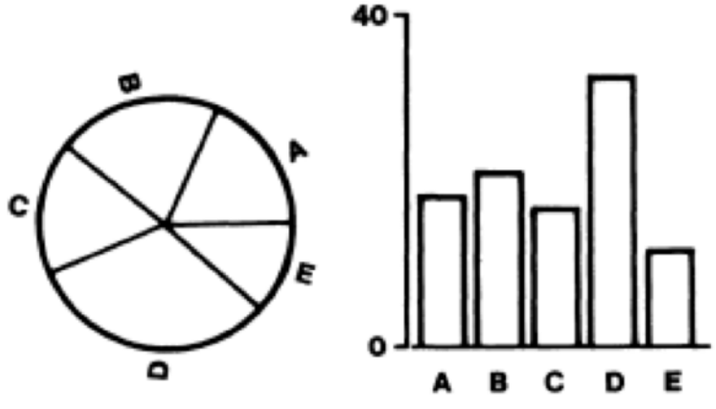
(Sometimes) use pie charts to show cumulative proportions

Pie charts are useful for showing cumulative proportions. For example, the nested pie chart below shows that Chrome 48 is by far the biggest browser, and that Chrome accounts for more than half the market.

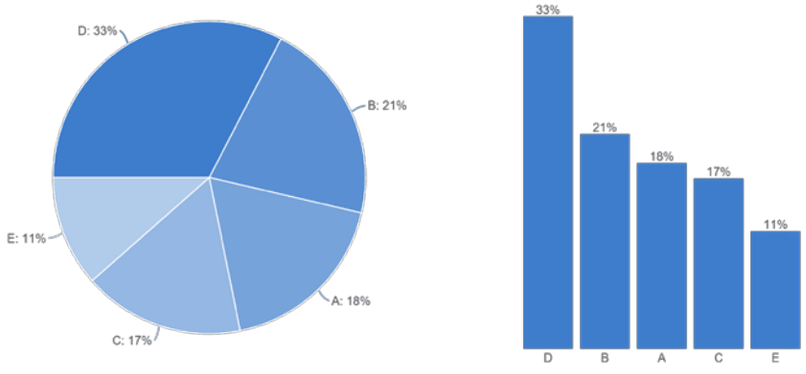


Despite this obvious strength, there is far from complete agreement about the strengths of pie charts and the closely related *donut charts* (pie charts with a hole in the middle). Some visualization experts criticize them heavily. This is when it is useful to go back to the original scientific research, as the conclusions attributed to it – that pie charts are never useful. However, this criticism is frequently unwarranted.

Cleveland and McGill's study[6] makes a compelling case that pie charts like the one on the left below are typically inferior to column charts like the one on the right, as people's estimates of the size of the bars are considerably more accurate than estimates of pie slices. Presumably back in 1984 when the study was published it was commonplace for people to create pie charts like the one on the left, and we can credit the study with having led to the eradication of such visualizations.



A more relevant comparison for today is to compare the charts below. As the values are shown on the visualizations, and the categories are sorted, it is difficult to believe that the finding of the original study would recur if the study was repeated with these more modern designs.



[6] William S. Cleveland and Robert McGill. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods", Journal of the American Statistical Association, Vol. 79, No. 387 (September), pp. 531-554.

Does that mean that one visualization is as good as the other? Not really. The column chart is usually preferable when the goal is to compare one category with another. However, if we are instead trying to make conclusions about cumulative proportions, the pie chart is usually the better choice. The pie chart allows us to see readily that together D and B account for more than 50%. This cannot be seen on the column chart; instead, the user is forced to look at the value labels and do the math.

One case in which the pie is better than the bar is if comparing two proportions. Where the data align to halves, quarters, or eighths, the pie chart will likely outperform a bar chart, as shown in the image below. These two visualizations show the same information, but only the pie chart makes the conclusion obvious.

It is hard to compare these lengths precisely

It is easy to see that 25% is removed

Use waffle charts when presenting to pie chart-haters

The previous section qualified the recommendation to use pie charts to show part/whole relationships with "sometimes". This is because, among some in the visualization community, it is an article of faith that pie charts are bad. Most who hold this view do not appreciate the irony that their views are based on poor data. Most books dealing with visualization advise against pie charts, either citing the original research or seeking to demonstrate the unsuitability of pie charts by presenting poorly constructed pie charts. For example, *The Big Book of Dashboards* presents the example shown below,[7] which is unsorted, shows no value labels, uses poor colors, and is not the type of pie chart that even a semi-numerate pie-chart devotee would use. In situations where a pie chart is desirable but not practical due to the audience, an acceptable alternative is often to use a waffle chart.[8]

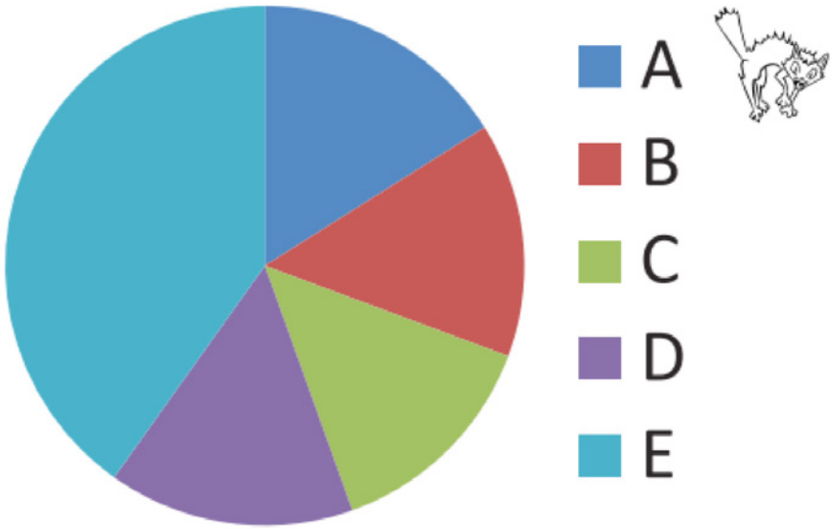
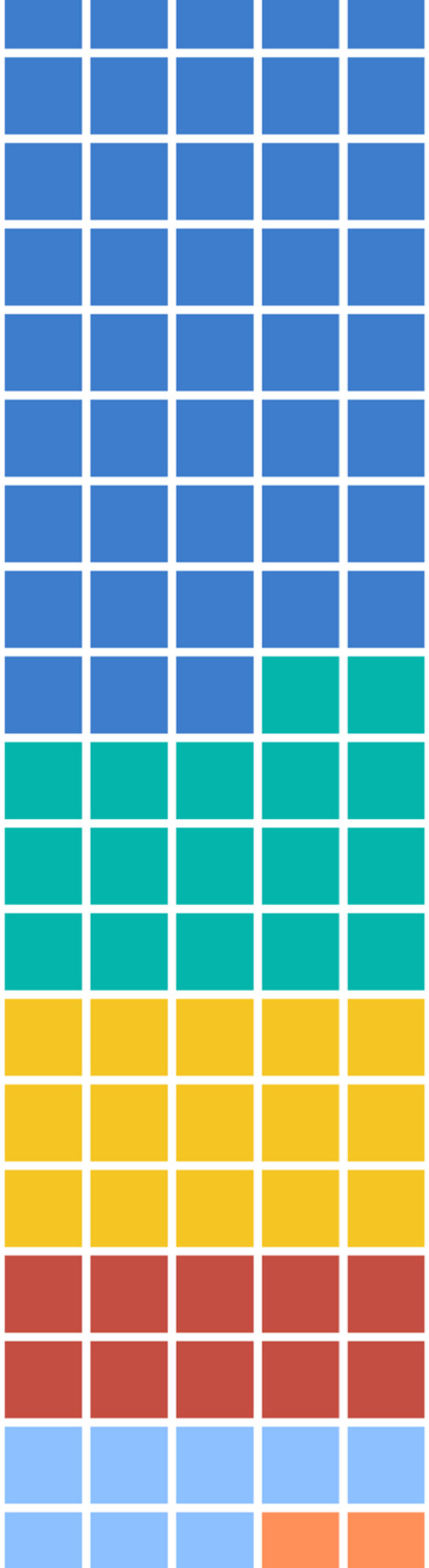


FIGURE 1.45 Can you order the slices from biggest to smallest?

[7] Steve Wexler, Jeffrey Shaffer, and Andy Cotgreave (2017), Big Book of Dashboards: Visualizing Your Data Using Real-World Scenarios, p. 32.
[8] For example, waffle charts are used throughout Cole Nussbaumer Knaflic (2015): Storytelling with data, and pie charts are described as being "evil" (p. 61).

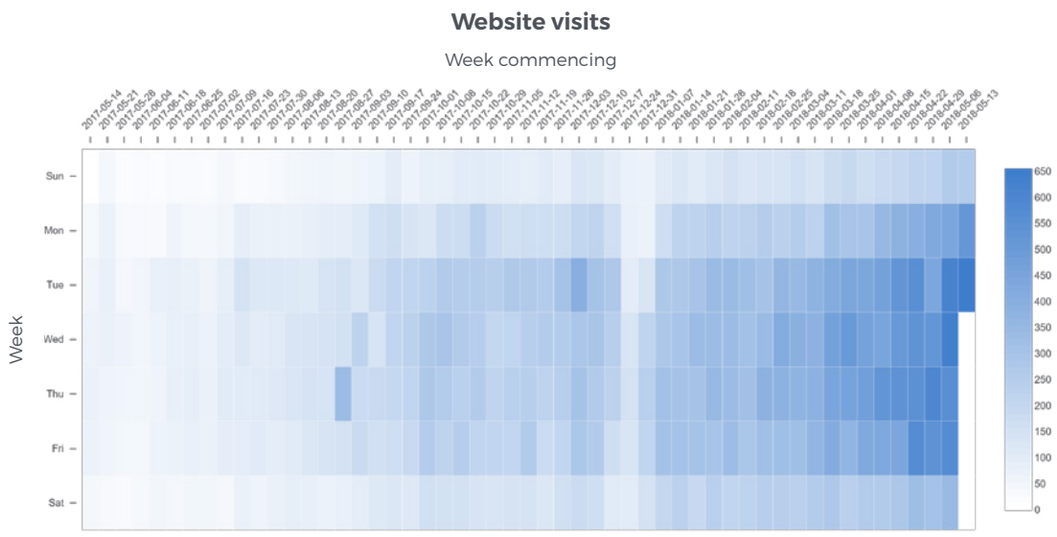


The *waffle* has two advantages over the pie: the elements are countable — which provides a useful form of redundant encoding — and perhaps most importantly, nobody famous has criticized them.

Start with heatmaps when viewing large tables of comparable numbers

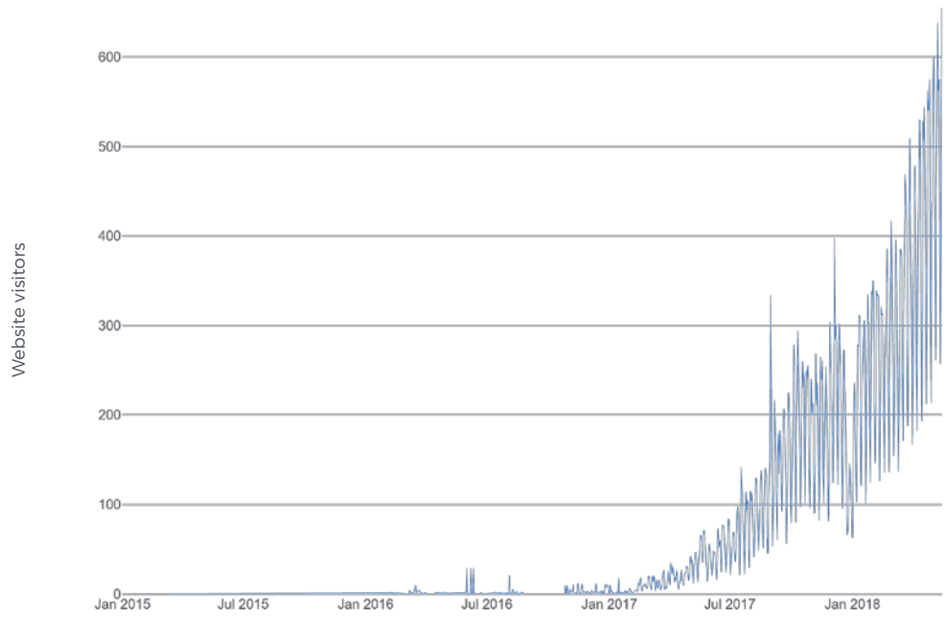
Most visualizations involve a transformation of data, from a representation as a table of numbers into a visual with some other coordinate system. For example, the pie chart converts numbers into angles and slices of a circle, and the bar chart converts numbers into a two-dimensional space with numbers as rectangles. Of the widely-used visualizations, heatmaps change the structure the least, retaining. the table's original structure of rows and columns, with the only change being replacement of numbers by a color. Consequently, heatmaps are usually the best visualization for communicating a basic understanding of patterns when the data consists of a large table of comparable numbers.

As an example, the visualization below, showing visits to a website over a year, plots 360 data points in a simple way that allows us to see that the visits have been steadily rising, with most of the traffic on weekdays.



When in doubt, use line, column, or bar charts

Column charts, bar charts, and line charts, are safe default charts. They are widely understood and for many problems they are the optimal visualization, as their encodings of data are, typically, most accurate. For example, while the heatmap above does a great job of communicating the basic story in the data, its encoding – color – is not one in which humans excel at finding patterns.[9] Basic questions, such as the percentage of growth to have occurred over the past year, cannot really be read from such a visualization. The line chart below provides us with much more precision about the quantities of the data above (and shows two extra years of data at the beginning). Note, though, that this line chart is better than the heatmap only because of the knowledge we gained by first plotting the heatmap (i.e., that the variation we see in the line chart between days is due to the effect of weekends).



[9] Jacques Bertin's (1967), Sémiologie Graphique. Les diagrammes, les réseaux, les cartes, Translation 1983. Semiology of Graphics by William J. Berg

Other rules

There are other rules of thumb, but there are many exceptions to these rules. Other rules include:

- Use radar charts (spider charts) to compare cycles, such as usage by hour across days.
- Use heatmaps to show data where there are patterns in two dimensions (e.g., by month and within month).
- Use violin plots to compare distributions (see Chapter 16, Create symmetry).

De-clutter

By reducing the amount of clutter on a visualization, we increase the chance that the user can see the key patterns in the visualization.

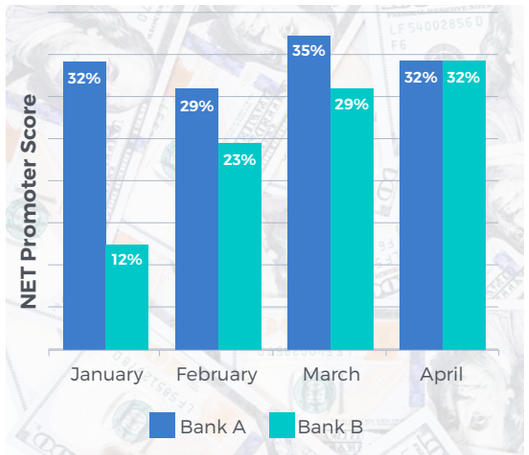
07

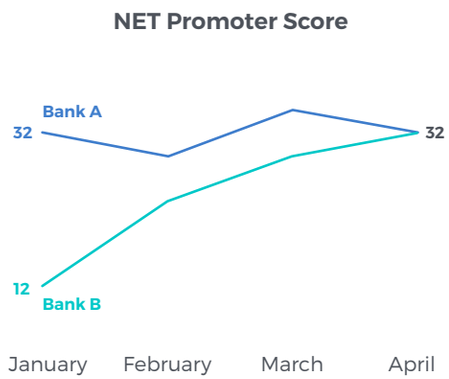
A core principle of visualization is to avoid distracting the user. We want the signal to be clear. This requires us to minimize the amount of noise. Statisticians sometimes quantify the amount of clutter as the data-ink ratio, where data is anything that cannot be erased without altering the meaning.[11] For example, most of the ink in the table below is uninformative, whereas the next table has a much higher data-ink ratio.

%	Aldi	Costco	Coles	Woolworths	Harris Farm	Foodland	IGA
Price	78%	46%	49%	40%	29%	29%	23%
Access	40%	15%	63%	61%	31%	37%	41%
Range	25%	38%	68%	70%	31%	47%	30%
Fresh	36%	25%	60%	64%	54%	41%	35%
Quality	45%	47%	71%	69%	60%	51%	43%

%	Aldi	Costco	Coles	Woolworths	Harris Farm	Foodland	IGA
Price	78	46	49	40	29	29	23
Access	40	15	63	61	31	37	41
Range	25	38	68	70	31	47	30
Fresh	36	25	60	64	54	41	35
Quality	45	47	71	69	60	51	43

This same technique can also be applied to charts. In the visualization below, for example, there are at least two kinds of noise. The background is pure noise — *chart junk*, to use the pejorative favored by statisticians. And most of the columns themselves are noise: we could erase everything except their tops and still interpret the data correctly.





By contrast, the line chart shown at right is much cleaner. Only the beginning and end value is shown for each series, as the intermediate numbers add nothing. Lines are used instead of bars, thus revealing much more whitespace. The background is plain white so that nothing distracts the viewer.

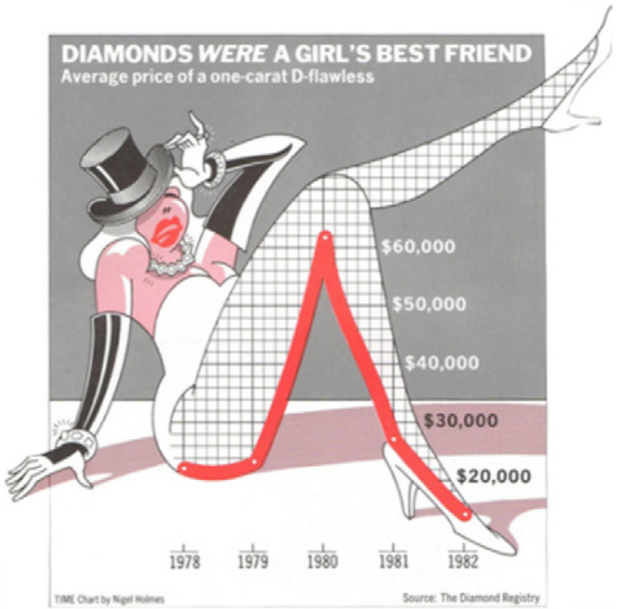
[11] Edward R. Tufte (1983), The Visual Displayr of Quantitative Information, Graphics Press.

Attract attention

Visualizations are often created with the intent of attracting attention. However, this technique is often at odds with other standard techniques (e.g. de-cluttering).

08

A visualization can only be interpreted if it is noticed, which is why graphic artists often put a lot of work into making visualizations more appealing. A famous and controversial example of this is shown below.



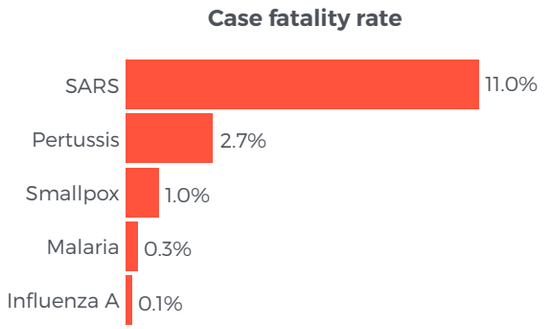
Statistician and artist Edward Tufte describes this visualization as "unsavory ... chockablock with cliché and stereotype, coarse humor, and a content-empty third dimension. ... contempt both for information and for the audience. ... who would trust a chart that looks like a video game?"[12] His conclusion is that the simpler visualization is the better, like the one shown below, in which all chart junk has been removed.

[12] Edward R. 1990Tufte, *Envisaging Information*, p. 34

However, as discussed in the next chapter, the sexist visualization is the superior one — a lot smarter than Tufte acknowledges.

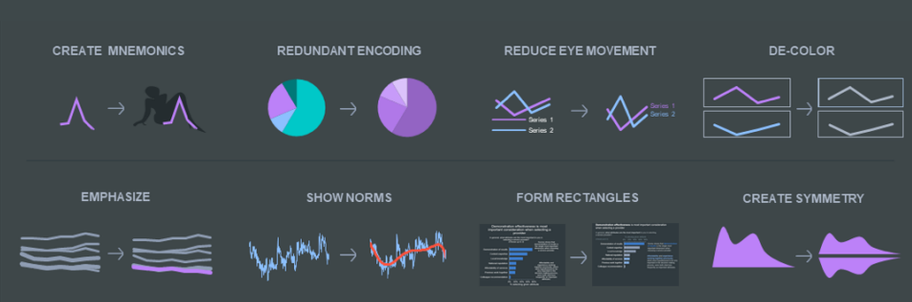
A simple way to gain attention is to use pictures and images to construct *pictographs*, such as the cola market-share chart shown to the left. When creating such visualizations there is a conflict between the accuracy with which the data can be perceived and the goal of attracting attention. Among the ultraconservative (such as Tufte) there are also issues of credibility.

Nevertheless, it is possible to create visualizations that are both attractive and have desirable perceptual properties. Even pictographs can be laudable — for example, the chart to the right. It is hard to imagine a more traditional visualization doing a better job. The use of rows of 10 make the data countable, which reinforces the data, crystalizing the somewhat abstract concept of a portion of a percent. The part/whole relationship of a percentage is communicated by the gray. The overall design makes the comparison between the numbers straightforward. A bar chart of the same data, like the one below, has much less impact.



Formatting

The previous eight chapters have reviewed the standard and widely known techniques used for creating visualizations. This section focuses on lesser-known techniques for improving visualizations through formatting.

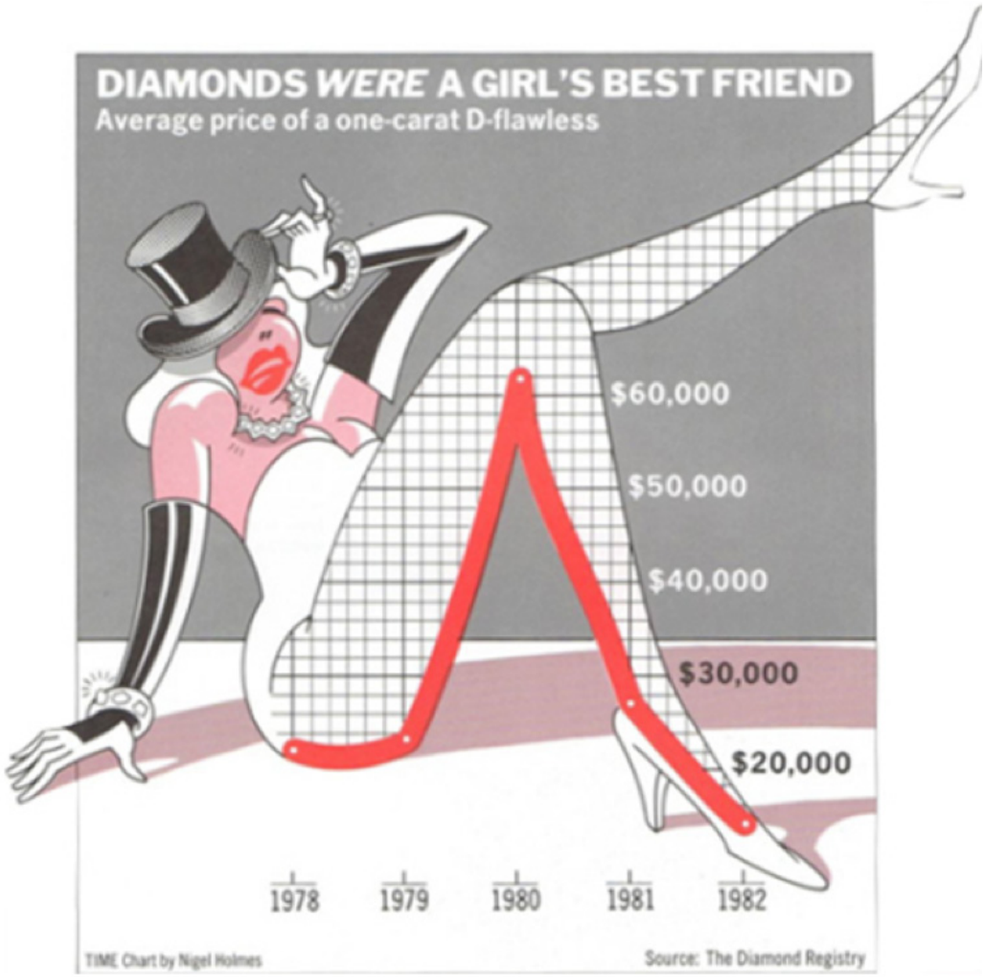


Create mnemonics

A *mnemonic* is something that assists viewers to recall information. A visualization that helps users recall the story in the data is preferable to a forgettable one.

