sciendo

Social Communication
Volume 2 (2019), pp. 11-20
DOI: 10.2478/sc-2019-0005

DISPLAYING LABELED QUANTITATIVE DATA

Marcin Kozak[1]

Abstract

The information world is full of labeled quantitative data, in which a number of qualitative categories are to be compared based on a quantitative variable. Their graphical representations are various and serve different audiences and purposes. Based on a simple data set and its different visualizations, we will play with the data and their visual representation. We will use well-known charts, such as a regular table, a bar plot, and a word cloud; less-know, such as Cleveland's dot plot, a fan plot, and a text-table; and new ones, constructed for the very aim of this essay, such as a labeled rectangle plot and a ruler-like graph. Our discussion will not aim to choose the best graph but rather to show the different faces of visualizing labeled quantitative data. I hope to convince the readers that it is always worth spending a minute on pondering how to present their data.

Key words: visualization, bar plot, labeled rectangle plot, ruler-like graph

INTRODUCTION

Scientific and popular science literature, mass media newspapers and magazines, and the Internet exploit various data using various representations. Such data can represent practically all areas, like economy, politics, biology; be of various size, from very small (three or four, and sometimes even two values) to so-called big data (e.g., of billions of values); have various aims, from presenting several values to inform or entertain, through exploring data to learn more about the underlying processes, to complex scientific analysis; and so on.

Data can vary, and their representations can vary as well. Sometimes it will be a sentence giving several values. It can be a table—a small one, with just several values; or of medium size, covering half a page; or quite a big one, like those in statistical tables published by official statistics agencies. It can be a simple graphic, like a pie chart showing proportions or a bar chart showing raw values (like costs) or a line plot showing trends over time, etc. It can be a complex graphic, showing many values. It can be a map. It can be a hybrid design, like a text-table or an infographic, the latter aiming to present sometimes quite complex information in an interesting and eye-catching way.

Here, I wish to invite you to join me in a discussion on one particular situation: graphical displaying one-way labeled data. Such data are grouped by one factor, and so such displays graph several quantities representing several groups. Scientific examples are commonplace. For mass media, two prominent examples are official statistics data (e.g., the unemployment rate in the EU countries) and opinion polls. When discussing

1 University of Information Technology and Management in Rzeszow
ORCID: https://orcid.org/0000-0001-9653-3108, e-mail: mkozak@wsiz.rzeszow.pl

such data, I will use two standard terms, typically used for one-way data: a variable (the unemployment rate in the example above) and groups (the EU countries). It will not be a scientific discussion aiming to choose the best type of visual representation of the data we will use. Instead, we will discuss various existing tools and propose new ones, not necessarily effective in terms of visual perception, but maybe effective in fulfilling other aims a graph's creator can have.

AIMS OF VISUALIZATION

What is the best way of graphing such data? In fact, is there any single best technique for one-way data?

To answer these questions, we should first try to address another question: How would we define the aims of displaying one-way labeled data?

The answer seems easy: To show the data. But this answer is way too general. Being more inquisitive, we might say: To facilitate group-to-group comparison. True, but how should we approach such a comparison? You can make a direct value-to-value comparison (this candidate had 54% of votes and thus beat the other one who had 13% of votes...) or a relative comparison (... is about four times as large). An appropriate graphical layout can facilitate seeing the groups in order of the characteristic of interest, but it can also facilitate the quick localization of a particular group, for example thanks to an alphabetical ordering. A graphical layout can also facilitate a table look-up of values (for example, by adding numeric labels to the graphical elements representing the corresponding values, or by adding a grid). For skewed data, the comparison can be simplified by transforming a numeric variable, for example using logarithms.

Note, however, that we are discussing a particular situation here: We have one variable and several groups, and we do not need to consider any error or uncertainty in the data. If we did, we would also have had to discuss how to include such additional information on a graph.

But outside of the scientific realm, graphs often aim to draw the attention of potential readers. For instance, glancing through a magazine, the readers having a number of texts to choose from might single out one that catches their eyes, and a fancy graph could be this eye-catching detail. An attractive graph in this context is not necessarily easy to read and, actually, does not have to convey the message, but it does have to direct the reader's attention. Multiple pie chart offers one example of such a graph [Kozak et al. 2015], but Tufte [1983] offers more examples.

