Recognizing manipulation and avoiding it yourself

We all watch the news and are flooded with surveys and charts, sometimes with little science behind them or with data that may be distorted in some way to make their point. In this week, we look at ways data is manipulated. We discuss practices that lead to distortion as well as review visual examples of flawed charts to reinforce the points.

Common design missteps that lead to distortion

Broken scales

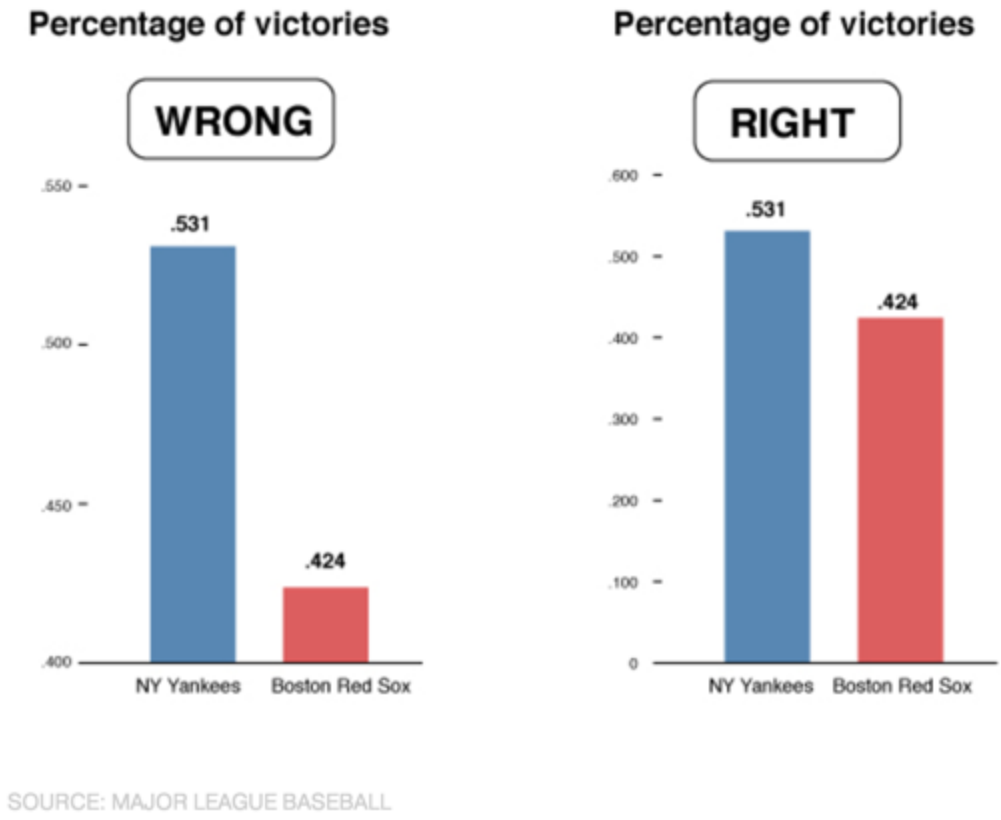
The most common way to manipulate the story that the data tells is through a visualization that causes the effect of exaggerated differences. For example, broken scales can imply disparities that don't really exist.

What we typically see is a wow statistic and visual and then on closer inspection now armed with knowledge from this class you notice what is going on.

The designer has depicted that data on a broken scale. Basically they didn't start with a zero Y-axis for example allowing you to see the rest of the bar chart or data in other visuals. What this does is make the audience think there is a big difference when really it is very small when you see the big picture or the real picture with a zero based scale.