09

In addition to attracting attention, *Diamonds were a girl's best friend* operates as a mnemonic. It plants in our memory that diamond prices trace the line from the reclining woman's buttock, up to her knee, and down to her instep. In doing so it associates the trajectory of diamond prices, (something we didn't know prior to seeing the visualization) with the shape of a bent leg (something we know well). Because our memories work through association, this improves our chance of remembering the visualization. Sexual associations are known to be particularly strong,[13] which further improves the quality of this visualization as a mnemonic.



[13] Joshua Foer (2012), Moonwalking with Einstein: The Art and Science of Remembering Everything, Penguin Books.

Another classic example by the same designer, Nigel Holmes — albeit one that clearly fails in terms of accurately representing the data — is shown below. A study has found that visualizations like this are more likely to be recalled after three weeks.[14]

Considerable skill and time is required to create visualizations like these, but a more straightforward approach is the liberal use of icons to create pictographs, such as in the simple visualization to the right. Although it is paint-by-numbers creativity compared to the earlier examples, it still is likely to be more memorable than a more conventional visualization.[15]

[14] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, Christopher Brooks (2010), Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts. ACM Conference on Human Factors in Computing Systems (CHI).
[15] This visualization was inspired by a similar waffle chart in Cole Nussbaumer Knaflic (2015): Storytelling with data, Wiley. The data is from https://www.businessinsider.com.au/warren-buffett-berkshire-hathaway-historical-returns-2017-5.

When a visualization attracts attention, it is more likely to be remembered. People can't recall things they don't notice. However, the technique of creating mnemonics is about more than just attracting attention: it involves creating some relevant association between the data and the visualization. In *Diamonds were a girl's best friend* the association relates to the shape. In *Monstrous costs* the association relates to the monstrous implication. With the visualization above, associations are created by the image of a wallet and the color of money.
Some general advice from the designer of the first two visualizations is to use only images that are instantly recognizable. Generally useful ideas include:[16]

- Sports. If the goal is to show performance, sports imagery such as jumping over, getting to a line first, hurdling, diving, scoring a goal can be successful.
- Tools. Where the goal is to show force of some kind — such as crack down, apply pressure, cut, saw, squeeze, or measure precisely — a relevant tool works well.
- Domestic appliances. Refrigerators freezing prices, vacuum cleaners sucking away profits, beds for sick economies, plants for growth, windows for looking through to the future, etc.

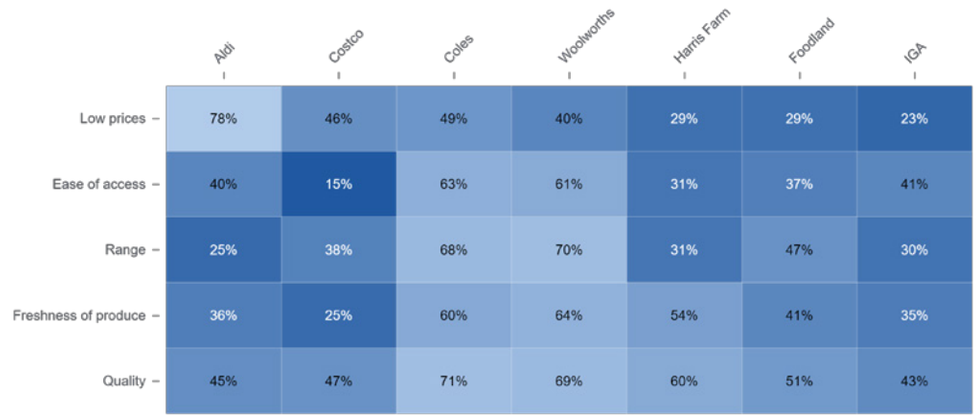
[16] Nigel Holmes (1984), *Designer's Guide to Creating Charts & Diagrams*, Watson-Guptill Publications.

Redundant encoding

Encoding refers to how data is represented in a visualization. *Redundant encoding* is when the same information is represented in multiple ways. With a few exceptions, it is generally desirable to use redundant encoding.

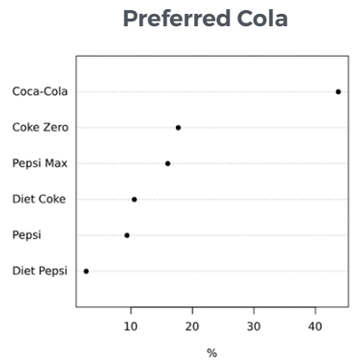
10

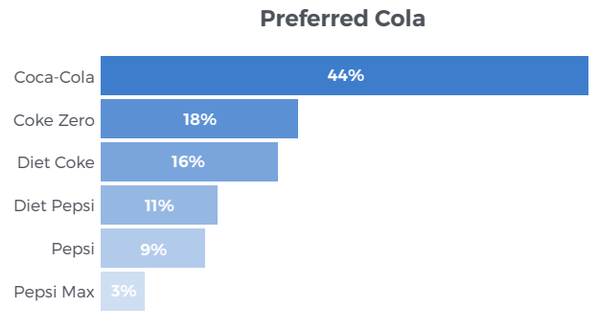
The heatmap below shows a rudimentary example of redundant encoding. The data is encoded both by the numbers shown in the cells and by the colors. In other words, the information appears twice. The benefit in this is straightforward: color allows us to see the pattern easily, and numbers allow us to quantify the patterns.



In academic visualizations, dot plots like the one below are often presented as being good practice. They maximize the data-ink ratio, which, as discussed in Chapter 7, De-clutter, is a way of reducing extraneous information.

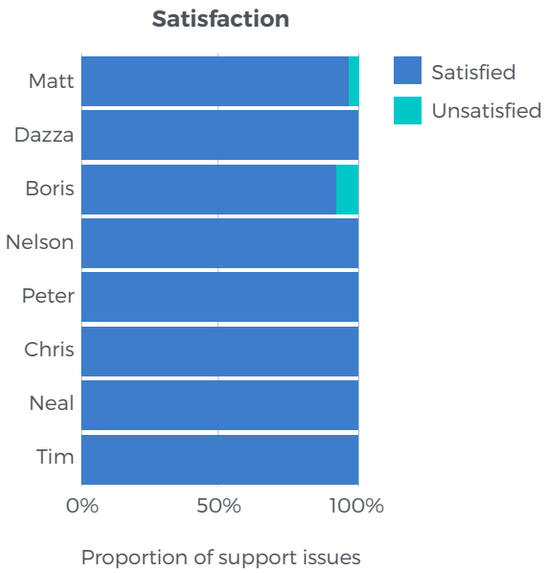
Dot plots are rarely used outside academic and government statistics. The bar chart is much more popular. With the dot plot, the viewer must deduce that the dots are in two-dimensional space. With a typical bar chart, such an interpretation is also possible if the viewer looks at the ends of the bars. However, there are two additional encodings that are not present in a dot plot: the width of the bar and the area of the bar. It is also possible to add more redundant encodings to a bar chart. In this example, the color of the bars also encodes the key findings, as do the order of the bars, the position of the labels, and the labels themselves. By encoding equivalent information in lots of ways, the bar chart makes it almost impossible for the viewer to miss its key patterns, which makes it the default chart of choice.

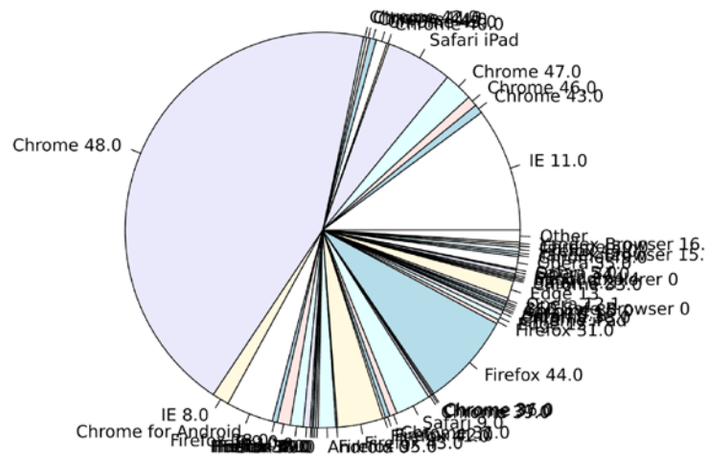




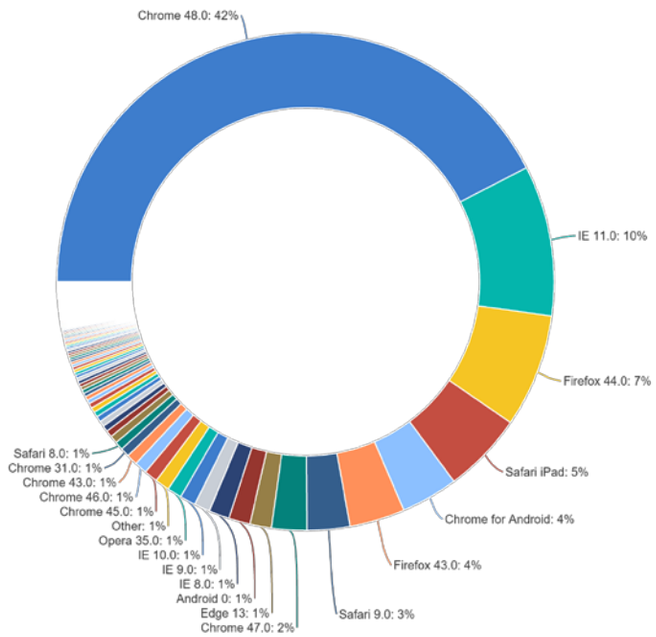
It has been argued that the redundant encoding is itself an example of clutter and should be avoided.[17] However, there is no data to support this conclusion: the data shows either that redundant encoding is desirable and "almost always a benefit"[18] or that it makes no difference at all.[19]

In Chapter 5, Represent accurately, we examined the stacked bar chart on the previous page, concluding that showing the both the unsatisfied and satisfied data on the same visualization was poor because the information was redundant. Redundancy in that context meant something different. The problem with using a stacked bar chart to represent only two categories is that the two series represent the same information in the opposite way. As a result, the two series cancel each other out, and do not add any more information. Even with the addition of a legend, the chart can only be interpreted with some effort and time on the part of the user.



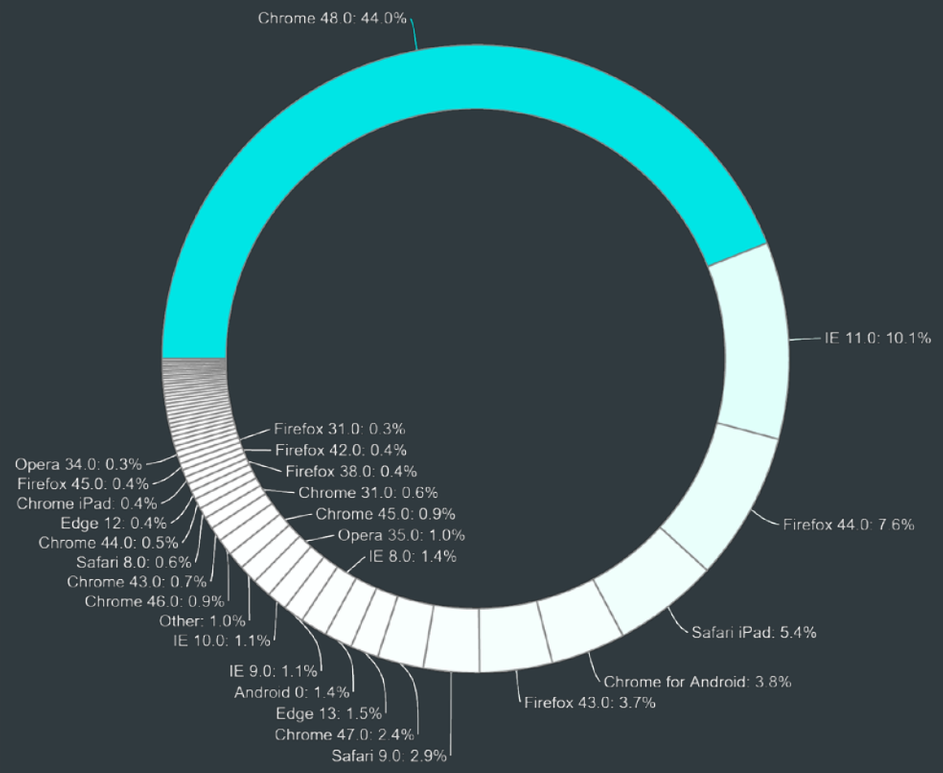


The chart above is a car crash — presumably it is examples like this which have given pie charts such a poor reputation among visualization pundits. By contrast, the donut chart below is extremely effective, in part because it has been designed to avoid the overplotting of labels. However, it also benefits from the use of redundant encoding. In the pie chart above, data is encoded in three ways: by the slice angles at the origin, by the slice area, and by the length of the arcs on the outside of each slice. In the donut chart below, the same information is also encoded by order, font size, and value labels.

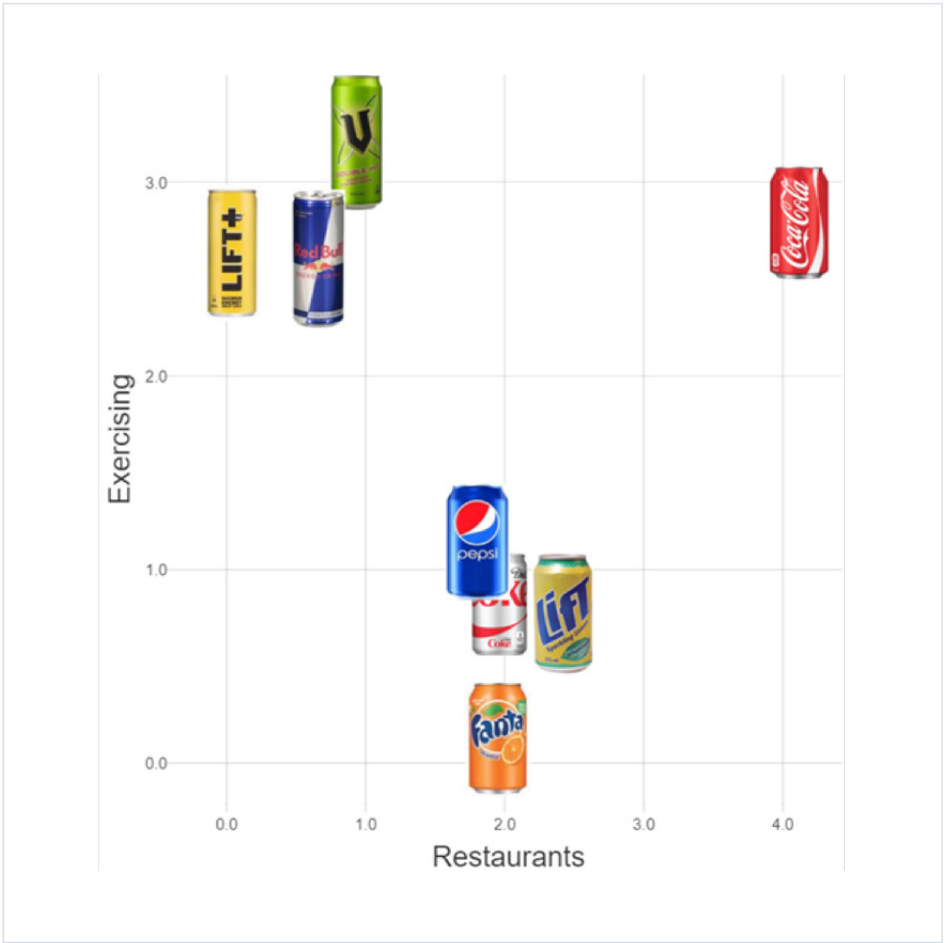


[17] Edward R. Tufte (1983), *The Visual Display of Quantitative Information*, Graphics Press.
[18] Colin Ware (2012), *Information Visualization: Perception for Design, 3rd Edition*, Morgan Kaufmann, Kindle Edition, p. 159.
[19] Russell Chun (2017), "Redundant Encoding in Data Visualizations: Assessing Perceptual Accuracy and Speed", Visual Communication Quarterly, Volume 24 (3), pp. 135-148.

We can introduce further redundancy by coloring the slices proportional to the values in the data, as done below. This visualization best communicates the overall dominance of Chrome 48. But it is not as pretty as the one above, which raises the question of how well it will attract attention.



The use of cans in the visualization scatterplot below is also an example of redundant encoding, as these cans, like most branded products, employ extensive redundant encoding (i.e., the branding is communicated by name, colors, font, and overall design).



One strategy for redundant encoding that typically also pays dividends in terms of attracting attention is the use of icons or images for countable data. In addition to the redundancy achieved via the branding, the pictograph below can be counted by the viewer.

Coca-Cola per day: 12.5

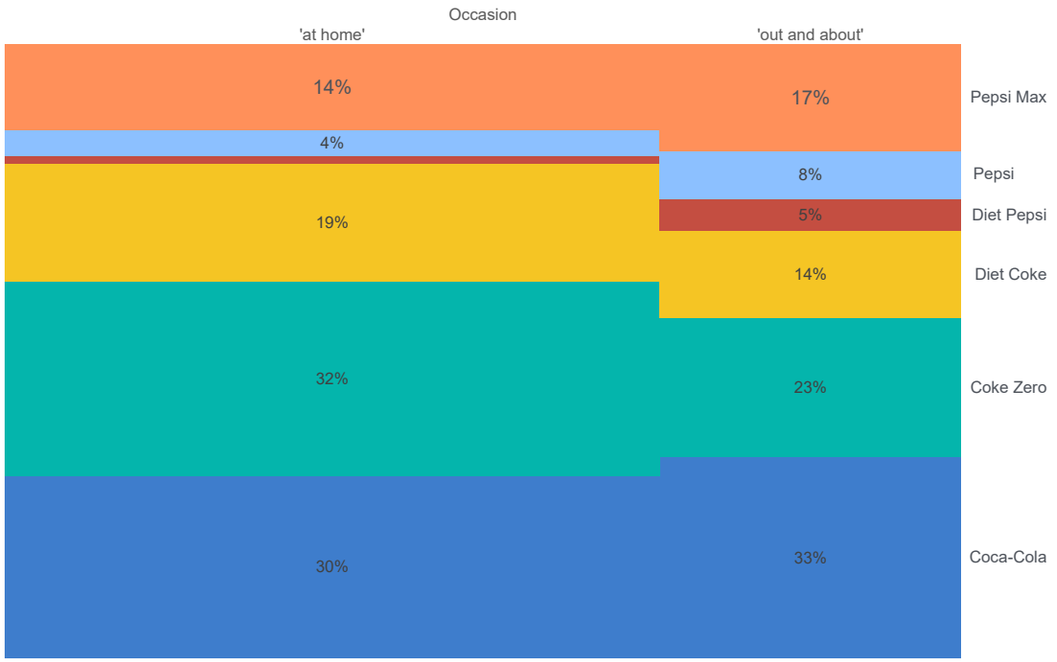
The strategy of using countable pictographs is achievable with many different types of visualizations, from pie and bar charts all the way through to *small multiples of treemaps* as in Otto Neurath's visualization of the First World War, in which the number of soldiers and casualties of the armies can be compared based on area and number of icons.

Reduce eye movement

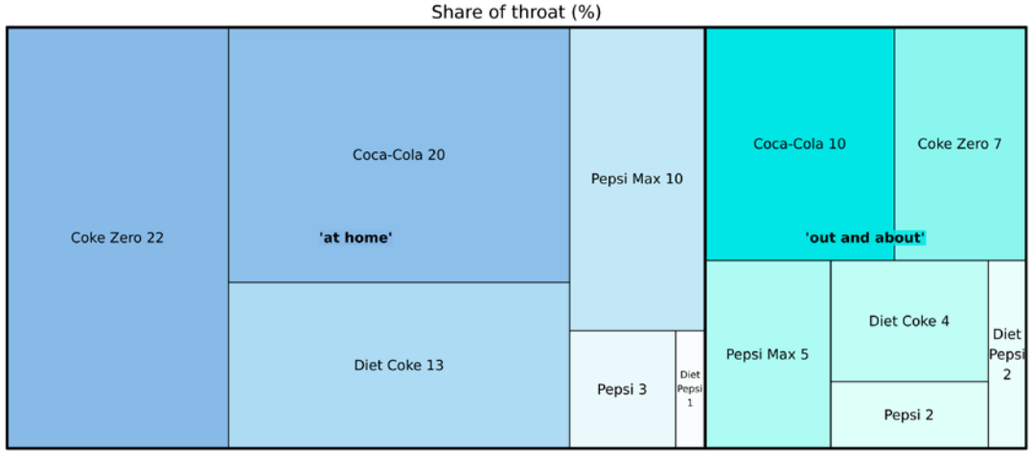
In an ideal world, a visualization can be instinctively understood with a glance. Usually, however, the user must repeatedly scan the visualization looking for key information. For example, they may have to look up the meaning of a color in a legend.
All else being equal, the fewer eye movements required for a viewer to interpret a visualization, the better.

11

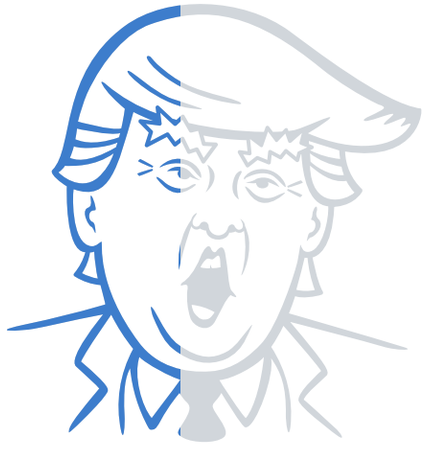
Consider the *mosaic chart* (also known as a *Marimekko* or *"mekko"*) chart below. What does the gold cell in the left column mean? To work this out the viewer must read to the top to work out that it means "at home", then trace the light orange to the right to deduce it is diet Pepsi. Because there is no number in the cell, they then need to compare its area to the other areas to deduce its size. It is not surprising that mosaic charts typically need to be explained.



By contrast, the *treemap* below arranges the tiles in such a way that there is room to fit in labels, so the user needs less eye movement to deduce the 1% of consumption is of Diet Pepsi “at home.” An added advantage of this is that color is no longer needed to assist in looking up the brand, so it can instead be used to show both occasion (color) and consumption (intensity of color). That is, in this example, the treemap has more redundant encoding than the mosaic chart.