Sometimes, it is important to help the viewer remember the data. But note that to remember the data does not mean to remember the graph. While helping the viewer remember the data can be difficult, helping him or her remember the graph can be simple: replace the bars in a bar plot with smiling devil faces of appropriate heights, or, even better, put in the background a fabricated photograph of a naked famous politician, preferably with added big hairy belly. Many will remember such graphs, but who will remember the data they showed? Borkin et al. [2015], in their fascinating study on the memorability of visualizations, observed that well-designed pictograms helped the viewers to remember a visualization—so maybe it is not true that the only thing one would remember is the belly and not the data? Without devoted research, however, it is difficult to say. Bateman et al. [2010] opposites the claim that all such "visual junk" is junk indeed. The experiment they provided showed that graphs with such visual junk were not read with less accuracy than were their plain alternatives—and that they were better recalled!

Again: The point is to make a graph in such a way that your aims are fulfilled. Some of these aims can be achieved at cost of other aims. For instance, attention can be attracted at a cost of making the graph difficult to read, by adding a lot of color that results in undesirable clutter; or the graph can be made easy to read at cost of boring the audience.

We could list more aims here, but there is no point in multiplying them. The point is that every graphic lives its own life, has its own aims, and is directed to a specific—broader or wider—audience. So, let us forget about a dream of specifying one general set of rules for graphing labeled quantitative data (let alone visualization in general). Instead, let us play with an example data set and use it to discuss what might be the best way to display such data. We can use such a study to shed some light on a broader context of data visualization.

A DATA SET: VISUALIZING NUMBERS OF THREATENED SPECIES IN VARIOUS AREAS OF THE WORLD

We will use a small data set (Table 1) made available in the help for the fan.plotfunction from the plotrix package[Lemon 2006] of R[2019], used by Lemon and Tyagi [2009]. The data show the numbers of threatened species in six areas of the world, estimated by the International Union for Conservation of Nature and Natural Resources (IUCN).

Table 1. The number of threatened species per continent. Data come from Lemon and Tyagi [2009].

Continent	No. of threatened species
Asia	7737
Africa	5994
Europe	1987
N&C America	4716
Oceania	2093
S America	5097

Assum that we want to graph the data to enable the viewer to compare the values both exactly and relatively. Displaying such a simple data set is usually no problem if the number of groups is not too big, like here. Is it indeed that simple?

Figure 1 shows a bar plot Lemon and Tyagi [2009] used to claim that this type of graph did not work well with these data. I agree with them—but we will get back to the bar plot later. First, let us consider what many users would likely choose to graph the data: the pie chart (Figure 2). While many people think of pie charts as an effective technique for displaying the relative sizes of various data values, many others disagree. It is sufficient to recall crude words that the help page for the pie function of R uses: “Pie charts are a very bad way of displaying information.” Since the pie chart does not help compare values relatively, it does not suit our aim, and so we need another type of display.

A lot has been written about the choice of visual elements on graphs [Tufte 1983, Few 2006, Kozak 2009], and Lemon and Tyagi’s bar plot does not follow these rules. I am not saying that one should always follow such rules—it all depends on the situation. But such was the reason behind Lemon and Tyagi’s [2009] proposal of the fan plot: Using the bar plot similar that in Figure 1, they claimed that the bar plot was inefficient to present the data from Table 1, hence the need for a new graph type.

The fan plot aims to display one-way labeled data with a dual purpose of facilitating both direct and relative comparisons of values. The technique is offered in the function `fan.plot` in R’s plotrix package [Lemon 2006].

