General Principles for Creating Effective Data Visualizations

Kim Hochstedler

11/19/2020

Alternate title: I can do statistics, but graphing is where I draw the line

Introduction

Statistical graphics visually display data using points, lines, coordinate systems, color, and other symbols. The goal of constructing a statistical graphic is to display the data in such a way that we (analysts) and other viewers may draw appropriate conclusions about the nature of our measurements and results. In this newsletter, we will discuss the importance of data visualization and provide guidelines to achieve the goals of statistical graphics. We begin with a motivating example of why graphical summaries are an important analysis step. The general principles for making informative statistical graphics are discussed in the remainder of this newsletter. The general principles are ...

1. Graphs should stand alone
2. Avoid distortion
3. Use data ink effectively

By illustrating the importance of data visualization and describing these guidelines, we hope to aid scientists and researchers as they proceed through data analysis and present their results to broad audiences.

The importance of data visualization

Data visualization is an important part of both exploratory data analysis and the presentation of final results from a research project. Graphs and figures often provide richer information than tabular, numerical summaries. For example, consider the summary statistics and regression analysis results from a study involving occupational data, presented in **Table 1** and **Table 2**. **Table 1** presents a measure of central tendency (mean) and a measure of spread (standard deviation) for two variables in the dataset, salary (thousands of dollars, $) and driving time to work (minutes) for 142 employees at a company. Most employees earn about $54,000 per year, and have a 48 minute commute. There is considerable variability, however, in both the salary and driving time variables.

We conducted an analysis to see if there was an association between salary and driving time, and the results are presented in **Table 2**. In a simple linear regression analysis predicting driving time to work (minutes) from salary (thousands of dollars, $), we found that there was a weak, negative association between the variables. We expect that for every additional $10,000 an employee makes per year, they probably live 1-2 minutes closer to work.

Table 1: *Summary statistics for salary (thousands of dollars, $) and driving time to work (minutes) for 142 employees of a company.*

Variable	Mean	Standard Deviation
Salary (thousands of dollars, $)	54.27	16.72
Driving time to work (minutes)	47.83	26.86

Table 2: *Results of a regression analysis predicting driving time to work (minutes) from salary (thousands of dollars, $) for 142 employees of a company.*

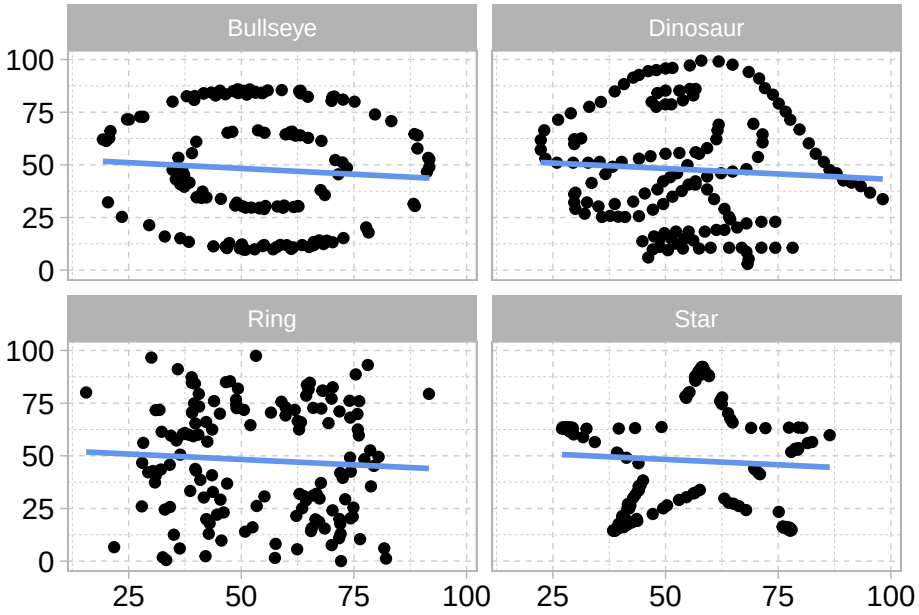
Intercept	Slope	Correlation
53.3	-0.11	-0.063

We could draw a lot of conclusions from our summary statistics and regression analysis results. With the information provided in **Table 1** and **Table 2** alone, we might make comments about the high variability in salaries at this company. Why are some employees paid so much and others paid so little? We might also notice that higher paid employees tend to live closer to work. Is cost of living higher near the company, but lower farther away, so only the highest-paid employees can afford to live nearby?

