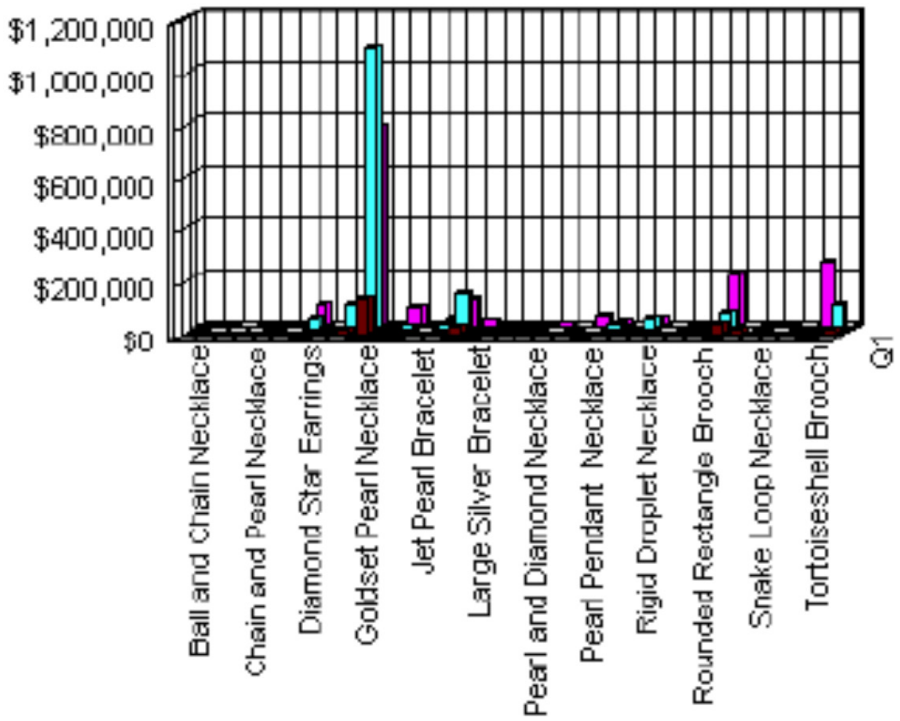
Visualizing Wide-Variation Data

Nick Desbarats, Sr. Educator and Consultant, Perceptual Edge
Visual Business Intelligence Newsletter
January/February/March 2016

Among the more horrific examples of poorly designed graphs to which Steve and I subject participants in our Show Me the Numbers course, this one tends to prompt a fair amount of discussion:



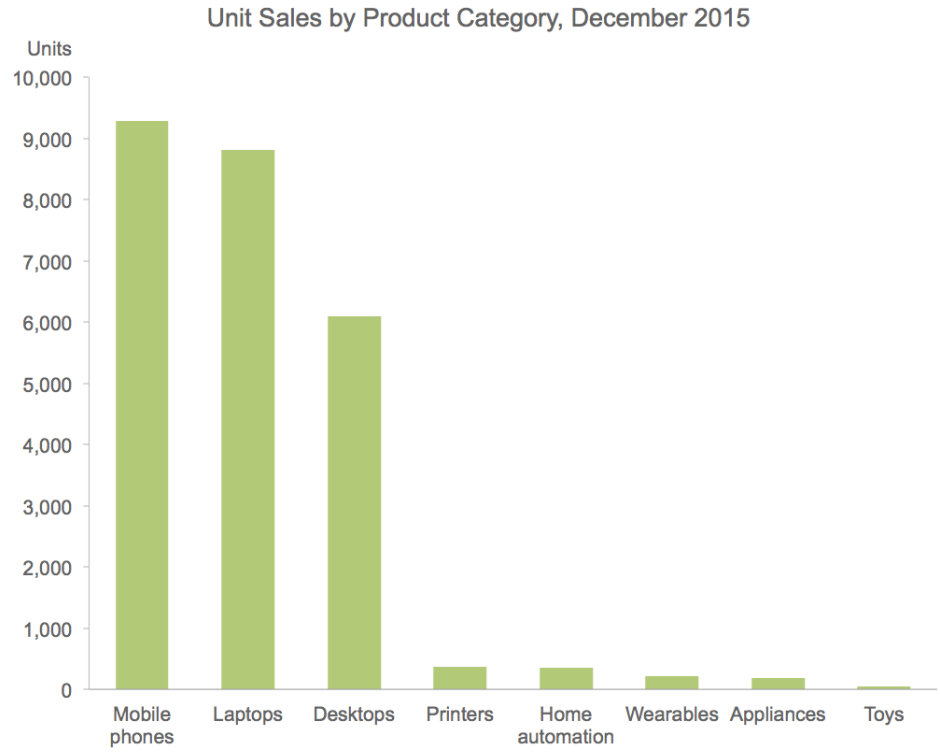
Source: User documentation of Business Objects

There are certainly plenty of teachable moments that arise when discussing this graph, but the one that comes up most often during workshops involves how to deal with extreme variation in a data series. In the data shown in this graph, for example, there are two values that are many times higher than most of the other values, making it a wide-variation data series. This particular graph obviously does a poor job of displaying this data since many of the values are encoded as vanishingly small bars that are effectively invisible to the viewer, even though there may be important insights that would be gleaned from comparing those smaller values to one another. How could the graph creator have avoided this problem, though? Is there a way to show very large values and enable viewers to easily and visually compare the smaller values in a data series as well?