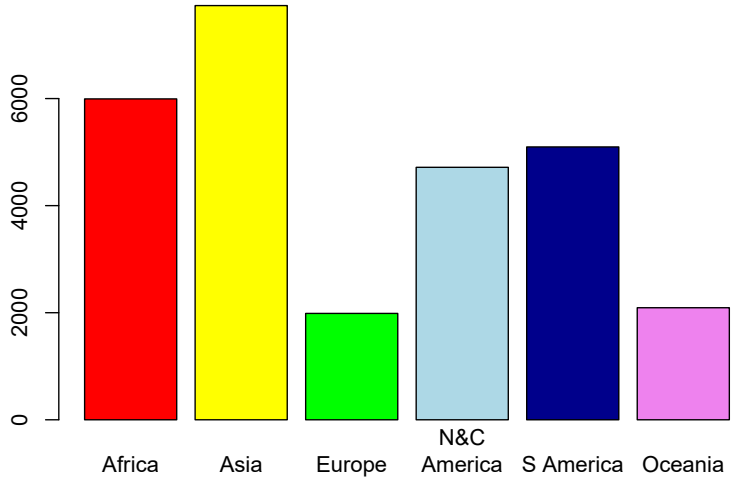

Figure 1. The bar plot showing the data from Table 1, similar to the one used by Lemon and Tyagi [2009].

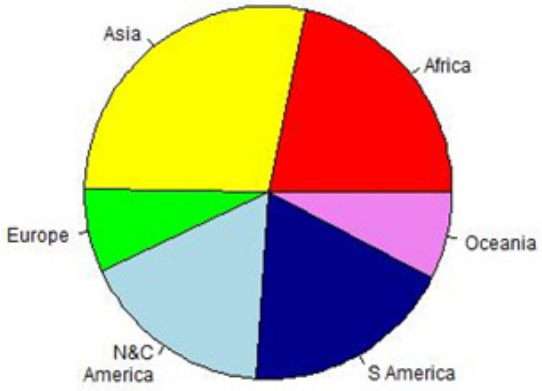

Figure 2. The pie chart showing the same data as does Figure 1.

Theoretically, the idea behind the fan plot is to show the exact values for the groups (like "Africa has around six thousand threatened species") and facilitate comparing the groups in terms of the relative values (like "N&C America has over twice as many threatened species as does Oceania"). It might be a subjective thing, but I have problems with both these aspects while reading Figure 3. Reading raw values is easy indeed, but not thanks to the graph's design but thanks to the data labels directly shown on the graph. Assessing relative differences is easy only in reference to Asia, which had the biggest number of threatened species and hence—because of how the fan plot is constructed—served as the visual benchmark for the other regions. Analyzing most (though not all) of the other relative differences is difficult.

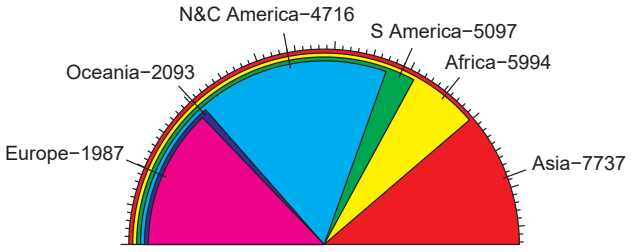

Figure 3. The fan plot for the studied data.

So, I think Table 1 is at least as effective in displaying the data as the fan plot, although the fan plot can be considered much more attractive. But there must be other visualization techniques which would do better than both table and fan plot. So, let us try to play with the data. In doing so, we should take account of the various contexts in which such data might be graphed.

PLAYING WITH THE DATA

Let us try to improve Table 1, using a text-table [Tufte 1983, Kozak 2007]. So, the number of threatened species per continent is:

```
Asia          7,737
Africa        5,994
S America     5,097
N&C America   4,716
Oceania       2,093
Europe        1,987
```

This tabular display exploits visual effects to make the data easier to read and compare, like the monospaced font face and the order of groups by the value. Many readers would consider this text-table more effective than Table 1. We might try to make it even easier for visual comparison, like here [Kozak and Krzanowski 2010]:

```
Asia          7737
Africa        5994
S America     5097
N&C America   4716
Oceania       2093
Europe        1987
```

Let us move on to a graphical comparison of the number of threatened species per continent. Lemon and Tyagi claim that the bar plot (Figure 1) does not work here [Lemon and Tyagi 2009]. I agree the *bar plotconstructed that way* does not work, but generally bar plots can do quite a nice job to present such data. Few [2006] writes, "The truth is, I never recommend the use of pie charts... Bar graphs are a much better way to display this information." Study Figure 4. Still a bar plot, it differs from that in Figure 1 in terms of three aspects. Firstly, the order by the number of threatened species is better for the group-to-group comparison than is the alphabetical order. Secondly, all the bars are red—what purpose does differentiating them by color serve? Thirdly, why not use horizontal instead of vertical version of the graph? Less often used, the horizontal version of the bar plot makes it easier to visually compare the bars' lengths (representing the groups' sizes); reading horizontal labels is easier, too, especially (unlike here) long ones.

