Chapter 5
Data Visualization

John F. Tripp

1 Introduction

Data analytics is a burgeoning field—with methods emerging quickly to explore and make sense of the huge amount of information that is being created every day. However, with any data set or analysis result, the primary concern is in communicating the results to the reader. Unfortunately, human perception is not optimized to understand interrelationships between large (or even moderately sized) sets of numbers. However, human perception is excellent at understanding interrelationships between sets of data, such as series, deviations, and the like, through the use of visual representations.

In this chapter, we will present an overview of the fundamentals of data visualization and associated concepts of human perception. While this chapter cannot be exhaustive, the reader will be exposed to a sufficient amount of content that will allow them to consume and create quantitative data visualizations critically and accurately.

J. F. Tripp (✉)
Clemson University, Clemson, SC, USA
e-mail: jftripp@clemson.edu

B. Pochiraju, S. Seshadri (eds.), *Essentials of Business Analytics*, International Series in Operations Research & Management Science 264,
https://doi.org/10.1007/978-3-319-68837-4_5

2 Motivating Example

A Vice President of Sales wishes to communicate his division's performance to the executive team. His division has performed very well in year-over-year sales, but is still rather small compared to other divisions. He is concerned that when his current year numbers are presented alongside other divisions, it will not accurately reflect the excellent performance of his team.

He wishes to provide the sales figures as well as show the growth in sales, year over year, and compare his rate of growth to other divisions' rates. Providing this information using only tales of numbers would be difficult and time consuming. However, this information can be provided in one or two simple but intuitive graphs.

3 Methods of Data Visualization

3.1 Working with (and Not Against) Human Perception

Consider Fig. 5.1. When you see this graph, what do you believe is true about the "level" of the variable represented by the line? Is the level greater or less at point 2 compared with point 1?

If you are like most people, you assume that the level of the variable at point 2 is greater than the level at point 1. Why? Because it has been ingrained in you from childhood that when you stack something (blocks, rocks, etc.), the more you stack, the higher the stack becomes. From a very early age, you learn that "up" means "more."

Now consider Fig. 5.2. Based on this graph, what happened to gun deaths after 2005?[1]

Upon initial viewing, the reader may be led to believe that the number of gun deaths went down after 2005. However, look more closely, is this really what happened? If you observe the axes, you will notice that the graph designer inverted

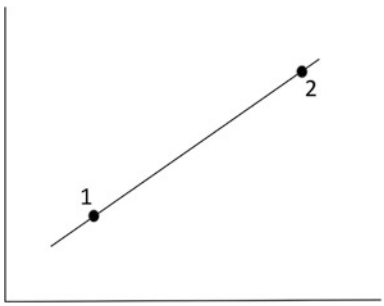


Fig. 5.1 A graph

[1] The "Stand Your Ground" law in Florida enabled people to shoot attackers in self-defense without first having to attempt to flee.

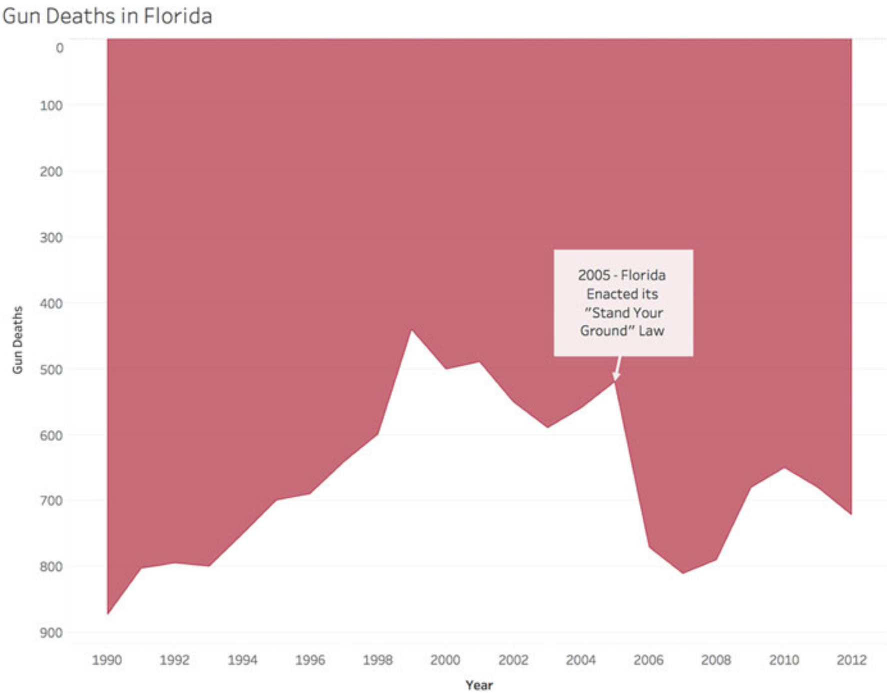


Fig. 5.2 A graph representing gun deaths in Florida over time

the y-axis, making larger values appear at lower points in the graph than smaller values. While many readers may be able to perceive the shift in the y-axis, not all will. For all readers, this is a fundamental violation of primal perceptive processes. It is not merely a violation of an established convention; it is a violation of the principles that drove the establishment of that convention.[2]

This example is simple, but it illustrates the need for data visualizers to be well trained in understanding human perception and to work with the natural understanding of visual stimuli.

3.2 Why We Visualize Data

Visual and cognitive processing in humans is not optimized to process relationships between large sets of numbers. While we are good at comparing several values against each other, we are simply not good enough at drawing inferences using sets or tables in order to communicate and compare trends for large groups of numbers. Let us look at a famous example.

[2]This example was strongly debated across the Internet when it appeared. For more information about the reaction to the graph, see the Reddit thread at http://bit.ly/2ggVV7V.

Table 5.1 Anscombe's Quartet

Set 1		Set 2		Set 3		Set 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

In Table 5.1, the classic example of "Anscombe's Quartet" is presented. In 1973, Anscombe created these sets of data to illustrate the importance of visualizing data. When you consider these four data sets, what sense can you make of them? How much are the sets the same? How much are they different? These data sets are also referred to in Chap. 7 on linear regression.

> *Information, that is imperfectly acquired, is generally as imperfectly retained; and a man who has carefully investigated a printed table, finds, when done, that he has only a very faint and partial idea of what he had read; and that like a figure imprinted on sand, is soon totally erased and defaced*—William Playfair (1801, p. 3).

As Playfair notes in the quote above, even after a great deal of study, humans cannot make clear sense of the data provided in tables, nor can they retain it well. Turning back to Table 5.1, most people note that sets 1–3 are similar, primarily because the X values are the same, and appear in the same order. However, for all intents and purposes, these four sets of numbers are statistically equivalent. If a typical regression analysis was performed on the four sets of data in Table 5.1, the results would be *identical.* Among other statistical properties, the following are valid for all four sets of numbers in Table 5.1:

- N = 11.
- Mean of X = 9.0.
- Mean of Y = 7.5.
- Equation of regression line: Y = 3 + 0.5X.
- Standard error of regression slope = 0.118.
- Correlation coefficient = 0.82.
- $r^2 = 0.67$.

However, if one compares the sets of data visually, the differences between the data sets become immediately obvious, as illustrated in Fig. 5.3.