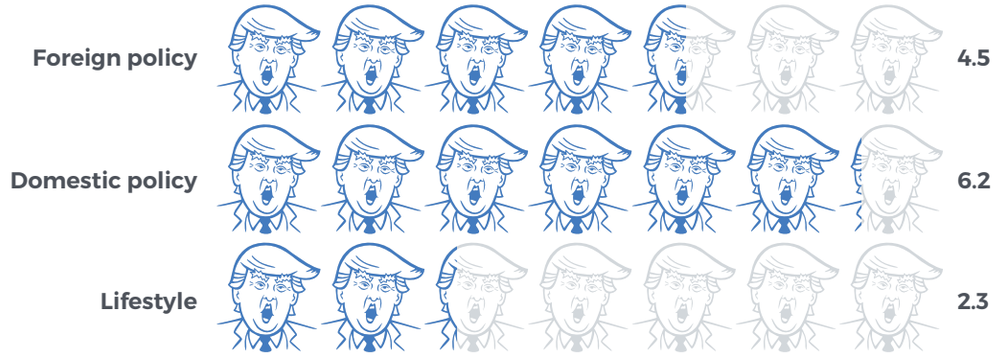


The *pictograph* below also reduces eye movement. By using an image of President Trump to create the image, one need not even label the visualization, eliminating the need for the viewer’s eyes to move from the image to the title and back.

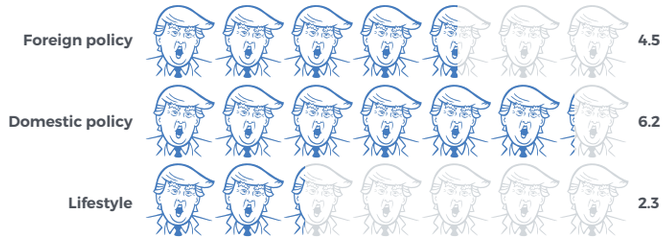


43% approval

The *pictograph* below shows data for a seven-point rating scale. In this case, the number of images performs the role traditionally performed by the user having to look up numbers on an axis or methodological note describing the number of scale points.



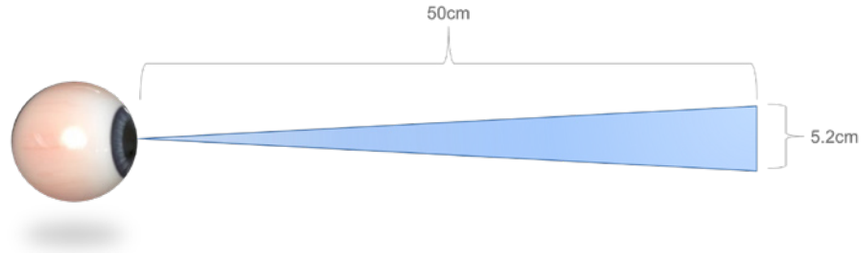
A simple mechanism for reducing eye movement is to reduce the size of any visualization. The smaller the visualization, the less time taken to scan it. The goal to strive for is to make all distinctions as small as possible but still clear to the viewer.[20]



Lean back away from the screen and look at the image below.[21] What do you see? A woman?

Now lean in and look at the image again. What do you see? You will likely notice her mouth is a gremlin and the right-side of her hair a toucan.

We can view an image as a whole only if it falls within about a 6˚ focal arc from our eye.[22] If we are 50 cm (20 inches) from an image that means the image needs to be less than 5.2 cm (2 inches) wide and high. If the image is larger we need to move our eyes to assemble the image within memory. Small visualizations are better.



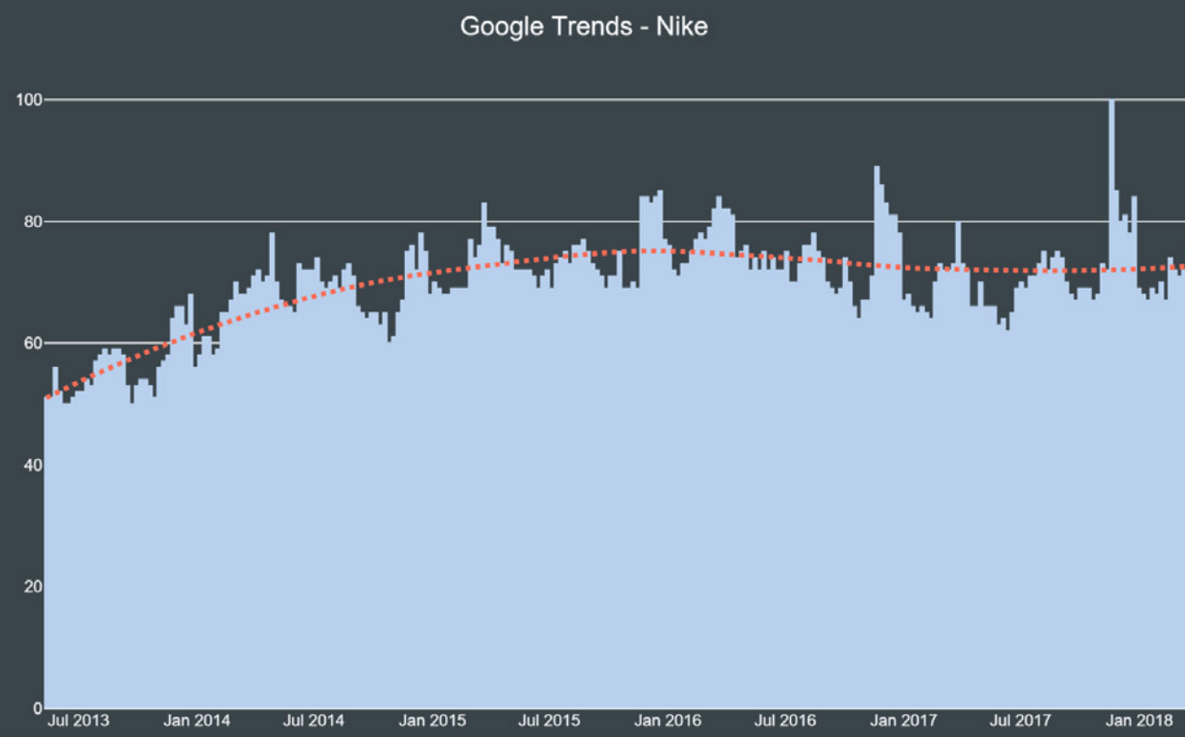
[20] Edward R. Tufte (1977), *Visual Explanations*, Graphics Press.
[21] From Colin Ware (2012), *Information Visualization: Perception for Design*, 3rd Edition, Morgan Kaufmann, Kindle Edition, p. 294.
[22] From Colin Ware (2012), *Information Visualization: Perception for Design*, 3rd Edition, Morgan Kaufmann, Kindle Edition.

Show norms

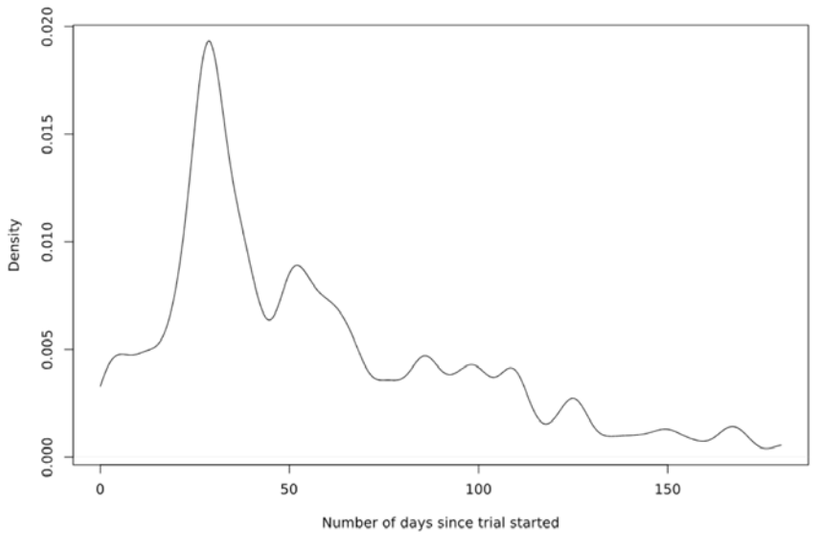
Many visualizations can be improved by showing norms (e.g., typical, average, median, or “normal” results).

12

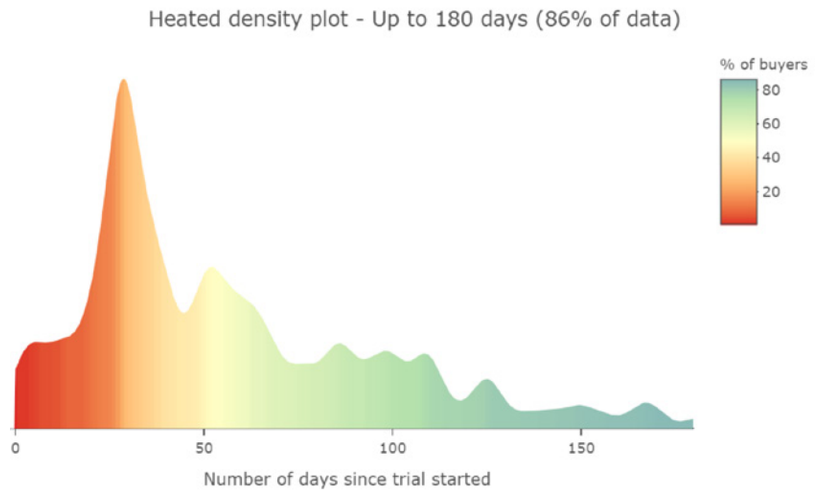
The most widely used *norm* in visualization is to show a line of best fit, such as in the visualization below. This allows the viewer to get an appreciation for both the overall trend and the difference of individual data points from that trend.



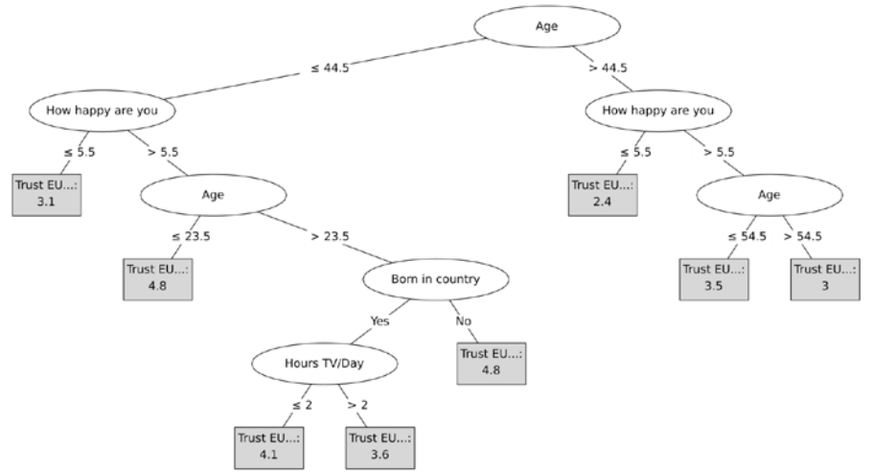
Density plots, like the one shown below, are commonly used for displaying the distribution of data. Although they can do a good job of showing the overall shape of a distribution, along with modes and the shape of tails, they cannot communicate simple statistics, such as medians and quartiles.



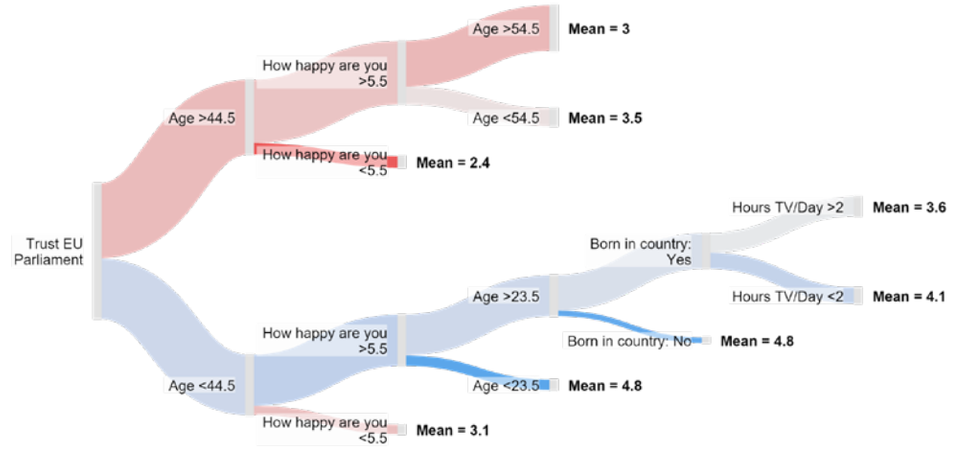
One solution is to overlay heatmap shading. In the example, we can see that the typical number of days since the trial started is a little over 50, which is not at all clear from the standard density chart.



In the *decision tree* below, which shows trust in the European Union Parliament by UK people prior to Brexit, there is no normative data. To interpret the tree, the viewer needs to read and remember all its content.



By contrast, when the tree is represented as a *sankey diagram*, with the branches colored according to the average values, interpretation becomes much more straightforward. In this example, the average level of trust is shown in gray and the width of the branches shows the proportion of people. So, we can quickly see that the key divide in the UK relates to age, with older and unhappier people having lower levels of trust in the EU parliament.



Emphasize

Use visual cues and comments to make the intended focus clear to the viewer (e.g., summaries, callout boxes, highlighting, arrows, color).

13

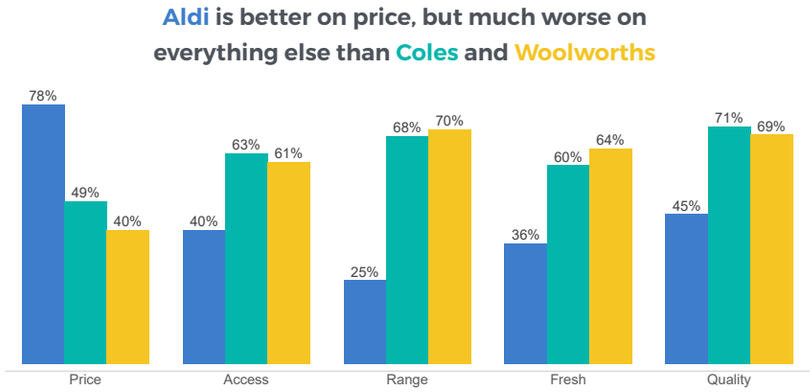
It is almost painful to count the threes below —

75639506847365866303757686037265860284658910783 0

— but much easier when they are emphasized.[23]

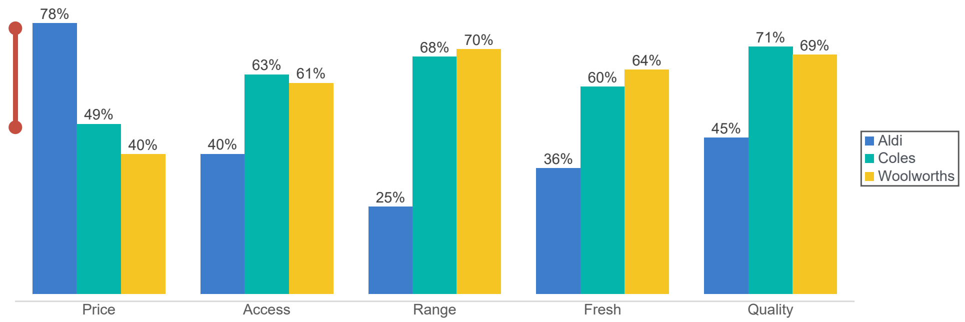
75639506847365866303757686037265860284658910783 0

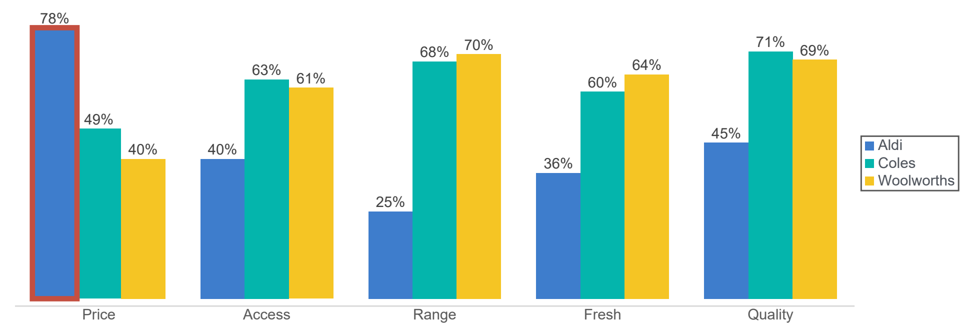
The simplest way to add emphasis to a visualization is by adding commentary.

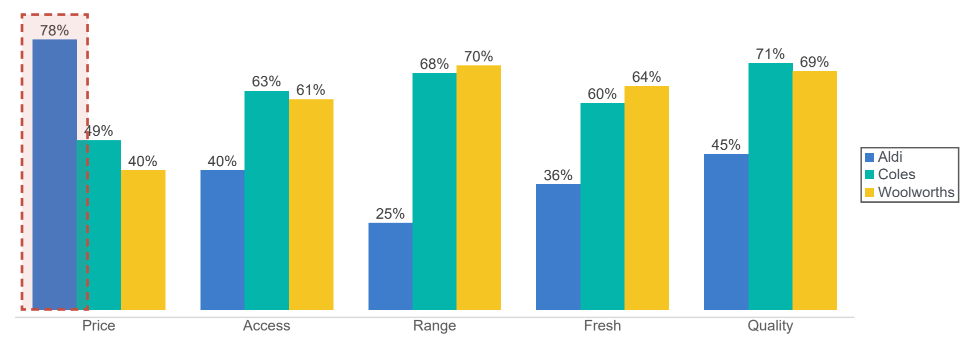


[23] Adapted from Cole Nussbaumer Knaflic (2015): Storytelling with data, Wiley, p. 103.

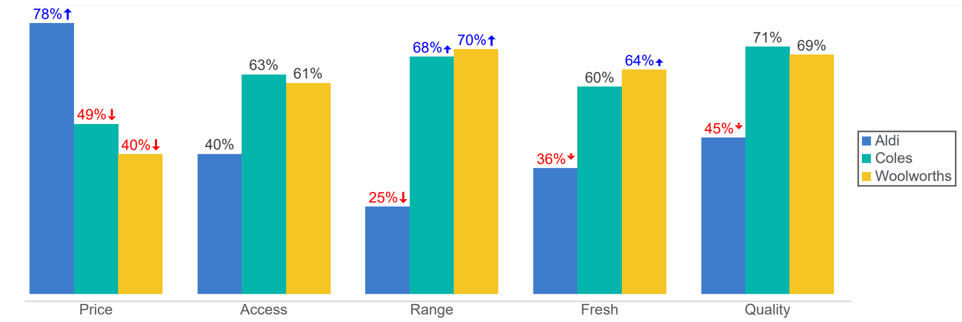
Adding additional graphical devices to emphasize key contrasts is also effective, as in the three examples below.

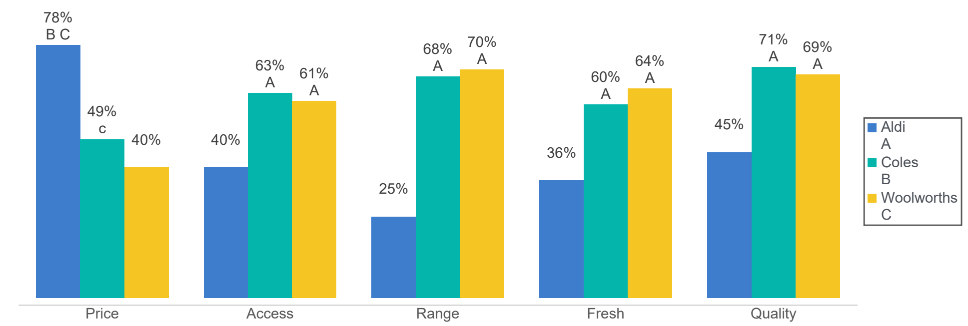






With more technical audiences, emphasis can be added by using automated statistical tests, as in the two examples below.





Reduce color

Remove as much color as you can. Where possible, make everything gray (exceptions to this rule to follow).

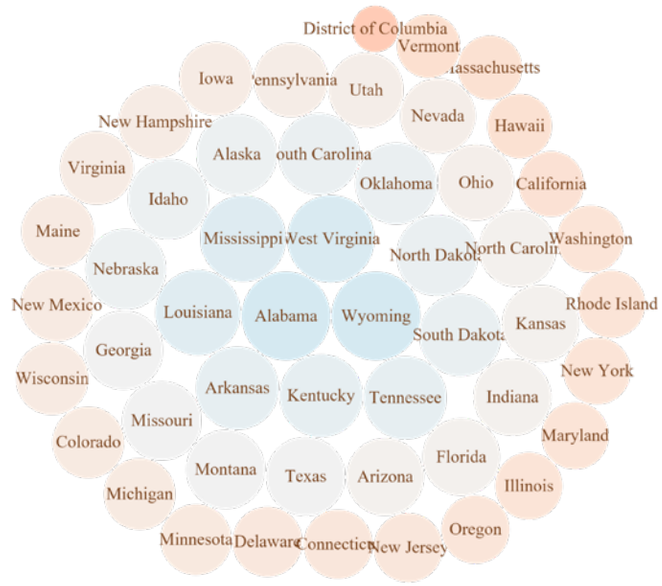
Anything you would ordinarily do in black will look better in dark gray.

Any peripheral information should be in a light gray so that it does not detract.

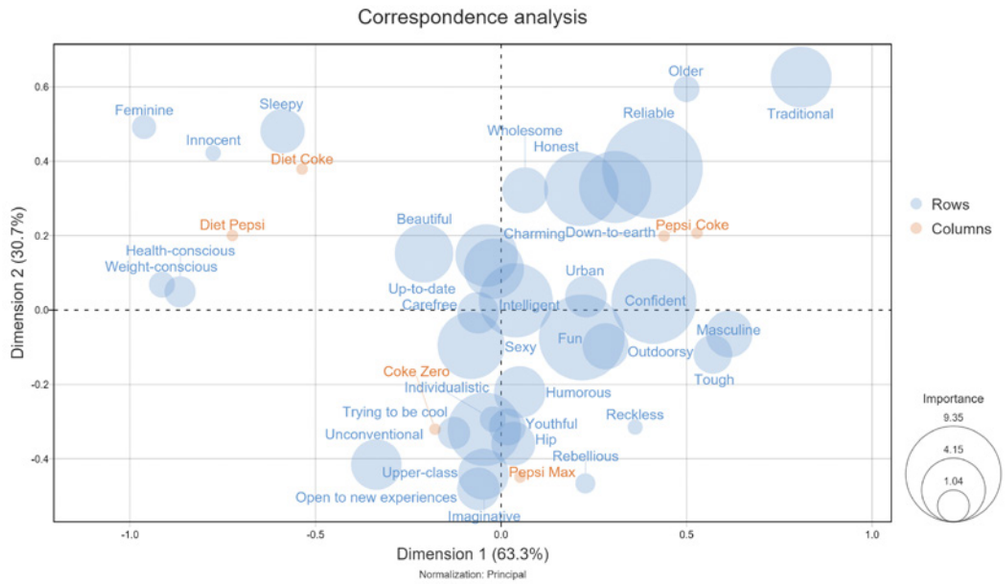
14

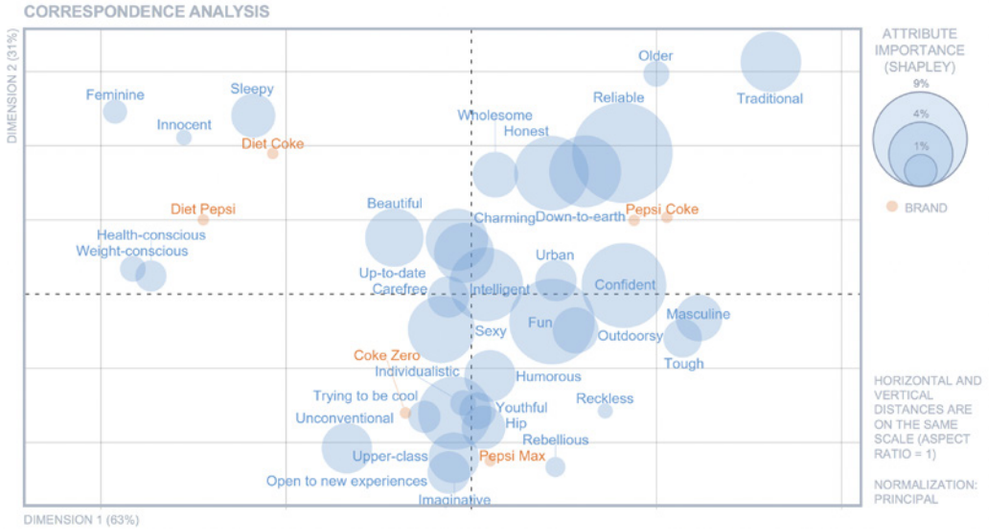
The *word cloud* above sets the font sizes proportional to approval of President Trump as measured in March 2018. An unfortunate aspect of this word cloud is that the most interesting result — that the District of Columbia has the lowest approval — is the hardest insight to find. This is due partly to our eyes focusing on larger words over smaller, but it is also due to how color has been used. In a word cloud, color is used to help the viewer disambiguate words. However, this results in a riot of color that makes it hard to focus on the meaning of the words.

