Avoiding Misleading Data Visualisations

Sam Cottrell

Innovation Design Engineering

Supervisors: Prof. Stephen Boyd Davis (Royal College of Art) and Dr. Sonia Ranade (The National Archives)

One of the key benefits of data visualisation is the ability to communicate information, even large sets of data, clearly and efficiently. Humans are visual creatures and as such diagrams and graphics can convey information and engage users in ways that textual methods cannot. Therefore, it is hard to overstate the importance of transparency in data visualisation.

While it can be said that in producing a graphic that attempts to convey a specific point, the creator has inherently imparted their own bias, care must still be taken to ensure that it is the data itself that delivers the message and not any misleading characteristics of the graphics.

Visualisations have become a tool of politicians, journalists and businesses, intended to guide the viewer toward a preordained conclusion. As visualisations have become more commonplace it is important to understand the ways in which they can be used to mislead, both intentionally and unintentionally. It is foolish to think that those that benefit will cease using these tactics when shown the error of their ways, so instead, this is intended as an overview so that as viewers we can identify when these devices have been employed, and as creators we know how to avoid them.

The methods that are covered here are not new, in fact many of them were covered in the seminal books "How to Lie with Statistics" (Huff, 1954) and updated in "The Visual Display of Quantitative Information" (Tufte, 1983). I will attempt to apply these principles, update and add to them to be more relevant to the ever increasing field of interactive data visualisations.

This piece is not intended to look at the merits of a particular aesthetic, but the more general graphical tricks that are employed.

Misleading Features

Tufte provides the following principles for maintaining graphical integrity:

- The representation of numbers, as physically measured on the surface of the graphic itself should be directly proportional to the numerical quantities represented
- Clear, detailed and thorough labelling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data
- Show data variation, not design variation
- In time series displays of money, deflated and standardised units of monetary measurement are nearly always better than nominal units

- The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data
- Graphics must not quote data out of context

The following are examples of how data visualisations can be made to be misleading, and subvert the messages within the data, through graphical methods and not adhering to the guidelines above. The examples have been created using arbitrary data and unless otherwise stated, Microsoft Excel 2013 (any modifications to the default produced graph are described in the image labels).

Truncation of Axes

When making comparisons, it is important to ensure that axes are not shortened in a way that distorts the data. This is especially important in a bar chart, where it is generally accepted that the size of the bar is in direct relation to the values. When axes are truncated, differences are exaggerated and can create the impression of important change when in reality there is relatively little.