Tables of data are excellent when the purpose of the table is to allow the user to look up specific values, and when the relationships between the values

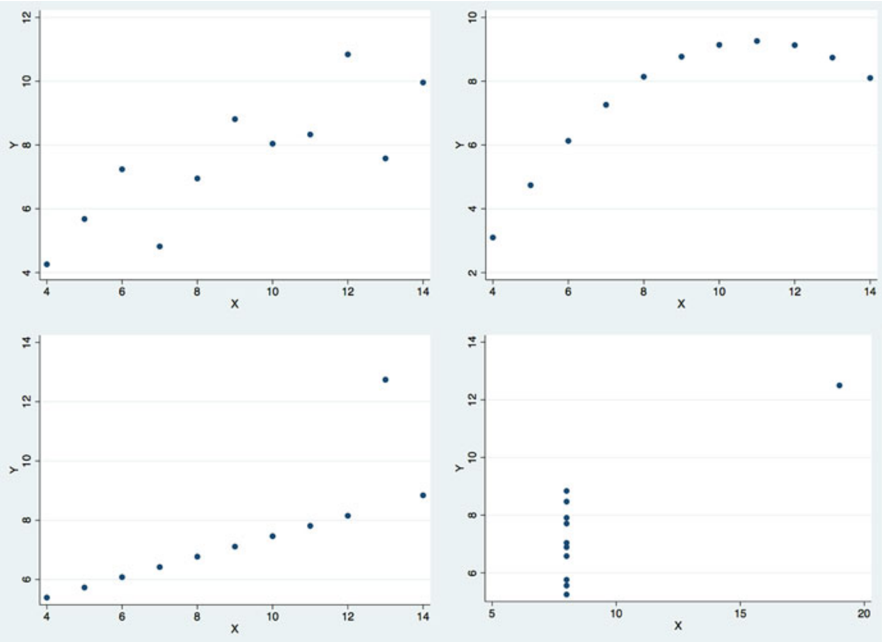


Fig. 5.3 Anscombe's Quartet

are direct. However, when relationships between sets or groups of numbers are intended for presentation, the human perception is much better served through graphical representations. The goal of the remainder of this chapter is to provide the reader with enough background into human visual perception and graphical data representation to become an excellent consumer and creator of graphical data visualizations.

3.3 Visualization as a Cognitive Aid

Humans are bombarded by information from multiple sources and through multiple channels. This information is gathered by our five senses, and processed by the brain. However, the brain is highly selective about what it processes and humans are only aware of the smallest fraction of sensory input. Much of sensory input is simply ignored by the brain, while other input is dealt with based on heuristic rules and categorization mechanisms; these processes reduce cognitive load. Data visualizations, when executed well, aid in the reduction of cognitive load, and assist viewers in the processing of cognitive evaluations.

When dealing with data visualization, likely the most important cognitive concept to understand is working memory. Working memory is the part of short-term memory that is concerned with immediate perceptual processing. Working memory is extremely limited, with the average human having the capacity to hold at most four "chunks" of information at once (see, e.g., Cowan 2010). For this reason, visualizers must build graphs that do not rely on the user to hold multiple chunks of information in order to "generate" understanding. Instead, visualizers should "do the work" for the user and visualize the relationships in question directly.

Practically, this means that visualizations are not effective when built for general purposes, for a generalized audience. For instance, in my experience with real-world corporate "dashboards," they tend to be designed in a manner that presents a generous amount of data that users might be able to leverage for different cognitive tasks. However, users would have to extrapolate the useful data themselves and perform additional steps to generate a result that assists in achieving their cognitive task. Instead, if visualizations were created with particular cognitive tasks in mind and constructed in a manner that would assist the viewer in the completion of the task, then the user would not have to do much in facilitating the understanding of the data.

Designing and building visualizations that are fit for purpose—for use in the processing of a defined cognitive task (communicating the status of sales in a region or the need for managerial intervention in a process, etc.)—is the key responsibility of any data visualizer. This will usually result in the need to create visualizations that are more tightly focused than those that already exist in businesses today. In doing so, we re-emphasize the need to reduce cognitive load for a particular cognitive task. A good visualization that is focused on a task, highlights key information, and provides computational assistance for viewers will allow them to build an understanding more quickly, and is of significantly more value than creating a single, multipurpose visualization that fails to assist in any of the viewers' cognitive tasks.

3.4 Six Meta-Rules for Data Visualization

For the remainder of this chapter, we will review and discuss six meta-rules for data visualization. These are based upon the work of many researchers and writers, and provide a short and concise summary of the most important practices regarding data visualization. However, it is important to note that these rules are intended to describe how we can attempt to represent data visually with the highest fidelity and integrity—not argue that all are immutable, or that trade-offs between the rules may not have to be made. The meta-rules are presented in Table 5.2.

By following these meta-rules, data visualizers will be more likely to graphically display the actual effect shown in the data. However, there are specific times and reasons when the visualizer may choose to violate these rules. Some of the reasons may be, for example, the need to make the visualization "eye-catching," such as for an advertisement. In these cases, knowing the effect on perceptive ability of breaking the rules is important so that the visualizer understands what is being

Table 5.2 Six meta-rules for data visualization

1. The simplest chart is usually the one that communicates most clearly. Use the “not wrong” chart—not the “cool” chart
2. Always directly represent the relationship you are trying to communicate. Do not leave it to the viewer to derive the relationship from other information
3. In general, do not ask viewers to compare in two dimensions. Comparing differences in length is easier than comparing differences in area
4. Never use color on top of color—color is not absolute
5. Do not violate the primal perceptions of your viewers. Remember, up means more
6. Chart with graphical and ethical integrity. Do not lie, either by mistake or intentionally

lost. However, in some cases, the reasons for violating the rules may be because the visualizer wishes to intentionally mislead. Examples of this kind of lack of visualization integrity are common in political contexts.

These rules are made to be understood, and then followed to the extent that the context requires. If the context requires an accurate understanding of the data, with high fidelity, the visualizer should follow the rules as much as possible. If the context requires other criteria to be weighed more heavily, then understanding the rules allows the visualizer to understand how these other criteria are biasing the visual perception of the audience.

Simplicity Over Complexity

> ***Meta-Rule #1***: *The simplest chart is usually the one that communicates most clearly. Use the “not wrong” chart—not the “cool” chart.*

When attempting to visualize data, our concern should be, as noted above, to reduce the cognitive load of the viewer. This means that we should eliminate sources of confusion. While several of the other meta-rules are related to this first rule, the concept of simplicity itself deserves discussion.

Many data visualizers focus on the aesthetic components of a visualization, much to the expense of clearly communicating the message that is present in the data. When the artistic concerns of the visualizer (e.g., the Florida dripping blood example above) overwhelm the message in the data, confusion occurs. Aside from artistic concerns, visualizers often choose to use multiple kinds of graphs to add “variety,” especially when visualizing multiple relationships. For instance, instead of using three stacked column charts to represent different part to whole relationships, the visualizer might use one stacked column chart, one pie chart, and one bubble chart. So instead of comparing three relationships represented in one way, the viewer must attempt to interpret different graph types, as well as try to compare relationships. This increases cognitive load, as the viewer has to keep track of both the data values, as well as the various manners in which the data has been encoded.

