Dos and Don’ts of data visualization

A few sample cases to understand the importance of proper data visualization

Marco Dalla Vecchia

2024-06-12

Table of contents

Introduction

So, you have run your big analysis and you are ready to create the figures for your great paper that will surely win you the Nobel Prize? Great!

> *How hard could it be, am I right??*

Here are a few things to consider next time you make your figures for a scientific publication. Remember, these are not strict rules but gentle reminders and recommendations. As always, to every rule, there is an exception. Just use your .

Using appropriate graphs for the message

It might sound obvious, but you would be surprised to see how many cases of simple data can be overcomplicated by the wrong graph, or how often a message can be biased by choosing a type of plot that doesn't fit its own data. Let's take a few examples.

Pie-charts

There are many articles explaining why pie charts are evil, Schwabish [6] does a particularly good job of discussing both sides of when you could and must not use pie charts to visualize your data. Pie charts have many issues, and Figure 1 has pretty much all of them. Take a look at Figure 1; here we are representing the population of each country in Europe as a fraction of the total population (i.e., the sum of all the pie slices equals the total European population). It's pretty clear how, in both sides of the figure, **there are way too many slices, too many labels, and too many colors**. The slices are hard to compare, and there are slices that are too different from each other. The pie chart on the right side is a little easier to read because it has been descendently sorted by population size but still has too many data points.

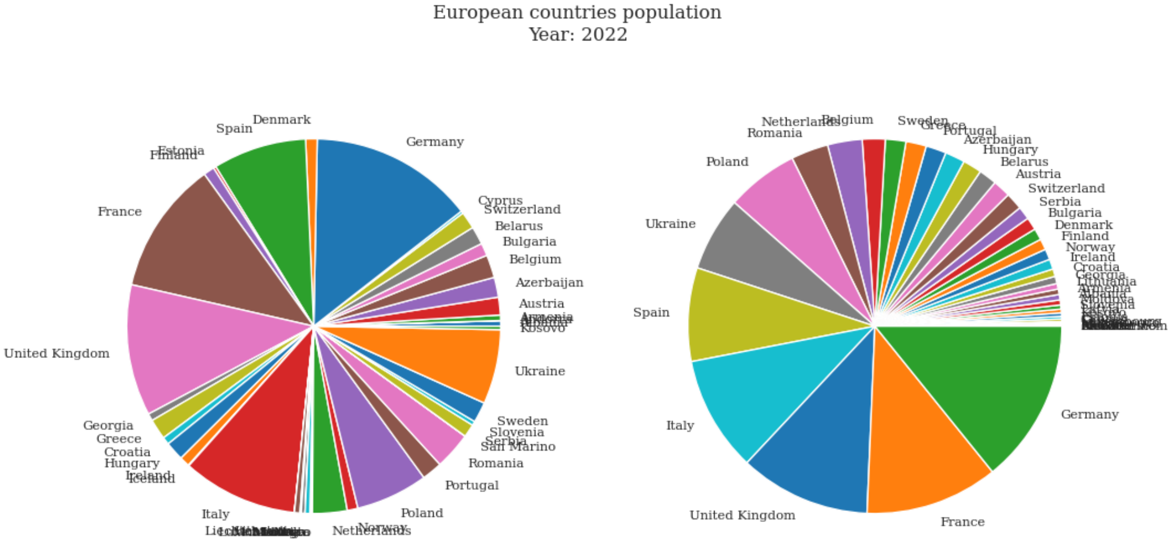

Figure 1: Pie charts for the visualization of european countries population

I know, it might sounds silly to display all that data in a pie chart, but Figure 1 still shows the point of how simple data can be incredibly overcomplicated with the wrong chart type. Luckily, pie charts are not often used in scientific publications.

Bar plots

Imagine displaying the same data in a bar plot like in Figure 2 instead. Here we can appreciate the data much better. There are still a lot of countries, but it's not as overwhelming; visually, the contrast between bars makes the interpretation much easier and there is no need for different colors. There are always good and bad ways to create the same figure, for example you can appreciate how in the right panel of Figure 2 comparing data is much harder simply because the data **is sorted alphabetically instead of numerically**.

But please remember that **bar plots are meant to display single values**. As we will see in Figure 10, using bars for showing grouped mean values, is a quick way to misleading data visualization.