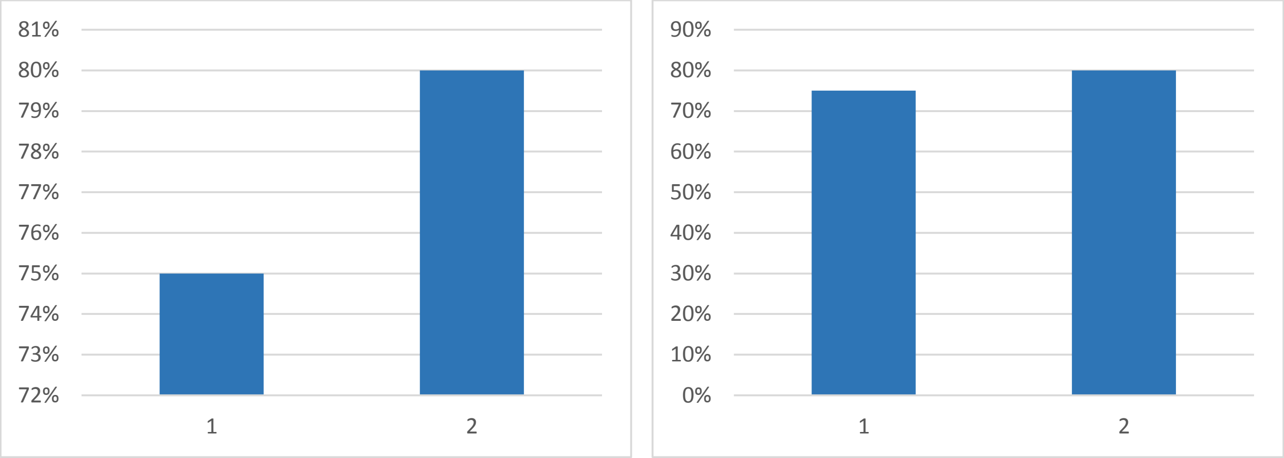

Figure 1: Two bar charts showing the same data, the chart on the left has been truncated by default in Microsoft Excel 2013, on the right the axis has been modified so that its minimum value is 0

When a break or truncation cannot be avoided it should be clearly indicated on the axis using one of the widely accepted symbols, and as always, axes should be clearly labelled.

Representation of Values by Area

When representing a value by using different sized images (e.g. in a bubble graphic) it is the area that must be taken into account and not the height of the image. Similarly, if the values are represented in 3 dimensions, the volume of the object must accounted for so not to mislead.

Another drawback of using this method to display values is that the viewer cannot draw comparison as well as when the values are depicted in other ways (Cleveland & McGill, 1984). This has become less of an issue as these type of graphics seem to be less prevalent than they were in the past.

Cumulative graphs

Cumulative graphs indicate a running total of values; this can obscure real performance. When this is used to display important metrics, such as monthly revenue, this will display the data in a more favourable light.

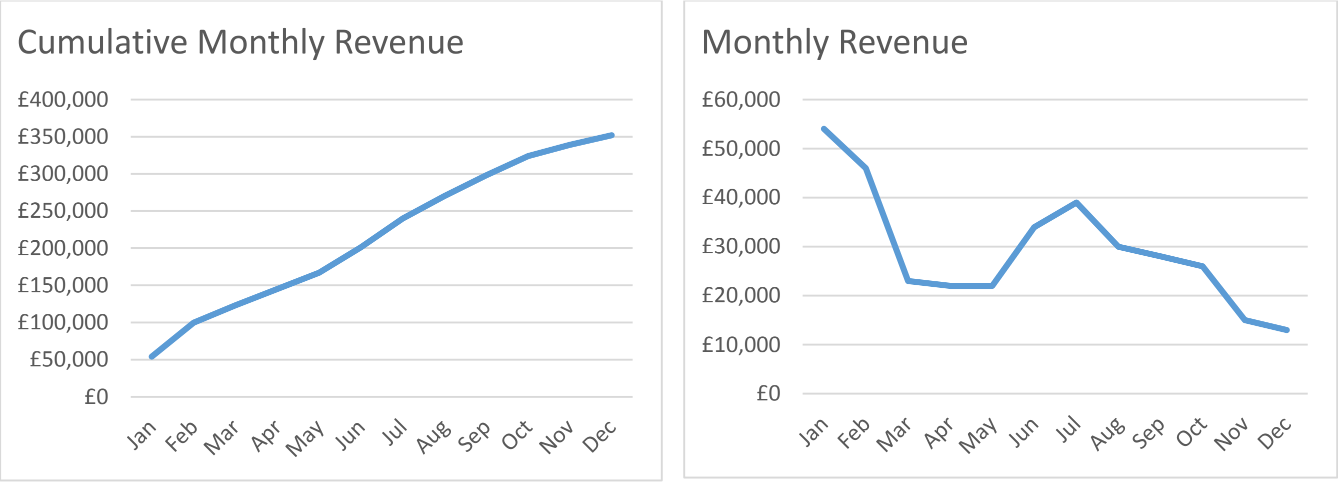

Figure 2: Two line charts showing the same data, the chart on the left showing cumulative revenue by month, on the right showing monthly values.

Occasionally the first plotted point will be a cumulative value from data that is not visible on the graph, this can further mislead the viewer.

Misuse of 3D

The use of a superfluous third dimension, which does not contain additional information, often makes it difficult to estimate the values represented.