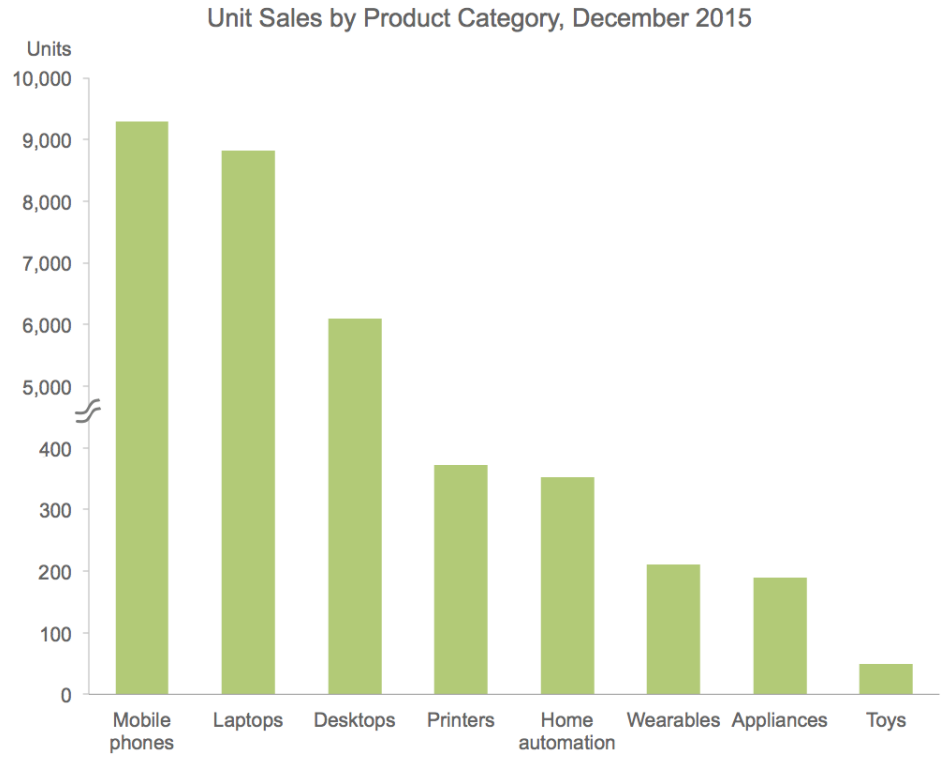
Being the creative and astute people that they are, workshop participants often come up with several solutions, a number of which I will discuss below. I will then conclude with some recommendations that reflect how I deal with wide-variation situations in practice.

Broken Quantitative Scale

Consider the following graph of another wide-variation data series:



As with the graph shown at the beginning of this article, the wide variation in this data causes the bars that encode small values to be too short to allow useful visual comparisons with one another. By introducing a break in the quantitative scale, however, the bars that encode the smaller values become tall enough to be seen easily:

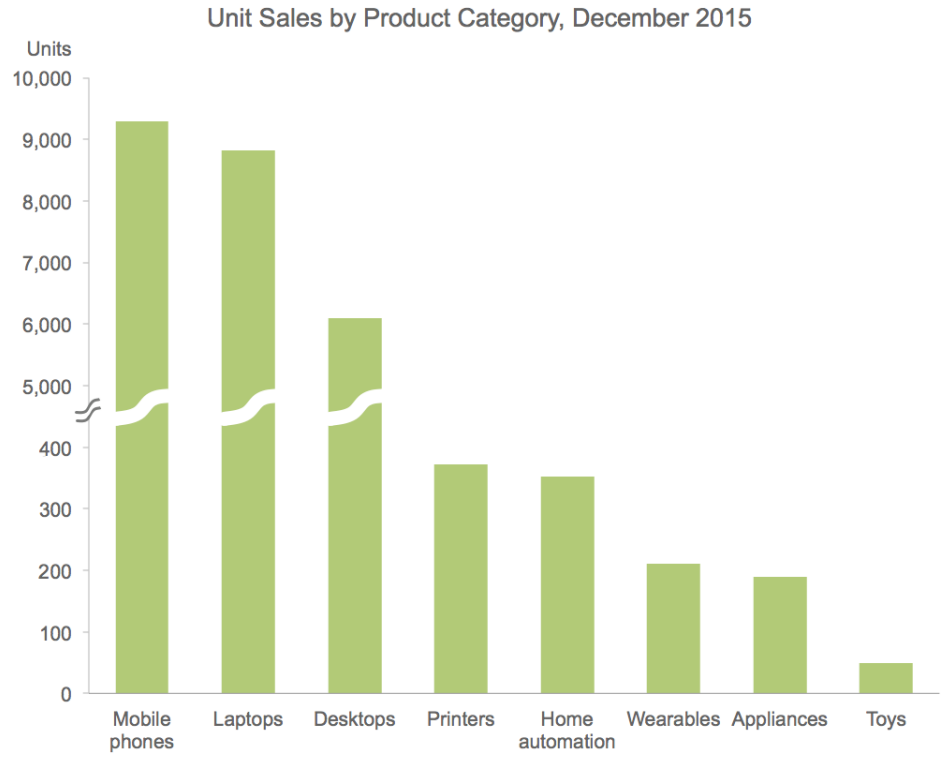


This seems like an elegant solution to the wide-variation problem, since the bars for both the small values and the large ones are clearly visible and their numerical values can be readily and accurately determined by visually comparing the tops of the bars with the quantitative scale on the left. This change, however, also means that comparing the heights of the bars with one another no longer yields accurate magnitude comparisons. Comparing the bar heights in this graph would, for example, indicate that Printers represent about 60% of Desktops, whereas they actually represent only about 6%. The purpose of displaying data in a bar graph in the first place, however, is to enable viewers to make visual magnitude comparisons of values based on the heights of the bars. The above solution, therefore, defeats the most basic purpose of a bar graph. The graph creator might as well simply show a table of numbers instead.

Another reason to avoid this solution is that it can lead the viewer to misinterpret the underlying data. There are at least two ways that this can happen. The first is that the viewer may miss the fact that the scale is broken and conclude that, for example, Wearables are about 25% of Desktops, which would be wildly incorrect since they are actually about 2%. The viewer may miss the break in the scale simply because they do not look for it, or because they have never seen a graph with a broken scale before and end up reading it incorrectly.

If the viewer notices the break in the scale and understands what it means, they will need to continually remind themselves that the bars' heights cannot be used to compare values across the data series. If they succeed in doing this, they are left to do mental math in order to make those comparisons. "How much larger are unit sales of Mobile phones than those of Wearables? Well, I can't visually compare their bar lengths so, let's see, Mobile phones looks like about 9,200, Wearables about 200, so 9,200 divided by 200, that's, um, um, about 45 times larger." That is a lot of work.

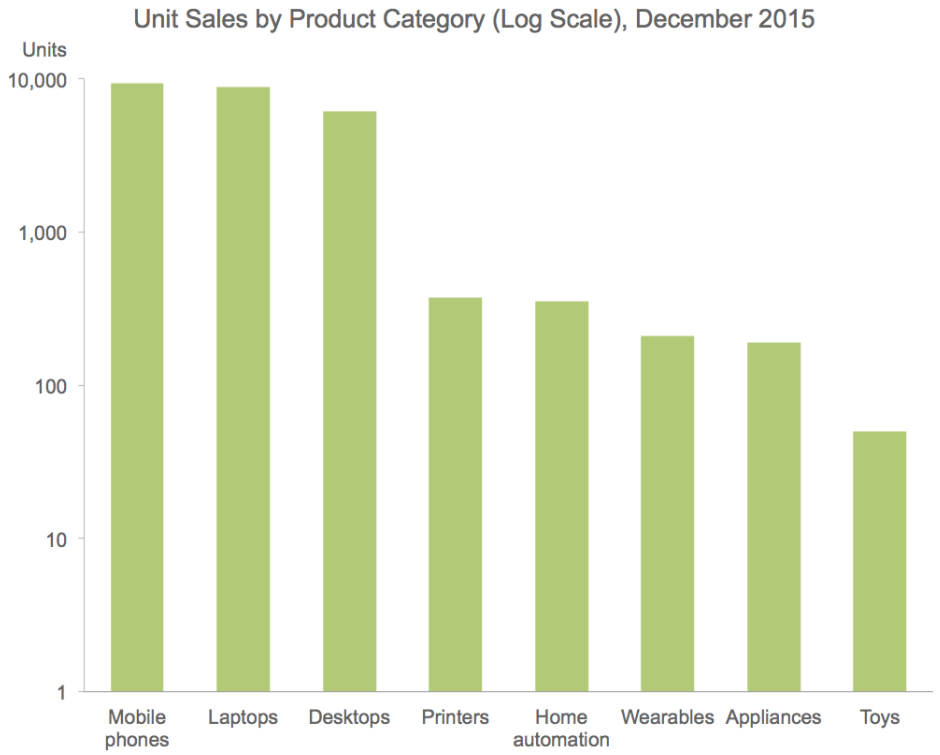
Some workshop participants have suggested breaking not just the scale, but the bars as well:



While this would make it more obvious that the graph features a broken scale, it, unfortunately, does not entirely mitigate the other issues discussed above.