The *circle packing* visualization is much more effective, in part because it is substantially less colorful. It is also because circle packing uses a norm (gray is 50% approval) and lots of redundant encoding, with approval ratings indicated by the size of the circles, the color, and their order.



A subtler example of reducing color is the *labeled bubble charts* below. In some ways the version at the top is more appealing. However, the black framing around the edge of the visualization detracts attention from the bubbles, which is mitigated by making them all gray and aligning them in rectangles (the focus of the next chapter). More examples of reducing color are presented in the remaining chapters.

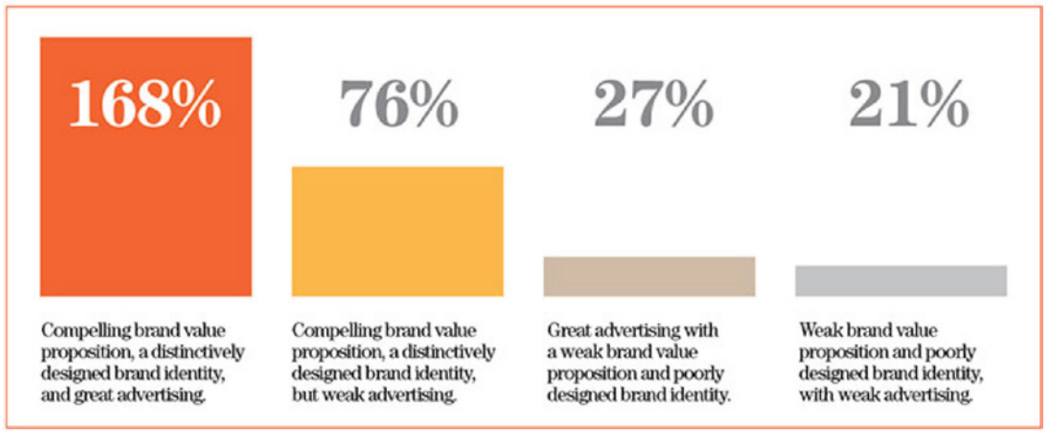




Form rectangles

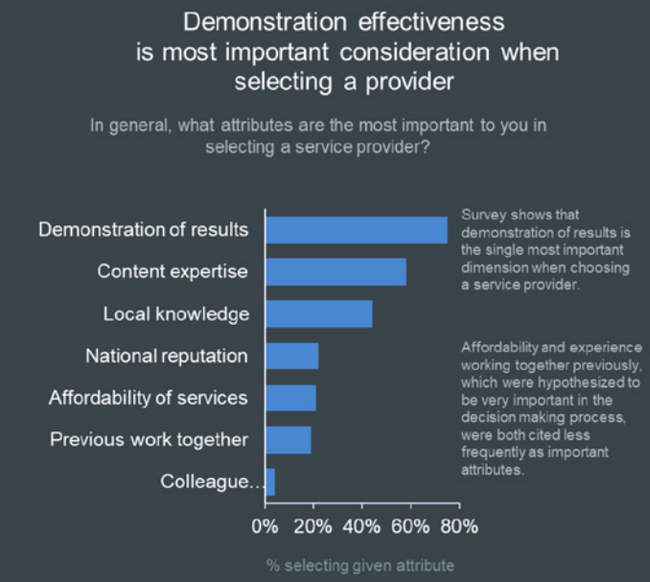
A visualization can often be improved by placing subjectively located elements and spaces so that they form rectangles. While this is a useful hack, a professional designer will be able to improve on this, as they have a much richer repertoire of design principles from which to draw.

What makes the *column chart* below[24] pleasing? Yes, the colors and fonts are nice. Less obvious is that the elements have been laid out to form rectangles.



[24] http://www.trulydeeply.com.au/2015/06/investing-in-brand-beats-advertising-study-branding-agency/

The commentary below makes the resulting visualization look quite messy. The mess is a problem on two levels. First, who wants a mess? Second, what we perceive as mess distracts us from looking at the data.



The same visualization with elements aligned into rectangles and colors directing eye movement is now more appealing and less distracting.[25]

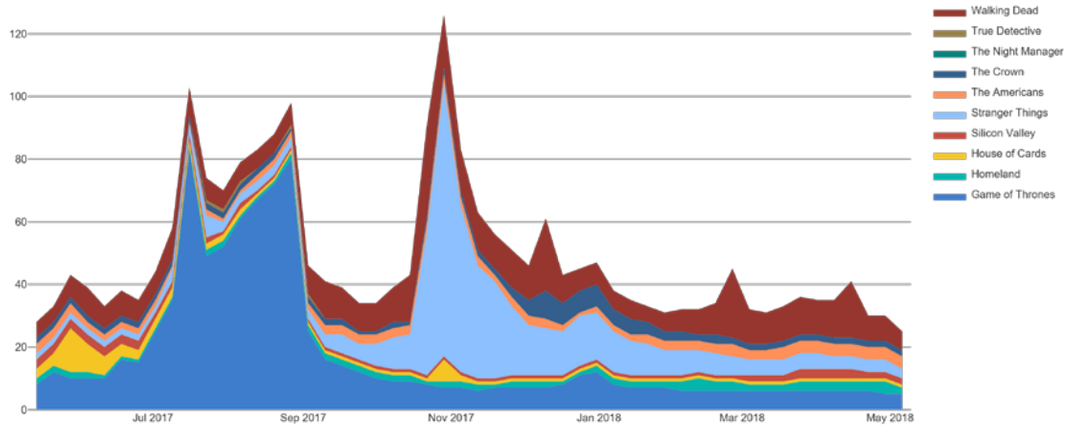
[25] Adapted from Cole Nussbaumer Knaflic (2015): *Storytelling with data*, Wiley, p. 81-82

Create symmetry

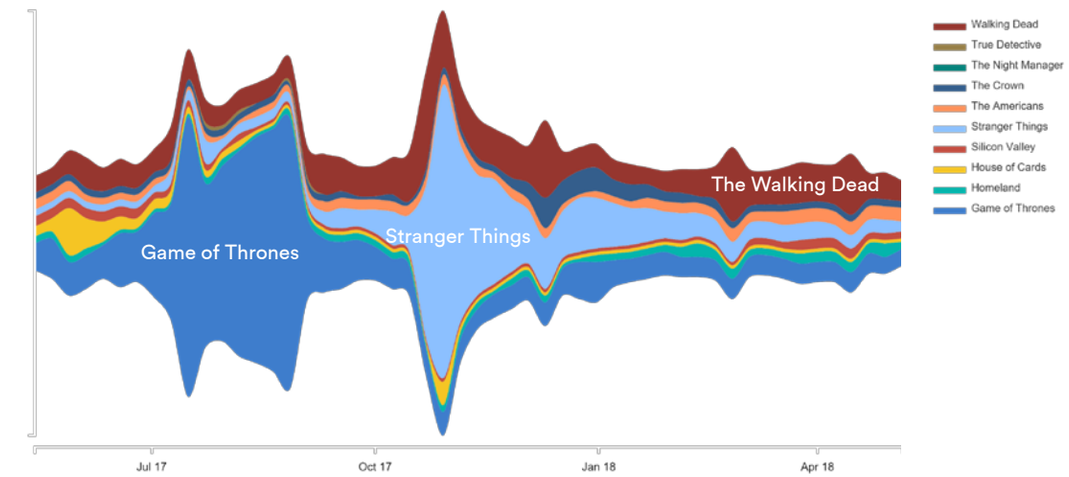
Forcing visualizations to have symmetry can be surprisingly effective.

16

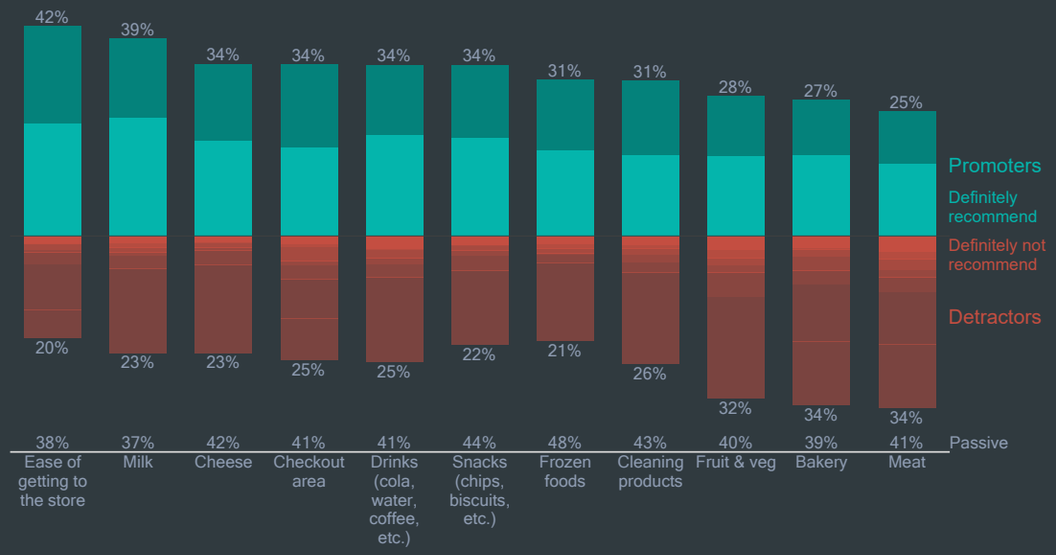
The *stacked area chart* below is interesting to look at but tells us little other than the peak times of popularity for *Game of Thrones* and *Stranger Things*. This is because any spikes in the lower series make the upper series harder to interpret.



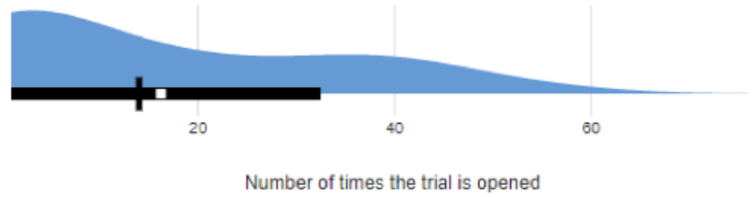
Stream graphs improve on area charts by making the visualization symmetrical around its middle. This does not affect our ability to interpret the cumulative data (which is now shown at the top and the bottom) but does make all the other series easier to read. For example, it is now clear from this visualization that *The Walking Dead* has relatively continuous appeal throughout the data. The reason that the stream graph is often better than the area chart is that by making the chart symmetrical, the height of the spikes is halved, which in turn makes the degree of distortion on the other series smaller.



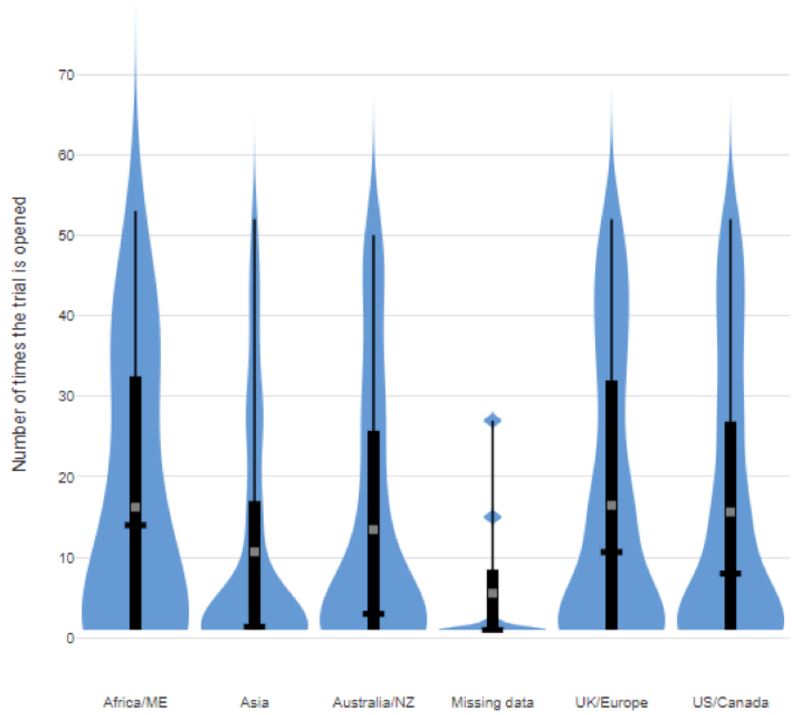
These benefits are also the reason why many market researchers to summarize rating scale data using stacked column charts with positive and negative series.



The visualization below seeks to improve upon *density plots* (see Chapter 12, Show norms) by overlaying a form of box plot over them, where the one dot shows the mean, the bar the median, and the thick black line shows the quartiles.

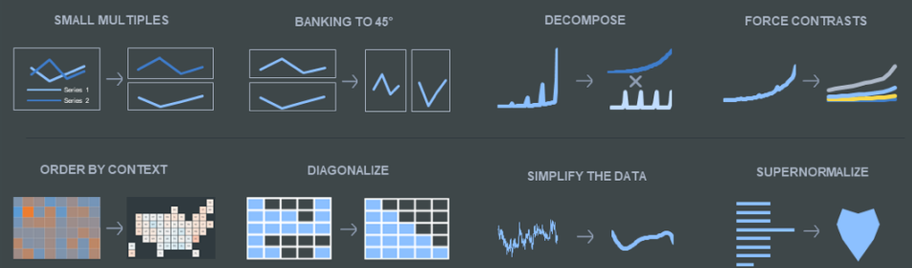


The additional information somehow makes the visualization less appealing. However, by mirroring the density plot and optionally flipping it onto its side, we obtain the more popular *violin plot*. Here the symmetry is ultimately redundant: it provides no real information but nevertheless makes the visualization more appealing.



Reshaping

This section focuses on techniques that improve visualizations by fundamentally altering the shape that appears in the data.

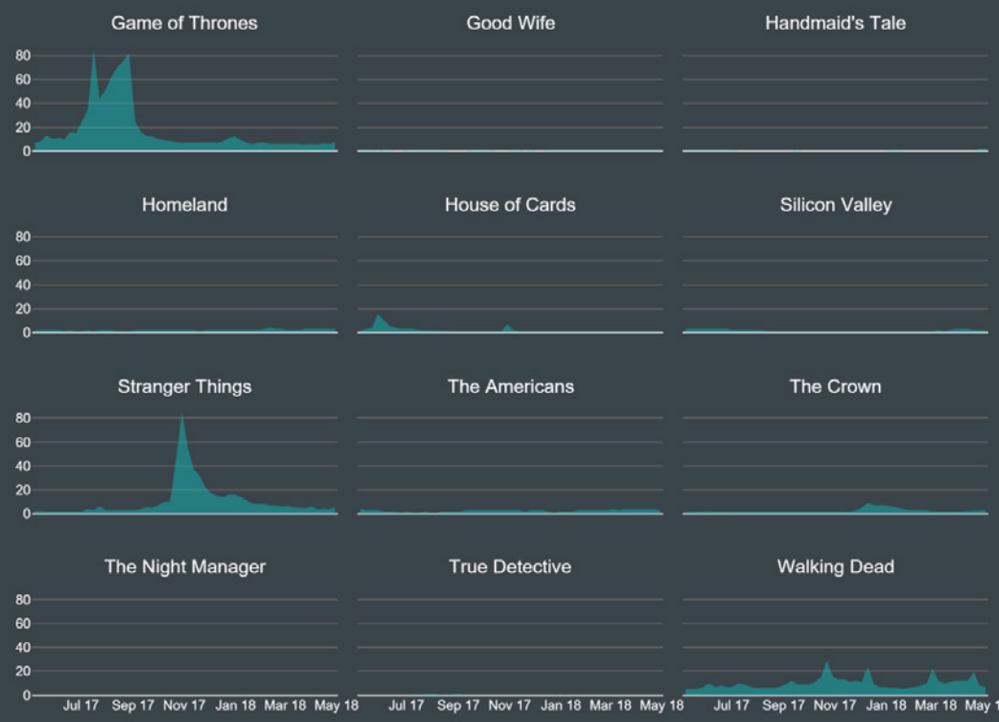


Small multiples

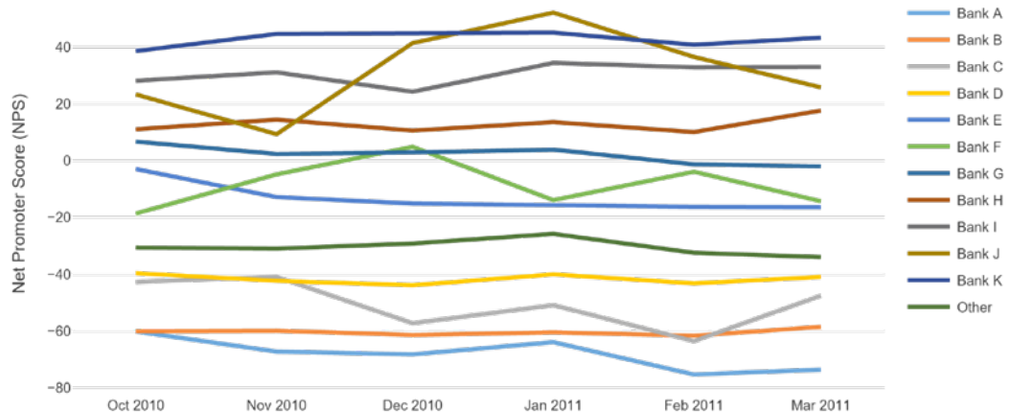
A *small multiple* design creates a separate visualization for each series of data. They are best for revealing the range of variation in the data and for comparing series of data.

17

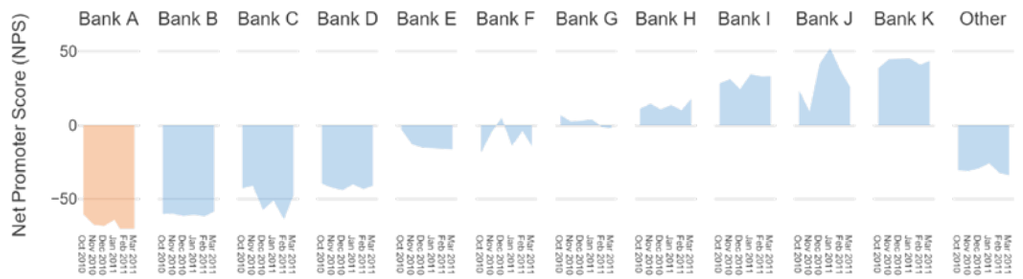
The visualization below shows the TV data from the previous chapter. Even though each data series is now given less than 10% of the space from the previous visualization, the visualization is substantially more revealing. The data for *Game of Thrones* is almost as detailed here, but the data for all the other series is displayed significantly better.



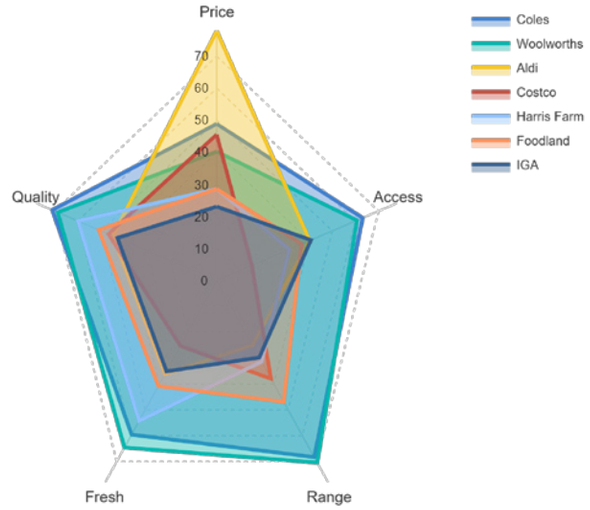
Both the stacked area chart and the stream graph suffer from obvious problems when it comes to representing the data for different series, due to the distorting effect caused by the stacking. No such distortions exist in the *line chart* shown below. Nevertheless, this visualization is still very difficult to read.



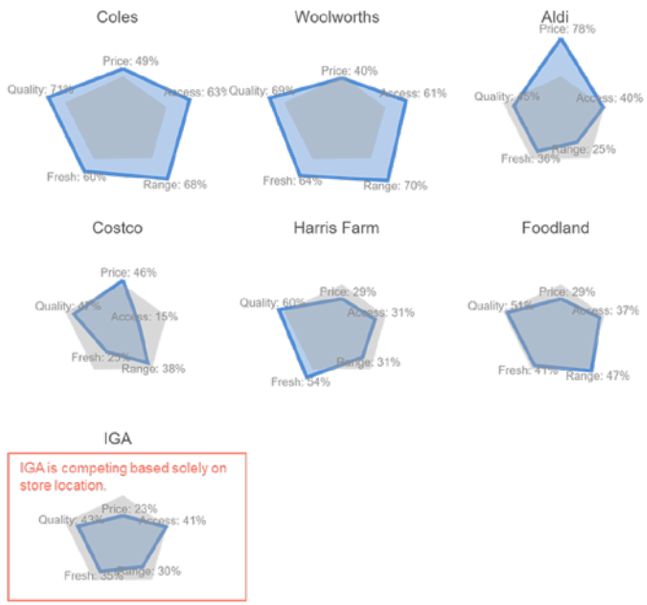
Small multiples of *area charts* are shown below. Even though each part of the visualization is very small, the key patterns are much easier to see. This is also aided by reduced color, added emphasis, and sorting. This visualization now makes it easy to compare the banks in terms of their overall levels, as well as their trends. The visualization also reveals the downside of small multiples: the number, size, and neatness of the labels.



Radar charts, also known as *spider charts*, are problematic when there are more than a small number of series being plotted. In the example below, which plots only seven series, we can extract insight, but it takes effort.



By contrast, in the small multiples version it is much easier to compare the different brands' strengths and weaknesses. The use of the small multiples has allowed us to improve the visualization by sorting, showing the norm (the averages, represented by the gray), reducing colour, and emphasizing a result that is of interest to the target viewer.



In the examples on the left, small multiples have been used as a way of disentangling more complicated visualizations. They can also be used more generally. For example, the table below shows the output from a model looking at brand preferences, where the norm, 0, is shown by the change of red to blue. This visualization allows the viewer to see readily that Apple is the most divisive of brands, with people who variously dislike it, are ambivalent, and like it more than any other brand, whereas Google is widely liked but less loved than Apple.