Fig. 5.4 Example of unnecessary visualization

Instead, we should focus on selecting a "not wrong" graph.[3] To do this, one must understand both the nature of the data that is available, as well as the nature of the relationship being visualized. Data is generally considered to be of two types, quantitative and qualitative (or categorical). At its simplest, data that is quantitative is data that is (1) numeric, and (2) it is appropriate to use for mathematical operations (unit price, total sales, etc.). Qualitative or categorical data is data that is (1) numeric or text, and (2) is not appropriate (or possible) to use for mathematical operations (e.g., Customer ID #, City, State, Country, Department).

Is a Visualization Needed?

An initial analysis of the data is required to determine the necessity of a visual representation. For instance, in the graphic shown in Fig. 5.4, a text statement is redundantly presented as a graph.

When a simple statement is enough to communicate the message, a visualization may not be needed at all.

Table or Graph?

Once you decide that you need a visual representation of the data to communicate your message, you need to choose between two primary categories of visualization—tables vs. graphs. The choice between the two is somewhat subjective and, in many cases, you may choose to use a combination of tables and graphs to tell your data story. However, use the following heuristic to decide whether to use a table or a graph:

> *If the information being represented in the visualization needs to display precise and specific individual values, with the intention to allow the user to look up specific values, and compare to other specific values, choose a table. If the information in the visualization must display sets or groups of values for comparison, choose a graph.*

Tables are best used when:

1. The display will be used as a lookup for *particular values*.
2. It will be used to compare *individual values* not groups or series of values to one another.
3. Precision is required.
4. Quantitative information to be provided involves more than one unit of measure.
5. Both summary and detail values are included.

[3]By using the term "not wrong" instead of "right" or "correct," we attempt to communicate the fact that in many cases there is not a single "correct" visualization. Instead, there are visualizations that are more or less "not wrong" along a continuum. In contrast, there are almost always multiple "wrong" visualization choices.

If a Graph, Which Graph?

The exploration of choosing which graph to use is a topic that requires more space than is available in this chapter. However, the process of choosing the correct graph is fundamentally linked to the relationship being communicated. For each kind of relationship, there are multiple kinds of graphs that might be used. However, for many relationships, there are a small group of "best practice" graphs that fit that relationship best.

How Many Dimensions to Represent?

When representing data visually, we must decide how many dimensions to represent in a single graph. The maximum number of data dimensions that can be represented in a static graph is five and in an interactive graph is six. Table 5.3 provides a list of the most likely graphs to be used for various relationships and numbers of dimensions.

Table 5.3 Data relationships and graphs

Relationship	Most likely graph(s)	Keywords	Max. # of dimensions
Time series	Trend line Column chart Heat map Sparklines	Change Rise Increase Fluctuation Growth Decline/decrease Trend	4
Part to whole	Stacked column chart Stacked area chart Pareto chart (for two simultaneous parts to whole)	Rate or rate of total Percent or percentage of total Share "Accounts for X percent"	4
Ranking	Sorted bar/column chart	Larger than Smaller than Equal to Greater than Less than	4
Deviation	Line chart Column/bar chart Bullet graph	Plus or minus Variance Difference Relative to	4
Distribution	Box/whisker plot Histogram	Frequency Distribution Range Concentration Normal curve, bell curve	4
Correlation	Scatterplot Table pane	Increases with Decreases with Changes with Varies with	6
Geospatial	Choropleth (filled gradient) map	N/A	2

1. X-axis placement
2. Y-axis placement
3. Size
4. Shape
5. Color
6. Animation (interactive only, often used to display time)

However, many graph types, because of their nature, reduce the number of possible dimensions that can be displayed (Table 5.3). For instance, while a scatterplot can display all six dimensions, a filled map can only display two: the use of the map automatically eliminates the ability to modify dimensions 1–4. As such, we are left with the ability to use color to show different levels of one data dimension and animation to how the level of that one dimension changes over time.

While the maximum possible dimensions to represent is six, it is unlikely that most visualization (or any) should/would use all six. Shape especially is difficult to use as a dimensional variable and should never be used with size. Notice in Fig. 5.5 that it is difficult to compare the relative sizes of differently shaped objects.

One of the biggest issues in many visualizations is that visualizers attempt to encode too much information into a graph, attempting to tell multiple stories with a single graph. However, this added complexity leads to the viewer having a reduced ability to interpret any of the stories in the graph. Instead, visualizers should create

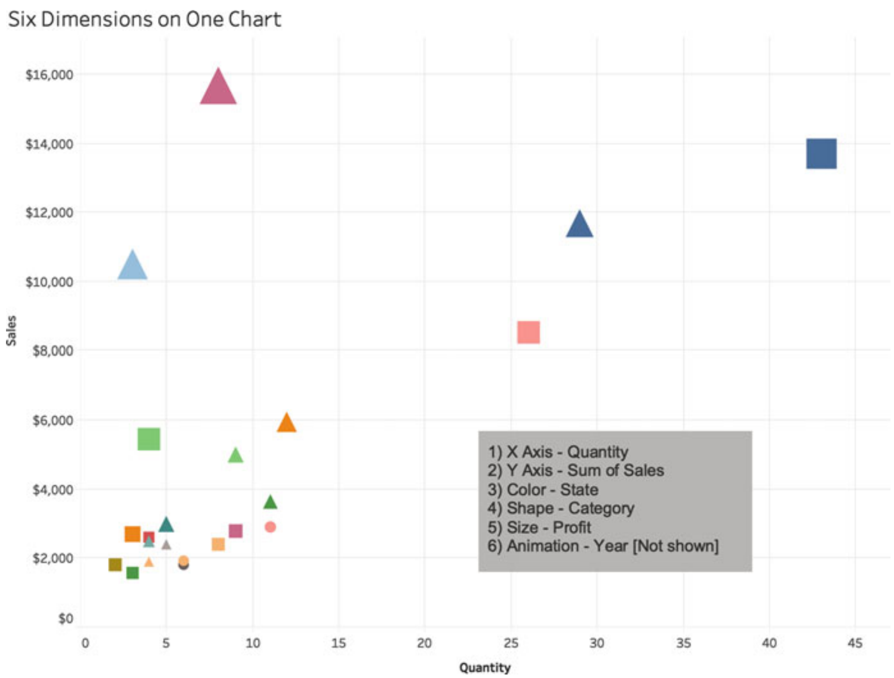


Fig. 5.5 Illustration of use of six data dimensions in a single graph

simpler, single story graphs, using the fewest dimensions possible to represent the story that they wish to tell.

Direct Representation

Meta-Rule #2: *Always directly represent the relationship you are trying to communicate. Don't leave it to the viewer to derive the relationship from other information.*

It is imperative that the data visualizer provide the information that the data story requires directly—and not rely on the viewer to have to interpret what relationships we intend to communicate. For instance, we often wish to tell a story of differences, such as deviations from plan, and budgets vs. actual. When telling a story of differences, do not rely on the viewer to calculate the differences themselves.

Figure 5.6 illustrates a visualization that relies on the viewer to calculate differences. Figure 5.7 presents the actual deviation through the profit margin, allowing the viewer to focus on the essence of the data set.

Again, the goal of the visualizer is to tell a story while minimizing the cognitive load on the viewer. By directly representing the relationship in question, we assist the viewer in making the cognitive assessment we wish to illustrate. When we "leave it to the viewer" to determine what the association or relationship is that we are trying to communicate, not only do we increase cognitive load, but we also potentially lose consistency in the way that the viewers approach the visualization.