By asking these questions and drawing these conclusions, however, we are getting ahead of ourselves. We never even checked if our simple linear regression was appropriate for our data! Does our data follow a relationship that appears linear, graphically? Are our predictor and outcome variables skewed? These questions can only be answered by looking at our data. We might make histograms to examine the shape of the distributions of the salary and driving time variables. It would also be wise to construct a scatterplot of salary vs. driving time, to examine the relationship between the variables prior to fitting a linear model.

The importance of such steps is illustrated in **Figure 1**. **Figure 1** presents scatterplots from 4 different datasets. In each dataset, the summary statistics for variables presented on the x- and y-axes are identical to those presented in **Table 1**. In addition, all of the scatterplots include the linear regression model fit by predicting the y-axis variable from the x-axis variable. All of these linear models have the exact same slopes and intercepts, identical to those presented in **Table 2**. The correlation for all four datasets is also -0.063.

Figure 1: *Scatterplots and linear models for 4 datasets that have the same summary statistics and regression analysis results as the data presented in Table 1 and Table 2.*



We probably did not expect to see a dinosaur, star, ring, or bullseye shape in the scatterplot of our data, yet each of these visualizations is plausible given our results in **Table 1** and **Table 2**. If we produced a scatterplot of employee salary ($) vs. driving time to work, and it looked like a dinosaur, we would likely have some pause before proceeding with our analysis! This example motivates why we should visualize our data. Numerical summaries simply do not supply enough information to make sense of our data.