SOURCE: BUREAU OF LABOR STATISTICS

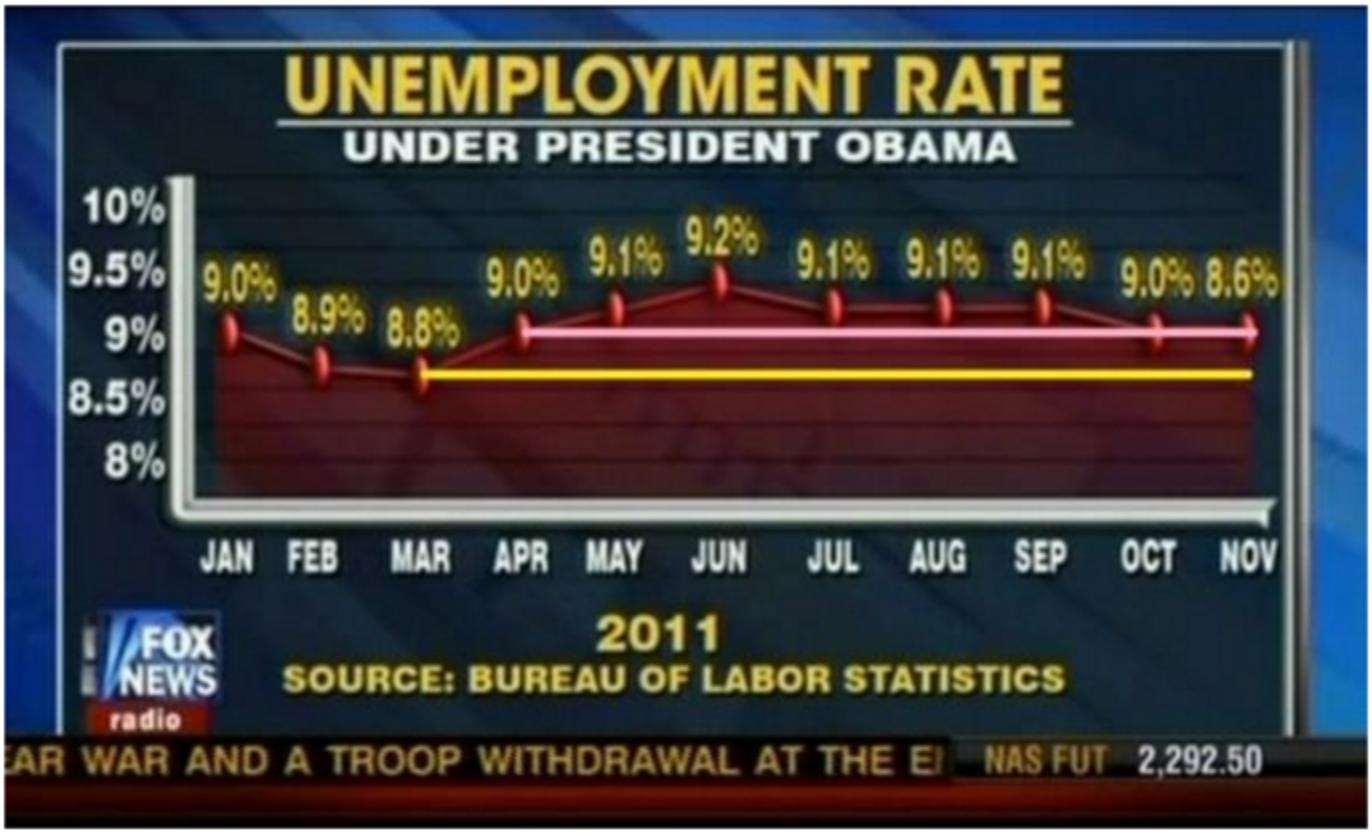
Being a Red Sox fan, this bar chart comparison is especially concerning.



Another way that exaggerated differences are often achieved while maintaining a zero-based scale is to use really small units in their visualization to magnify the effect. I've seen this in my own job currently where product managers use basis points (which is 1/100 of a percent) to show the impact that a new feature or feature update has caused to a KPI-like conversion.