MaxDiff: Hierarchical Bayes

Prediction accuracy (leave-1-out cross-validation): 70.5%

	Respondent Coefficients	Mean	Standard Deviation
Apple		3.2	6.1
Microsoft		1.1	2.3
IBM		-2.4	1.8
Google		2.3	2.2
Intel		-1.1	1.5
Samsung		2.0	3.8
Sony		0.9	2.2
Dell		-1.6	2.4
Yahoo		-2.6	1.6
Nokia		-1.7	2.8

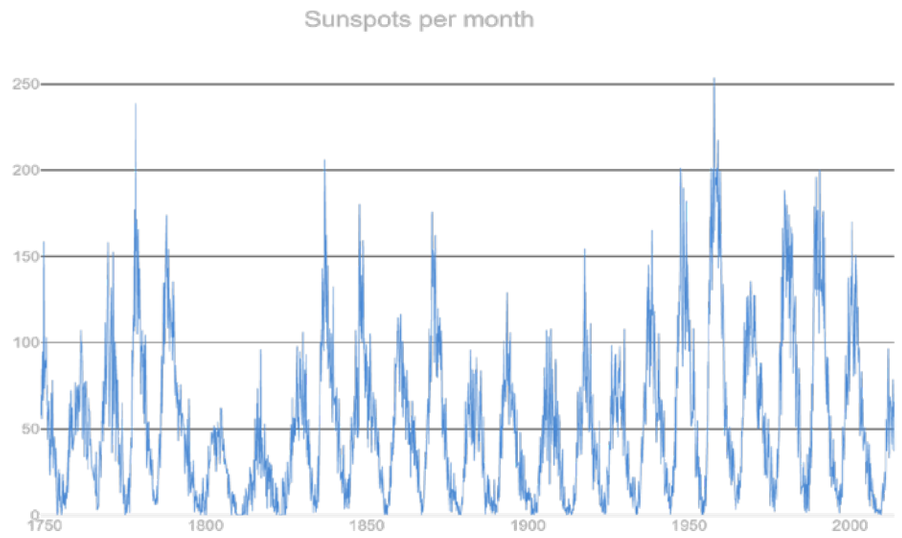
Each of the examples in this chapter illustrates the effectiveness of small multiples. It is a powerful technique. However, the use of small multiples does involve trade offs. As discussed earlier, the small size can make the labels messy and difficult to read. A second limitation is that they trade an improvement in ease of seeing the pattern for a decrease in ease of comparing series. In situations where there is minimal overlap between series and no need to use color to disambiguate series, small multiples will tend to detract. For example, data in the multiple-line chart shown in the next chapter would be less accessible if rendered as small multiples.

Banking to 45°

The vertical or horizontal orientation of visualizations should be modified so that the average “slope” is around 45°.

18

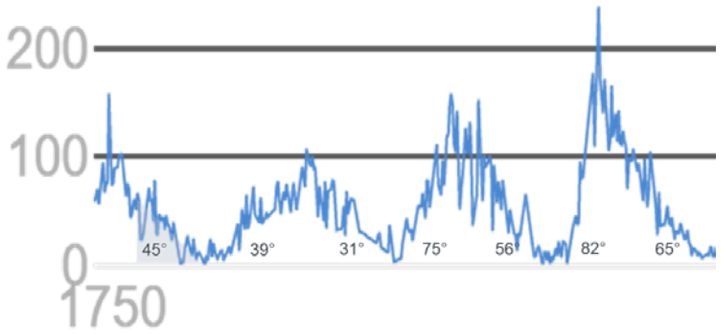
The *line chart* below shows the number of sunspots per month for the last 270 years.



The same data is shown below, except that the visualization’s *aspect ratio* — height to width — has been changed. The new visualization is vastly superior. We can still see all the patterns evident in the larger visualization, but one new pattern is now much easier to spot: the rate at which sunspot activity increases is typically much faster than that at which it decreases — that is, most of the spikes are skewed, with the peak to the left.[26]

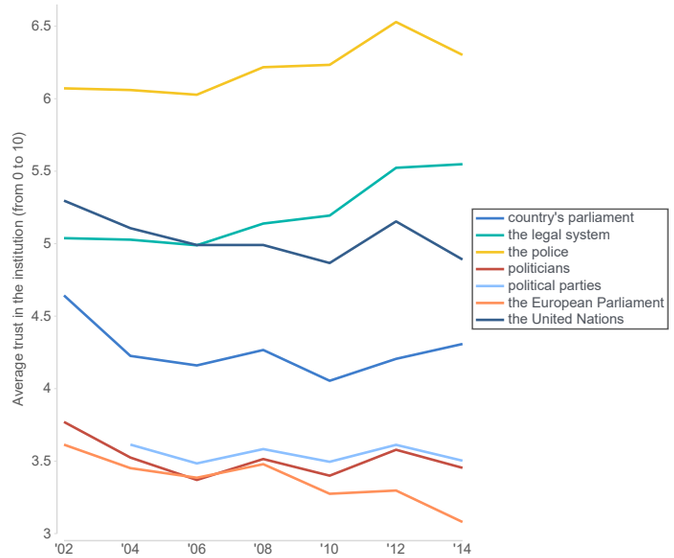
[26] Both the idea of banking, and this application, come from William S. Cleveland (1994), *The Elements of Graphing Data*, Hobart Press.

For some reason, our brains find it easy to grasp patterns that are near to 45˚ angles. In the sunspot data, the patterns relate to cyclicality rather than individual data points; by changing the aspect ratio so that the average is around 45˚, we make the trend much easier to see.

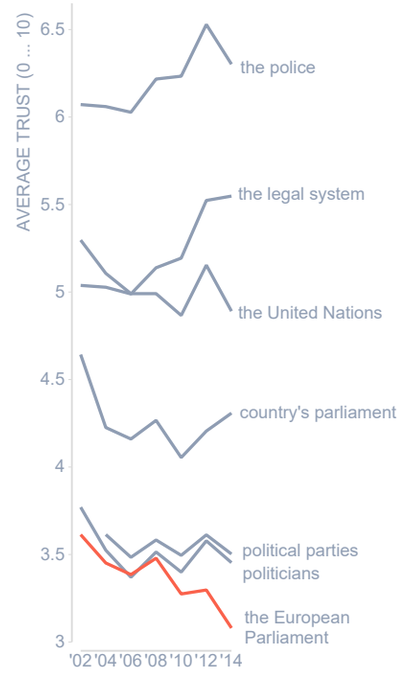


While it is possible to use algorithms to attempt to bank the data optimally,[27] eyeballing seems both easier and likely better, due to the difficulty in defining which aspect of the data to bank. In the sunspot data, for example, we do not want to bank the actual line but rather its cycle aspect. Also, the "45˚ rule" is not based on rock-solid science,[28] so the quick and easy approach of changing the aspect ratio to accommodate the eye is sufficient.

[27] Jeffrey Heer, Maneesh Agrawala (2006), "Multi-Scale Banking to 45 Degrees", IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 12(5), 701–708.
[28] Justin Talbot, John Gerth, Pat Hanrahan (2012). "An Empirical Model of Slope Ratio Comparisons", *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis).*



The *line chart* above shows attitudes to various institutions over time.[29] It is redrawn below with the data banked to 45˚, less color, clear emphasis, and a smaller size.



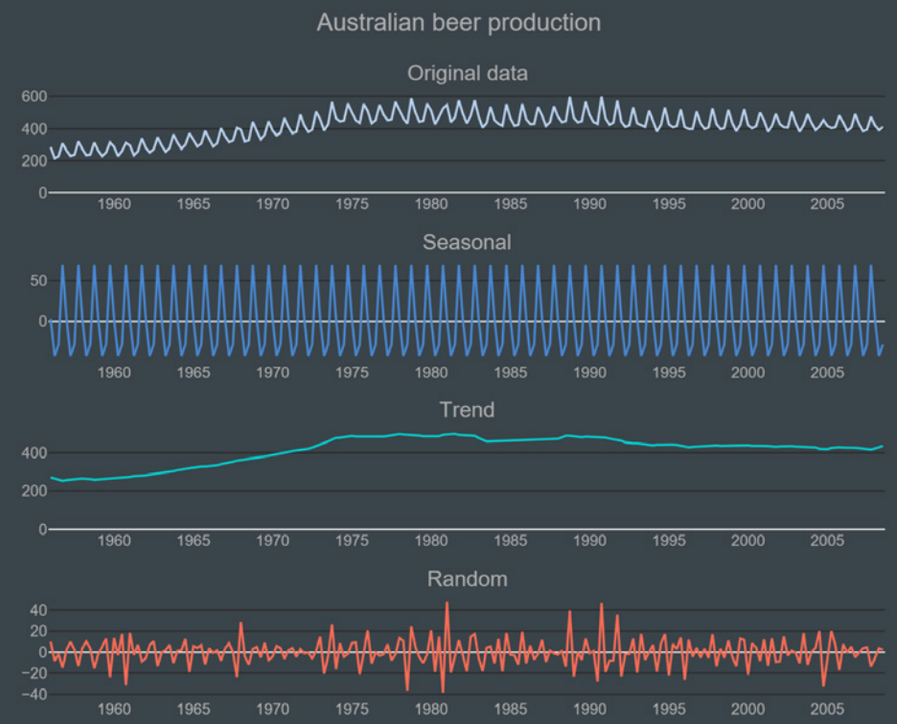
[29] European Social Survey, Data file edition 2.1. NSD - Norwegian Centre for Research Data

Decompose

Where there are multiple different patterns of interest, it is often useful to break down the data into parts, so that the viewer can concentrate on each part separately.

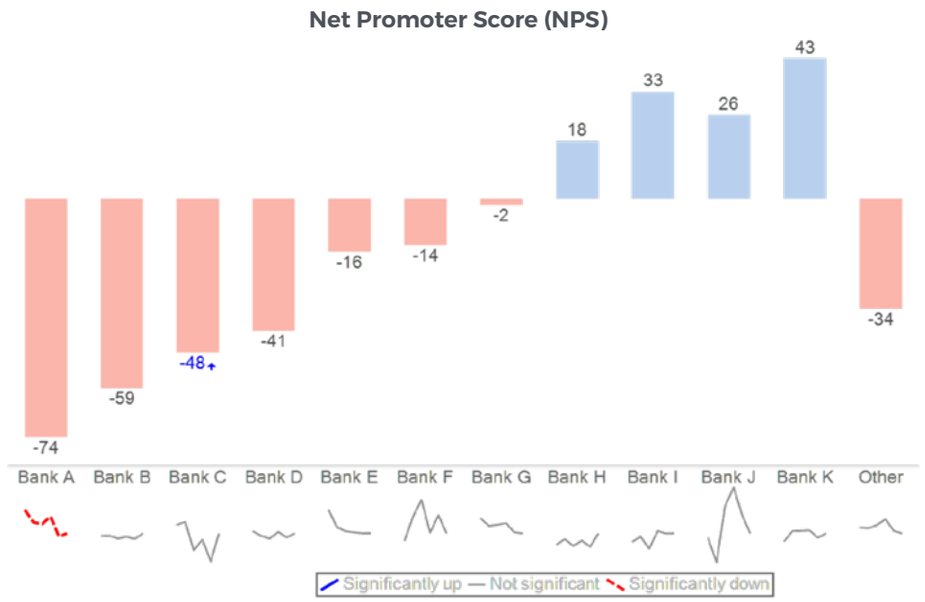
19

The most well-known type of decomposition in data visualization is the *seasonal decomposition*. In the example below, Australian beer production[30] is shown by month in the plot at the top. The three plots underneath are the components of beer sales, which collectively add up to the data shown at the top. By isolating the different components of the original data, interesting results — the precise point where growth stopped, and the various step points of decline — become easier to see.

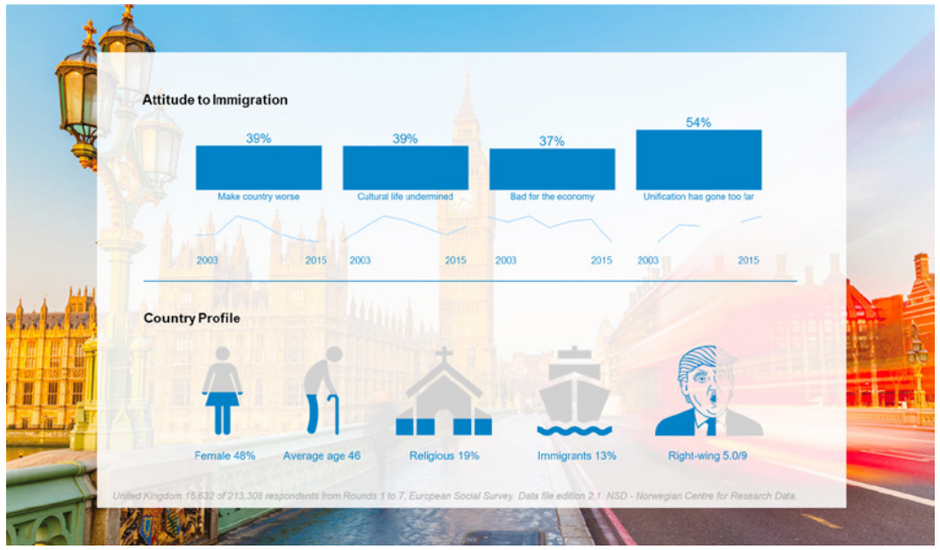


[30] Australian Bureau of Statistics. Cat. 8301.0.55.001.

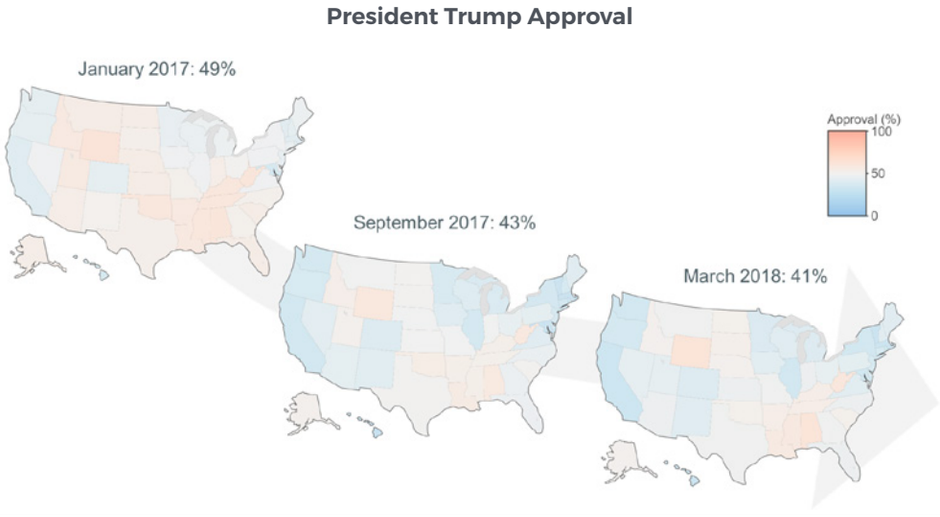
The visualization below is a *decomposition* of the banking net promoter score previously examined in Chapter 17, Small multiples. In this visualization, each column represents the most recent month's scores, as these were of greatest interest to the users of the visualization because they were used to determine bonuses. The *sparklines* below each column show the trend data. Statistically significant results have been emphasized, showing Bank C's most recent month as significantly higher than the preceding month, and Bank A as the worst and in freefall.



The decompositions shown so far have all been quite technical in their design, but the basic principle can be used in more attractive visualizations, such as the one shown below.



The visualization below decomposes President Trump's approval in terms of overall trend and trend by state.



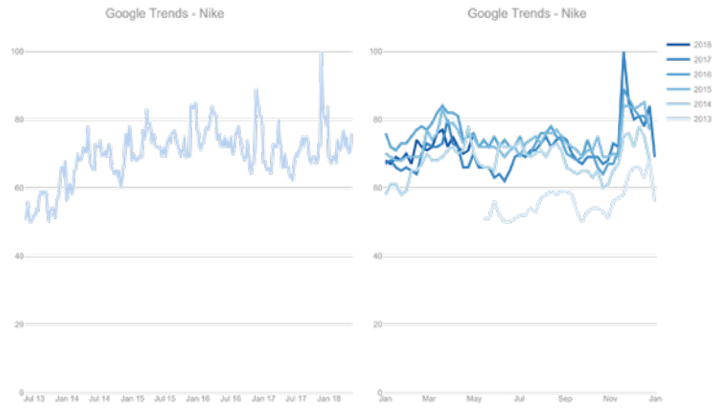
Force contrasts

Contrasts are the key differences relevant to the viewer. Visualizations are improved by making these contrasts obvious.

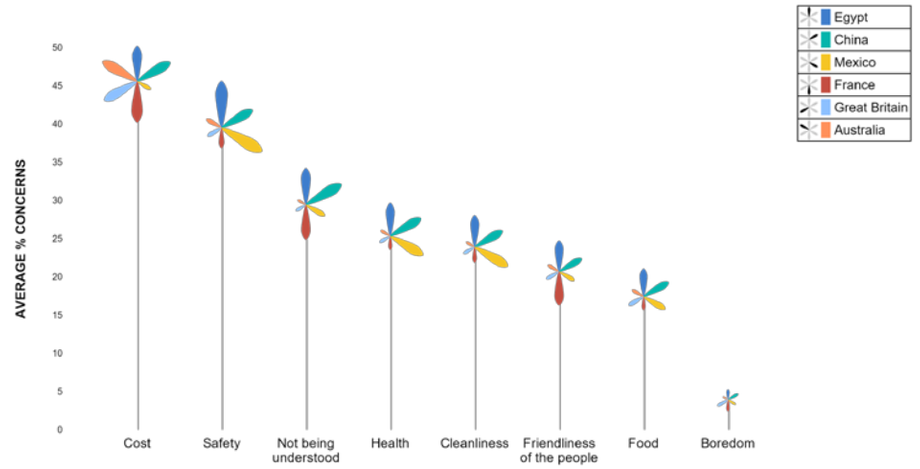
20

The *line chart* on the right shows the same data as the one on the left, but each year's data has been shown as a separate line, making the comparisons between different years easier. The visualization on the right makes two contrasts much clearer to the viewer:

- The main growth occurred from 2013 to 2014.
- There is strong seasonality, with Nike interest peaking in March/April, August, and at Christmas.

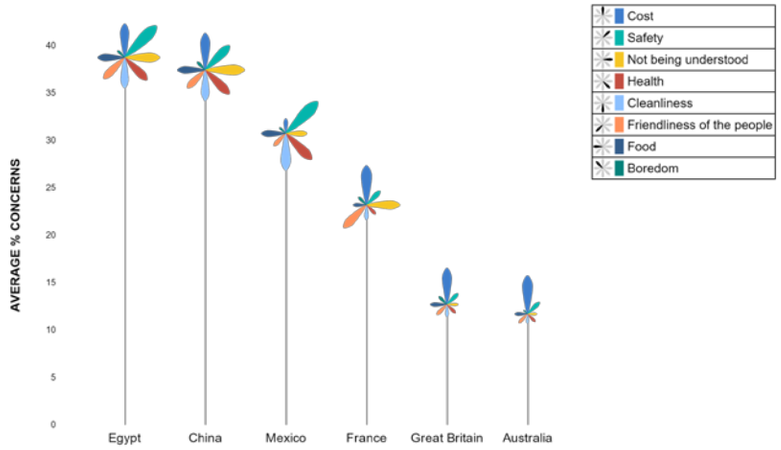


The visualization below is a *palm trees* visualization, showing the concerns that Americans have about different countries when travelling abroad. The height of each palm tree shows the average concern across the countries, with cost being the biggest concern, followed by safety. The length of the fronds shows the extent of association for each country. For example, Cost is a concern regarding all countries except Mexico, and Safety is a concern in Mexico, Egypt and, to a lesser extent, China.

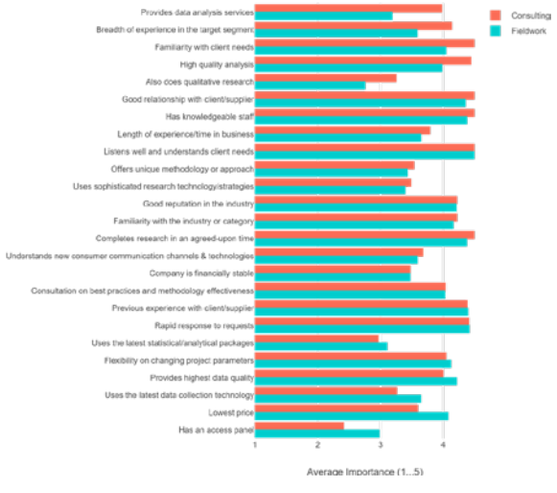


Although these *palm trees* are an interesting visualization, we must work reasonably hard to extract insight from it, because it shows poor contrasts. Below the visualization has been created after first transposing the data, making the contrasts easier to see.

In the visualization below, contrasts between the countries are much more accessible. The height of each tree shows us the levels of concerns about each country: Egypt and China, followed by Mexico, are the countries where people have the greatest concerns. The shapes formed by the fronds allow us to contrast the countries. Concerns about Australia and Great Britain are almost identical, and largely relate to Cost, whereas France is constrained both by Cost and concerns about Friendliness and Not being understood.

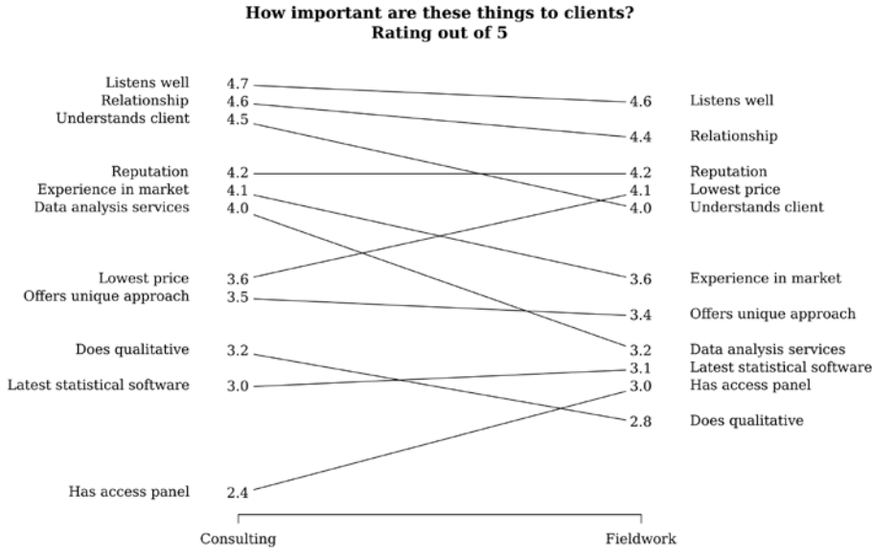


The *clustered bar chart* below is an excellent example of how not to show contrasts. The key comparison — between consulting and fieldwork — is shown by the difference in length of the red and green bars. That is, the key thing that the viewer should look at is communicated by the absence of a visual element: instead, attention is focused on a wall of bright, uninteresting color.

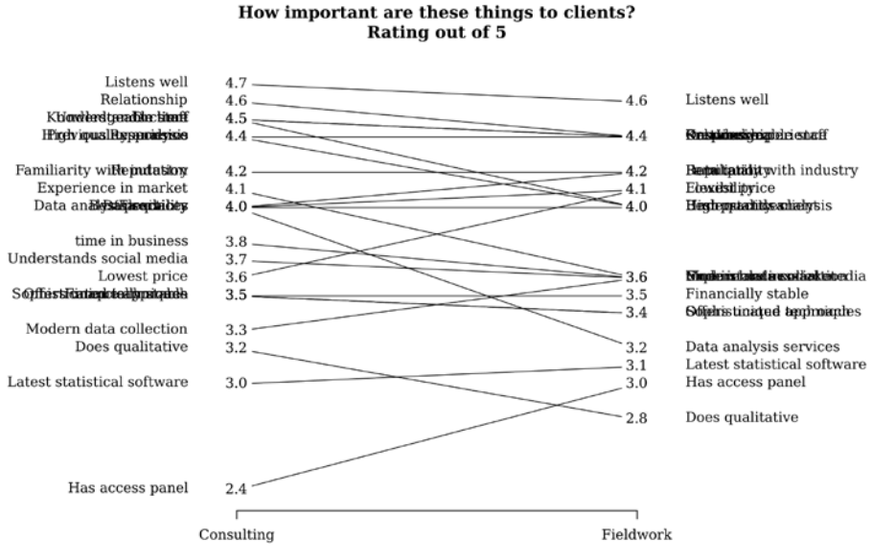


By contrast, the *dumbbell plot* below highlights and draws focus to the differences, which are also shown in a way that accommodates color-blindness.