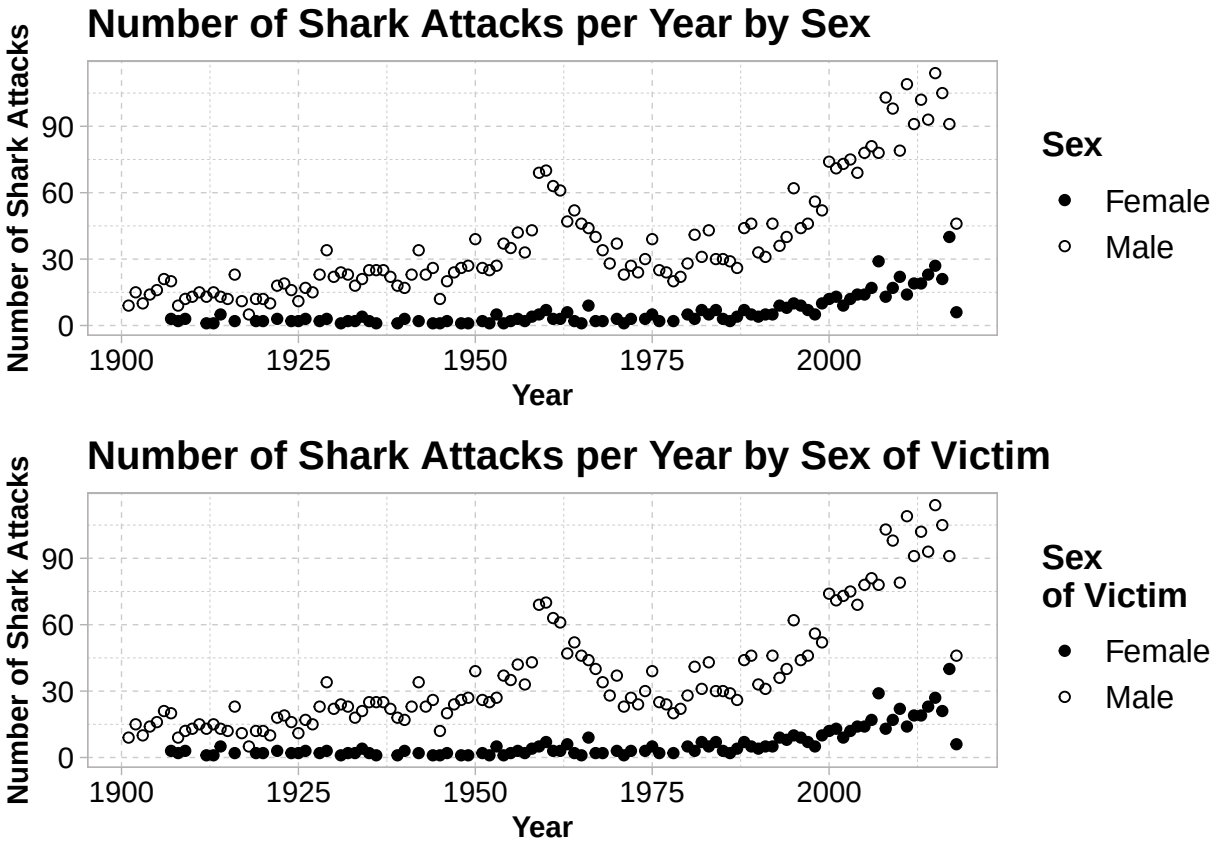
General principles for making graphs

1. Graphs should stand alone

A statistical graphic should contain enough information that a viewer can interpret what the graph is presenting, without having to refer back to other materials [2]. This can be accomplished by paying careful attention to the titles, labels, and captions associated the graph. An example of the level of specificity that can be required to achieve self-contained graphs is presented in **Figure 2**.

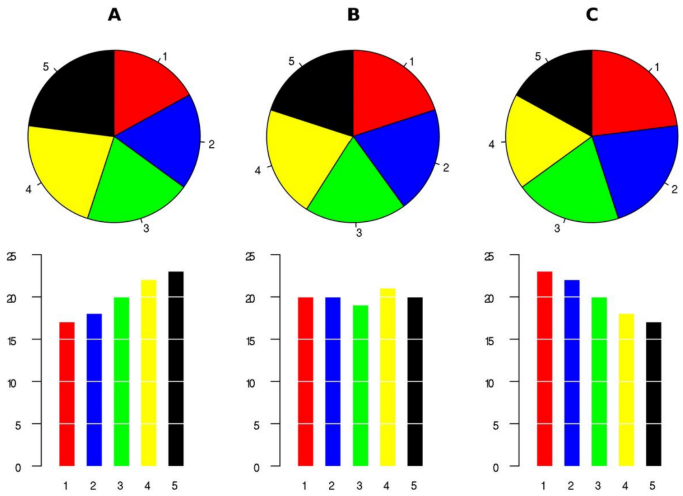
Both graphs in **Figure 2** display the number of recorded shark attacks across the globe per year since 1900, with points colored by the sex of the attack victim. In **Figure 2A** (top), however, the labeling is imprecise. The viewer may read **Figure 2A** and conclude that male sharks are much more aggressive than female sharks, as they account for more attacks per year than female sharks. Such a conclusion is plausible based on the presentation of **Figure 2A**. In contrast, **Figure 2B** (bottom) clearly specifies that the sex grouping variable applies to the attack victims, not the sharks involved in the attack.

Figure 2: *Number of recorded shark attacks per year since 1900, divided by sex of the attack victim. Figure 2A (top) demonstrates imprecise labeling and Figure 2B (bottom) demonstrates more precise labeling.*



Effective labeling is one necessary step to ensure a statistical graphic stands alone. Another key consideration is whether viewers can observe the main trends or takeaways in results, based on a presented graph. **Figure 3** provides an example where the primary figure (a pie chart) must be supplemented by secondary information (in this case, a bar chart), in order to obtain meaning from the graphic [8]. All 3 pie charts presented in **Figure 3** appear to show equal representation in the data across 5 categories. Upon closer inspection, however, a bar chart in panel A reveals a clear trend in the percentage of each group, such that higher group numbers have higher percentages than lower group numbers. Similarly, a bar chart in panel C reveals that the pie chart actually shows higher percentages for lower numbered groups. Only panel B aligns with our expectation based on the pie chart, that all groups have nearly identical representation.

Figure 3: *Pie charts and associated bar charts, presenting the percentage of observations in a data set belonging to one of five unique groups. Because there are many groups and the group percentages are numerically close, the pie charts appear to represent uniform percentages across the groups. We need auxiliary information, in the form of bar charts, to see the true percentages by group.*



If the pie charts in the upper row of **Figure 3** were to be presented in isolation, we would probably not detect the relationships between the groups that are observed in panel A and panel C. Thus, auxiliary information was necessary to make sense of the data displayed in the original pie charts in **Figure 3**.