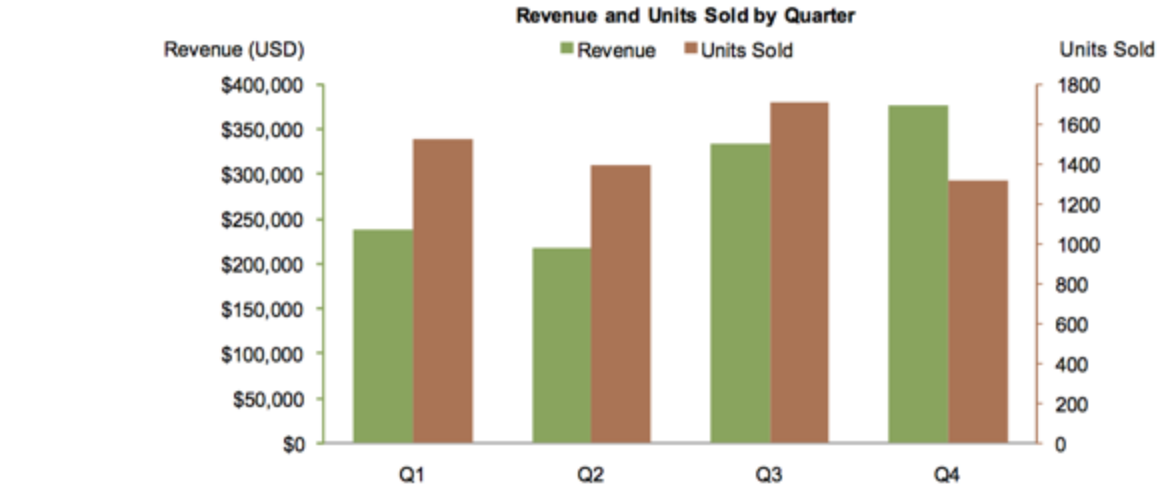
Once in a great while there may be a case where using a non-zero starting point is appropriate. For example, if one variable has data points that do not deviate far from the mean, a graph starting at zero would be too large to show the details within this variable. A linear graph would represent the data as a straight line. Visualizing this data in a graph that starts at a non-zero number would allow one to "zoom in" and view the small variations that make the data meaningful.

Misplotted/Inconsistent plotting of data points



Simply Statistics

Showing data on two different scales can make for an apples-to-oranges comparison.



Copyright © 2008 Stephen Few, Perceptual Edge

As Stephen Few says "Magnitude comparisons between values with different units of measure and scales are not appropriate, and should therefore be discouraged. Because bar graphs are designed for magnitude comparisons, a graph with a dual-scaled axis should never exclusively encode values as bars." (source: https://www.perceptualedge.com/articles/visual_business_intelligence/dual-scaled_axes.pdf)

What happens is the viewer will inherently make comparisons between the height of the bars, which makes no sense since each bar is using a different scale.

In earlier weeks' readings, it was mentioned that dual Y-axes should be avoided. Also the statistician Peter Ellis has pointed out the following five points with dual axis time series charts:

1. The designer has to make choices about scales and this can have a big impact on the viewer

2. In particular, "cross-over points" where one series cross another are results of the design choices, not intrinsic to the data, and viewers (particularly unsophisticated viewers) will not appreciate this and think there is more significance in cross over than is actually the case

3. They make it easier to lazily associate correlation with causation, not taking into account autocorrelation and other time-series issues

4. Because of the issues above, in malicious hands they make it possible to deliberately mislead

5. They often look cluttered and aesthetically unpleasing

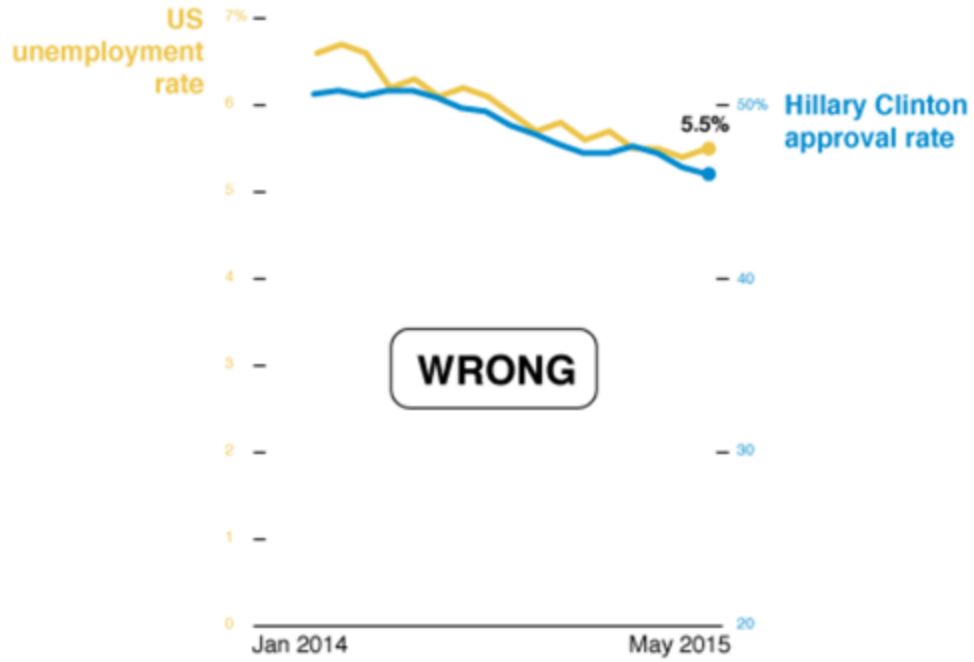
(source: http://ellisp.github.io/blog/2016/08/18/dualaxes)

It is so rare that dual Y-axes can be used appropriately. So when you run into the temptation to use a dual Y axes, it is best to avoid it and create two separate charts as shown below.