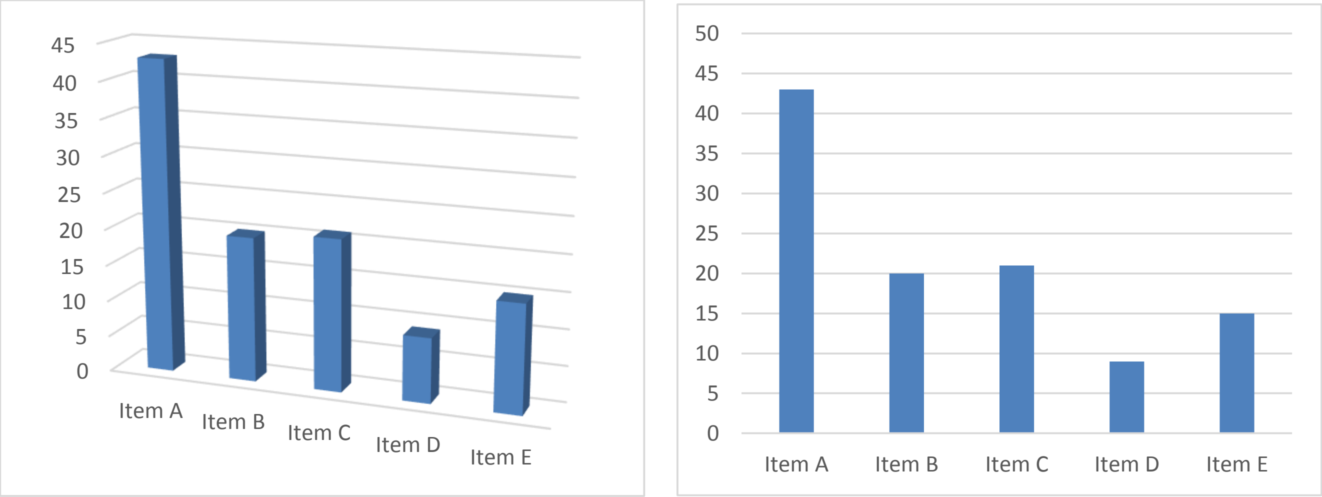

Figure 3: Two bar charts showing the same data, the chart on the left is using the default 3D settings of Microsoft Excel 2013

These issues are further exaggerated by extreme rotation and perspective changes.