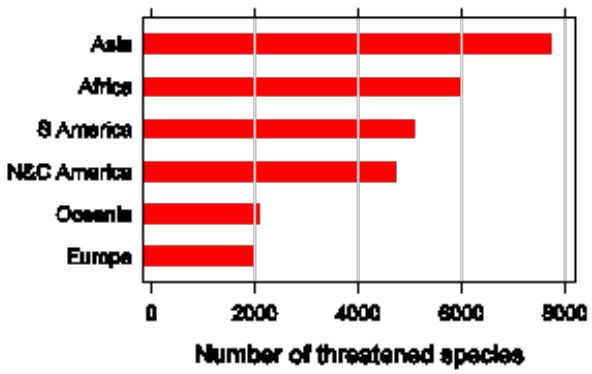

Figure 4: A horizontal bar plot for the data studied.

Maybe Cleveland's dot plot (Figure 5) can do better [Cleveland 1994]? In its basic version, the quantitative axis does not include zero, and the lines do not join the label with the dot. In Figure 5, however, these two elements facilitate grasping relative between-group differences [Jacoby 2006]. A choice for a scientific publication, this graph is not particularly attractive for nonscientific outlets. In science we do not need to draw the readers' attention to the graphic, but in other outlets we often want to. In fact, Figures 3 and 4 were much more attractive than is Figure 5, but they were so at cost of difficult interpretation.

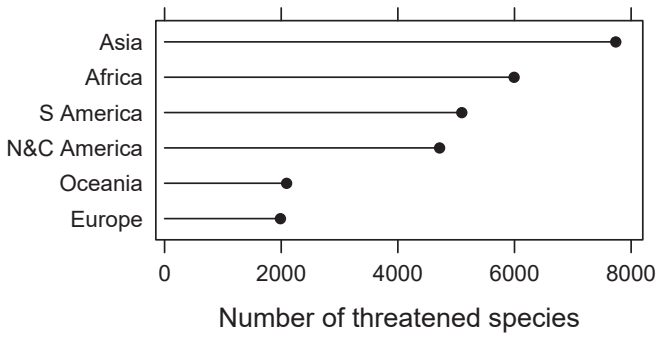

Figure 5: Cleveland's dot plot for the data studied.

Figure 6 shows a different version of the dot plot. Instead of putting the group names at the y-axis, I put them within the data rectangle. So, I ignored Cleveland's advice to put only necessary text within the data rectangle [Cleveland 1994]. Red points together with the labels (which are in monospaced font on purpose—I consider it a nice visual effect) make the plot look differently—not so "scientific"—from the one in Figure 5. I had to deal with the difficulties with the relative assessment of the groups' sizes, hence the rug (the short lines added to the horizontal axis, representing the exact group values). This graphic may attract attention, also by being atypical, and yet it seems to convey the desirable information in an easy to understand form.

Contemplating the fan plot, I came up with an idea of a similar plot which would be free of the drawbacks of the fan plot. See Figure 7 for the *labeled rectangle plot*. I have problems believing this is a novel type of graph, but I donott remember seeing it anywhere else.

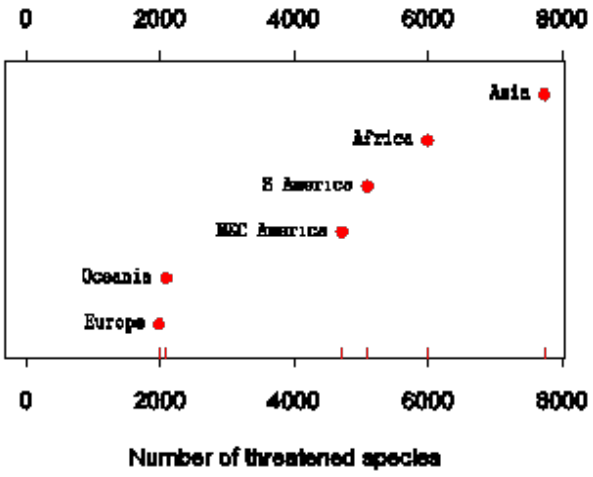

Figure 6: A modified version of Cleveland's dot plot for the studied data.