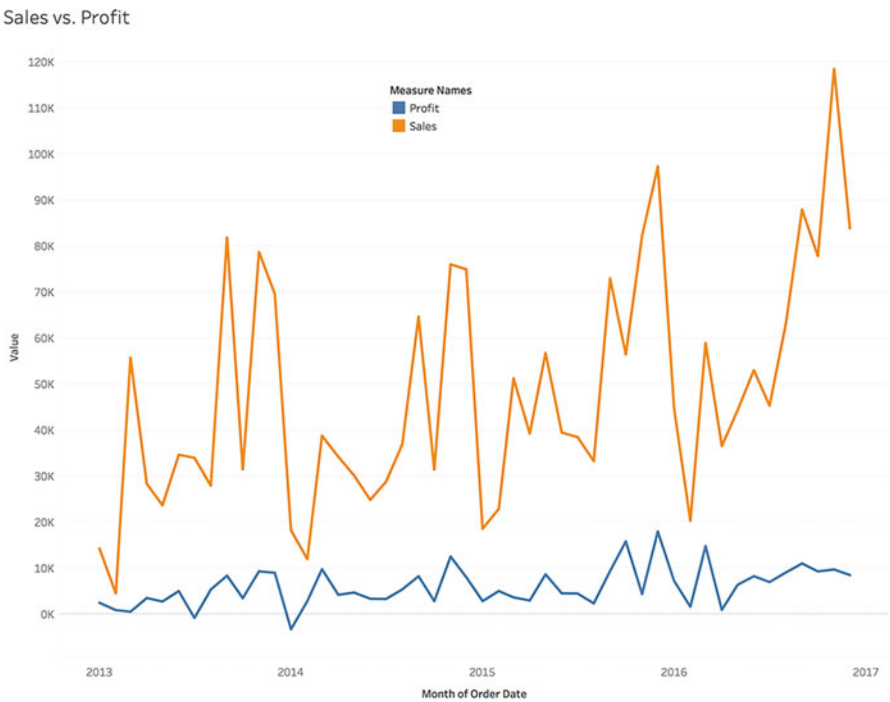


Fig. 5.6 The difference between sales and profit is not immediately clear

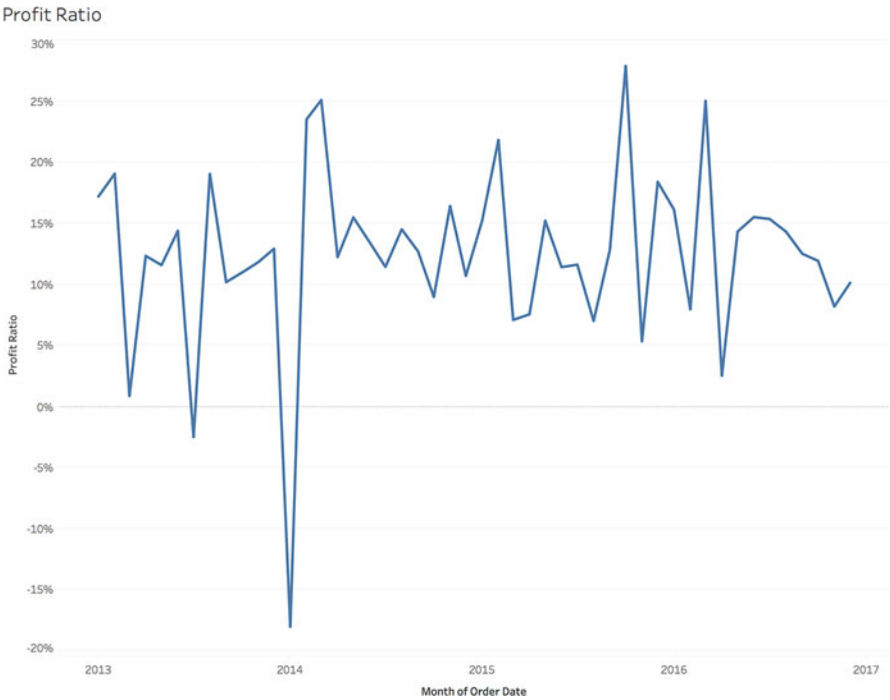


Fig. 5.7 The profit ratio is clear and intuitive

By properly visualizing relationships directly, we not only reduce cognitive load, we also reduce the potential for the viewer to misunderstand the story we are trying to tell.

Single Dimensionality

Meta-Rule #3: *In general, do not ask viewers to compare in two dimensions. Comparing differences in length is easier than comparing differences in area.*

When representing differences in a single data variable (e.g., sales by month, profit by product), visualizers should generally use visual comparisons on a single visual dimension—it is much easier for viewers to perceive differences in length or height than differences in area. By doing this we also avoid some common issues in the *encoding* of data.

While most modern software packages manage the encoding of data very well, whenever a visualizer chooses to use differences in two dimensions to represent changes in a single data dimension, visualizations tend to become distorted. Figure 5.8 illustrates the issue when visualizers directly represent data differences in two dimensions. In this example, the visualization is attempting to depict the ratio between two levels of a variable. However, the ratio of the circles being 1:2 is not

Fig. 5.8 Difference between circles and lines

Table 5.4 Sales of fictitious salespeople

Salesperson	YTD sales (in $)	Share of sales
Deepika Padukone	1,140,000	37%
George Clooney	750,000	25%
Jennifer Garner	740,000	24%
Danny Glover	430,000	14%

immediately apparent from the visual. Using $\pi\left(\frac{d}{2}\right)^2$ we see that the "smaller" circle has an area of $\frac{\pi}{4}$ and the "larger" circle has an area of $\frac{\pi}{2}$, hence 1:2.

While both examples are properly encoded, the area of each exactly maintains the proportion of 1:2 that is found in the data, which of the visualizations more clearly communicates what is present in the data? The simple bars—these only vary on one dimension and therefore more clearly illustrate the relationship that is present in the data.

To 3D or Not to 3D

One of the software features that many people enjoy using, because it looks "cool," is 3D effects in graphs. However, like the example above, 3D effects create cognitive load for the viewer, and create distortion in perception. Let us look at a simple example. Table 5.4 presents some YTD sales data for four fictitious salespersons.

In this data set, Deepika is well above average, and Danny is well below average. This same info is presented in 2D and 3D, respectively, in Fig. 5.9a, b. However, when viewing the 2D representation, the pie chart is less clear than the table (from the chart, can you identify the #2 salesperson?). Moreover, the 3D chart greatly distorts the % share of Deepika due to perspective.

In fact, with 3D charts, the placement of the data points has the ability to change the perception of the values in the data. Figure 5.10a, b illustrate this point. Note that when Deepika is rotated to the back of the graph, the *perception* of her % share of sales is reduced.

Pie Charts?

Although the previous example uses pie charts, this was simply to illustrate the impact of 3D effect on placement. Even though its use is nearly universal, the pie chart is not usually the best choice to represent the part-to-whole relationship. This is due to the requirement it places on the viewer to compare differences on area instead of on a single visual dimension, and the difficulty that this causes in making comparisons.