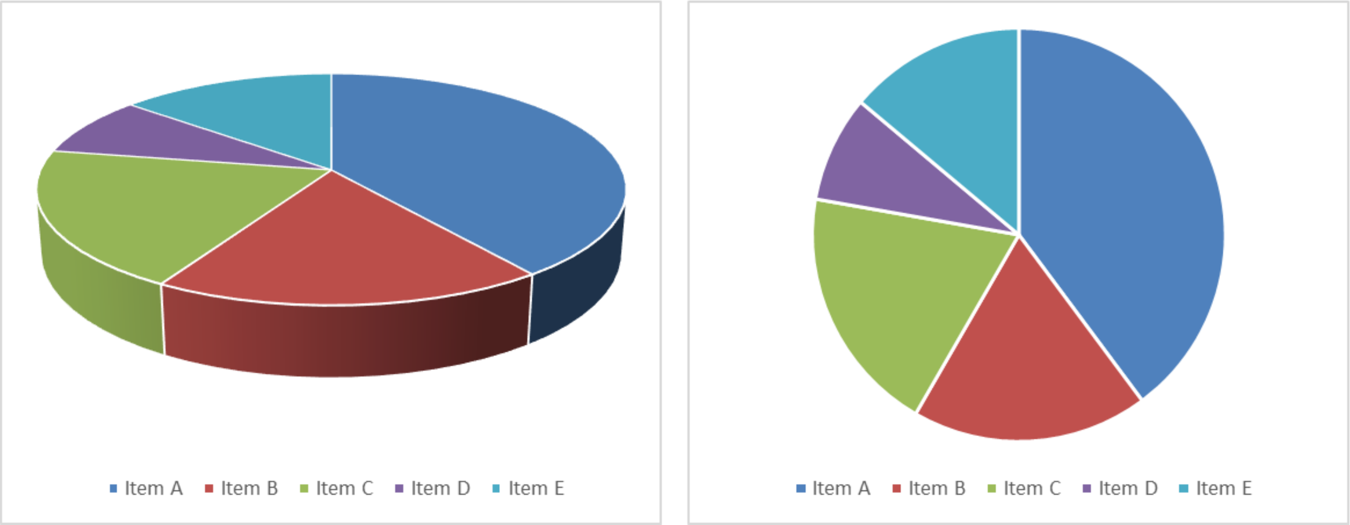

Figure 4: Two pie charts showing the same data (the same data is also used in Figure 3), the chart on the left has been rotated 25° on the Y axis. Note how "Item B" appears much larger in the chart on the left.

Pie charts should generally be avoided as it is more difficult to determine the relative size of pie slices than it is to compare relative lengths on a bar chart (Cleveland & McGill, 1984). Note how it is difficult to determine which is larger between "Item B" and "Item C" in Figure 4, and it is visible in the 2 dimensional chart in Figure 3 that "Item C" is slightly larger than "Item B". It may be that certain visualisations purposefully rely on these weakness to obscure relationships.

Disregarding Standard Practices

This is a particularly deceptive practice whereby a visualisation is created that ignores common conventions.

Unfortunately there are many examples of this in the media, ranging from inverted and non-continuous axes, to pie charts that do not add up correctly and changes to the units of comparison. Without further inspection, these deceptions can cause a completely different interpretation of the data due to our trust and familiarity with these conventions.

Detracting Attention from the Data

It may be the intent of the designer to distract from a message in the data through the use of elaborate graphics, downplaying the data containing areas or over-emphasising accompanying datasets.

While it is not my intention to suggest a particular aesthetic, a cleaner, simpler graphic with minimal embellishments does minimises this.

Another issue related to this is that of introduced or artificial complexity. The graphic may be made to appear more complex than it really is either to discourage interest, or to put a particular concept into question by stating that it is so complicated to be incomprehensible, rather than presenting a neutral schematic that accurately depicts a complex system.

It is not fair to say that all charts should be understandable in a glance, this would depend on its intended use and audience, for example a graphic designed to be visible for a few seconds as part of a television news item should be quickly comprehensible, but a large item in a newspaper or book may contain information that invites the user to study it for some time.

Statistics and Manipulation of the Data

There are many ways in which the data itself can be altered to better suit the message that the visualisation creator wishes to impart. Many of these are particularly insidious because, without having access to the original data, it is unclear that the data has been manipulated. These methods include, but are not limited to:

- Selective sub sampling (cherry picking)
- Omission of data/key variables
- Failure to put data in context (e.g. not price adjusting for inflation)
- Incorrect extrapolation
- Fabrication of data (lying)

A comprehensive and accessible introduction to this can be found in “The Tiger That Isn’t” by Blastland and Dilnot (2007).

We must also be careful in what we are implying by creating the visualisation and remember some of the fundamentals such as correlation does not imply causation, which is a misunderstanding of statistics where erroneous claims regarding an idea by using statistics that have little or no ability to make such claim.

Interactivity

Interactivity offers the ability for users to explore datasets dynamically by manipulating mappings through actions such as zooming, filtering and other forms of interrogation. This can aid in identifying trends, patterns and clusters otherwise hidden in the datasets. The predefined choices of how and what data is represented defines the limitations to which the users can explore and make up their mind about the data.

All of the examples that precede this section are still applicable to visualisations that incorporate interactivity, indeed some of the issues can be alleviated through the correct usage of interactivity (e.g. options can be given to the user to manipulate the axes, or switch between cumulative and individual depictions of the data, or even to change between types of graph entirely).

However, they can be corrupted in the same way that static visualisations have been, and interactivity brings additional issues to the forefront (some of these are difficult to describe textually in a static medium, further indicating the benefits of interactive media).