Graphs are self-contained when they have precise titles, labels, and captions. Additionally, viewers should be able to draw appropriate conclusions based on statistical graphics. Practice reading your plot as if you are an "outsider", without knowledge of the data displayed in the graph. Better yet, ask someone who is not involved in your project to explain the conclusions they draw from your work. If an "outsider" can interpret your graph correctly without extensive supplementary text, your graph can likely stand alone.

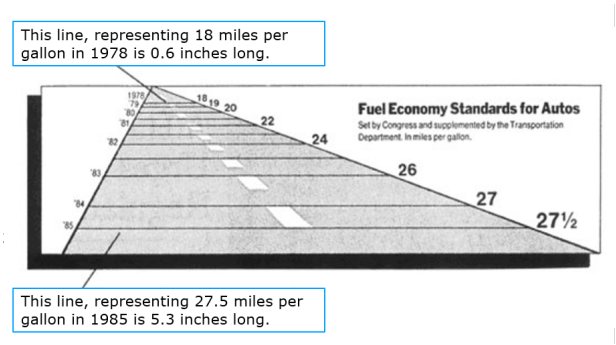
2. Avoid distortion

Distortion occurs in a data visualization when the physical representation of data in a graph is not proportional to the quantities that are represented. A graph does not distort the data when the size of an effect in the visualization matches the size of the effect in the data. We can actually numerically represent distortion through the *Lie Factor* [3].

$$\textbf{Lie Factor} = \frac{\text{size of effect in graph}}{\text{size of effect in data}}$$

A *Lie Factor* of 1 indicates no distortion, because the size of the effect in the graph matches that of the data. If the *Lie Factor* is above 1, the size of the effect in the graph is inflated, compared to the true effect. If the *Lie Factor* is below 1, the size of the effect in the graph is deflated, relative to the effect in the data.

Figure 4: *Fuel economy standards for automobiles in the United States. Graph originally printed in the New York Times.*

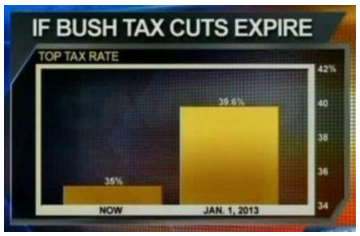


The graphic displayed in **Figure 4** can be used to demonstrate a *Lie Factor* calculation [4]. For this calculation, we consider how the percentage change in fuel economy standards from 1978 to 1985 is represented in the graph, compared to the true values of the data. We can compute the true percentage change in fuel economy standards between 1978 and 1985 as $\frac{27.5-18}{18} * 100 = 52.8\%$. We can compare this to the percentage change in the horizontal lines representing fuel economy standards in the graph, $\frac{5.3-0.60}{0.60} * 100 = 783\%$. The *Lie Factor* is the ratio of these effects, $\frac{783}{52.8} = 14.83$. Because the *Lie Factor* is above 1, we can conclude that **Figure 4** inflates the true difference in fuel economy standards between 1978 and 1983, by distorting the size of the effect in the data through the graphical representation.

$$\textbf{Lie Factor} = \frac{(5.3 - 0.60) \,/\, 0.60}{(27.5 - 18) \,/\, 18} * 100 = \frac{783}{52.8} = 14.83 > 1$$

Distortion can occur even when using more standard data visualization techniques than that of **Figure 4**, like the bar chart in **Figure 5**. **Figure 5** presents a misleading graph presented by Fox News that displayed the top tax rates of President Bush's and President Obama's tax plans [9].

Figure 5: *Tax rates under President Bush's (left) and President Obama's (right) tax plans. Graph originally presented by Fox News.*



This graph, importantly, begins the y-axis at a value of 34%, rather than 0%. While the absolute difference in the tax rates is accurately depicted by the difference in the bar height of the graph, most viewers judge the disparity in tax rates by the area of the bars. In this case, the bar associated with President Obama's tax rate appears about 5 times larger than the bar associated with President Bush's tax rate (we will assume this approximate area difference for our calculations). Given that the true absolute difference in the tax rates is only 4.6%, we can compute the *Lie Factor* for **Figure 5** as follows:

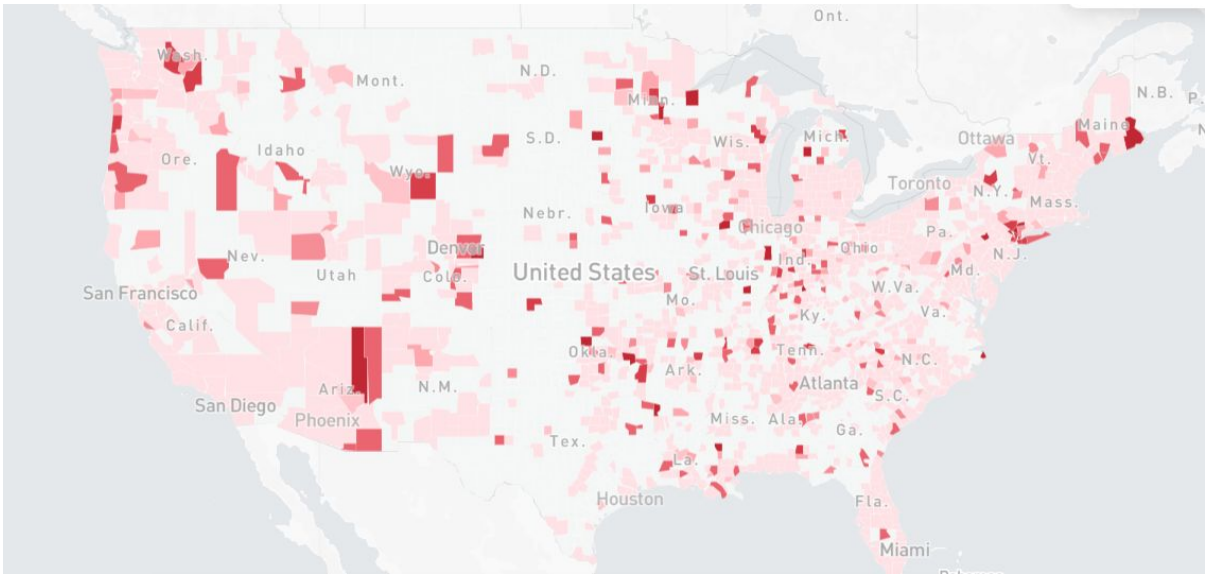
$$\textbf{Lie Factor} = \frac{(5 - 1) \,/\, 1}{(39.6 - 35) \,/\, 35} * 100 = \frac{4}{0.1314} = 40.44 > 1$$

Because the *Lie Factor* is greater than 1, the size of the difference in tax rates between the two presidents is much larger in the graph (as measured by bar chart area) than in the data. This example of distortion illustrates why it is important to begin axes at 0 for most plots, especially those that rely on area to display effects in the data.

There are some graphs where the *Lie Factor* cannot be computed directly, even though there is evidence of distortion. In **Figure 6**, for example, distortion occurs in the choice of color that was used to represent missing data. **Figure 6** represents the summarized results of COVIDcast, a COVID-19 symptom survey

conducted by researchers at Carnegie Mellon University [10]. The survey included respondents from across the United States. Results were summarized at the county level, and a color was assigned to each county to indicate the average severity of symptom reports. In this figure, light red indicated low symptom reports. Darker shades of red indicated more severe symptom reports.

Figure 6: *COVID-19 symptom reports by county, from April 2020. Symptom reports were obtained from national Facebook surveys conducted by Carnegie Mellon University's Delphi Research Group. Darker red shading indicates more severe symptom reports. Areas in white did not have enough data to calculate symptom severity estimates.*



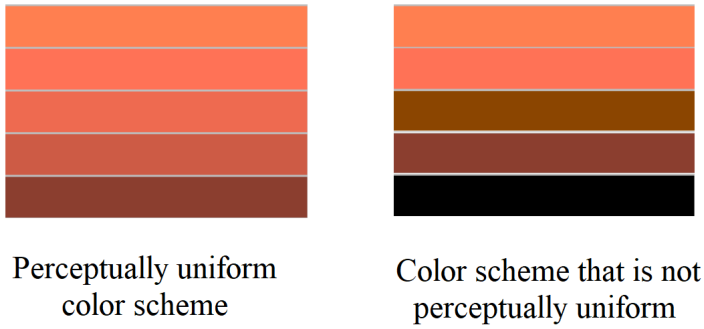
It is notable that much of the country is shaded in white. Without looking closely at a figure legend explaining the color distribution of the graph, one might assume that white indicates very low or virtually no symptom reports. In reality, however, areas shaded in white had too little data to compute a symptom summary metric. Thus, counties in white were instances of missing data.