Logarithmic Scale

More mathematically inclined workshop participants sometimes suggest changing the quantitative scale from a linear scale to a logarithmic scale to encode wide-variation data:



While a linear scale has intervals that are the same across the entire scale (the interval in the original version of this graph was 1,000 for all intervals), a logarithmic scale has unequal intervals. In this example, each interval is 10 times greater than the interval just below it, although bases other than 10 can be used as well. One upshot of this change is that short bars are lengthened, making them tall enough to be seen and visually compared without a break in the scale.

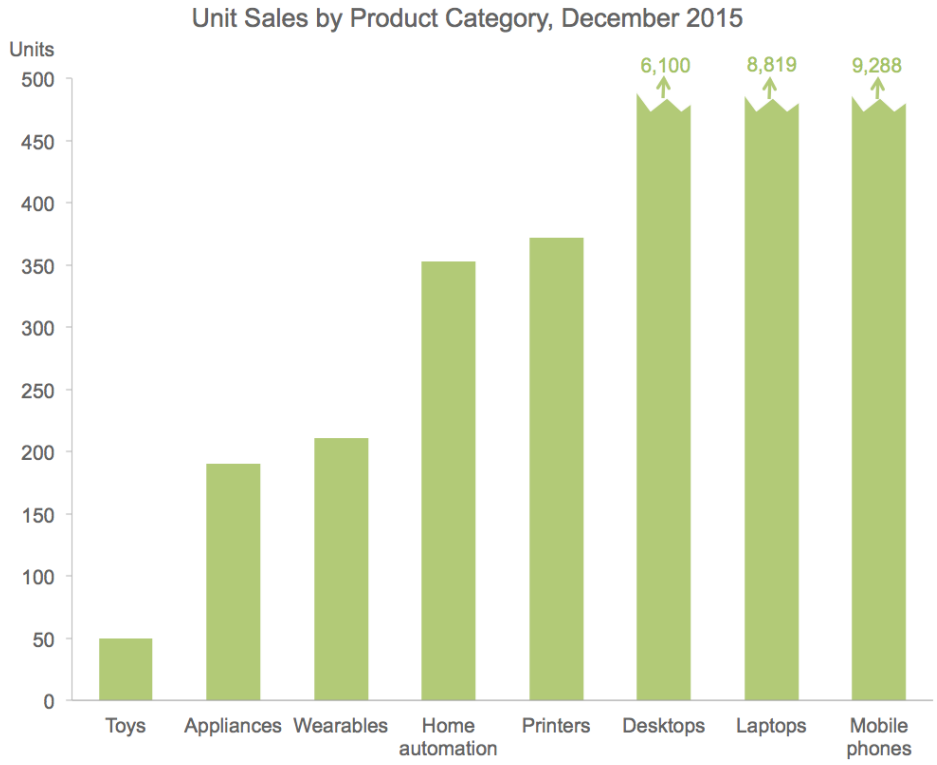
In practice, however, I avoid this solution for reasons that are similar to those that relate to the Broken Quantitative Scale solution discussed previously. Firstly, this solution assumes that viewers will notice that the scale is not linear, and also that they will know how to read a logarithmic scale, both of which may or may not be true, depending on the audience. Secondly, even if the viewer notices that the scale is logarithmic and knows how to read it, they will still be prone to making errors such as perceiving Printers to be about 60% of Mobile phones, which is not even remotely accurate (they are actually about 4%). Thirdly, even if the viewer succeeds in avoiding these errors, they will still not be able to compare values easily. How much greater is the Printers category than Toys? Three times greater? Five times? Ten? Making even rudimentary visual comparisons between the values will be almost impossible for most people. Printers have about eight times the unit sales of Toys in this example, by the way. Fourthly, when using logarithmic scales to communicate quantitative data visually, even people who are familiar with logarithmic scales will have difficulty accurately "eyeballing" values in graphs that feature them. What is the actual, numerical value of Printers in the above example? Probably somewhere between 300 and 600? That is a pretty wide range. The actual value for Printers is 372.

If the audience consists entirely of experienced scientists, mathematicians, or the like, they may have trained themselves over a number of years to read graphs with logarithmic scales somewhat accurately, but even they will not be able to do so nearly as easily or accurately as they could with graphs that use linear scales.