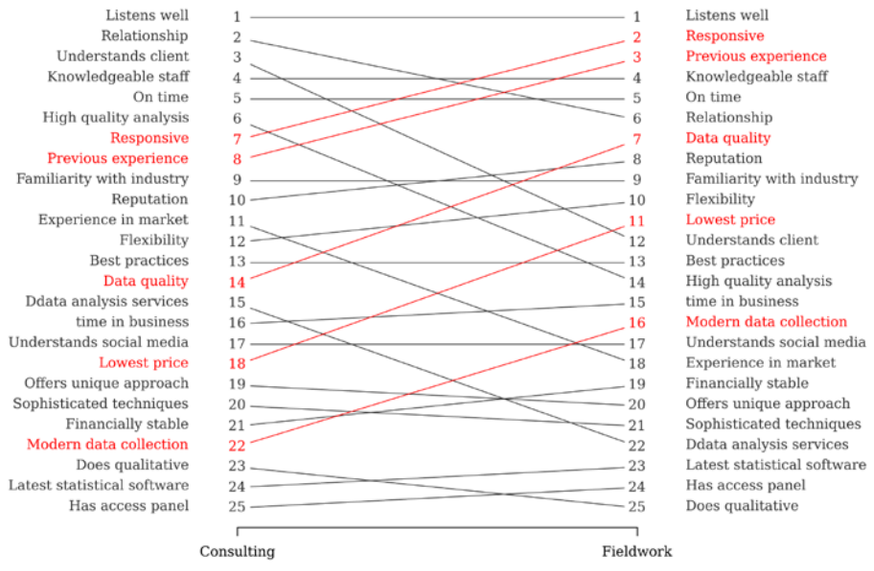
Another way to emphasize contrasts is to use a *slope chart*, where steepness of slope is proportional to size of difference.



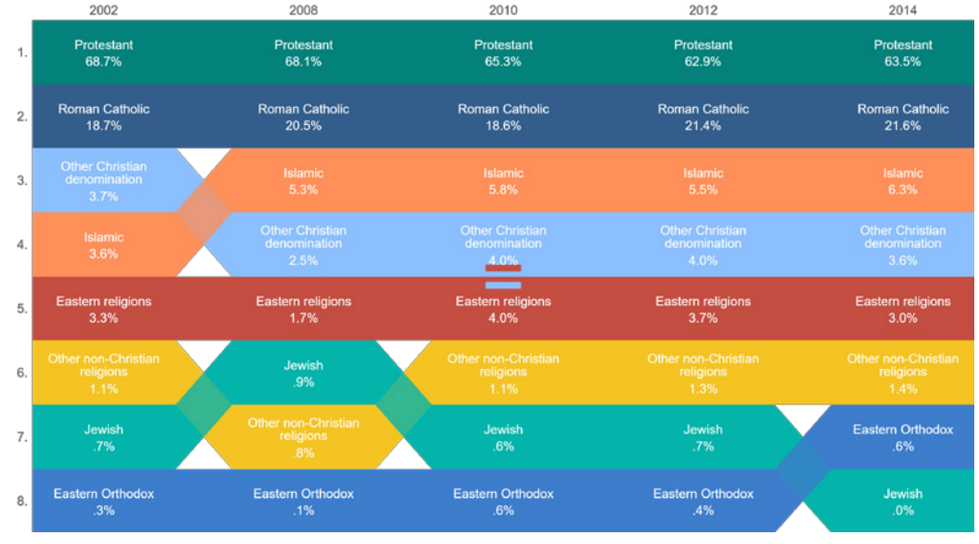
However, a practical problem with *slope graphs* is that often they suffer from severe overplotting problems...



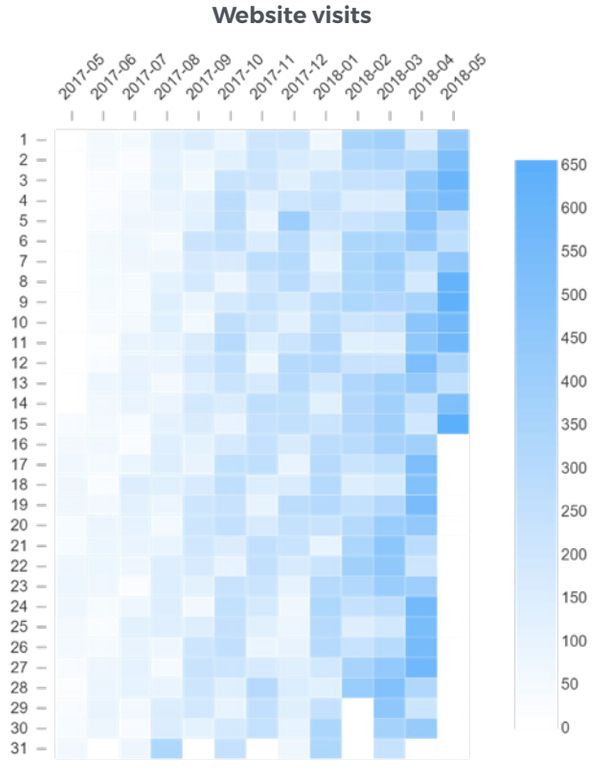
A solution to the overplotting is to plot the ranks rather than the original values. The resulting visualization is known as a *bump chart*.



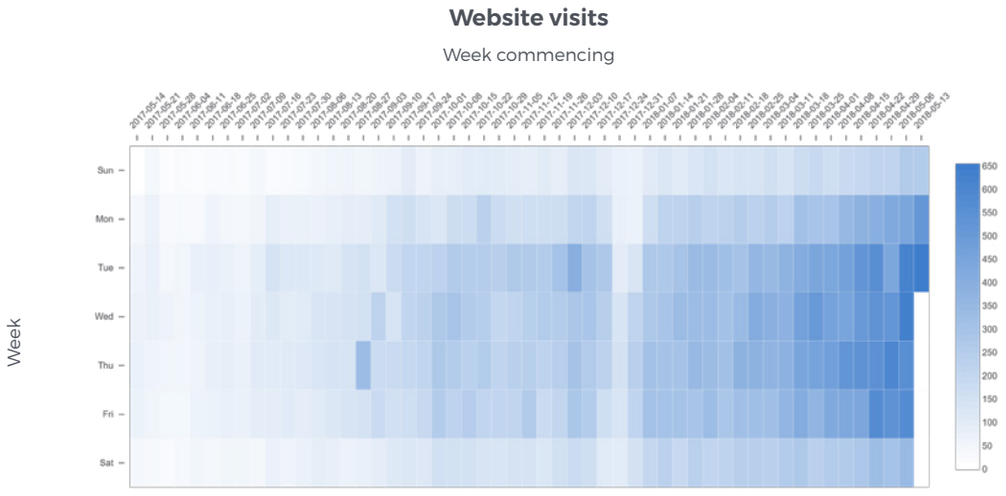
These can be made more exciting, and original numbers can be shown as well as rank.



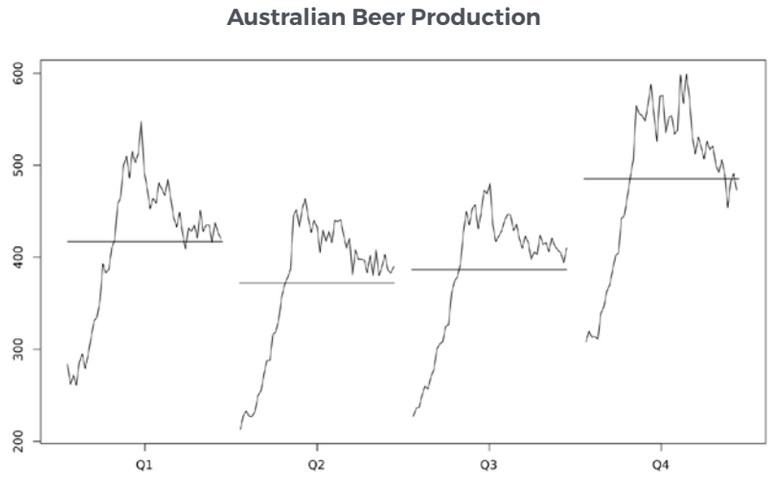
The *heatmap* below shows a year's worth of website visitation data. It reveals a growth in website traffic, with the overall level of blue linked to each month increasing from left to right.



The same website data is shown in the *heatmap* here, but the data has been reorganized to focus on days of the week rather than days of the month. This visualization makes obvious something that is hidden in the monthly visualization: the strong day-of-week effects, with low visitations on Saturday and especially Sunday.



An alternative decomposition of the beer production data in the previous chapter is provided by the *cycle plot* below. This plot emphasizes the level of sales by quarter, showing us that the peak is in Q4 (which seems a bit odd, as Q1 is the hottest quarter in Australia). This pattern cannot readily be discerned from the seasonal decomposition. Whether or not this is better than the earlier seasonal decomposition depends on the question of interest: The process of choosing a contrast is determined by the story to be communicated.

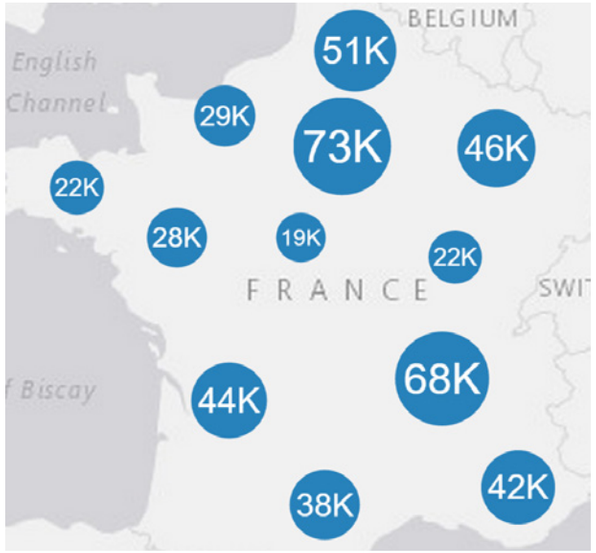


Order by context

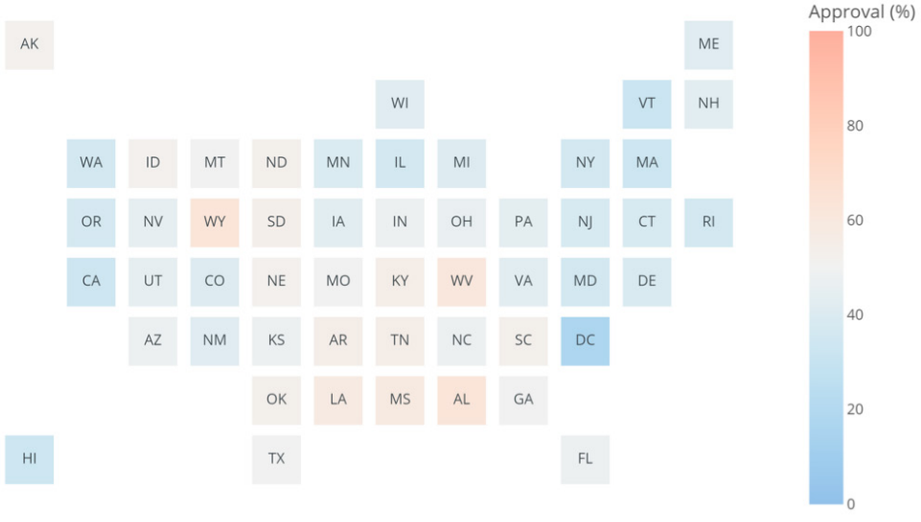
Rearrange the values or series of data, using additional variables that explain or contextualize the results.

21

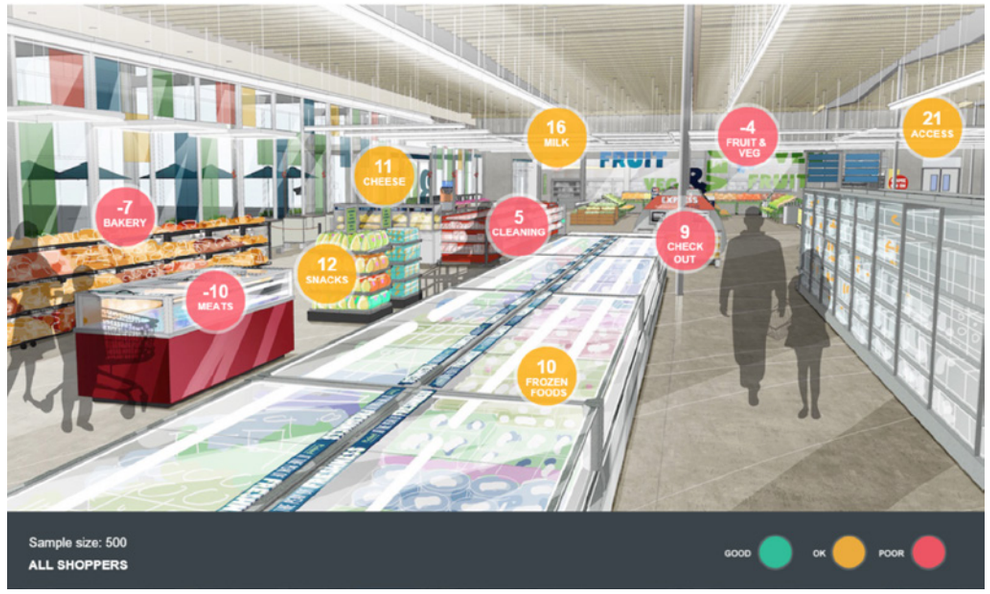
Where data is geographic in nature, context can be added by plotting the values directly onto a map, as shown below.



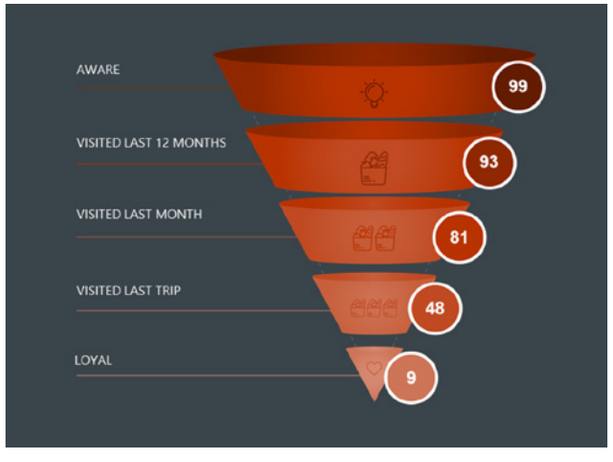
Chapter 12, Show norms, illustrated the use of packed circles to show President Trump's approval by state. In the visualization below, the circles have been replaced by squares, ordered to align roughly with the geographical positions of the states. This is known as a *state bin*, and it solves a problem of accuracy of representation in *choropleths* - that the emphasis given to a state is determined by that state's geographical size.



In the visualization below, the net promoter score for different departments of a supermarket are communicated using a traffic-light system and overlaid on an image representing a supermarket. In some instances, the associations are literal – such as with frozen foods – while in others the correspondence between the data and the visual elements seems tenuous at best. The point to providing this kind of pseudo-context is that it offers the viewer a framework to coordinate the data. Further, associating data with a visual map is another useful strategy for aiding memory.[31]

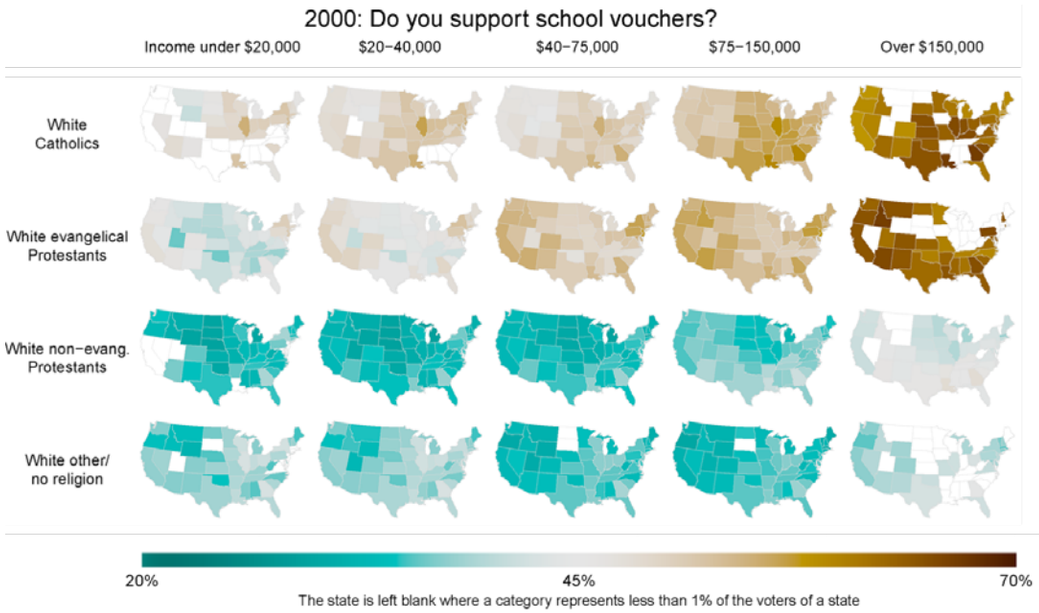


Another type of context is to order the data chronologically or by process, such as with *funnels* like the one below.



[31] For example, a standard strategy used in memory competitions is to associate specific items with rooms in a "memory palace" or some other geographic area. Joshua Foer (2012), *Moonwalking with Einstein: The Art and Science of Remembering Everything*, Penguin Books.

Ordering by context can be effectively combined with small multiples, as was shown in Chapter 19, Decompose, and also in the *coplot (condition plot)* below.[32]



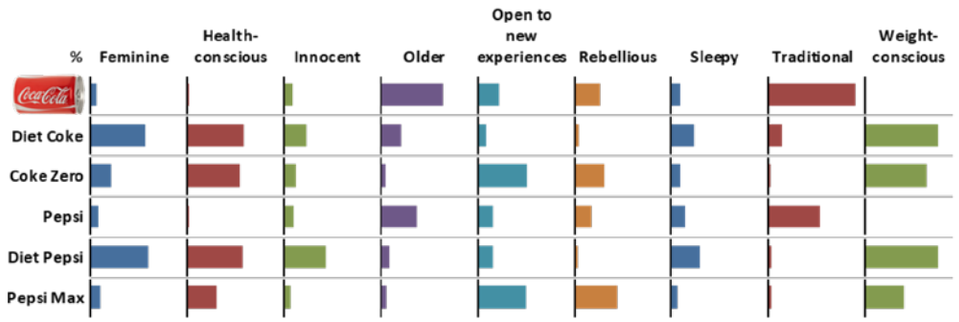
[32] From http://andrewgelman.com/2011/04/04/irritating_pseu/; some rows removed.

Diagonalize

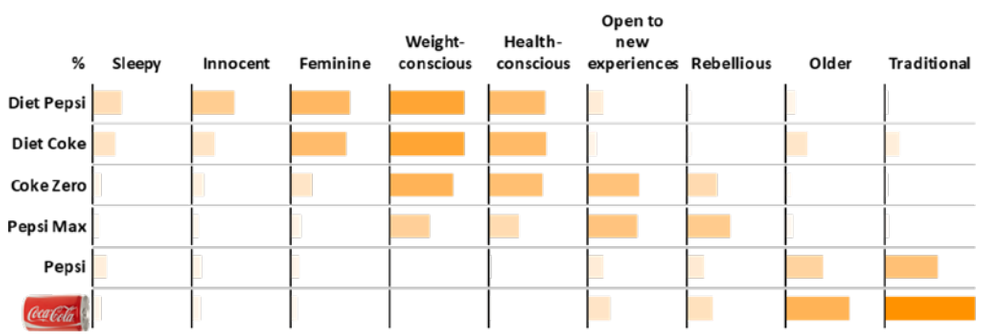
Reordering the data so that key patterns appear as diagonal lines makes the visualization more accessible.

22

Consider the visualization below. How does Coca-Cola compare to the other brands?



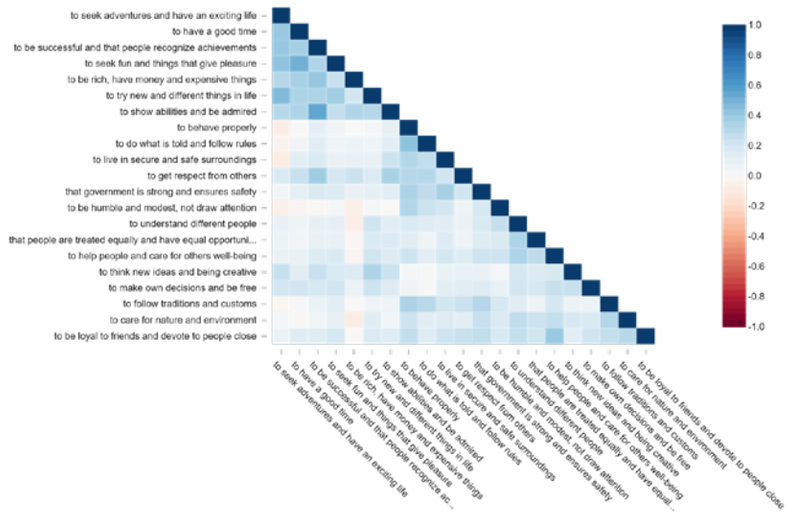
The visualization below shows the same data, but with color reduced, redundant encoding, and *diagonalized*, it is dramatically easier to see how Coca-Cola differs.



Diagonalization,[33] like banking to 45°, is powerful and easy to do: rearrange the rows and columns so that something approximating a diagonal appears.

[33] Jacques Bertin's (1967), Sémiologie Graphique. Les diagrammes, les réseaux, les cartes, Translation 1983. Semiology of Graphics by William J. Berg

The technique can also be applied with tables and heatmaps, although in huge tables it can be useful to use algorithms to assist in the process.[34] The heatmap below is for a correlation matrix. The pattern that appears is typically one of "steps", where each step can be interpreted as a factor from factor analysis.



The *loadings* from a factor analysis (principal components analysis) are shown below. Here, diagonalization has been combined with the use of bars (redundant encoding) to clarify both the factor structure and the ambiguity of the structure for the fourth component.

Principal Component Loadings: Important

4 components explaining 47.8% of the variance

	Component 1 (16.4%) Eigenvalue: 3.44	Component 2 (13.3%) Eigenvalue: 2.79	Component 3 (11.2%) Eigenvalue: 2.36	Component 4 (6.94%) Eigenvalue: 1.46
to seek adventures and have an exciting life	0.730			
to have a good time	0.717			
to be successful and that people recognize achievements	0.676			
to seek fun and things that give pleasure	0.675			
to be rich, have money and expensive things	0.598			
to try new and different things in life	0.593			
to show abilities and be admired	0.573			
to behave properly		0.718		
to do what is told and follow rules		0.692		
to live in secure and safe surroundings		0.607		
to get respect from others		0.551		
that government is strong and ensures safety		0.541		
to be humble and modest, not draw attention				
to understand different people			0.609	
that people are treated equally and have equal opportunities			0.688	
to help people and care for others well-being			0.568	
to think new ideas and being creative			0.468	
to make own decisions and be free			0.428	
to follow traditions and customs		0.492		0.527
to care for nature and environment				0.503
to be loyal to friends and devote to people close				0.441

Rotated loadings: Input: Correlation matrix; Missing data setting: Use partial data (pairwise correlations); Sample size: 14540 to 15270; Rotation: Varimax; Unrotated eigenvalues: (1) 4.51, (2) 2.74, (3) 1.73, (4) 1.07, (5) 0.97, (6) 0.92, (7) 0.87, (8) 0.80, (9) 0.74, (10) 0.70, (11) 0.66, (12) 0.64, (13) 0.61, (14) 0.61, (15) 0.58, (16) 0.55, (17) 0.52, (18) 0.51, (19) 0.49, (20) 0.43, (21) 0.40

[34] Ordering by the first eigenvalue from correspondence analysis or factor analysis often provides a satisfactory solution.