The idea behind it is simple: Construct a bar plot and join the bars so that they are superimposed one on another; the highest bar will be the widest while the smallest will be the narrowest. My first try with the labeled rectangle plot consisted of squares instead of rectangles, but it wasted too much space and looked a little clumsy. Even worse, seeing squares, the viewer might, even subconsciously, compare bar areas instead of heights. The use of rectangles is free from this issue. Because all the bars are next to the y-axis, a table look-up of values is as easy as in the regular bar plot, if not easier (cf. Figure 4). Assessing a ratio for any pair of values is simple. The graph attracts attention by the use of colors. It should also be easy to understand for most readers: Its similarity to the classical bar plot along with the numerical vertical axis should be enough for them to see that bars' heights—not their area—are to be compared.

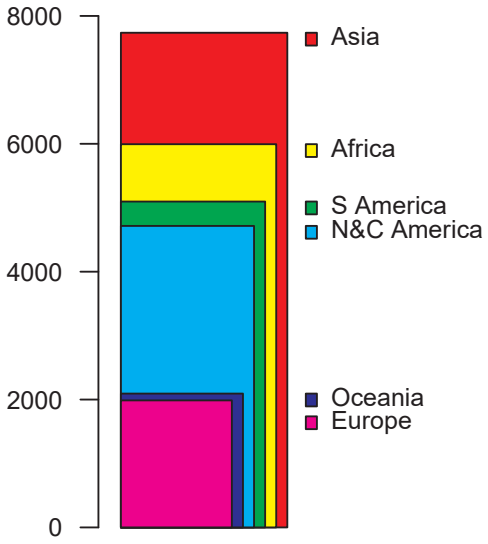

Figure 7: The labeled rectangle plot for the studied data.

Figure 8 shows the *ruler-like graph*. (Unfortunately, I cannot use the term "ruler graph" because it's been used in other scenarios.) Is it better than the previous graphs? It depends on the criteria, but in general it's not. The readers should encounter no problems with reading the raw values and comparing their relative sizes. Its main advantage is its design, which is likely to attract the readers' attention, likely much more than any other graph we have discussed so far.

Figure 8: The ruler-like plot for the data studied.

Figures 9 and 10 show two versions of a data cloud, also called a word cloud or a tag cloud. Of course, because of their design, data clouds do not convey any of the messages we have dealt with in this paper—one will be able to neither read the values nor assess their relative sizes. But this visual technique is so effective in showing where the number of threatened species is the biggest that I could not resist mentioning it.

Figure 9: The data could for the studied data.

Figure 10: Another version of the data could for the studied data.

Conclusion

Bar plots, pie charts, dot plots, fan plots, labeled rectangle plots, ruler-like plots, data clouds, and other types of graphs can be used to visualize one-way labeled data—depending on the type of data, the number of groups and their grouping, the audience, the aims, and the imagination of the graph's author. In this particular situation, we might also present the data on a map showing the numbers of threatened species on the continents. This would be a different type of visualization, difficult to compare with those we have discussed: It would aim to show the geographical distribution of the problem.

We could create many more types and versions of graphs, which would make sense in practice but not necessarily in this paper. This is because I hope I have made my point: Consider your data, consider your aims, consider your audience and the other circumstances, and make the plot that best fits this triad. Indeed, we could create many more graphs, and some of them might be better than those we worked with. Sometimes one type of display will do better; other times, another will. In certain situations, even a sentence providing several numbers can do better than any tabular or graphical display. It all depends on the context.

The point is, always be aware of what you are doing, and be ready for your work to be criticized; in the real world, *the objectively best data display does not exist*.

