PAPER PENCIL PRINCIPLE FOR DATA VISUALISATION

“Although there are no “golden rules”, one should keep in mind some basic guidelines that work most of the time. They are useful as long as we use them wisely, always having in mind that above all else, we want to show the data. So, here is a list of all the guidelines I was able to collect from different sources. You may find these referred at the end of this post.

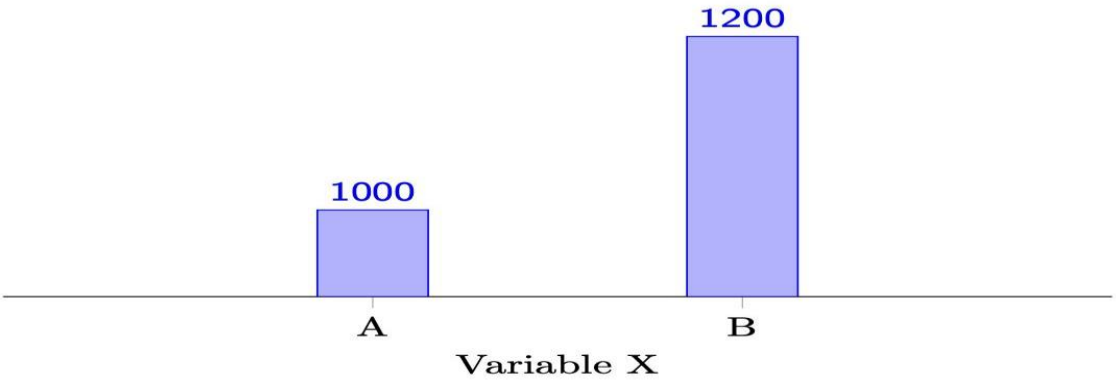
Guideline #1 – Visualizations ought to be self-explanatory. If you require the reader to go back and forth between the text and the figure, you are asking for an extra effort. Simple things help fix this. For example, having a clear title or caption is half-way to having an easy-to-read figure. Moreover, all axis should be carefully labelled and presented with the respective units of measurement.

Guideline #2 – Turn off the box around the figure. The same for the boxes connecting the axes. This is so common, and I actually find it strange that most libraries draw these boxes by default (e.g., python visualization library Matplotlib).

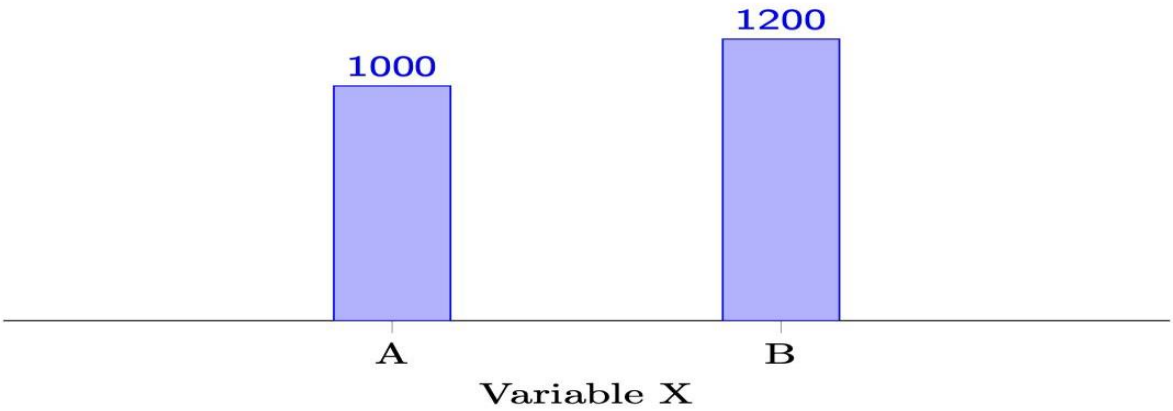
Guideline #3 – Only have one x- and one y-axis. It is very tempting to use two y-axes to compare different variables and to present data patterns from different angles. Not only it makes the visualization overwhelming and difficult to understand, but also it may inadvertently pose spurious correlations. Tyler Vigen’s work brilliantly exposes this issue, using this strategy to establish correlations between variables that are totally unrelated.

Guideline #4 – Use visual variables (color, shape, shade) only for data variation.

Guideline #5 – Axes must start at a meaningful baseline. E.g., bar charts should start at zero (most of the time). When this guideline is not followed, some data patterns will most likely be distorted. See the example in the figure below.