Showing a correlation can imply causation.

Sometimes data will follow the same trend. Just like in real life when things happen at the same time we may think that one thing caused the other, similarly, when data follows a similar trend and is plotted

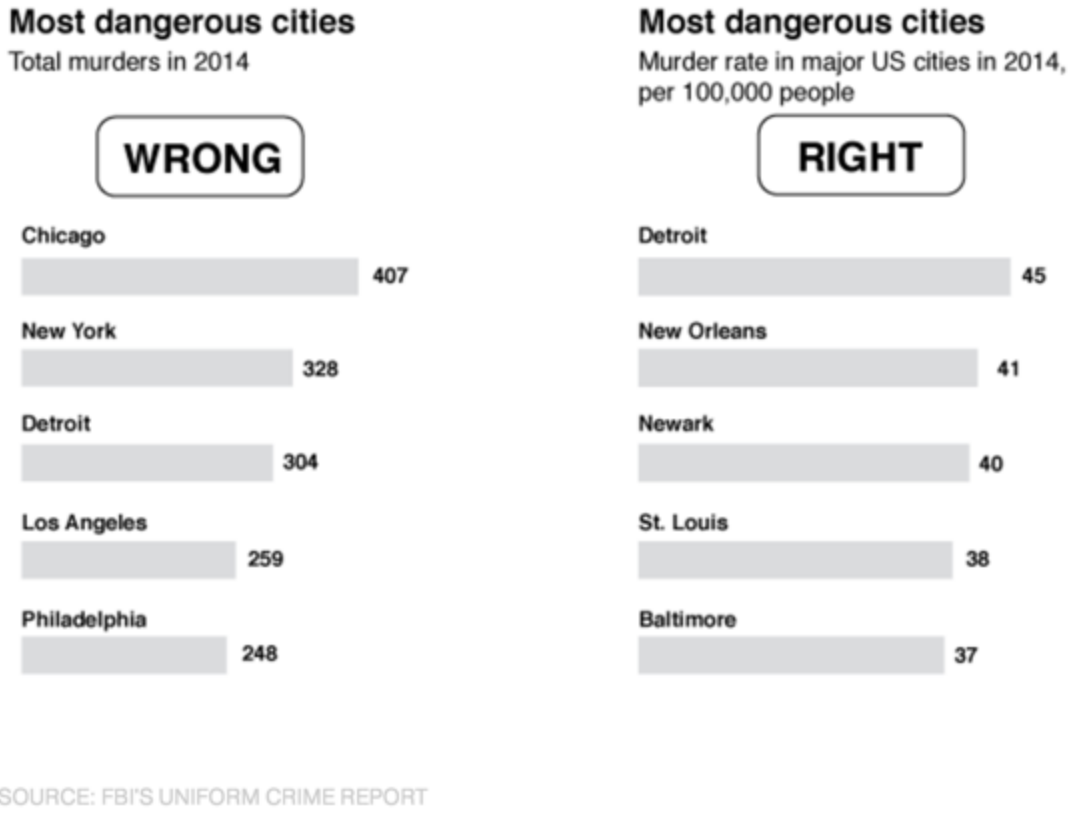
together in the same chart, users may think that this correlation of trend implies causation when it does not.



Ignoring population size makes rates impossible to compare.