This is not to say that log scales should never be used in graphs, only that they are not a good choice when the purpose is to enable the viewer to make linear magnitude comparisons of values, as is almost always the case with bar graphs. If the purpose is, for example, to compare the rates of change of several variables (see https://www.perceptualedge.com/blog/?p=24), or to make log magnitude comparisons of values (assuming that the audience understands what those are), then a log scale may be a good choice.

Truncated Large Values

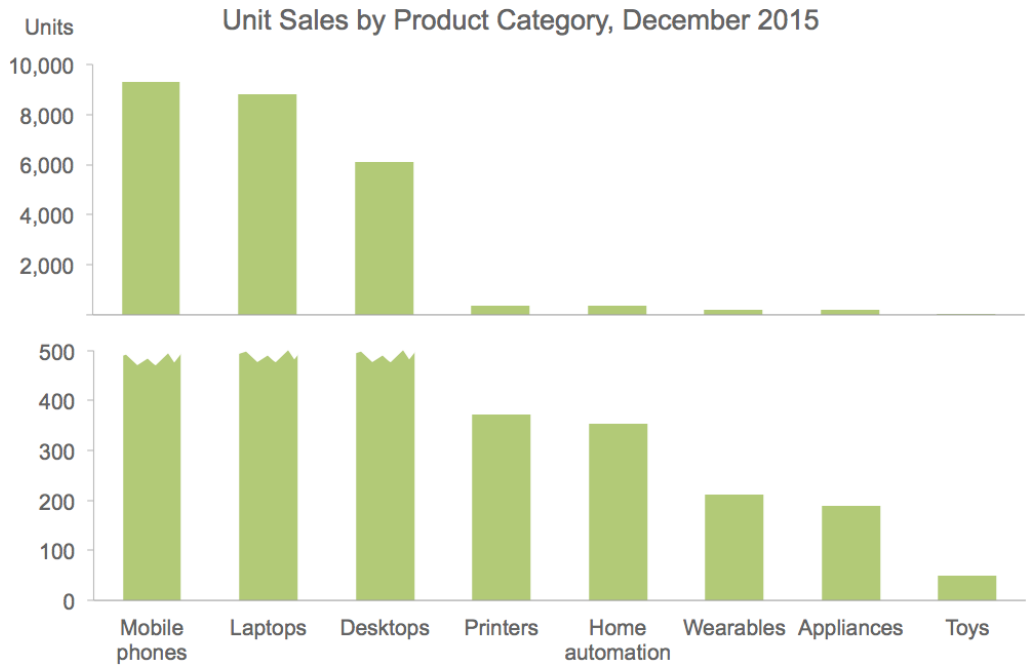
If the larger values are not an important part of the story, they can be truncated:



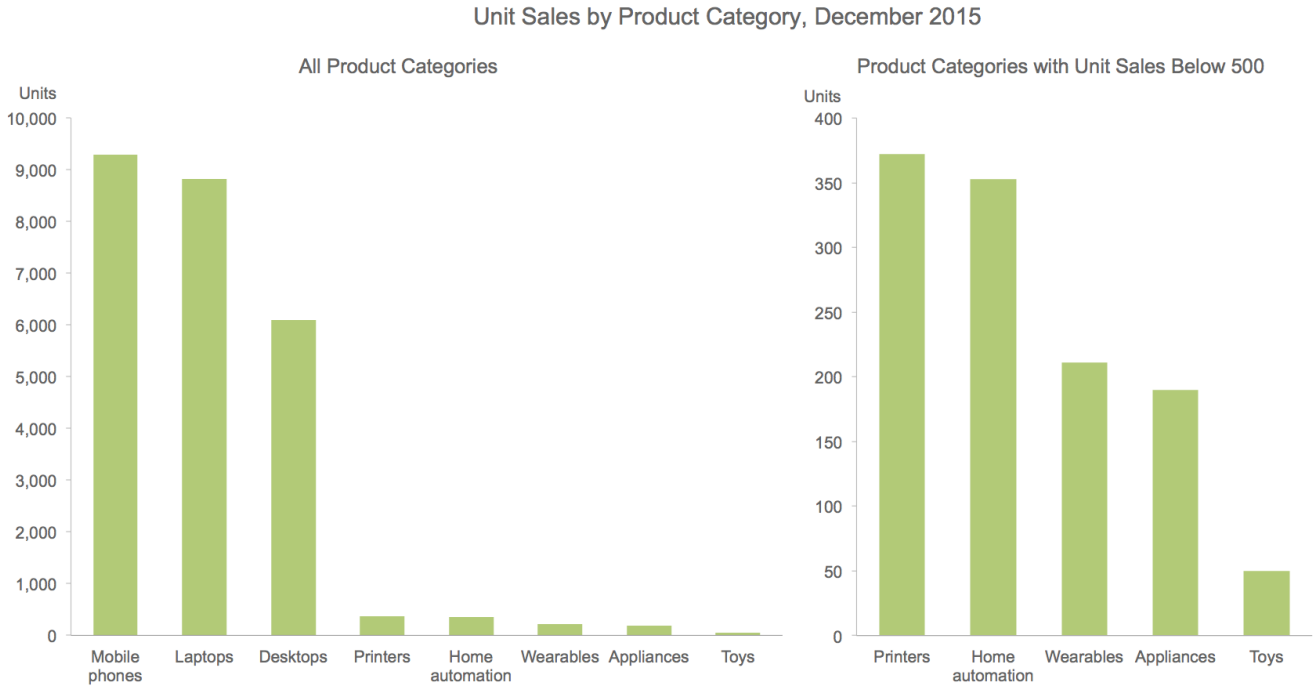
As shown in the above example, it is important to make the bars of large values very obviously truncated so that viewers do not mistakenly perceive the bar lengths as encoding the actual values. Again, this solution should only be considered when the small values are the story that needs to be told, and the larger ones are of little importance.