Going back to the 2D example in Fig. 5.9a, it is very difficult to compare the differences between George and Jennifer. The typical response to this in practice is

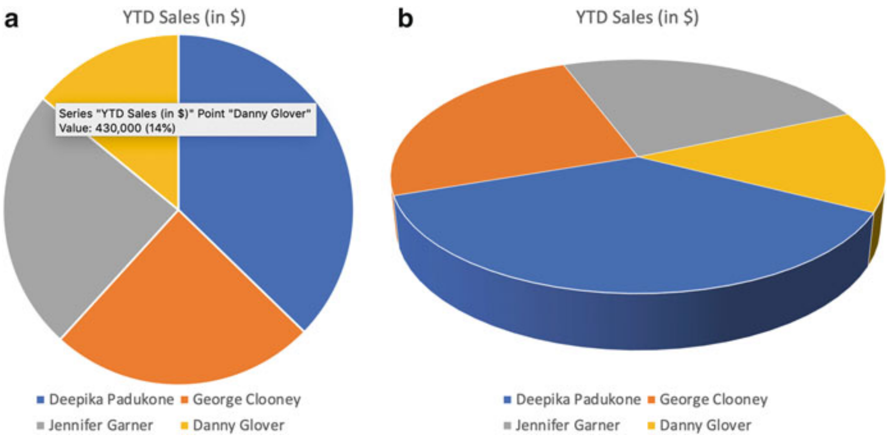


Fig. 5.9 (**a**) 2D pie chart of sales. (**b**) 3D pie chart of sales

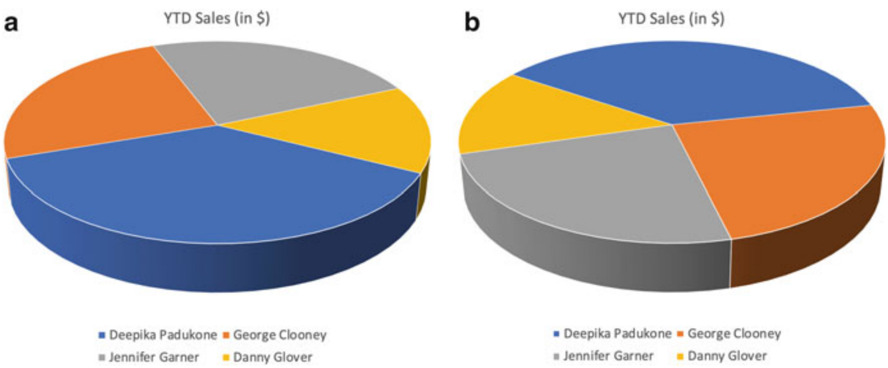


Fig. 5.10 (**a**) Chart with Deepika at front. (**b**) Chart with Jennifer at front

to add the % values to the chart, as Fig. 5.11 illustrates. However, at this point, what cognitive value does the pie chart add that the viewer would not have gained from Table 5.4?

Building a Fit-for-Purpose Visualization

When considering the visualization that is to be created, the visualizer *must* focus on the purpose of the viewer. Creating a stacked column, sorted bar chart, table, or even a pie chart, could all be "not wrong" decisions. Remember, if the user is interested in looking up precise values, the table might be the best choice. If the user is interested in understanding the parts-to-whole relationship, a stacked column or pie chart may be the best choice. Finally, if the viewer needs to understand rank order of values, the sorted bar chart (Fig. 5.12) may be the best option.

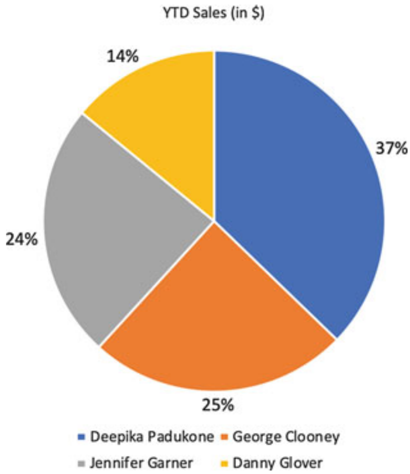


Fig. 5.11 Annotated pie chart

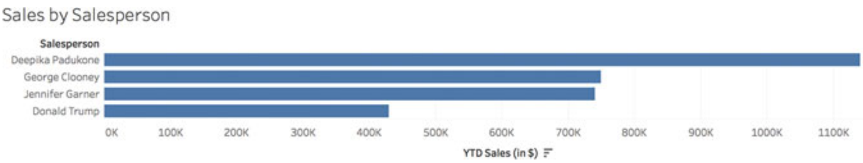


Fig. 5.12 Bar chart of sales

Fig. 5.13 Heat map of sales

Comparing Levels of a Dimension over Multiple Categories

While a stacked bar graph is excellent at showing levels of one variable over one set of categories, often we wish to compare levels of a variable over multiple categories. While we might choose a number of visualizations for this (including possibly a table, if precision is required), one visualization that is optimized for looking at the levels for a single variable at the intersection of two categorical dimensions is the heat map.

Heat maps use a color gradient (see next section for a discussion on the proper use of gradients), within a grid of cells that represent the possible intersections of two categories—perhaps sales by region by quarter (Fig. 5.13).

The heat map illustrates that the fourth quarter tends to be the strongest in all regions, and when compared to the other regions, the East region seems to perform

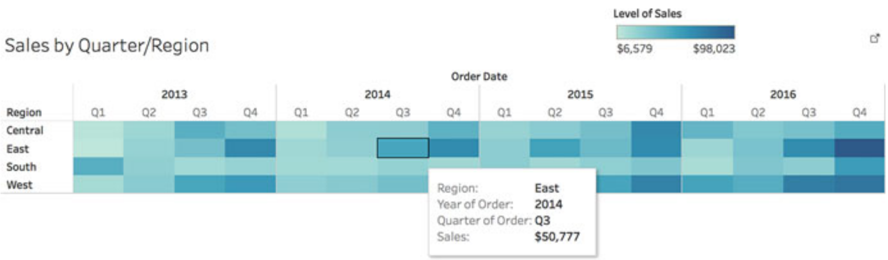


Fig. 5.14 Heat map of sales with interactive tool tip

Fig. 5.15 Perception of color on color

consistently highest in the fourth quarter. This type of data presentation is good for illustrating comparative performance, but only in a general way, as an overview. To add more value to the heat map, one might decide to add the actual number to the cells or, as a better choice, add a tool tip in an interactive visualization (Fig. 5.14).

In almost every case, there are multiple choices for visualizations that are "not wrong." The visualization that is "fit for purpose" is the one that properly illustrates the data story, in the fashion that is most compatible with the cognitive task that the viewer will use the visualization to complete.

Use Color Properly

Meta-Rule #4: *Never use color on top of color—color is not absolute.*

One of the most important concepts in visualization is the appropriate use of color. As with shape, color should be used to provide the viewer meaning. This means that color should be used consistently, and within several rules for human perception. In order to understand these rules, we must spend a few paragraphs on how humans perceive color.

First, color is not perceived absolutely by the human eye. Look at the example in Fig. 5.15. Which of the five small rectangles does your visual perception tell you is the lightest in color?

Most people (if they are honest) will quickly answer that the rightmost rectangle in Fig. 5.15 is the lightest in color. However, as Fig. 5.16 illustrates, all five small rectangles are actually the same color.