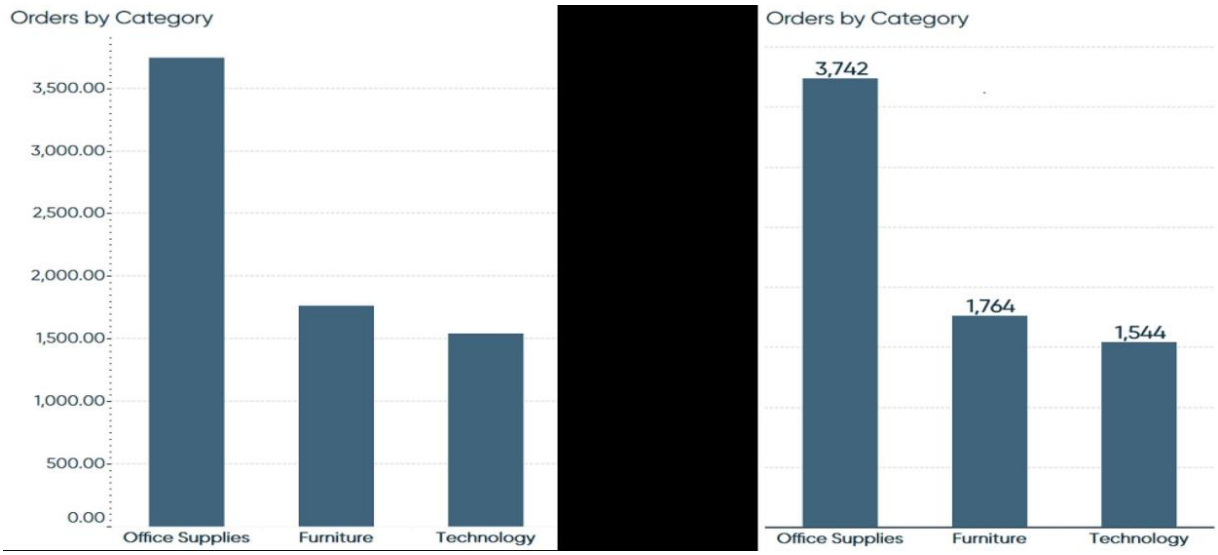


The height of the bar of B looks three times bigger than the bar of A – an increase of 300%. However, from A to B the data only increases 20% (my math: (1200-1000)/1000 * 100%). The y-axis is starting at 900 instead of starting from its natural baseline: 0. This distortion is often used to amplify results, misleading the reader. The same graph without distortion would look as follows:



Guideline #6 – Never use different colors to represent the same kind of data. It is tempting to use different colors only for aesthetic purposes. The downside is that it inherently asks our brain to understand the reason for that variation. This is a waste of energy you shouldn't ask from your readers.

Guideline #7 – Label elements directly, avoiding indirect look-up. Avoid requiring your reader to go back and forth through the different elements of your graph (and sometimes text) to understand your figure. Remember what I said about waste?… For example, avoid using a legend when you can label that information directly in the objects without making the graph more complicated. This is, of course, one of those rules with many exceptions. Never forget to be critical about guidelines! The figure below shows two bar plots: the first does not follow this guideline; the second applies this guideline by labelling the values directly on top of each bar. Which one do you think is easier to read?



Guideline #8 – Text labels should never be rotated (nor vertical). E.g., use a horizontal bar chart when category names are too long. This seems like a tiny detail, but I recommend you to try it the next time you create a visualization. Details matter.

Guideline #9 – Highlight what's important. An image is worth a thousand words. And your visualization is no exception. But you need to guide the reader to a few messages you find essential . You don't want your readers to waste energy processing irrelevant messages while overlooking the most important ones.

Guideline #10 – Use bold type/lines only to emphasize something. This one follows the previous guideline. Keep your reader focused and straight to the point.

Guideline #11 – Don't use 3D effects. I hope I don't need to explain this one.

Guideline #12 – Avoid pie charts (and donut charts). It is difficult to compare many slices in a pie chart. Very simple charts are the exception.

Guideline #13 – Sort data for easier comparisons. E.g., in a pie chart or bar chart, it is easier to compare the sizes of the different bars if they are sorted according to their size.

Guideline #14 – Don't be afraid of creating separate graphs. If your graph is getting overly complex, think of ways of dividing it in multiple graphs. A nice way of doing this is by thinking about the messages you really want to convey in the figure. Then, you need to divide those messages in two groups and illustrate these groups in separate graphs.

Guideline #15 – Use line plots only when variables are ordinal or numerical. Line plots connect data points sequentially and show something typically called trend between those data points. If the variable has no particular order but we show a trend between sequential data points, we might be sending a wrong message.

Guideline #16 – Care for colorblindness. It is estimated that colorblindness affects 8% of men worldwide. Thus, it is important to make sure that the color you use to highlight data patterns is perceptible by everyone. For most people affected by this condition, red and green are indistinguishable – use these colors with care. Sometimes it is hard to avoid using red and green, as they are typically used to denote good and bad, success and failure, etc. In these cases, you may want to pick red and green colors that differ in their saturation level. Test your images with tools like Coblis.