Two-Graph Solution

This solution involves splitting a graph with wide-variation data into two graphs. The first graph is a “macro” graph that includes all of the values in the data series, plotted on a scale that spans the entire range of the data series. The second graph is a “micro” view that shows the data plotted on a scale that spans a smaller quantitative range such that bars for the smaller values can be visually compared with one another easily and accurately.



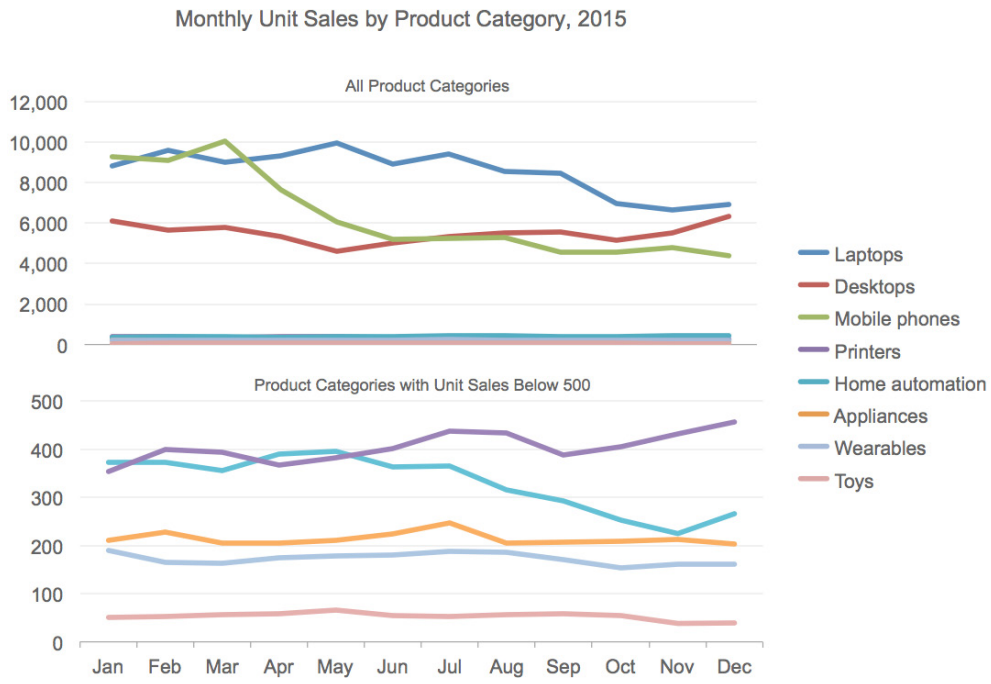
Here is another variation of this solution with a different arrangement:



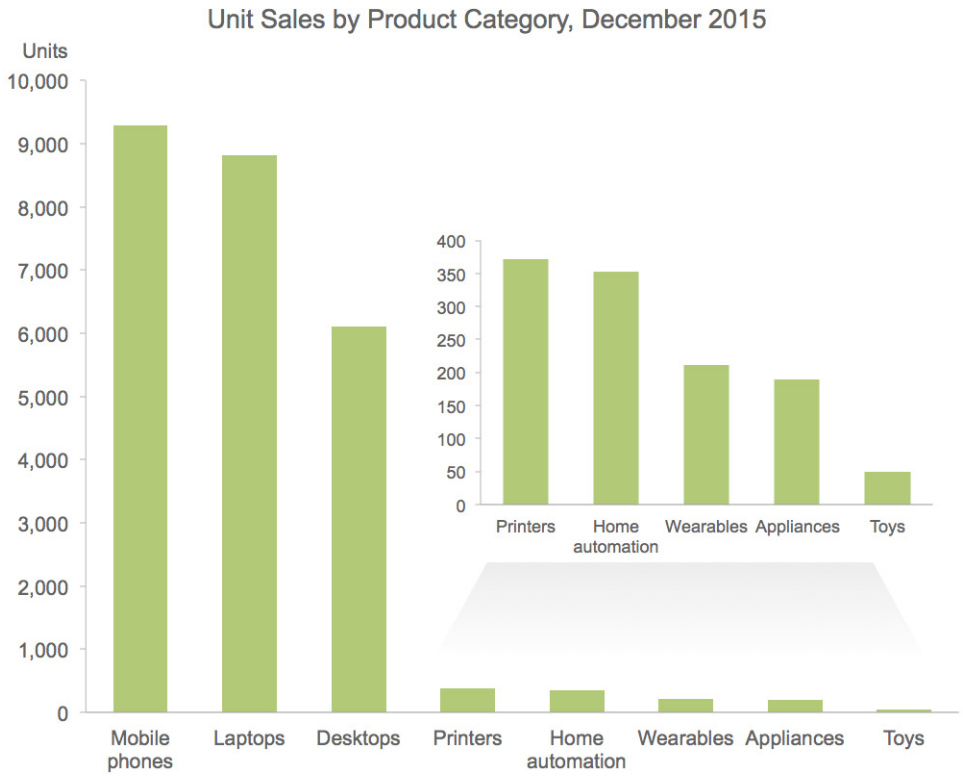
This solution has several strengths when compared with others that we have discussed thus far:

1. Viewers are less likely to make severely erroneous visual comparisons.
2. Viewers can easily see the “big picture”, enabling them to develop a general sense of the degree of variation in the data. In this example, the viewer can readily see that the large values are roughly tens of times larger than the smaller ones in this data series.
3. The smaller values can be easily and accurately compared with one another. If the smaller values are an important part of the story, this would be an important benefit.

This solution can be applied to other graph types, as well:

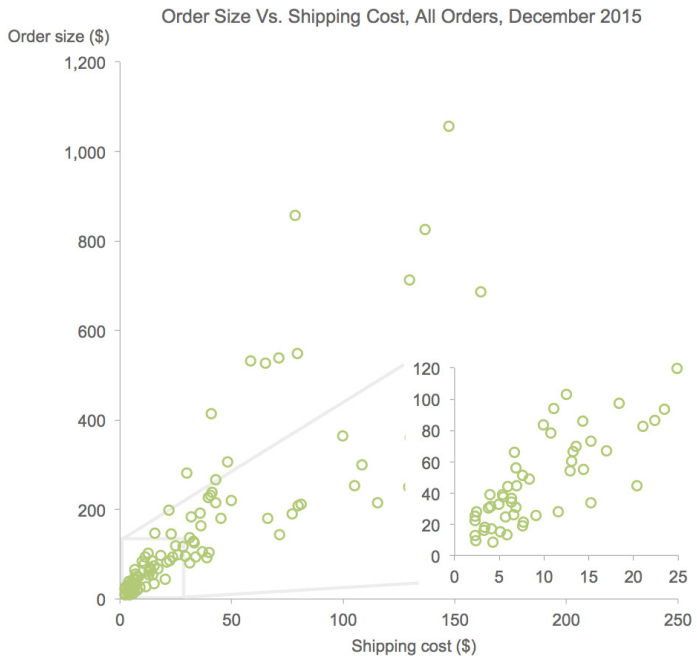


Another variant of this solution, which I will call the “magnified inset graph” solution, prescribes showing all values plotted on a scale that spans the entire range of the data series, and then adding an inset graph that shows a micro, or “magnified” view of the smaller values that enables those values to be easily compared with one another.



Personally, I find that the relationship between the two graphs is clearer when they are arranged in this manner,