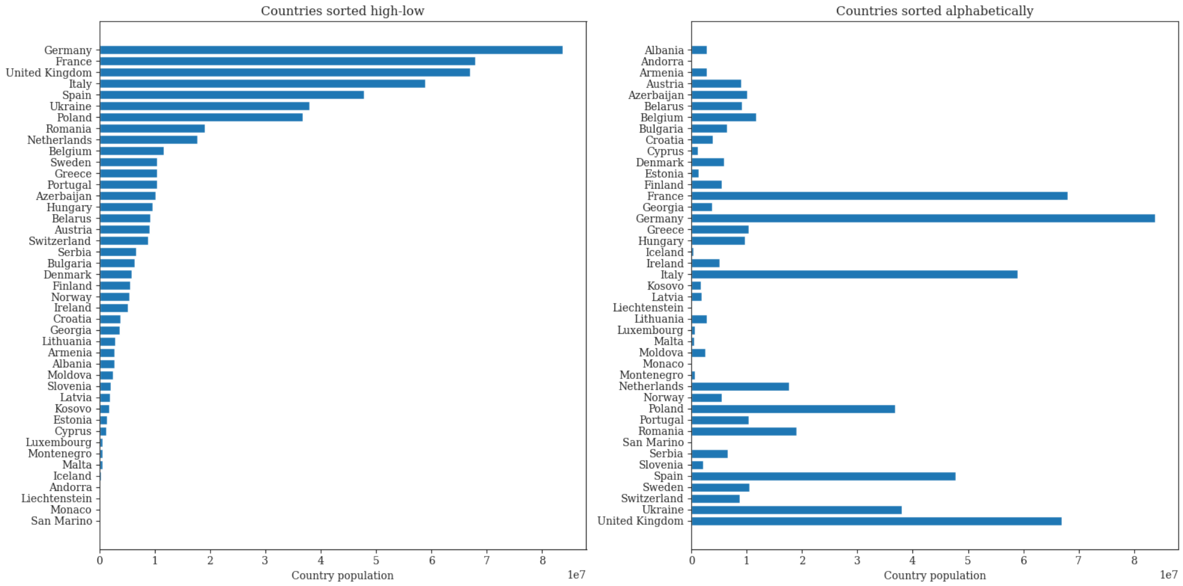

Figure 2: Bar charts for the visualization of european countries population

Heat map

Heatmaps are quite useful, especially when you have two-dimensional data to display and you have many categories for each. They are often used in scientific publications, but they can be tricky to interpret as well. Look at Figure 3, for example: it seems to be pretty straightforward at first, but how easily can you appreciate the **numerical difference between countries' populations or between the population of a country** throughout the years? Notice how having **a common scale across all countries** makes subtle differences in some countries less noticeable and how hard it is to have an exact estimation of the population.

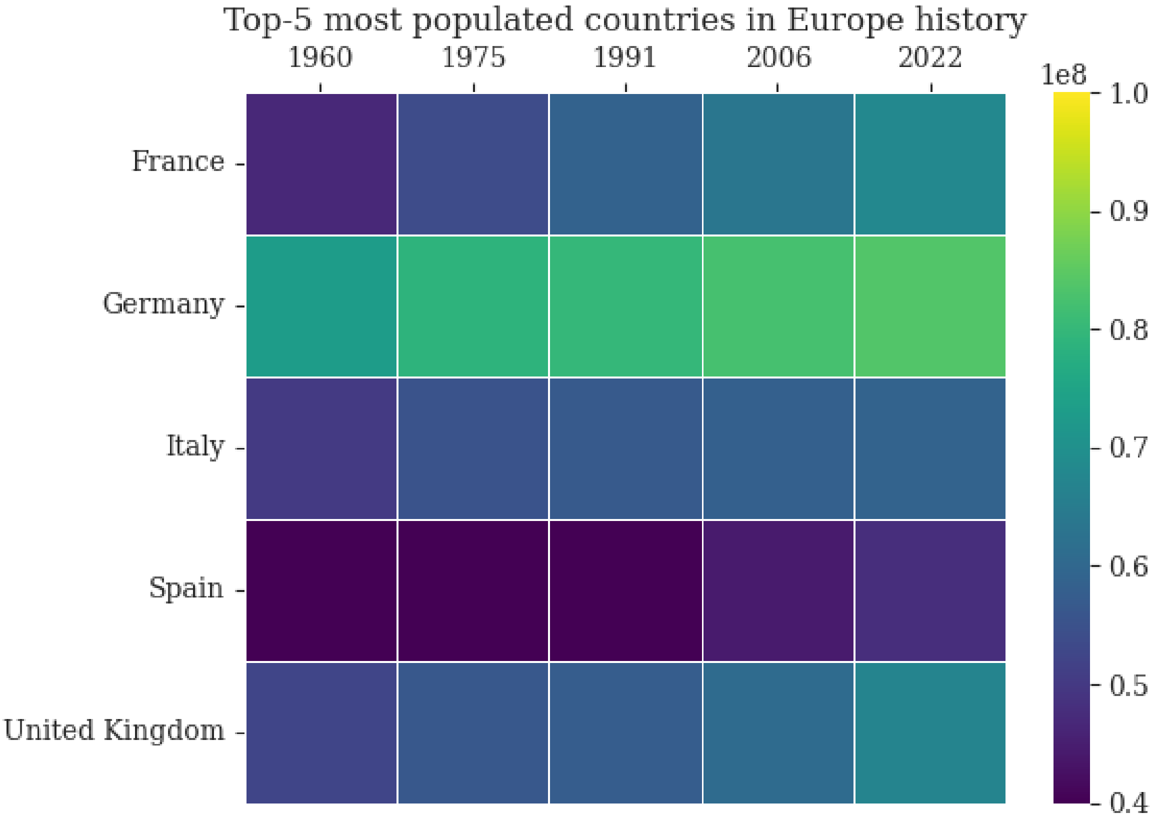

Figure 3: Heatmap for the visualization of european countries population in different years