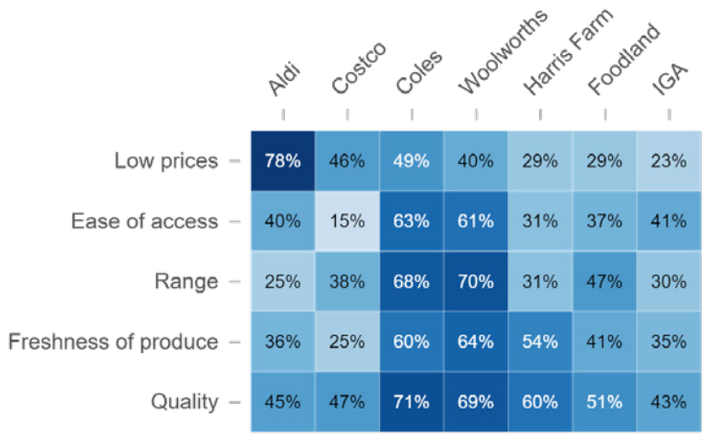
Simplify the data

Visualizations can be simplified — and usually improved — by reducing the quantity of information displayed through aggregating, smoothing, or filtering.

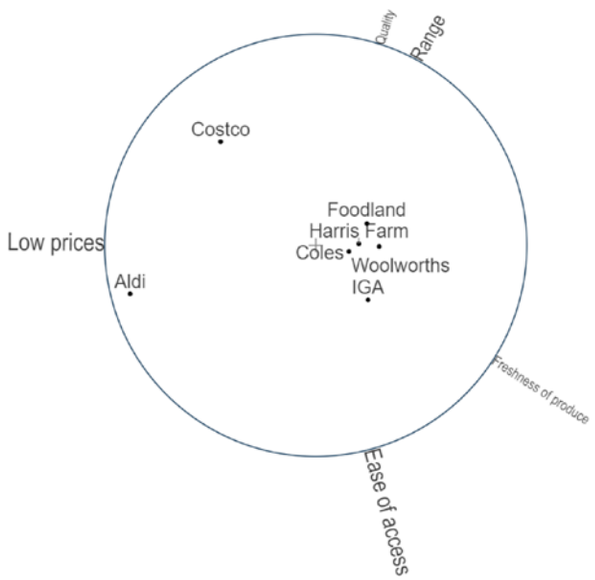
23

The visualization above shows the results of 25 years of opinion polls about the Australian prime minister. The simpler visualization below has been de-cluttered and banked to 45°, and the data has been *smoothed*.

Although the *heatmap* below is simple to interpret, we must work to extract meaning from it. We need first to recognize the pattern and then try to deduce what the pattern means by looking at the numbers and the associated row and column labels.



By contrast, the *moon plot* below is simpler to interpret because it shows much less information.



Supernormalize

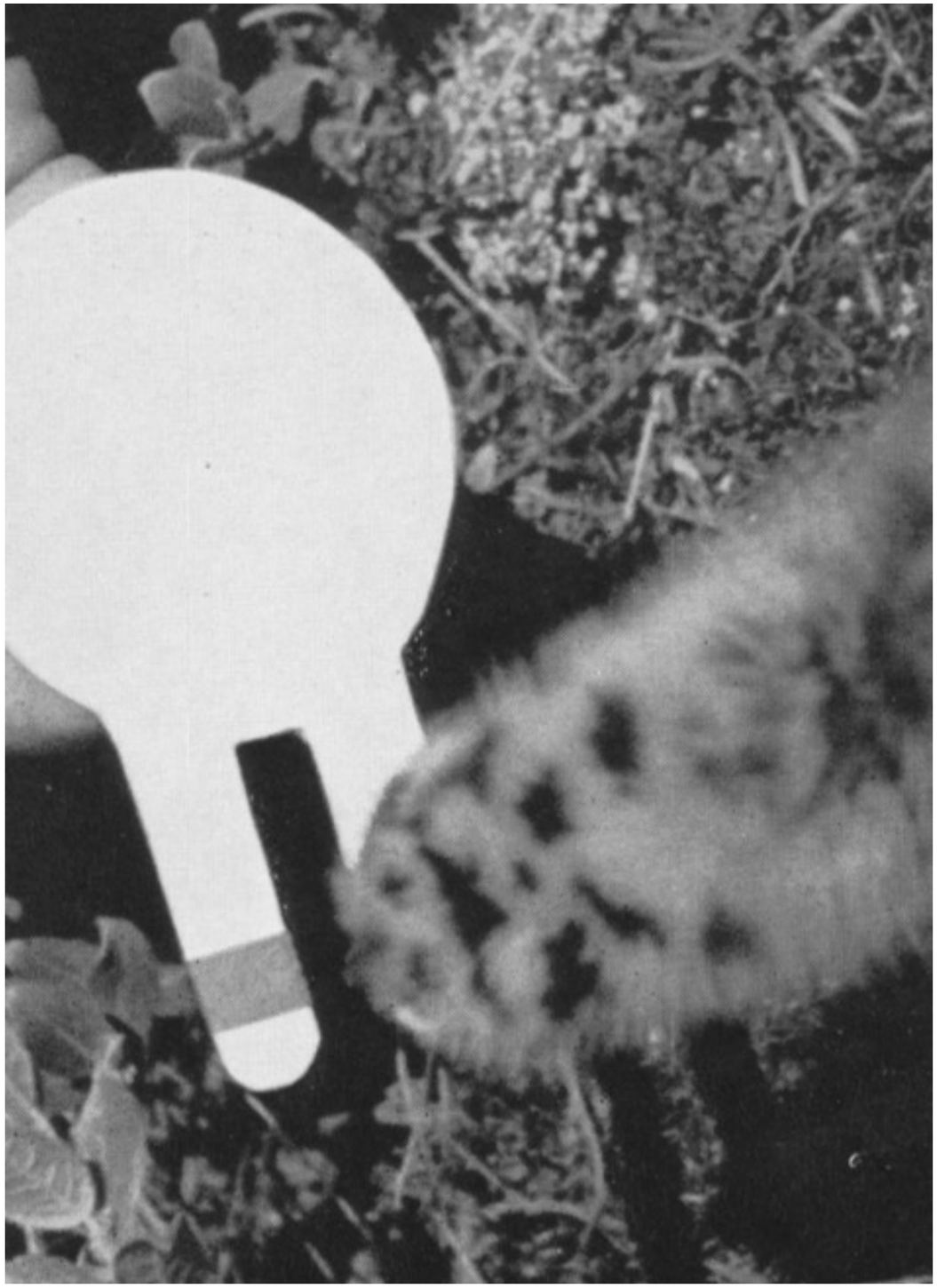
Visualizations work best when they resemble familiar shapes and patterns. The more we can exaggerate the visualizations to match those shapes, the better.

Supernormalization is:

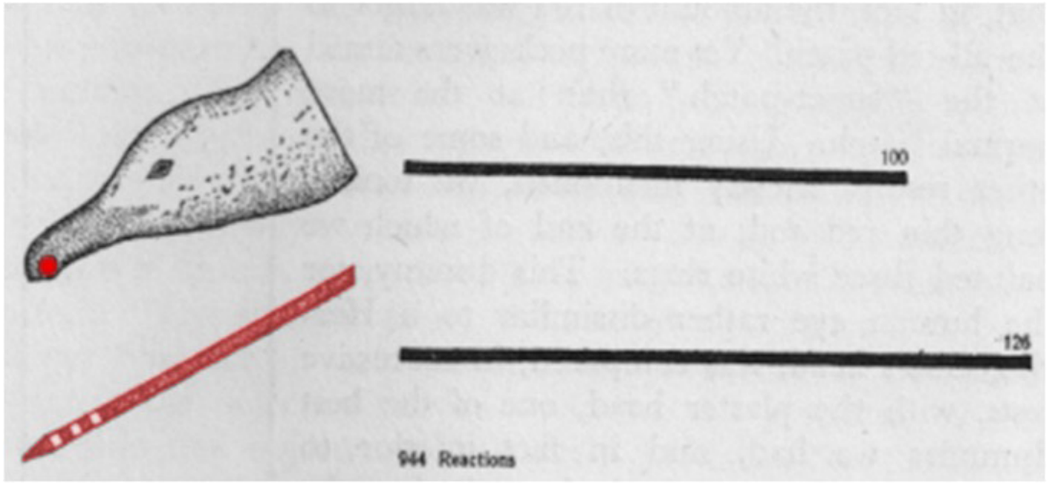
1. A meta-technique, subsuming all the other techniques.
2. An invented word.
3. Newish – to be treated as a useful framework rather than a rock-solid theory.
4. Defined as: creating visualizations that use shape and color in ways that tap into our instinctive skills at interpreting visual stimuli.
5. The last of the techniques presented in this book.

The Dutch Nobel laureate Niko Tinbergen noticed that when chicks hatch, they seemed to instinctively recognize their mother. He conducted a seemingly heartless experiment to understand this phenomenon. He took the mothers away and showed the chicks a cardboard cutout with a gull's face painted on. The chicks still pecked at it.

Tinbergen then showed the chicks a two-beaked monster. They pecked at that, too.



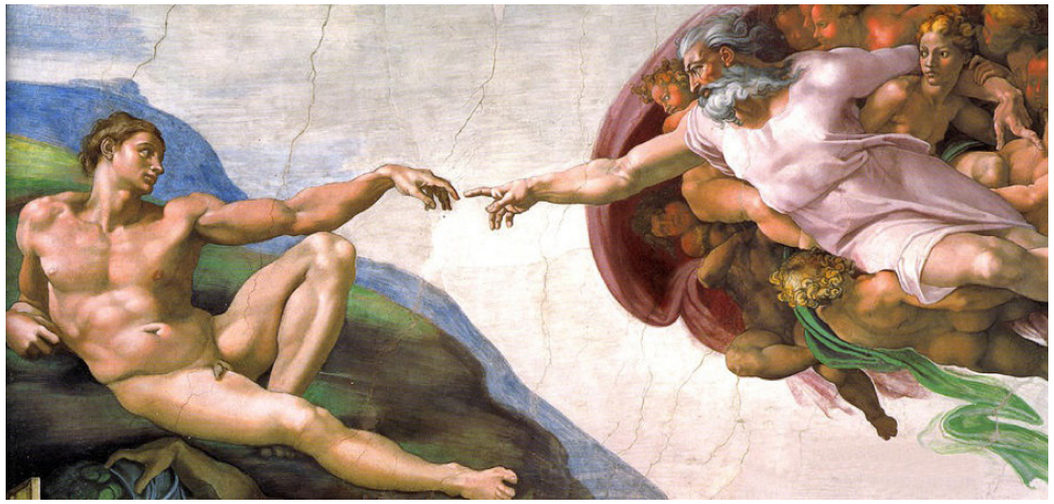
Tinbergen's experiment tells us both that birds are born with an instinctive understanding of "mother" (or perhaps "food"), and that this understanding is somewhat fuzzy. His excellent visualization of the results (which both reduces eye movement and attracts attention) illustrates a key finding: a red rod with three white marks was pecked at more than a plaster cast of a herring gull's head. Why? While evolution gifts chicks an instinctive understanding of their mother, the rod is both adequate and less detailed, and so easier to learn.



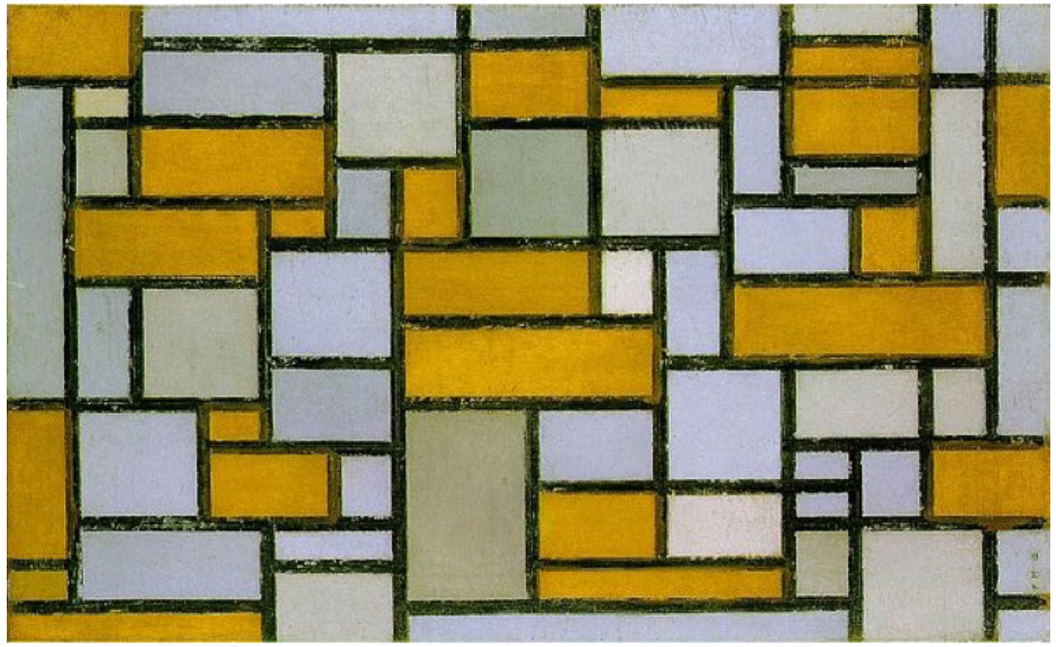
This basic idea should not come as a surprise. Toys and cartoons are usually *supernormal* in much the same way.

When viewing anything, if we can quickly get the gist of what we are looking at, we can also then quickly work out how best to explore and understand what we are looking at.[35] Any experienced driver can sit in a new car and usually work out very quickly how to drive it, because everything follows a logic that is clear to them. They know where to look. By contrast, if you put a computer-illiterate person in front of a computer, they have no framework with which even to deduce where to look.

When we see Michelangelo's *The Creation of Adam* we "get" it. An art expert might provide a much richer explanation, but we can instinctively interpret much of its design.

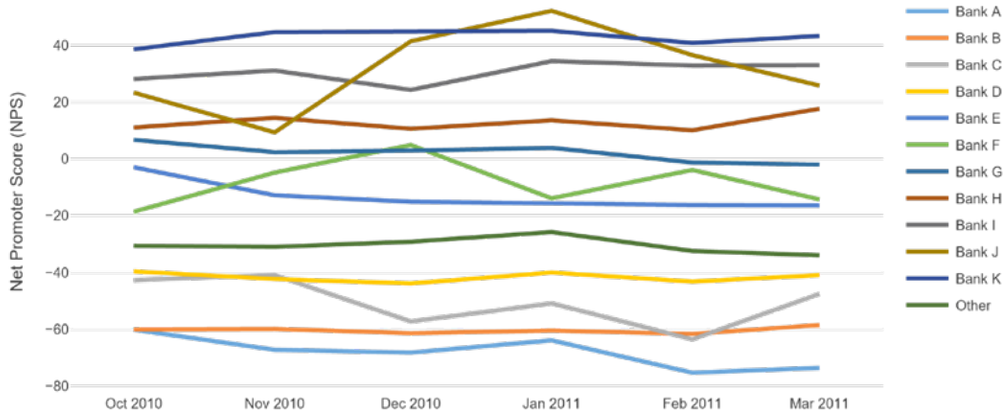


By contrast, our instincts tell us little about how to interpret Mondrian's *Composition with Gray and Light Brown*. Our instincts tell us so little that we cannot even work out which way is up. As discussed at the beginning of this book, a great visualization is one that can be quickly understood. Without casting aspersions on Mondrian, to be successful in data visualization we need to be more like Michelangelo.



A good visualization is one that makes it easy for the viewer to extract meaning. This requires that it use shapes and colors in familiar ways. At a crude level we can achieve this by using standard visualizations, such as the *line*, *bar*, and *pie charts* people have been trained to view. However, more generally, we should create visualizations that contain patterns and shapes, and use colors in ways that tap into the viewer's instincts. If viewers can work out the gist of what they are looking at, they can much more quickly extract meaning. Techniques like diagonalization, redundant coding, ordering by context, and small multiples all work by leading to the creation of visualizations that are easier to interpret, because the shapes are recognizable and comprehensible.

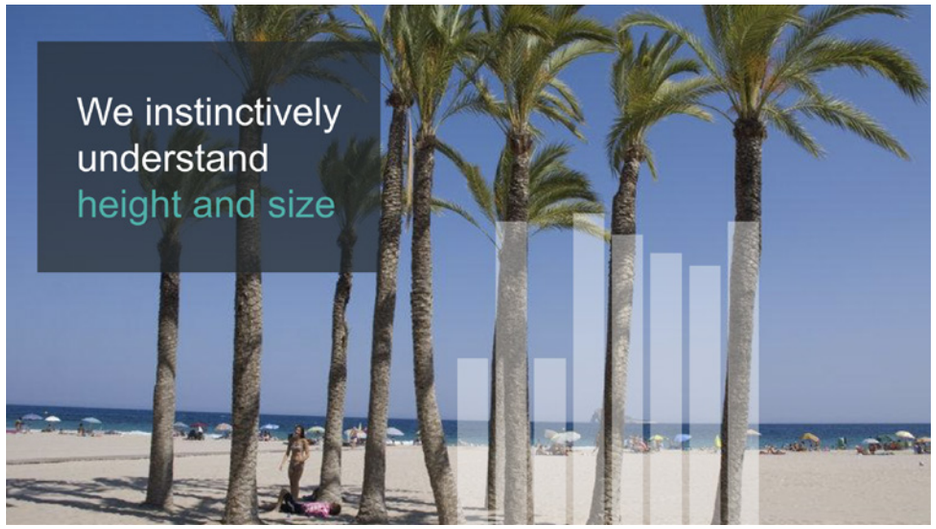
[33] Colin Ware (2012), *Information Visualization: Perception for Design*, 3rd Edition, Morgan Kaufmann, Kindle Edition.



The line chart from Chapter 17, Small multiples, is reproduced above. What makes it such a poor visualization? A basic problem is that it does not look like anything familiar. At best it looks like some colorful broken spaghetti. We humans didn't evolve looking for patterns in spaghetti, so it is no surprise that we find the visualization unsatisfactory.

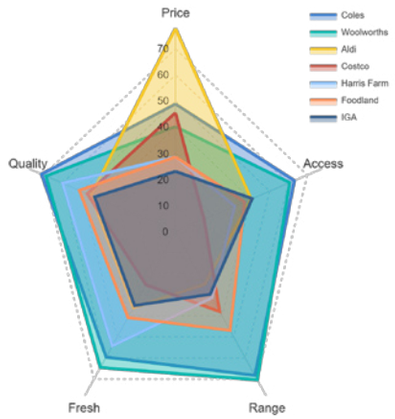
By contrast, the shapes of the small multiples are more recognizable.

We instinctively understand the concept of solid objects having height, which makes the small multiple of are charts easily understandable.

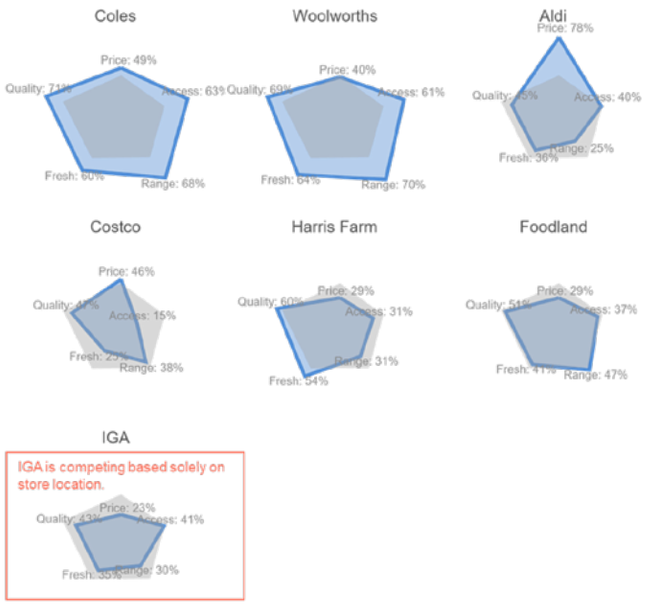


The profile of the tops and bottoms of each of the bars are familiar shapes that we know how to interpret. The lines that we see on tops and bottoms of these small multiples, which are banked to about 45°, reflect the hills and mountains of the natural world. Our brains have evolved to understand such shapes, as it was necessary for our basic survival: Can we run up that hill when escaping the lion? Or is it too steep to climb?

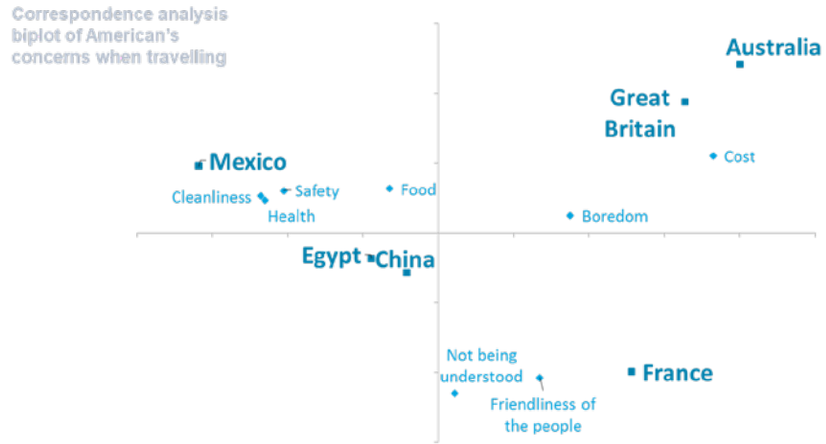
The traditional *radar chart* shown below looks like a witch's hat. Nothing in our evolution has primed our brains for looking at patterns in witches' hats.



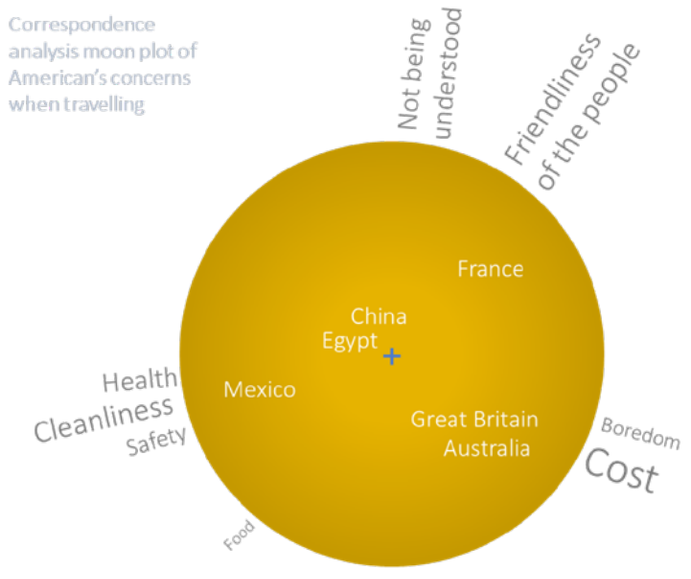
We have, however, evolved a good understanding of how to distinguish between shapes of objects. For example, without the ability to tell apart different shapes of leaves, we would have great difficultly in avoiding poisoning ourselves. So, when we construct small multiples of radar charts, we can be confident that the viewer will be able to interpret them by tapping into their instinctive ability to tell shapes apart.