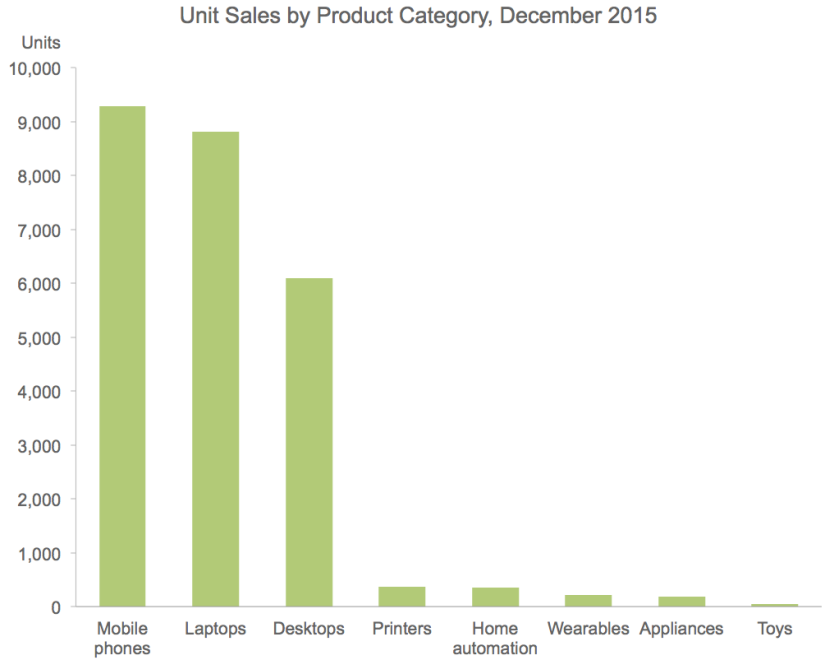
and this arrangement tends to take up less real estate on a screen or page. It can also be applied to other types of graphs, such as the example below:



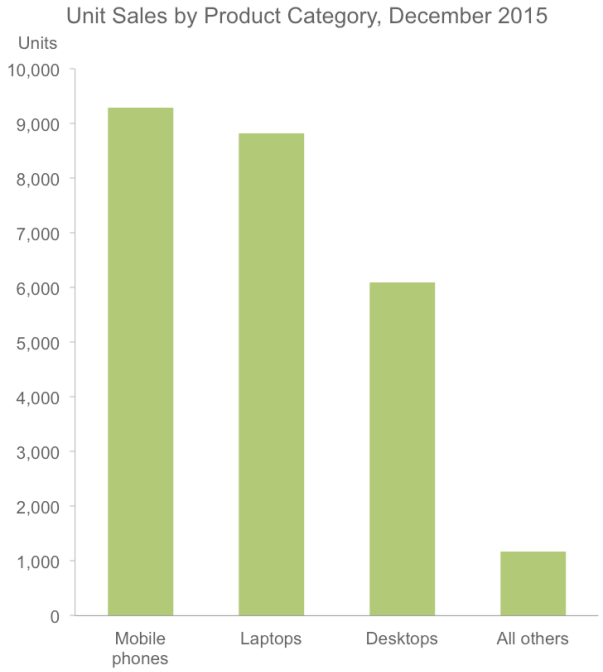
Note that this scatterplot example assumes that the larger values are not an important part of the story of this data since some of those points may be occluded by the inset graph. If this is not the case, and if there is not enough empty space in the macro graph for the inset graph, a stacked or side-by-side graph arrangement might be a better choice.

Final Recommendations

Before making any recommendations, I should underscore that none of these solutions are needed if the viewer is unlikely to be concerned with the smaller values in a wide-variation data series. If the story that the viewer should derive from a graph is that there are some large values that are important and oh, by the way, there are also some unimportant smaller ones, then a basic graph would be perfectly adequate:



In fact, if the smaller values are not important, it might make sense to group them together as “All others” which, in some cases, may obviate the wide-variation problem altogether:



If this is not the case and comparing the smaller values with one another is an important part of the story, then which of the solutions discussed herein should you use? In practice, when I need to encode wide-variation data graphically, I tend to use one of the two-graph solutions since these are the only ones discussed here that do not pose a significant risk that viewers will come to an incorrect understanding of the underlying data. Among the two-graph solutions discussed above, I tend to use the magnified inset graph solution if I can comfortably fit the inset graph into the macro graph without occluding important data in the macro graph. If this cannot be accomplished, I will then normally try a stacked or side-by-side two-graph arrangement.

In workshops, I often notice that participants feel compelled to show all of the data of interest in a single graph, although I do not believe that there is any particular reason to do so. If your objective is to communicate the data in the simplest, clearest and most accurate way possible, then sometimes that way will involve more than one graph. Indeed, multi-graph solutions can be very useful in addressing a variety of common data visualization challenges, beyond that of representing wide-variation data.

Discuss this Article

Share your thoughts about this article by visiting the Visualizing Wide-Variation Data thread in our discussion forum.

About the Author

Nick Desbarats, a Senior Educator and Consultant with Perceptual Edge, delivers private data visualization design training and consulting services. For over 20 years, Nick has been designing information displays that enable senior decision makers to make effective, data-driven decisions. A technology sector veteran, Nick has held senior technical, product, and marketing positions in four enterprise software companies. His visualizations have been featured in major television and online media outlets, and he is a contributor to the Perceptual Edge blog and the quarterly *Visual Business Intelligence Newsletter*. To learn more about Perceptual Edge’s work and to access an extensive library of articles, please visit www.perceptualedge.com.