The optical illusion presented in Fig. 5.15 is caused by the human visual perception characteristic that colors, and differences in colors, are evaluated relatively, rather than absolutely. What this means in practice is that colors are perceived differently depending upon what colors are around (or "behind") them. The gradient

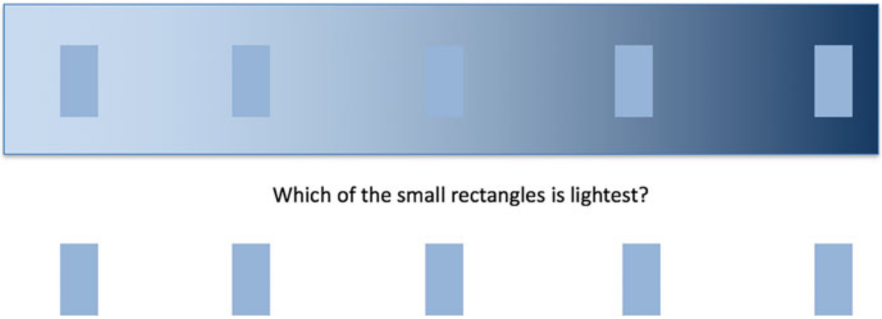


Fig. 5.16 True colors of all bars

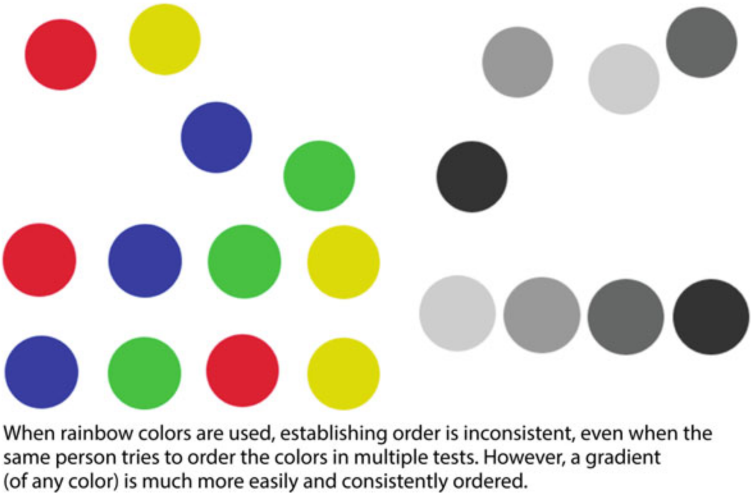


Fig. 5.17 Ordering colored chips

behind the small rectangles in Fig. 5.15 causes changes in perception as to the color of the rectangle, which, in the context of data visualization, changes the *meaning* of the color of the rectangles.

Second, color is not perceived as having an order. Although the spectrum has a true order (e.g., Red, Orange, Yellow, Green, Blue, Indigo, Violet: "ROY G. BIV"), when viewing rainbow colors, violet is not perceived as "more" or "less" than red. Rainbow colors are simply perceived as different from one another, without having a particular "order." However, variation in the level of *intensity* of a single color is perceived as having an order. This is illustrated in Fig. 5.17, and in the following quote:

> "*If people are given a series of gray paint chips and asked to put them in order, they will consistently place them in either a dark-to-light or light-to-dark order. However, if people are given paint chips colored red, green, yellow, and blue and asked to put them in order, the results vary,*" *according to researchers* David Borland *and* Russell M. Taylor II, *professor of computer science at the University of North Carolina at Chapel Hill.*

Using Color with Meaning

Based upon the previous discussion, we now turn to the use of color and its meaning in data visualization. In general, use the following heuristic when deciding on how to use color:

> *When representing levels of a single variable, use a single-color gradient,*[4] *when representing categories of a variable, use rainbow colors.*

Levels of a single variable (continuous data) are best represented using a gradient of a single color. This representation of a single color, with different levels of saturation, visually cues the user that, while the levels of the variable may be different, the color represents levels of the *same concept.* When dealing with categories (or categorical data), the use of different colors cues the users that the different categories represent *different concepts.*

Best Practices for Color

When building a visualization, it is easy to create something that is information-rich, but that does not always allow users to quickly zero in on what is the most important information in the graph. One way to do this is through the choice of colors. In general, colors that are lower in saturation, and are further from the primary colors on the color wheel are considered more "natural" colors, because they are those most commonly found in nature. These are also more soothing colors than the brighter, more primary colors.

For this reason, when designing a visualization, use "natural" colors as the standard color palette, and brighter, more primary colors for emphasis (Fig. 5.18).

By using more natural colors in general, viewers will more calmly be able to interpret the visualization. Using more primary colors for emphasis allows the visualizer to control when and where to drive the attention of the viewer. When this is done well, it allows the viewer to find important data more immediately, limiting the need for the viewer to search the entire visualization to interpret where the important information is. This helps to achieve the visualizer's goal of reducing the cognitive load on the viewer.

Use Viewers' Experience to Your Advantage

> ***Meta-Rule #5:*** *Do not violate the primal perceptions of your viewers. Remember, up means more.*

In the example provided in Fig. 5.2, we reviewed the disorientation that can occur when the viewers' primal perceptive instincts are violated. When viewing the graph with the reversed Y-axis, the initial perception is that when the line moves down, it should have the meaning of "less." This is violated in Fig. 5.2. However, why

[4]Based upon the discussion of the concept of non-order in the perception of rainbow colors, the use of a two-color gradient will not have meaning outside of a particular context. For instance, a red–green gradient may be interpreted as having meaning in the case of profit numbers that are both positive and negative, but that same scale would not have an intuitive meaning in another context. As such, it is better to avoid multicolor gradients unless the context has a meaning already established for the colors.

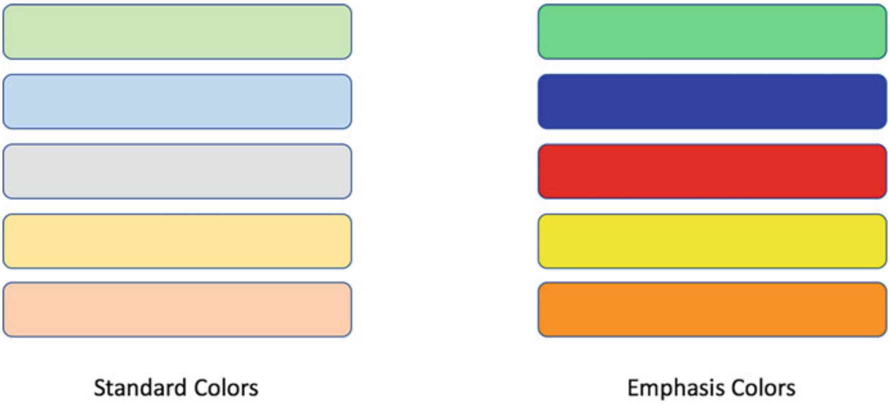


Fig. 5.18 Differences between standard colors and emphasis colors

is it that humans perceive "up," even in a 2D line graph, as meaning "more"? The answer lies in the experiences of viewers, and the way that the brain uses experience to drive perception.