Time series graph

In contrast with pie charts, it's quite common to find time series graphs in publications. However, displaying temporal data can still be challenging depending on the type of data, time resolution and quantity of data to display. Let's take a look at Figure 4, here we are displaying population in a few countries over time like we did in Figure 3, but this time in a bar plot format. You might appreciate how this arrangement favors more the **comparison between countries in each year rather than the population evolution across years**. Also, in my opinion, the number and colors of the bars are cluttering and needlessly complicating the figure.

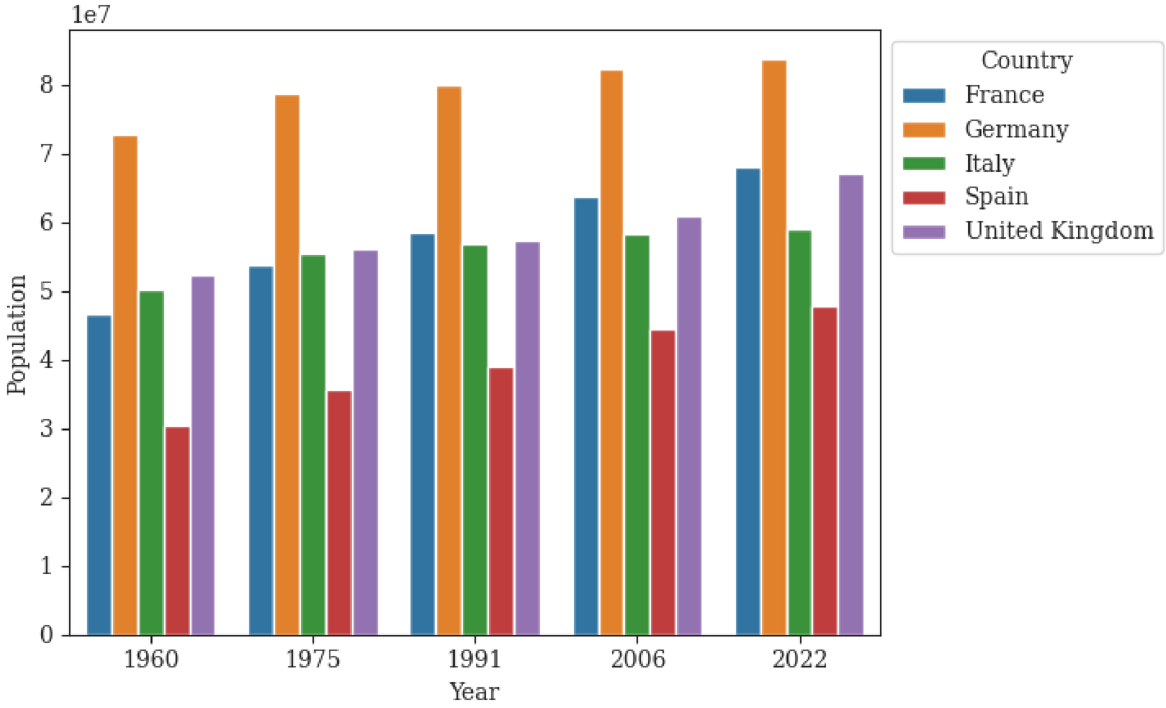

Figure 4: Barplot for the visualization of european countries population in different years

Let's compare exactly the same data displayed in Figure 5 in a line graph. In this case, the emphasis is more on the growth of the population of each country, but comparing across countries is also not too difficult. Notice how, in this case, **the use of color helps distinguish the lines and how replacing a legend with the direct labeling** of the data lines makes it much quicker to read the graph.