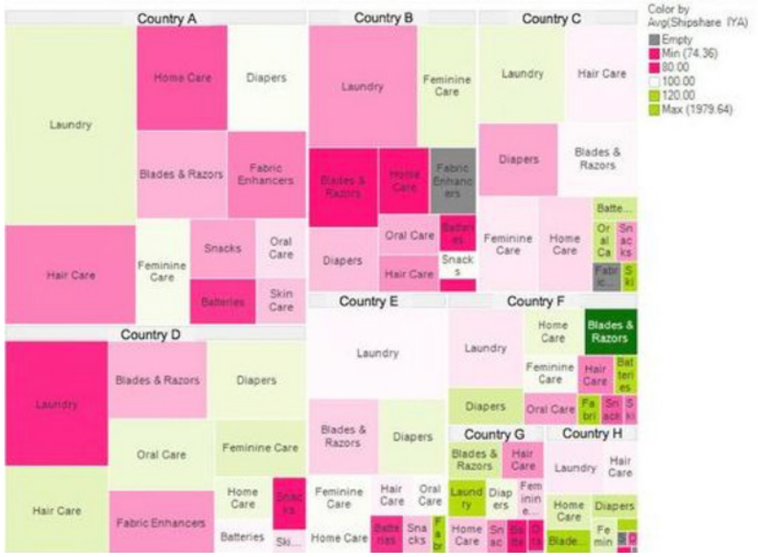
Evolution has gifted us excellent spatial awareness. It is for this reason that the standard *correspondence analysis biplots*, such as the one shown below, cause so much difficulty. Our instincts tell us that things placed close together are associated in some way. It is obvious to somebody new to correspondence analysis that the plot below implies that Food and China are associated.



It is precisely for this reason that the *moon plot* is superior to the *biplot*. The viewer can safely rely on proximity without needing to understand linear algebra and the concept of linear projection.



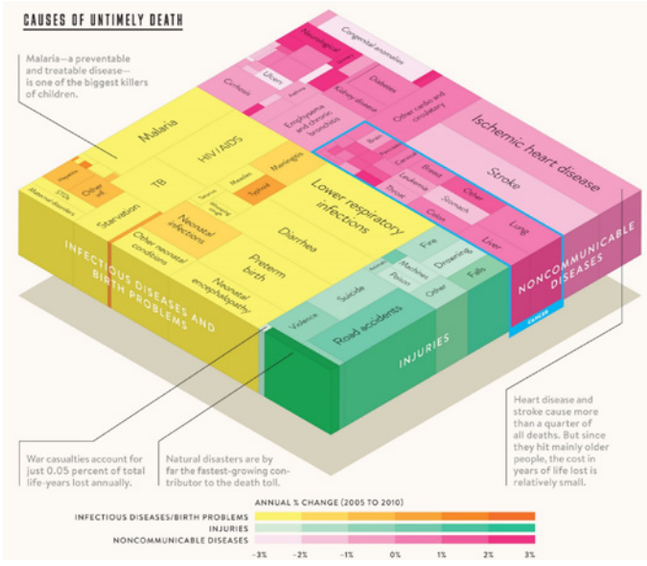
The use of color and shape in the visualization below is unnatural. The color scale of gray to bright pink to white to green does not exist in nature. With the right cues, we can interpret such a color scale with some ease (for example, see the *heated density plot* in Chapter 12), but the resulting visualization here is more Mondrian than Michelangelo. It takes effort to extract any meaning.



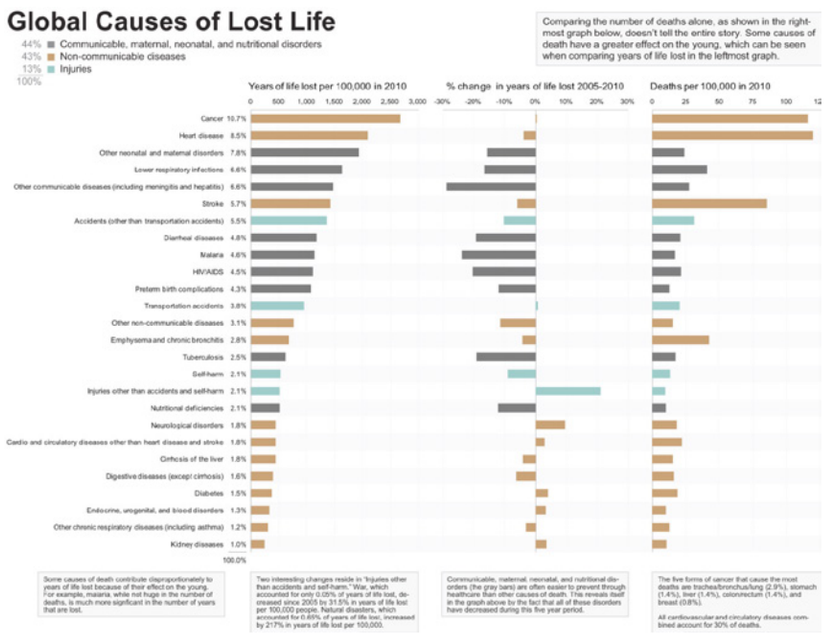
The basic conception of this visualization is not the problem. The colors are natural colors, which we can instinctively understand as they relate to important distinctions in nature such as those between flowers, leaves, and bark.

And the basic idea of having different tiles organized in such a visualization (*treemaps*) is also often very effective. We have evolved to be excellent at understanding proportions and size. As children, nobody needs to explain to us when half a cake has been eaten, or that a cupcake is smaller than a cake. We get it instinctively. We instinctively understand the concept of proportionality, which is why we can interpret pie charts.

The visualization below uses essentially the same design and coloring as the earlier one, but to much greater effect. Color is used to disambiguate large regions, much as occurs in nature. The degree of color is also much more consistent with how we perceive such intensity in nature (e.g., to signify the depth of water and the amount of rain in a cloud).

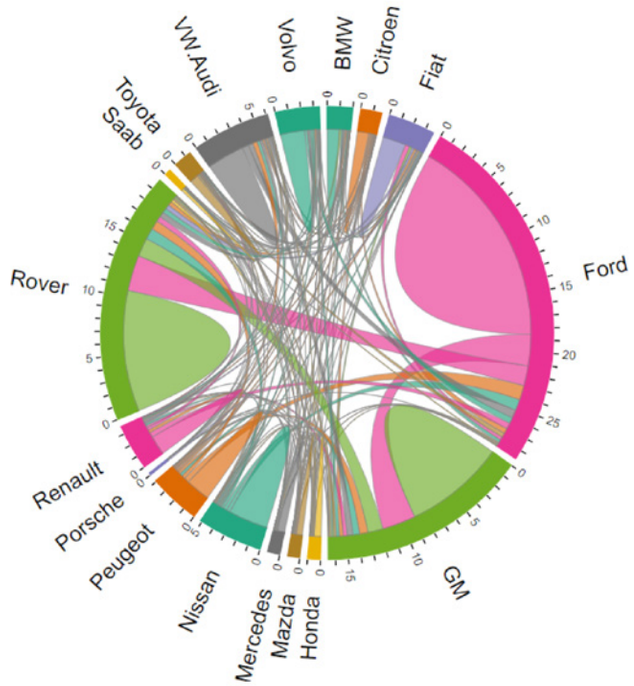


Bill Gates loves this visualization (below) because "it shows that while the number of people dying from communicable diseases is still far too high, those numbers continue to come down." Visualization writer Stephen Few hates it: "This is an important message and a noble goal. But how well does the graph above tell this story? Not very well, actually."[36] He hates it so much he created his own version (below), the superiority of which he explains as follows: "By using bar graphs, we've made it easier to interpret and compare the data, so that it's easy to focus on the stories contained in the data, rather than struggling to decode an inappropriate and ineffectively designed display."

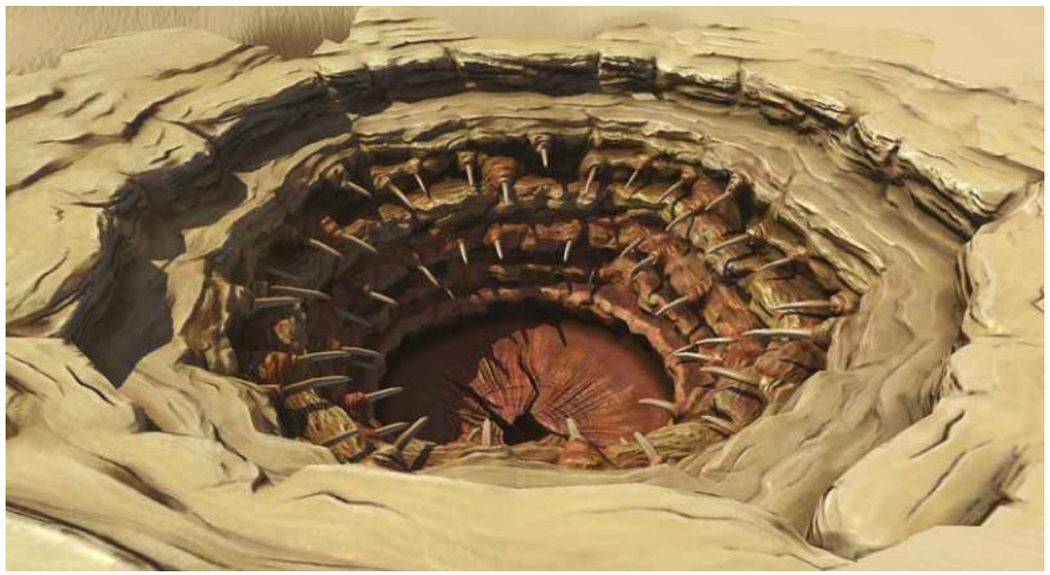


The original and the revised visualization are trying to achieve very different things. If the viewer has a commitment to understanding the detail of causes of death, then the second visualization is the better one, because it represents the data more accurately. However, if the goal is to educate people, the original visualization is much better. By using color and shape as they appear in nature, the visualization engages the viewer and makes them much more likely to take the time to extract information. Don't take anyone's word for this: Spend ten seconds looking at each and see what you learn. There's a good chance you just get bored by the second one and move on.

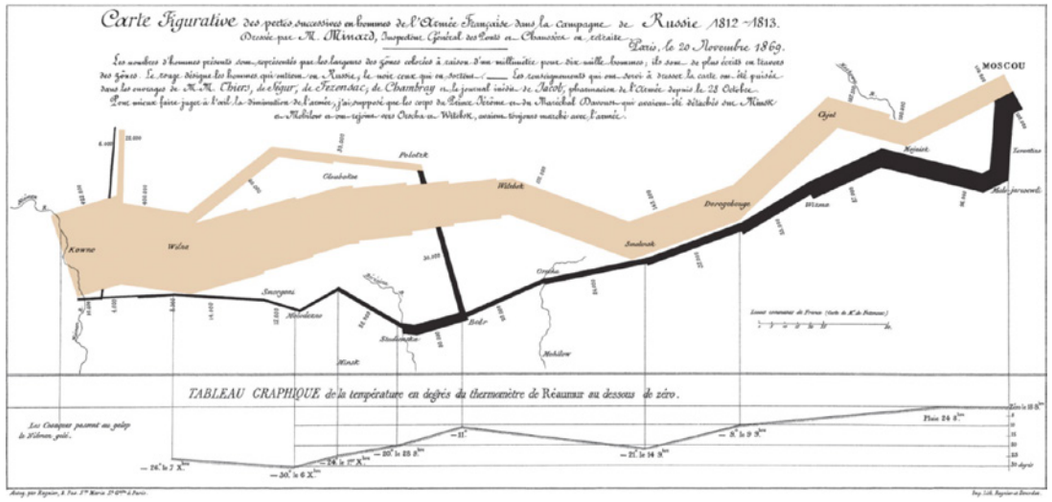
[36] https://www.perceptualedge.com/example20.php



Chord diagrams, such as the one above showing switching between different car brands, are fabulous to look at, but what do they mean? We can instinctively work out that they show proportionality, much like a pie chart, and that Ford, Rover, and GM, are "big". However, working out the meaning of all the "chords" of the chord diagram is harder. It looks like the Great Pit of Carkoon from *Return of the Jedi*. Fascinating yes, but not somewhere we expect to find a pattern.

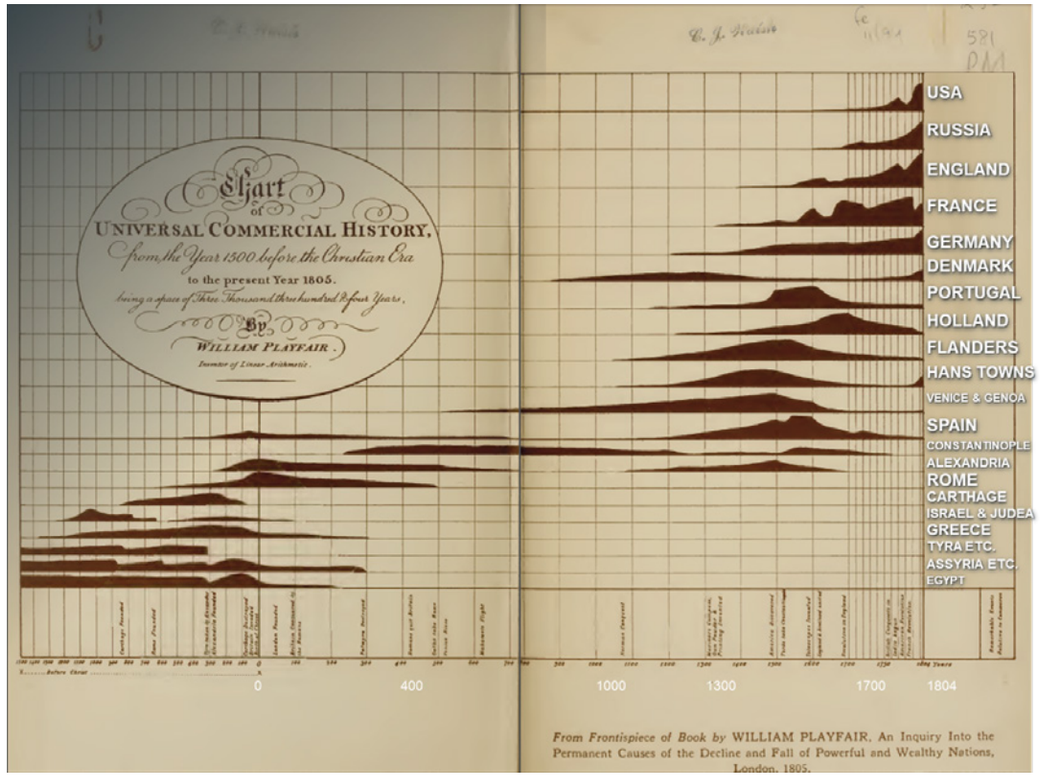


Charles Joseph Minard's 1869 chart (now known as a *Sankey diagram*) is sometimes described as the greatest visualization of all time. However, it needs to be explained to be understood. We do not instinctively get it. What do you see? A tree branch? Not an object from which insight is commonly extracted.



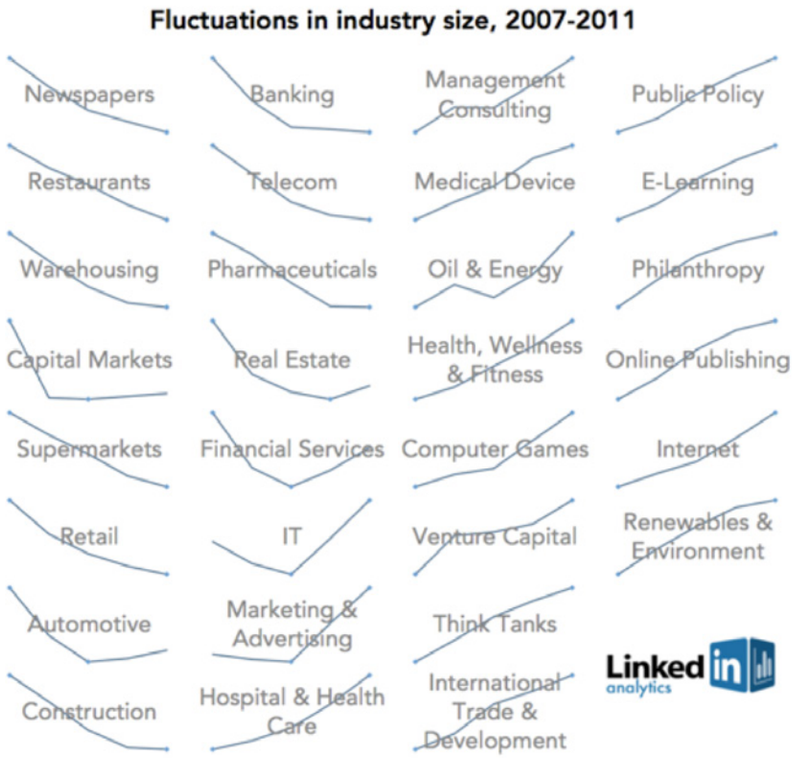
Minard's visualization shows Napoleon's ill-fated invasion of Russia in 1812. Once that has been explained it is appreciated more readily. The width of the line is proportional to the size of the Grande Armée and has been superimposed over its route, affording us a compelling and precise image of the army's decimation through attrition — over time and space — rather than actual war.

William Playfair's *Universal Commercial* has a yet more ambitious goal: to summarize 3,500 years of the world economy. It breaks many rules governing the accurate representation of data but nevertheless succeeds in communicating a vast amount of information. Each country's history appears as a mountain with smoothed silhouette/profile, allowing us to discern the pattern easily. For example, we can see that the USA's economy halved during the Revolutionary War, but by 1804 it was stronger than ever before. The diagonalization facilitates an easy comparison across the world, making clear just how dark the Dark Ages really were, and also directing our attention to many largely forgotten early European economies, such as the Hanseatic League and Flanders.



From Frontispiece of Book by WILLIAM PLAYFAIR, An Inquiry Into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations, London, 1805.

Creating visualizations like those of Minard and Playfair is rarely practical in real-world market research, but the same principles of story-telling apply. An ugly but magnificent visualization is the one below from LinkedIn, detailing industry performance during the Great Recession.[37] The small multiples have been reordered to form a wave pattern, something our brains recognize instantly and can use to search for information. It is easy to work out from this visualization which industries declined and then grew because we understand the pattern instinctively, allowing us to look in the right places.



Great visualizations are ones which tap into our instincts. The viewer should not have to work to find patterns. The patterns should jump out at us and make it easy for us to draw conclusions. Each of the preceding chapters has illustrated various tools for improving visualizations, but the ultimate technique is the one described in this chapter, of creating visualizations that are in line with the types of patterns we have evolved to see in nature.

[37] https://blog.linkedin.com/2012/03/08/economic-report

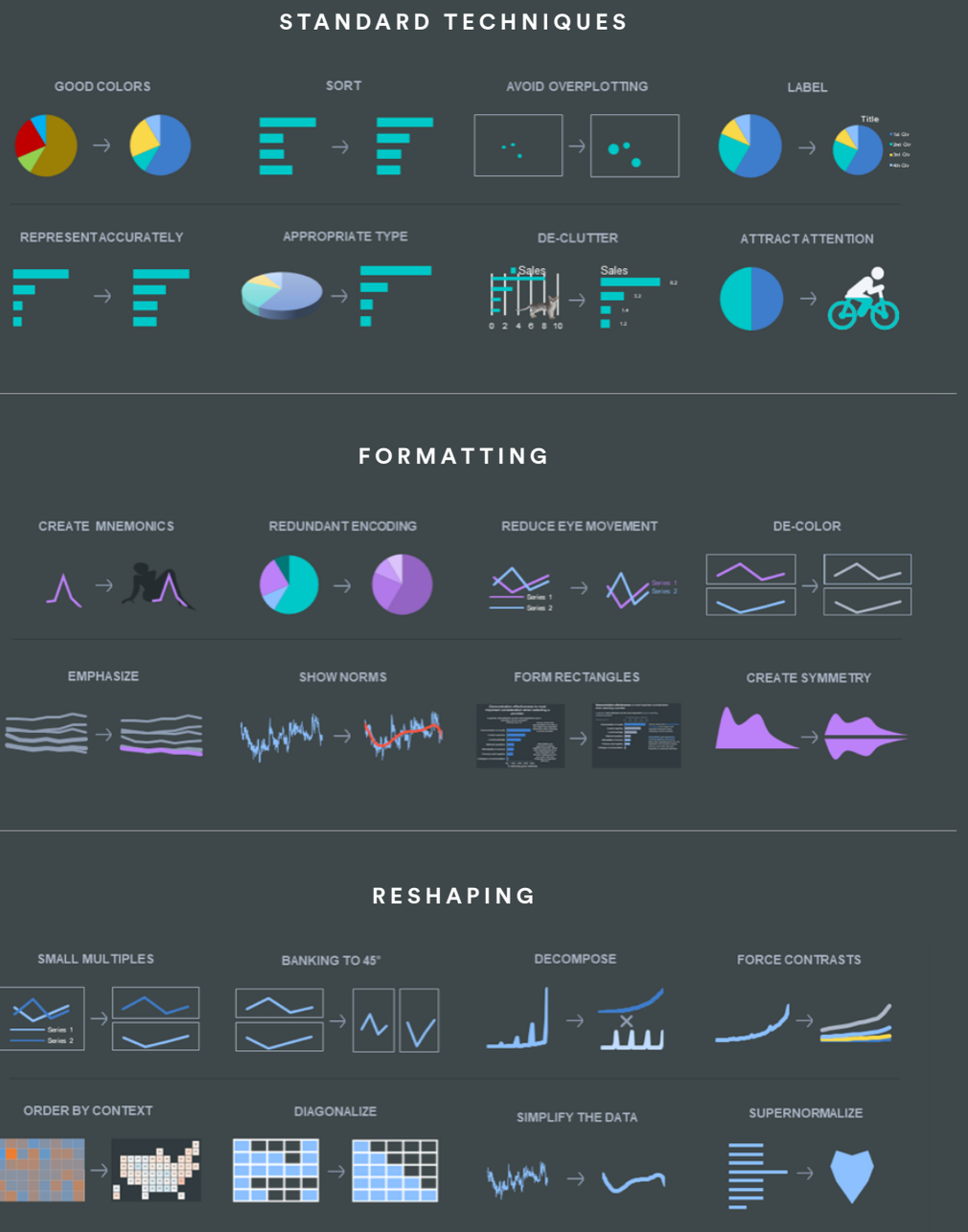
Software

All of the computer-generated visualizations in this book can be created using either:

- Displayr (www.displayr.com), or
- Q (www.q-researchsoftware.com) in conjunction with PowerPoint

Summary

The goal when creating visualizations is to allow people quickly to discover and remember the key stories in data. We do this by creating supernormal shapes, using the techniques below.



About the author

Tim Bock is the founder of Displayr *www.displayr.com*. Tim is a data scientist, who has consulted, published academic papers, and won awards, for problems/techniques as diverse as neural networks, mixture models, data fusion, market segmentation, IPO pricing, small sample research, and data visualization. He has conducted data science projects for numerous companies, including Pfizer, Coca Cola, ACNielsen, KFC, Weight Watchers, Unilever, and Nestle. He is also the founder of Q *www.qresearchsoftware.com*, a data science product designed for survey research, which is used by all the world's seven largest market research consultancies. He studied econometrics, maths, and marketing, and has a University Medal and PhD from the University of New South Wales (Australia's leading research university), where he was an adjunct member of staff for 15 years.

DISPLAYR

RESEARCHSOFTWARE
A DIVISION OF DISPLAYR

Want to cut your analysis and reporting time in half?

See Displayr in action →

Analysis and reporting software built to save you time