This color shading strategy is an example of distortion because it falsely creates the perception of low symptom reports in areas where we do not have data. The *Lie Factor* cannot be computed in this case because the effect in the data is non-existent - there is no data! But distortion is present because the graphic misrepresents the (lack of) data. Specifically, this graph most likely deflates the true COVID-19 symptom severity across the United States because such large regions of the country are shaded to imply that all respondents are in perfect health. In reality, there are probably non-zero symptoms in these counties. That being said, without knowing the true experiences of all respondents in those areas, we cannot compute the *Lie Factor* or determine the type of distortion precisely.

It is worth noting that the designers of the COVIDcast dashboard have since updated their methods for displaying missing data. Areas with too little data to compute symptom summaries are now shaded in grey hashed backgrounds that are clearly dissimilar from the remainder of the color scheme. Using patterned backgrounds that include colors that are not meaningful in the color scheme of the remainder of the graphic is an appropriate way to display missing data in a map, like that in **Figure 6**.

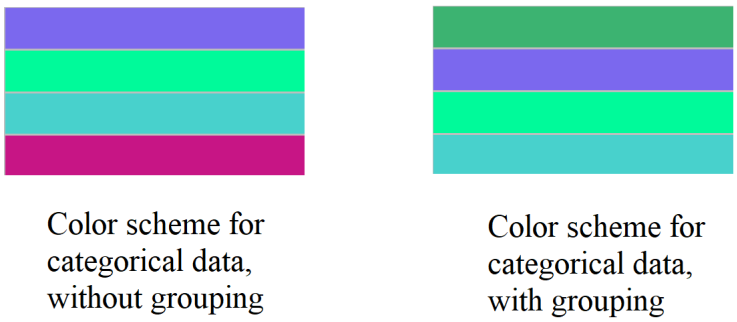
Distortion that occurs via color displays can also be avoided by using color gradients that are "perceptually uniform" for continuous data [11]. This means that the perceived distance between each unit change in color matches the distance between each unit change in the data. In the perceptually uniform color scheme in **Figure 7**, the equidistant "steps" between shades of orange match equidistant changes in data. In the color scheme that is not perceptually uniform, there is a higher contrast between the 4th and 5th colors, for instance, than the 1st and 2nd colors. Thus, even if both color changes represent the same size of effect in the data, the high contrast between the 4th and 5th colors would likely imply a larger change in the data than the true effect.

Figure 7: *Example of a perceptually uniform and non-perceptually uniform color scheme for a continuous variable.*



For categorical data, care should be taken in selecting color schemes that appropriately distinguish individual categories, or show groups present in the data. For example, if a graph displays the percentage of respondents that fall in one of four groups, and all four groups are represented by a unique color, viewers are generally able to distinguish the groupings in the data. If, however, two groups are displayed with different shades of the same color, viewers are likely to perceive those two groups as being related, and distinct from the differently colored other groups. The difference in these types of color schemes are displayed in **Figure 8**. Misuse of color schemes for categorical data has the potential to create distortion if a perception of relatedness between categories is portrayed by the graph, and that association is not present in the data.

Figure 8: *Examples of color schemes for categorical data, one without perceptions of grouping and one with perceptions of grouping.*



We need not usually worry about constructing statistical graphics with egregious *Lie Factors* like that of **Figure 4** and **Figure 5** if we construct data visualizations using well-known software packages. Data visualization tools (like those available in the `R`) automatically scale graphics to match the scale of the data [6]. Pre-existing color schemes, like those featured in `RColorBrewer` can be used to guarantee perceptually uniform and appropriate qualitative data displays [7]. Still, if one is experimenting with new visualization techniques, or strays from well-known software packages, *distortion* is a key consideration in evaluating the final statistical graphic.

3. Use data ink effectively

Data ink is defined as "non-erasable" and "non-redundant" core of the statistical graphic [3]. *Data ink* presents information about the data. Any other features of the graph that contain unnecessary information can be considered "decoration". Good graphics should only include data ink, and decoration should be removed whenever possible. By maximizing the amount of the graphic that is devoted to data presentation, we avoid drawing the attention of our viewers to irrelevant elements of our results or analysis.