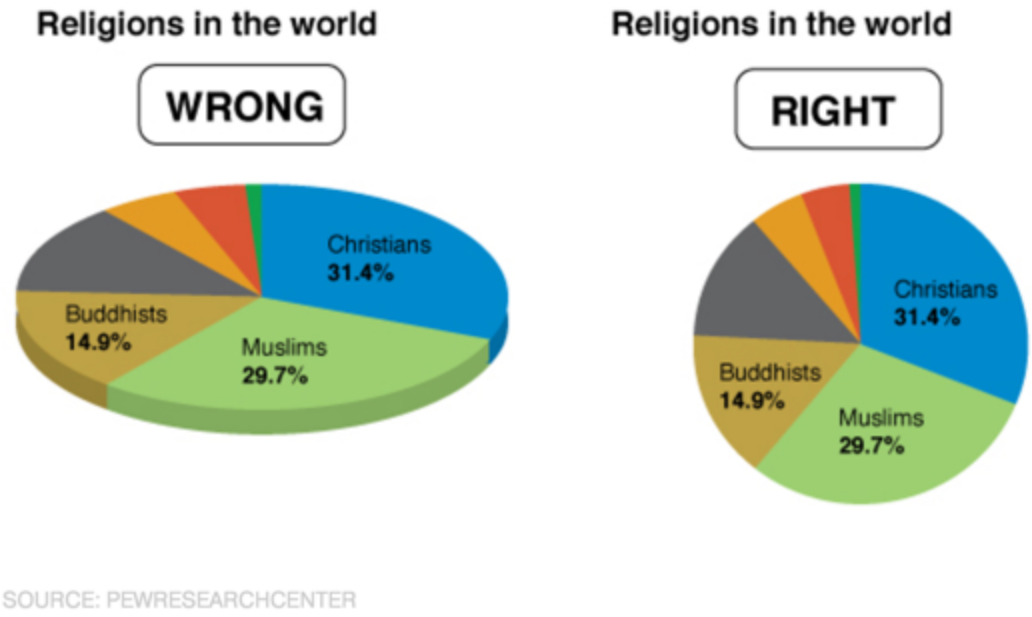
Now and then you may see charts that compare the absolute counts of an event happening in a particular region, such as a city or a state, but fail to recognize they are making a comparison across vastly different sizes of populations. For example, a graph might show the number of people who quit Microsoft compared to other companies that are a tenth the size of Microsoft without factoring the companies' size in the visual. The user might conclude that Microsoft is a bad place to work when in reality the percentage of employees at each company rather than the real number of employees might prove that Microsoft actually has very good retention rates.

In the following graphs, the first example doesn't factor in the population differences of the cities so it is misleading.



Use of 3-D and other effects can also be a source of visual deception.

In our readings previously, we saw how pie charts are misleading in general and should be avoided in most cases. However, look at the effect of the 3-D added to the pie chart below. Even though Christians have the largest percentage, the 3-D distorts the visual and makes it look like Muslims may be a greater amount on casual glance.



Why we should care

I've included this topic for several reasons. First of all, given the amount of data we encounter through the media, it is important to have some training to identify when a chart may be deceptive to avoid being misled. Secondly, we all are designers and will at some time be creating our own charts and it is important to be educated how others can manipulate data by what methods so we don't unintentionally design these elements into our representations and accidentally mislead folks.

Take into consideration some conclusions from a report put out by the University of California, UC San Diego Experts Calculate How Much Information Americans Consume.

The report suggests the average American consumes 34 gigabytes of content and 100,000 words of information in a single day. (Leo Tolstoy's "War and Peace" is only 460,000 words long.) This doesn't mean we read 100,000 words a day — it means that 100,000 words cross our eyes and ears in a single 24-hour period. That information comes through various channels, including the television, radio, the Web, text messages and video games.

Again, this is why we should care given the amount of information we consume in a given day. Masses can easily be swayed intentionally or unintentionally.

Just for fun

Here are some more links showing misleading charts that we can learn from

http://blogs.wsj.com/economics/2016/01/25/what-a-misleading-chart-on-u-s-trade-looks-like/

https://www.buzzfeed.com/katienotopoulos/graphs-that-lied-to-us?utm_term=.inbM8JOqE#.wry3Jyn8D

http://mediamatters.org/research/2012/10/01/a-history-of-dishonest-fox-charts/190225

https://www.baekdal.com/insights/misleading-types-of-graphs-for-the-media

Spurious Correlations: http://tylervigen.com/spurious-correlations.

Trumps Bar Charts: https://www.washingtonpost.com/graphics/politics/2016-election/trump-charts/

One final thing to look into, if you have 30-90 minutes someday, is this interesting interactive research project which discusses the amount of data we produce and stored in digital memory. For future generations to make the most sense of it, we should be mindful of how we represent data and follow best practices. http://inlimbo.tv/en/