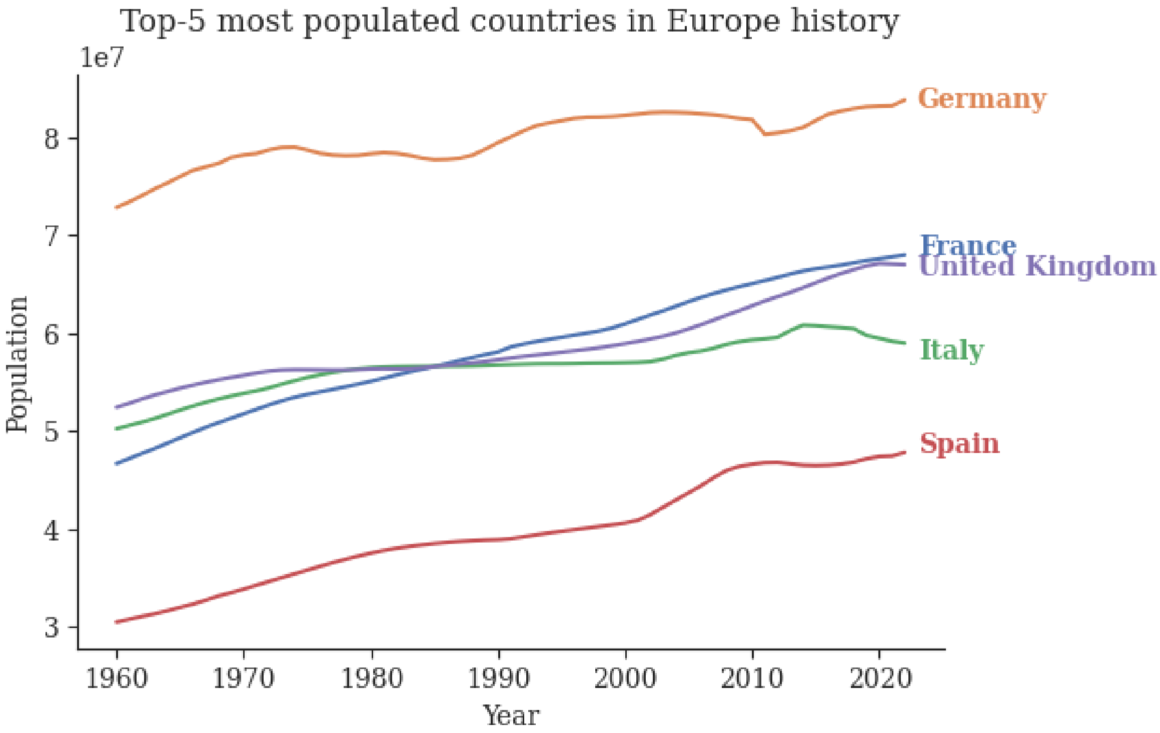

Figure 5: Lineplot for the visualization of european countries population in different years

Truncated axis (i.e. start your axis at zero)

A classic potential *'mistake'* in data visualization is the incorrect usage of scaling. For example in Figure 6, due to the nature of how we read bars in a bar plot, **we tend to compare bars amongst each other**, rather than comparing each bar with the Y-axis scale (or origin). For this reason, you can appreciate how the difference between the grades of our three (potentially randomly named) students look much different on the left side of Figure 6 than on the right side. This is risky! Whenever you can tweak the message simply by adjusting the Y-axis origin, it's a sign of bad data representation.

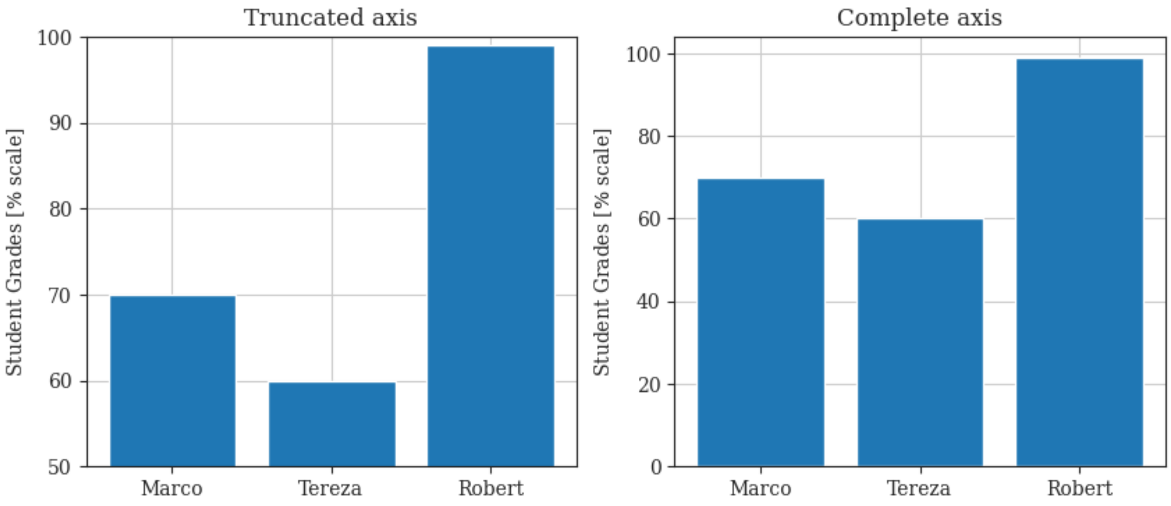

Figure 6: Comparing barplots with truncated or complete Y-axis.