Limitations to Interactivity

Although the creator of an interactive visualisation defines the ways in which the visualisation can be interacted with and interrogated, they should not limit this in a way that highlights one particular point of view (e.g. allowing users to view clusters of data in more detail that better support their hypotheses, but leaving the rest of the visualisation). They should instead allow all the data to be interrogated equally, and to reasonable extents.

“The techniques which aim for extrinsic meaning often explicitly limit interactivity, ensuring the communication of the creator’s predefined perspective rather than fundamentally unpredictable user interpretations of the data.” (Lau & Vande Moere, 2007)

Change Blindness

Change blindness is a perceptual phenomenon that occurs when people do not notice changes in visible elements of a scene (Nowell, Hetzler & Tanasse, 2001). It is exacerbated by a break, or flicker between the two similar images or states. Visual Intelligence can only detect changes in those parts of the image to which we explicitly attend, so a variation only becomes apparent if the area of the change is viewed at the time of transition.

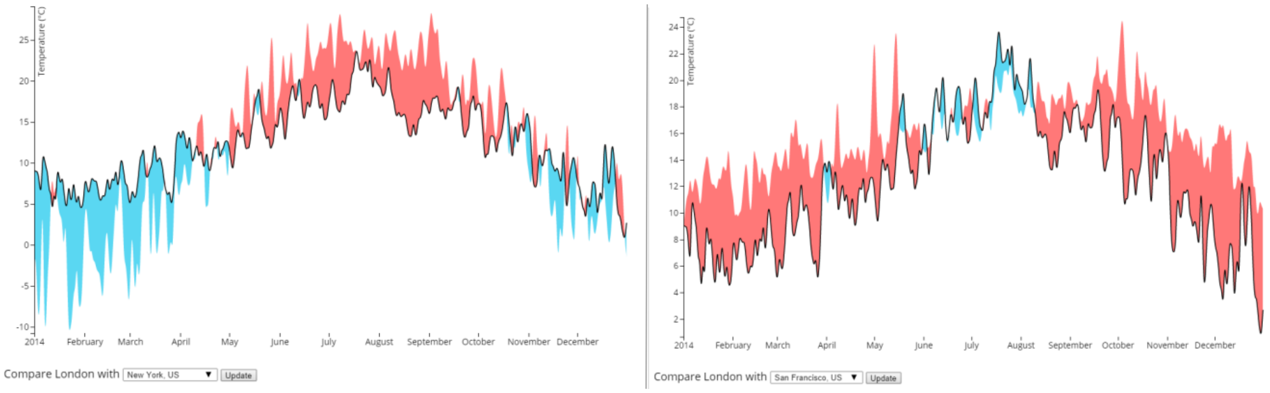

Figure 5: Difference in temperatures between London and other major cities in 2014 (New York on the left, San Francisco on the right) an example of 2 states of a simple interactive visualisation created using D3.js where in changing the data displayed also changes the scale and axis ranges, the black line (London temperature data) remains unchanged (data from NOAA NCDC climate data online)

In figure 5, the differences in colour clearly indicate that a change has taken place, and the entire "scene" has changed, however if a quick transition between the two images occurs, the subtle, but still important differences (e.g. the changes to the scale) may go unnoticed due to change blindness.

The risk of bringing about change blindness can be minimised through the use of smooth, graduated transitions between states. One of the most powerful features of modern data visualisation toolkits (e.g. D3.js) is their ability to transition smoothly between graphic types, layouts, and datasets displayed.

This can be further improved by staggering the transitions so that they happen one by one (e.g. transitioning the axes first to accommodate any further changes to the data.), as having many transitions happen all at once can still have the result that the user does not realise that these elements have changed.

Emphasising elements that have altered, using colour, size or other effects can also effectively indicate change. These effects should remain moderate so not to seem unprofessional, or detract from the data.