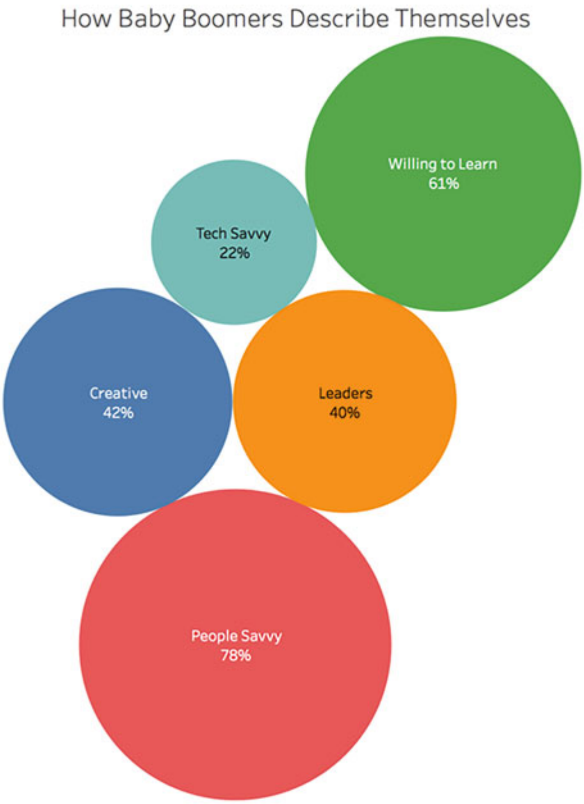
For example, most children, when they are very young, play with blocks, or stones, or some other group of objects. They sort them, stack them, line them up, etc., and as they do so, they begin the process of wiring their brains' perceptive processes. This leads to the beginning of the brain's ability to develop categories—round is different from square, rough is different from smooth, large is different than small, etc. At the same time, the brain learns that "more" takes up more space, and "more," when stacked, grows higher. It is from these and other childhood experiences that the brain is taught "how the world works," and these primal perceptions drive understanding for the remainder of our lives.

When a visualizer violates these perceptions, it can cause cognitive confusion in the viewer (e.g., Fig. 5.2). It can also create a negative emotional reaction in the viewer, because the visualization conflicts with "how the world works." The visualizer who created Fig. 5.2 was more interested in creating a "dripping blood" effect than in communicating clearly to her viewer, and the Internet firestorm that erupted from the publication of that graph is evident in the emotional reaction of many of the viewers.

Another common viewer reaction is toward visualizations that present percentage representations. Viewers understand the concept of percent when approaching a graph, and they know that the maximum percent level is 100%. However, Fig. 5.19 illustrates that some graphs may be produced to add up to more than 100%.

This is because the graph is usually representing multiple part-to-whole relationships at the same time, but not giving the viewer this insight. The solution to this perception problem is to always represent the whole when presenting a percentage, so that viewers can understand to which whole each of the parts is being compared (Fig. 5.20).

Fig. 5.19 Annotations sum to 243%



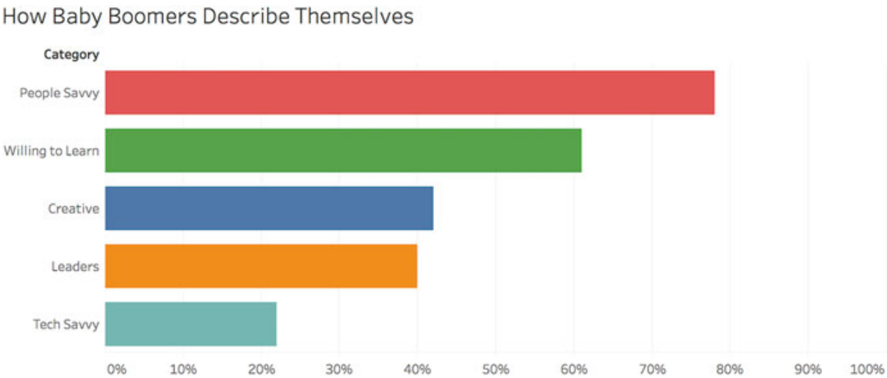


Fig. 5.20 Representing multiple parts to whole

There are obviously many more of these primal perceptions that viewers hold, and modern software packages make it rather difficult to accidentally violate them. In almost every case, these kinds of violations occur when the visualizer attempts to add some additional artistic dimension to a graph, instead of focusing on presenting the data with the highest possible fidelity.

Represent the Data Story with Integrity

Meta-Rule #6: *Chart with graphical and ethical integrity. Do not lie, either by mistake or intentionally.*

Finally, it is important that, at all times, visualizers work to accurately represent the story that is in their data. This means that the effect in the data must be accurately reflected in the visualization. Edward Tufte, in his classic book *The visual display of quantitative information*, provides a number of rules for charting graphical integrity. While we cannot cover them in detail here, we provide a summary.

First, Tufte introduces the "Lie Factor," which is the ratio between the effect in the visualization and the effect in the data. So if the effect in the data is 1, but in the visualization it is 2, the lie factor would be 2/1 or 2. In order for the visualization to accurately represent the data, the Lie Factor should be as close to 1 as possible. Figure 5.21 illustrates the Lie Factor.

Tufte's other rules for graphical integrity, when broken, create Lie Factor ratios that are higher or lower than 1. These other rules for graphical integrity are summarized below:

Use Consistent Scales. "A scale moving in regular intervals, for example, is expected to continue its march to the very end in a consistent fashion, without the muddling or trickery of non-uniform changes" (Tufte, p. 50). What this means is that when building axes in visualizations, the meaning of a distance should not

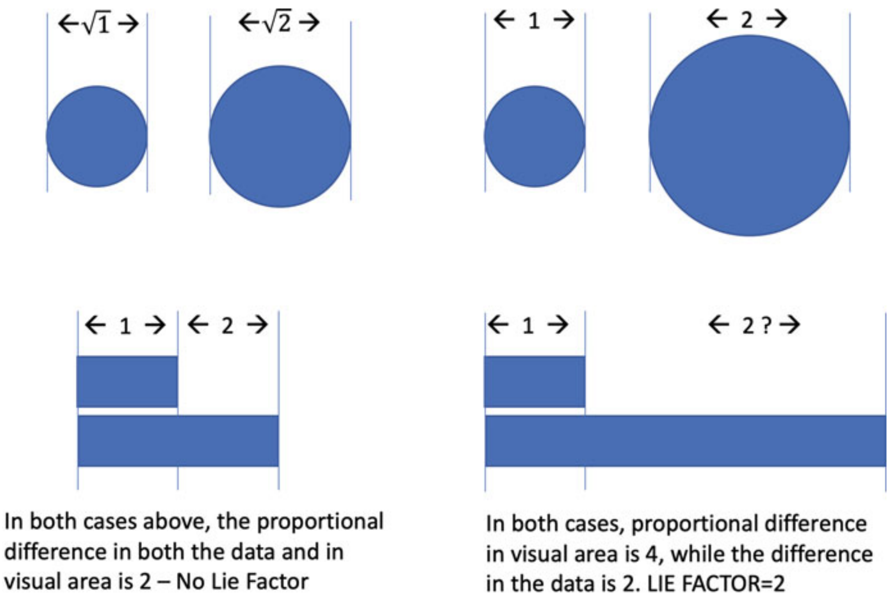


Fig. 5.21 Lie factors in proportional differences

change, so if 15 pixels represents a year at one point in the axis, 15 pixels should not represent 3 years at another point in the axis.

Standardize (Monetary) Units. "In time-series displays of money, deflated and standardized units [. . .] are almost always better than nominal units." This means that when comparing numbers, they should be standardized. For currency, this means using a constant unit (e.g., 2014 Euros, or 1950 USD). For other units, standardization requires consideration of the comparison being communicated. For instance, in a comparison between deaths by shooting in two states, absolute numbers may be distorted due to differences in population. Or in a comparison of military spending, it may be appropriate to use standardization by GDP, land mass, or population. In any case, standardization of data is an important concept in making comparisons, and should be carefully considered in order to properly communicate the relationship in question.