Figure 9 represents a figure that does not effectively use data ink [5]. Some elements of the graph are

redundant, and do not present information about the data. For example, the height of each bar in the plot is the necessary element of the graph that communicates the value of each label: A, B, C, and D. Thus, the color of the bars and the cone shape do not add any additional information. These elements are unnecessary, and examples of decoration. Moreover, the three-dimensional nature of the graph actually distort the data. The numbers plotted in the graph are intended to be 2, 4, 6, and 8, but the cones do not actually reach the lines corresponding to those values on the figure, due to the "decorative" depth element. **Figure 10** represents the same data as **Figure 9**, but omits the "decorative" components. In **Figure 10**, all elements of the graph correspond to an aspect of the data. Nothing could be "erased" without the graph losing it's meaning - everything we see is *data ink*.

Figure 9: *A figure that does not use data ink effectively. Decoration is present in the color of the bars, the shape of the cones, and the use of a 3-dimensional perspective.* [5].

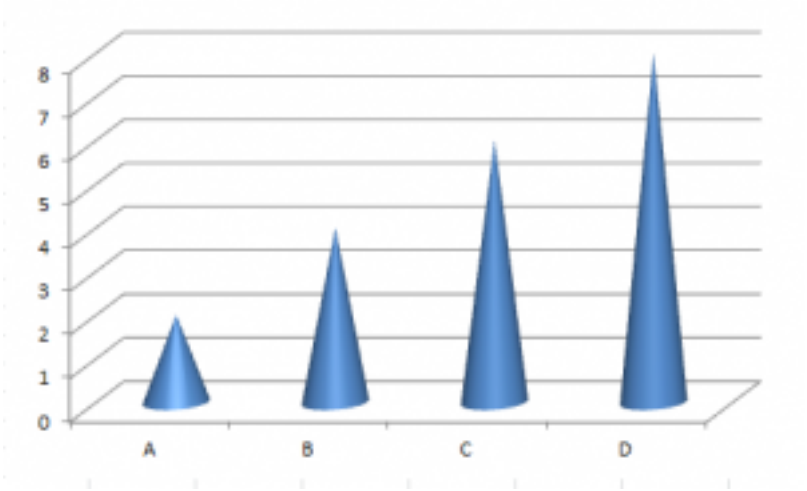
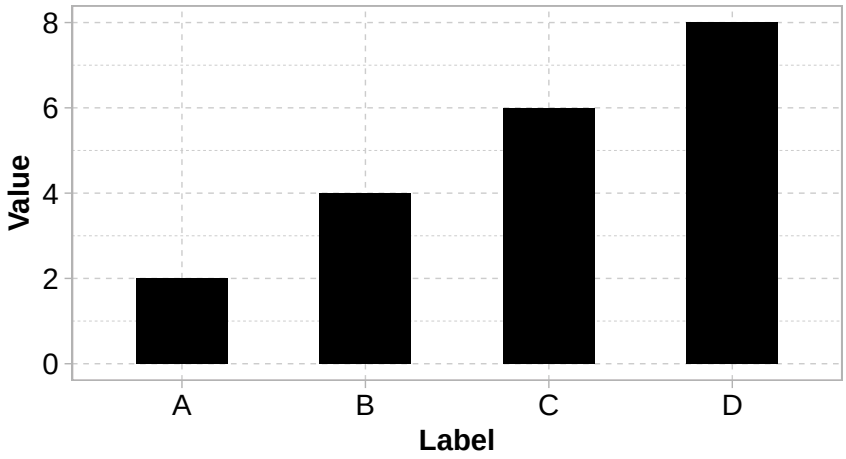


Figure 10: *A figure that uses data ink effectively. This figure presents the same information as Figure 9, but does so without extraneous decoration.*