Another option, particularly for interactive visualisations, for conveying change and being able to compare the data is overlaying the new state with a ghost (faint, or outline) of the previous state.

Detracting Attention from the Data

In addition to the ways that this occurs in static visualisations, covered previously, embellishments in the digital domain can also include animated elements that can further detract from the data, and these may not be part of the original design, it may be due to the site/medium on which the visualisation is shown.

The addition of interactive elements to improve the understanding of relationships can also lead to overcrowding in a dataset. In allowing datasets to be interrogated or different datasets to be displayed at the direction of the user we must also be careful to ensure that the visualisation remains clear in all circumstances and that elements scale and are interpolated accordingly. Ghosting, for example, is only useful in some cases, and its use can result in overcrowding or over complication of a graphic.

Ethics

"When we see a chart or diagram, we generally interpret its appearance as a sincere desire on the part of the author to inform. In the face of this sincerity, the misuse of graphical material is a perversion of communication, equivalent to putting up a detour sign that leads to an abyss" (Wainer, 2000).

Data visualisations do not lie, it is the designers of purposefully deceptive visualisations that do. Visualisations can mislead. Misleading is distinct from lying because a graphic can unintentionally lead readers astray without the conscious intervention of its designer. According to professional ethics codes, knowing the truth and hiding it, or conveying it in a way that distorts it is simply unacceptable (Cairo, 2014).

"Seek Truth and Report It, Journalists should be honest, fair and courageous in gathering, reporting and interpreting information" (Society of Professional Journalists, 2014).

So the intention of the creator defines whether the actions performed are ethically sound (right or wrong) and creating a misleading visualisation as a consequence of mistakes while gathering, analysing or representing data is ethically neutral (Cairo, 2014).

However it should be said that it is responsibility of the creator to understand the principles of good visualisation design and integrity, and as such producing visualisations without adhering to the documented guidelines could be deemed unethical, though more ethical than purposefully misleading (lying).

As illustrated in the earlier quote by Wainer, the consequences of lies and mistakes are both equally serious.

Summary

As the means and tools to produce and publish visualisations become ever more accessible, we must be careful to ensure that we are creating ethical graphics, and to understand how other creators may have crafted misleading visualisations.

The principles garnered through the development of static data visualisations should be applied to the ever increasing field of interactive data visualisation, and we should be careful to apply these new tools in a way that does not distort the inherent messages in the data.

One of the most important things that a viewer can do, is question the data source itself. Does the publisher or author have anything to gain by misrepresenting the data? Do they have a reputation for producing misleading visualisations? Although a reputation is not always a guarantee of untrustworthiness.

Some creators may be quick to state that their graphics are not intended to focus on accuracy and instead are more interpretive and so fall into the category of informative or visualisation art. In which case that should be clearly stated or it should be accompanied by the data, or an accurate, literal data visualisation that supports any statements made as transparently as possible, so not to undermine the message of that visualisation, the integrity of the author or the field as a whole.

References

Blastland, M., & Dilnot, A. (2007) *The Tiger That Isn't: Seeing Through a World of Numbers*

Cairo, A. (2014) Lying with Infographics and Visualization, retrieved February 14th, 2015 from http://www.thefunctionalart.com/2014/02/lying-with-infographics-and.html

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387), 531-554.

Lau, A., & Vande Moere, A. (2007). Towards a Model of Information Aesthetics in Information Visualization. *Information Visualization, 2007. IV'07. 11th International Conference* (pp. 87-92).

Nowell, L., Hetzler, E., & Tanasse, T. (2001). Change Blindness in Information Visualization: A Case Study *Information Visualization, IEEE Symposium on* (pp. 15-15). IEEE Computer Society.

Society of Professional Journalists (2014) Code of Ethics, retrieved February 14th, 2015 from https://www.spj.org/pdf/ethicscode.pdf

Tufte, E. R. (1983). The Visual Display of Quantitative Information (2nd Ed.). Cheshire, CT: Graphics press. 77

Wainer, H. (2000). Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot, Psychology Press, 2