There are exceptions however; notice how in Figure 7, having the truncated Y-axis on the left side helps to describe the difference between grades of the students across the year compared to the right side. Because we tend to read line plots as evolution across time and not so much as difference between the lines, we could argue that using a truncated axis in this case is acceptable.

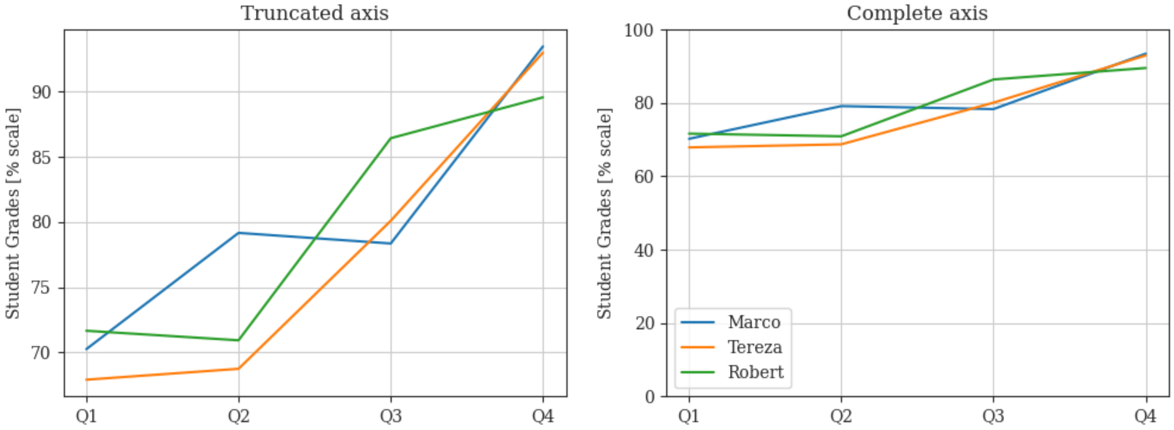

Figure 7: Comparing lineplots with truncated or complete Y-axis.

Comparing values of different scales

Broken axis

In Section , we have seen that changing the scaling by adjusting the origin of the axis of a figure can change the message a figure is displaying. In Figure 8, we can see population data across three different countries with vastly different scales. The issue here is that China's population is several factors higher compared to Italy's or Austria's population. In Figure 8, I showed three different ways to deal with data on such different scales. In the left panel, we can appreciate the raw data as it comes; it's a fair representation but it really makes it difficult to compare the countries with each other. A common solution is to use a **broken axis**, as displayed in Figure 8's right panel. Here, the Y-axis scale is not continuous but simply clipped in a way that data between 100 million and 1300 million people is removed from the graph. Although visually the comparison between bars is now much easier, this bar plot can be very misleading: at first sight, China's population appears to be a factor of 3 higher rather than a factor of ~20 in reality! Sure, the broken Y-axis is quite clear, but it's one step closer to misinterpreting the message of the figure. A much better way to deal with data of different scales is the middle panel in Figure 8, where the simple usage of a logarithmic scale on the Y-axis (log scale) makes the comparison between bars easier and fairer.

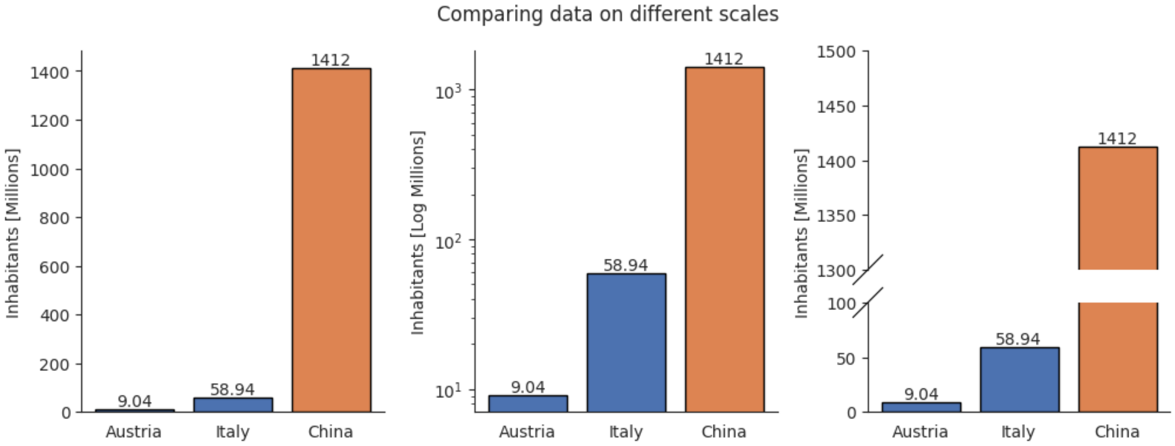

Figure 8: Comparing values of different scales

Double axis

Depending on the type and amount of data to be displayed, we can be tempted to use double axes in our figures. In Figure 9, we can see an example of how Austria's population is really on a whole different scale than the population of China. Although this is certainly the case, we could easily make it look like they are almost identical by using two independent Y-axes with different scales! Sure, it allows a closer comparison of the data between countries, but it can also mislead the reader into thinking that the data scales are very close to each other.