We have discussed both standard and non-standard types of display. Wherever I could, I referred to scientific evidence, for instance, from user studies conducted by Cleveland and his colleagues. Obviously, there can be no scientific evidence for newly created types of display, like the ruler-like graph. What is more, not always should we care about the perception effectiveness of the display tools we are considering. As already mentioned, sometimes we do not aim to provide the clearest picture of the data; instead, we might wish to direct the readers' attention or to convince them. (And to convince does not have to mean to show; sometimes to convince you have to *hide*; see Huff [1954])

So, this paperis not—and it is not *for purpose*—a valid scientific exploration of various types of display for such data. Instead, we can treat it threefold.

First, we brainstormed displaying one-way labeled data. Ideas gained that way may constitute a material for scientific research. Here, for instance, we might wish to pursue with conducting a user study comparing the fan plot with the bar plot or with developing the labeled rectangle plot.

Secondly, we can move outside of science and consider this brainstorm a way of finding the best type of display in a particular situation. (You may find a similar—though a much shorter one and in a different context—brainstorm here: https://www.datarevelations.com/tag/lollipop.) Imagine you are a visualization expert in a weekly magazine, and you are to support an author working on a text on threatened species on the globe. She gives you the data from Table 1 and asks to graph them. So, what you do, is brainstorm the problem, a process that should account for the magazine's audience, the article's level of formality, the text's style, and other circumstances. You are unlikely to follow the direct path I showed above, but you might do a similar thinking process, which would eventually lead you to the very best display of the data. Depending on the situation, you could either do it yourself or share the result with the author once you're done, or you could include the author in the brainstorm, often—though not always—a more effective approach. In the former approach, if the author does not like the display you proposed, you would have to iterate the process, striving for both perfection and her acceptance. In the latter, you two being in constant disagreement might make it impossible to reach the final solution that would satisfy the both of you.

Finally, I have shown a few new possibilities of graphing such data. Since we have only discussed but not studied them, these proposals can be further explored in in-depth, experimental studies. Who knows, maybe some of the ideas will be found valuable enough to become part of visualization toolkit? Some of them may become of more practical use, such as in mass media. So, feel free to take up these ideas and study them in detail.

References

BATEMAN, S., MANDRYK, R. L., GUTWIN, C., GENEST, A., MCDINE, D., & BROOKS, C. (2010), Useful junk?: the effects of visual embellishment on comprehension and memorability of charts.Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 2573-2582.

BORKIN, M. A., BYLINSKII, Z., KIM, N. W., BAINBRIDGE, C. M., YEH, C. S., BORKIN, D., PFISTER, H. OLIVA, A. (2015), Beyond memorability: Visualization recognition and recall.IEEE Transactions on Visualization and Computer Graphics, Vol. 22, No. 1, pp. 519-528.

CLEVELAND W.S. (1994), The elements of graphing data, 2nd ed. Summit, NJ: Hobart, USA.

FEW, S. (2006), Information dashboard design. The Effective Visual Communication of Data, Sebastopol: O'Reilly Media.

HUFF D. (1954), How to lie with statistics, W. W. Norton & Company.

JACOBY W. (2006), The dot plot: A graphical display for labeled quantitative values, The Political Methodologist Vol. 14, No. 1, pp. 6-14.

KOZAK, M. (2009), Text-table: an undervalued and underused tool for communicating information.European Science Editing, Vol. 35, No. 4, pp. 103-105.

KOZAK, M., HARTLEY, J., WNUK, A., TARTANUS, M. (2015), Multiple pie charts: Unreadable, inefficient, and over-used.Journal of Scholarly Publishing, Vol. 46, No. 3, pp. 282-289.

KOZAK, M., KRZANOWSKI, W. J. (2010), Effective presentation of data. European Science Editing, Vol. 36, No. 2, pp. 41-42.

LEMON, J. (2006), Plotrix: a package in the red light district of R.R-News, Vol. 6, No. 4,pp. 8-12.

LEMON J., TYAGI A. (2009), The fan plot: A technique for displaying relative quantities and differences.Statistical Computing and Graphics Newsletter, Vol., 20, No. 1, pp. 8-10.

R CORE TEAM (2019), R: A language and environment for statistical computing. R Foundation for Statistical

Computing, Vienna, Austria. URL https://www.R-project.org/.

TUFTE, E.R. (1983), The Visual Display of Quantitative Information. (1st and 2nd eds.). Cheshire, CT: Graphics Press.