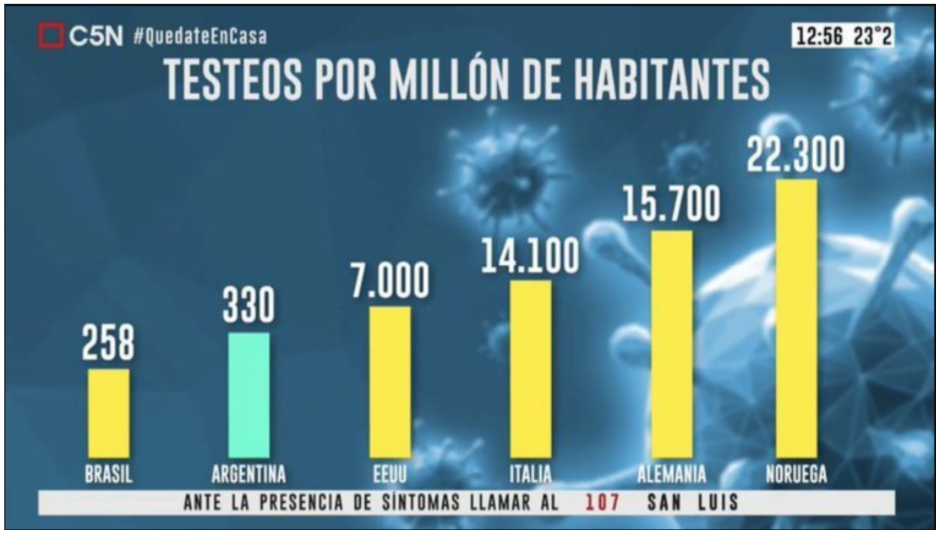


The misuse of data ink can be avoided by only assigning elements to the graph that map back to the original data. For example, the height of the bars in **Figure 10** is an element of the graph that is assigned to the count for each label in the original data. If we adjusted the shape aesthetic of the bars to be cones, as in **Figure 9**, we would have included an element of the graph (shape) that did not map back to anything in the data. Investigators should also take care to not include redundancy in their graphics. For example, if each group has a different label, like A, B, C, and D, in **Figure 10**, it is redundant to *also* assign each group to have a different color bar. In that case both the "label" and "color" elements map back to the same data element (group). These redundancies tend to draw viewers' attention away from the data-driven elements of

the graph and should thus be avoided.

The misuse of data ink is fairly common in media graphics, that are generally designed to be interesting and attention grabbing. For example, **Figure 11** presents a graph of COVID-19 test capacity per million residents of 6 countries, presented by Argentinian media [12]. The decorative background of the plot is unnecessary to convey the data. A more appropriate background would have included grid lines, to allow viewers to directly compare the height of bars in the graph.

Figure 11: *COVID-19 testing capacity per 1 million inhabitants of 6 different countries, with decorative elements (like the background). Graph originally printed by Argentinian media.*



Perhaps grid lines were omitted in this case because **Figure 11** does not actually include a y-axis. The heights of the bars in this graph are ordered by testing capacity, but the *differences* in the heights of the bars are completely arbitrary! Thus, in this example, distortion *and* a misuse of data ink combine to create a very decorative and misleading statistical graphic.

Conclusion

In this newsletter, we presented an example of why data should be visualized in both the analysis and writing stages of a research endeavor. We also presented three broad principles for creating effective statistical graphics. We hope this presentation motivates researchers to explore the principles of in your own work

References

[1] Locke S, Cairo A, Matejka J, Fitzmaurice G, McGowen LD, and Cotton R. "datasauRus: Datasets from the Datasaurus Dozen". 2018. https://cran.r-project.org/web/packages/datasauRus/index.html
[2] Bartee L, Shriner W, and Creech C. "Presenting Data - Graphs and Tables". *Principles of Biology*. OpenOregon: 2020.
[3] Tufte, ER. *The Visual Display of Quantitative Information*. 1986. Graphics Press: USA.
[4] Friendly, M. "Gallery of data Visualization". 2001. http://www.datavis.ca/gallery/lie-factor.php
[5] Robbins, N. "Winner of the Bad Graph Contest Announced". *Forbes*. 2012.
[6] Wickham H. "ggplot2: Elegant Graphics for Data Analysis". Second Edition. Springer: 2016.
[7] Neuwirth, E. "RColorBrewer: ColorBrewer Palettes". 2014.
https://cran.r-project.org/web/packages/RColorBrewer/index.html
[8] Fisher, M. "Graphics for conversation: How to illustrate your story". *Scalar*. 2019.
[9] "Fox News continues charting excellence" *Flowing Data*. 2012.
[10] DELPHI. "COVIDCast: Real-time indicators of COVID-19 Activity". 20 April 2020. https://delphi.cmu.edu/
[11] Kay, M. "Grammar of graphics". *Big Data Summer Institute*. Michigan: Ann Arbor. 1 July 2019.
[12] Kotsehub, N. "Stopping COVID-19 with misleading graphs". *towards data science*. 19 June 2020.