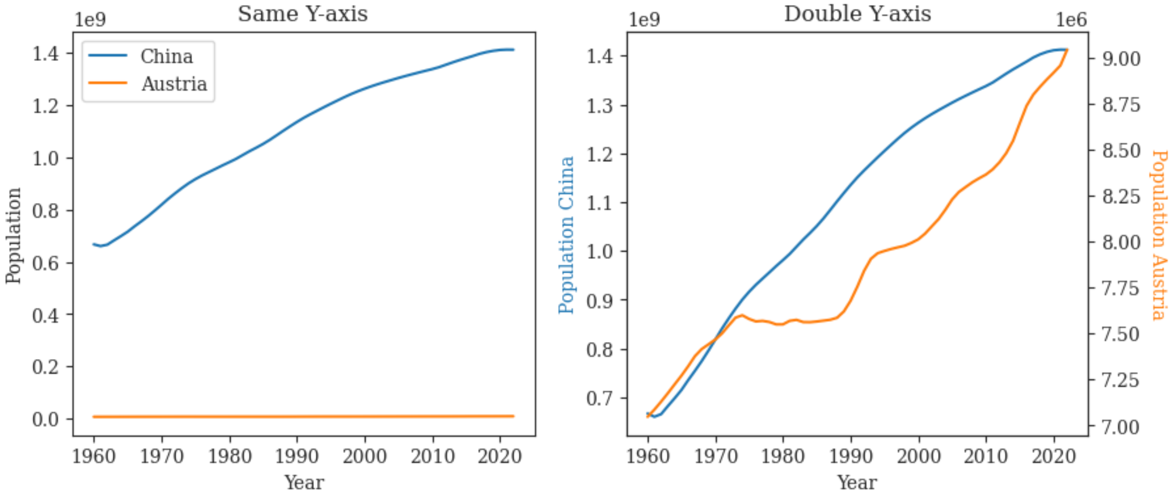

Figure 9: Visual effect of using double Y-axis on data of different scales.

I hope it's clear how you can tell different stories with your figures simply by adjusting the corresponding scales of the two Y-axes. Notice how I tried to make things clearer by coloring the Y-axis labels with the same color as the data lines.

Mean separation

Li [3] has collected an ever-growing list of bad visualization cases, but one particular one that we see over and over in scientific literature, is the the habit of comparing and visualizing sample means with each other.

In Figure 10, I replicated Li [3] 's original figure in Python to show this issue. On the left panel, we are using bars (which should be used only for single values) to represent the mean of a response for two groups of experimental treatments. Although adding the error bars makes the bar plot a little more accurate, it's a representation to avoid, mostly because it does not show how the underlying data of the two groups really behaves! **Always be careful to simplifying a whole population down to its mean and standard deviation**.

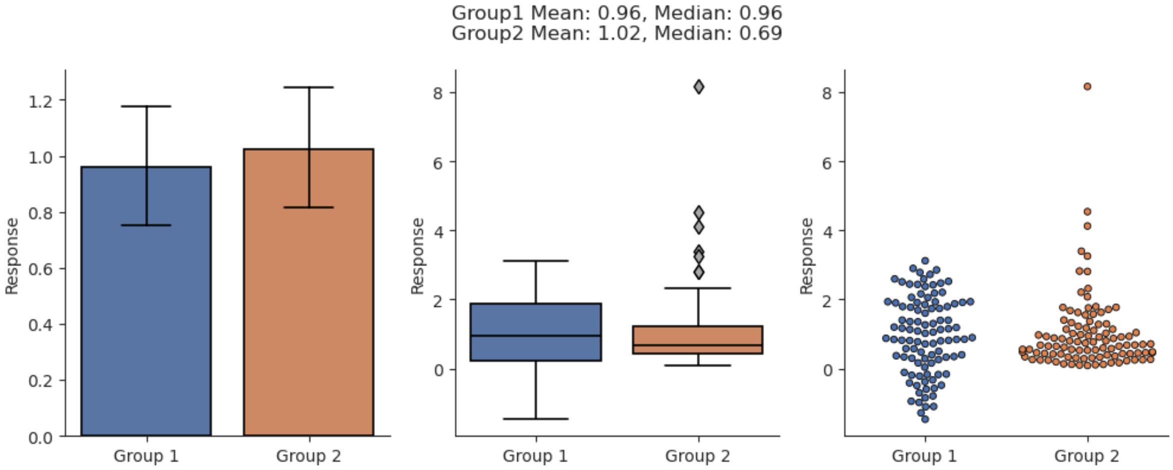

Figure 10: Replica of mean separation figure by Li [3]

In Figure 10 middle panel, we can appreciate how the usage of boxplots at least highlights slightly different median values, different data spreads (look at the interquantile range (IQR)) and shows that in the second group there are a few outliers that behave differently to the rest of the group.