Present Data in Context. "Graphics must not quote data out of context" (Tufte, p. 60). When telling any data story, no data has meaning until it is compared with other data. This other data can be a standard value, such as the "normal" human body temperature, or some other comparison of interest such as year-over-year sales. For instance, while today's stock price for Ford might be $30, does that single data point provide you with any understanding as to whether that price is high, low, good, or bad? Only when data is provided within an appropriate context can it have meaning.

Further, as Tufte illustrates, visualizers choose what they wish to show, *and* what they choose to omit. By intentionally presenting data out of context, it is possible to change the story of that data completely. Figure 5.22 illustrates the data from Fig. 5.2 presented out of context, showing the year after the enactment of the "Stand Your Ground" law.

From the graph in Fig. 5.22 it is not possible to understand if the trend in gun deaths in Florida is any different than it was before the law was enacted. However, by presenting this data in this manner, it would be possible for the visualizer to drive public opinion that the law greatly increased gun deaths. Figure 5.23 adds context to this data set by illustrating that the trajectory of gun deaths was different after the law was enacted—it is clearly a more honest graph.

Show the Data. "Above all else show the data" (Tufte, p. 92). Tufte argues that visualizers often fill significant portions of a graph with "non-data" ink. He argues that as much as possible, show data to the viewer, in the form of the actual data, annotations that call attention to particular "causality" in the data, and drive viewers to generate understanding.

When visualizers graph with low integrity, it reduces the fidelity of the representation of the story that is in the data. I often have a student ask, "But what if we WANT to distort the data?". If this is the case in your mind, check your motives. If you are intentionally choosing to mislead your viewer, to lie about what the data say, stop. You should learn the rules of visualization so that (1) you don't break them and *unintentionally lie*, and (2) you can more quickly perceive when a visualization is lying to you.

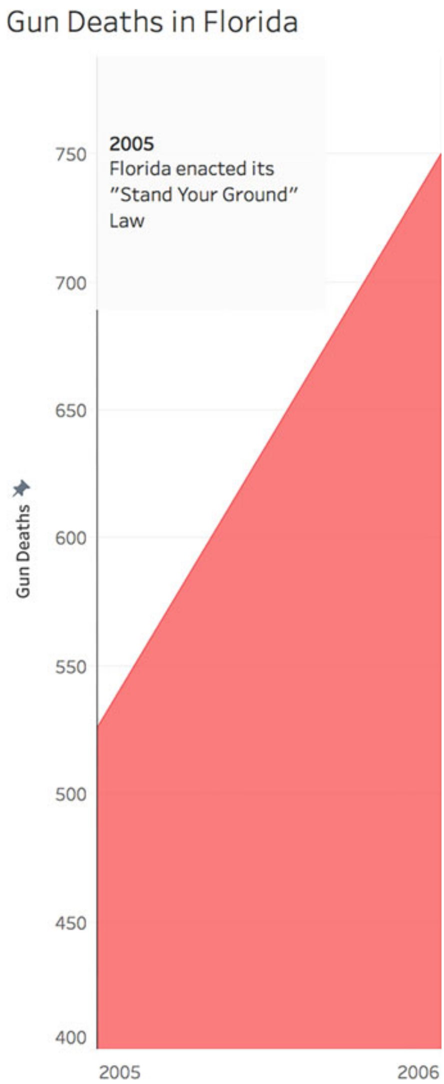


Fig. 5.22 Lack of illustration of context

4 Software and Data Visualization

Based upon the widespread recognition of the power of the visual representation of data, and the emergence of sufficiently inexpensive computing power, many modern software packages have emerged that are designed specifically for data visualization. Even the more generalized software packages are now adding and upgrading their visualization features more frequently. This chapter was not intended to "endorse" a particular software package. It stands to illustrate some of the rules

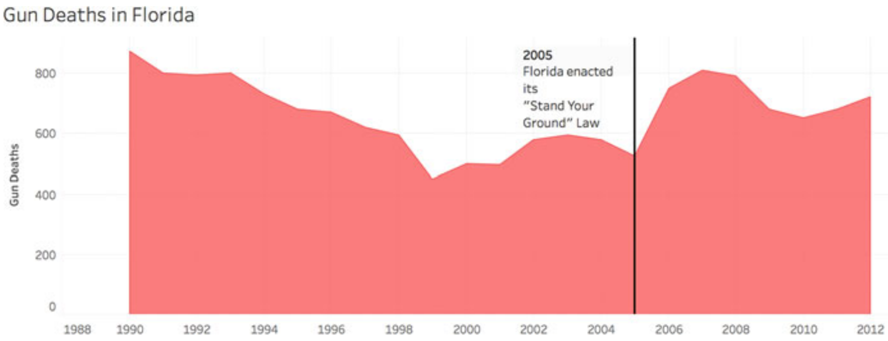


Fig. 5.23 Gun deaths in Florida, presented intuitively

that might explain why some software packages behave in a certain manner when visualizing particular relationships.

Because most visualizers do not have the freedom to select any software they choose (due to corporate standards), and because research companies such as Gartner have published extensive comparative analyses of the various software packages available for visualization, we do not recreate that here.

5 Summary and Further Reading

As stated above, a single chapter is far too little space to describe the intricacies of data visualization. Few (2009) and Tufte and Graves-Morris (1983) are good sources with which to broaden your knowledge of the "whys" of data visualization.

Exercises

Ex. 5.1 Answer the following conceptual questions:

(a) What is the key issue with using 3D graphs?
(b) When displaying differences in a single data dimension, explain why using differences in object area is suboptimal.
(c) How may data dimensions can you represent on a scatterplot graph?
(d) Which kind of color should you use when representing different levels of a single variable?
(e) What are some problems with gradients?
(f) Find an example of a quantitative graph in the media. Evaluate whether or not the graph is properly conforming to Tufte's principles described above.

Ex. 5.2 Scenarios

(a) You wish to provide a visualization that illustrates the rank and relative differences in various salespersons' results. Which graph would be the most "not wrong"?
(b) You wish to denote categories of customers using color. Should you use rainbow colors or gradient colors?
(c) You wish to illustrate the percentage of donations coming from a particular percentage of donors. What kind of relationship(s) are you attempting to illustrate? Which visualization would be the most "not wrong"?
(d) You wish to illustrate the amount that each of your product lines contributes to the percentage of total sales over a 5-year period. What would be your choice for the most "not wrong" graph?

Ex. 5.3 Two views of the same data

Go to the New York Times website, and view this interactive visualization.

http://www.nytimes.com/interactive/2012/10/05/business/economy/one-report-diverging-perspectives.html?_r=1&. Accessed on Feb 23, 2019.

While this visualization provides two views of the same data, critique its success at providing alternative viewpoints. Notice how the different visualizations utilize/break Tufte's rules in order to shift perception.

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician, 27*(1), 17–21.
Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57.
Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, CA: Analytics Press.
Playfair, W. (1801). "The statistical breviary; shewing, on a principle entirely new, the resources for every state and kingdom in Europe; illustrated with stained copper-plate charts, representing the physical powers of each distinct nation with ease and perspicuity. By William Playair".
Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 2, No. 9). Cheshire, CT: Graphics Press.