The right panel of Figure 10, finally reveals the real difference between groups, which, despite having very similar means and standard deviations, actually show to be dramatically different.

When comparing and displaying mean values by group, try to always show the underlying data points and pay attention when boiling data down to mean and standard deviation.

SuperPlots

In this section we go one step further to Section and take a look at a replica I created of Figure 1 of Lord et al. [4] below. As we can see from Figure 11, Lord et al. [4] not only highlights that using average bar plot to display your groupss data is not correct but it criticizes how simple mean-comparing tests (such as t-tests) are often misused by considering the sample size as large as all the data points that were acquired.

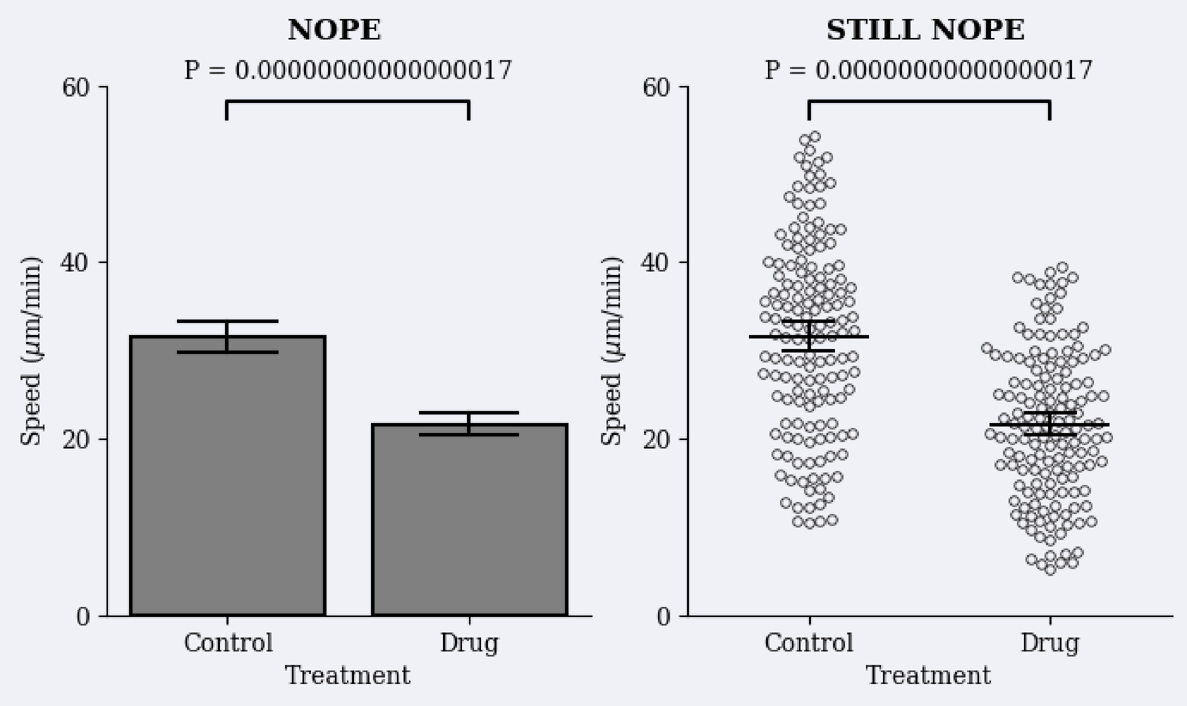

Figure 11: Replica of bad figure from Lord et al. [4]

Imagine comparing the average response of many cells (quantified a migration speed difference) between two treatments. We have repeated this experiment three times, which is (for some reason) classically considered the standard value for biological replicates. A common mistake is to treat each single cell as a separate independent measurement (i.e. $N = 100$), leading to **wrong and incredibly low p-values**!

What we should really be doing is to consider each independent experiment as a separate measurement as thus fairly comparing the data between treatments ($N = 3$) like shown in the right panel in Figure 12.

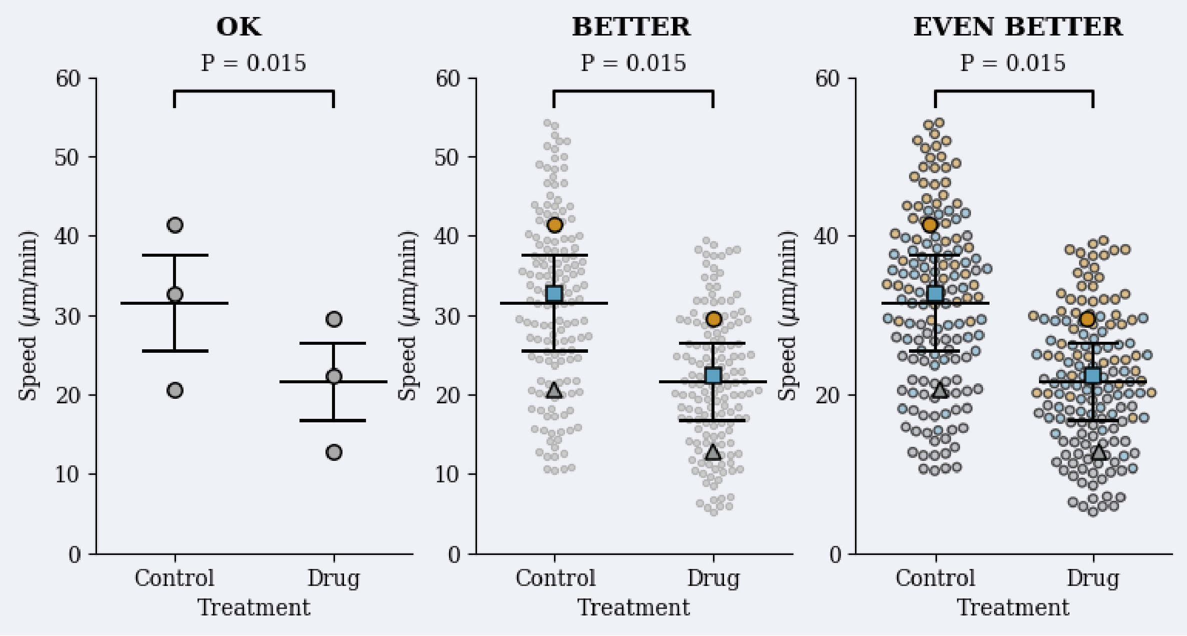

Figure 12: Replica of good figure from Lord et al. [4]

But we can do much better! Lord et al. [4] coined the concept of **SuperPlot**, which clearly showcases every single data point, grouped average and fair comparion of biological variability across measurements. In Figure 12 middle panel we are clearly showing the three independent measurements and on the right panel we are even providing information on which cell was measured in which experiment.

Here I only replicated the middle panel of Figure 1 of Lord et al. [4] (beause that was the only available dataset), but check the entire publication to understand the correct usage of SuperPlots and how they allow for better clarity and evaluation of variability and reproducibility of your data!

Do you want to create your own **SuperPlot**? Here are a few resources:

- Check out the original publication Lord et al. [4]: it contains examples on how to do simple SuperPlots in Excel, GraphPad Prism, R and Python
- Check out the source code for this article here
- Use this online tool by Joachim Goedhart

If you are confused about classical statistical testing, the famous N, and p-values, I cannot recommend enough Royle [5], in which he tackles everything from proper experimental design to appropriate data management and analysis.

Data visualization reading recommendations

Check out what's available in the ISTA library! For example:

- Kirk [1]
- Knaflic [2]
- Schwabish [6]
- Royle [5] → great book, strongly recommended book for all beginning PhD students (especially if will be doing Microscopy and Cell Biology)

I already cited Li [3] for Figure 10, but check out the whole series on GitHub.

Source code

Do you want to replicate the figures I show in this article? Do you want to reuse this material? Head over to the IOF GitLab repository, but make sure to cite it correctly following the license guideline.

Python plotting libraries comparison

Check out this other article for an extensive comparison of Pyton plotting libraries here.

References

[1] Andy Kirk. *Data visualisation: a handbook for data driven design*. SAGE, 2016.

[2] Cole Nussbaumer Knaflic. *Storytelling with data: a data visualization guide for business professionals*. Wiley, 2015.

[3] C. Li. *cxli233/FriendsDontLetFriends: FriendsDontLetFriends (v6.2)*. Version v6.2. Mar. 2024. DOI: 10.5281/zenodo.10802096. URL: https://doi.org/10.5281/zenodo.10802096.

[4] Samuel J. Lord et al. "SuperPlots: Communicating reproducibility and variability in cell biology". en. In: *Journal of Cell Biology* 219.6 (June 2020). ISSN: 0021-9525. DOI: 10.1083/jcb.202001064. URL: https://dx.doi.org/10.1083/jcb.202001064 (visited on 05/16/2024).

[5] Stephen J. Royle. *The digital cell: cell biology as a data science*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 2019.

[6] Jonathan A. Schwabish. *Better data visualizations*. Columbia University Press, 2021.