Data Visualization
Communicating Visual Stories

John F. Tripp, PhD
Baylor University

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 analyst must be concerned with communicating the results to the reader. Unfortunately, human perception isn't optimized to the understanding of 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, utilizing visual representations of data.

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

Working With (and not against) Human Perception

Consider Figure 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 of the variable greater or less at point 2 compared with point 1?

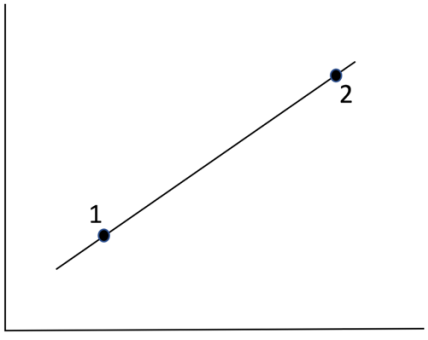


Figure 1

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 since you were a child, you've learned that when you stack something (like blocks, rocks, etc.), the more you stack, the higher the stack becomes. So, from

a very early age, you learn that “up” means “more”. Now consider Figure 2. Based on this graph, what happened to gun deaths after 2005[1]?

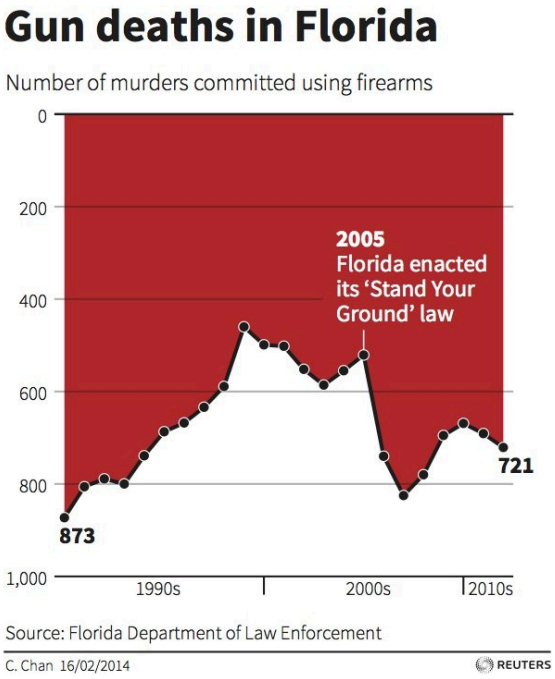


Figure 2

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 what happened? Observe the Y axis more closely. As you can see, the graph designer in this case inverted the Y axis, making higher values lower in the graph than lower values. While many readers may 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 and established convention, but is a violation of the principles that drove the establishment of the convention[2].

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

Why we Visualize Data

Human visual and cognitive processing is not optimized to process relationships between large sets of numbers. While humans are good at comparing several values against each other, humans

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

[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

are not good at drawing inferences using sets or tables of numbers to attempt to communicate and compare trends or large groups of numbers. Let's look at a famous example.

In Figure 3, the classic example of "Anscombe's Quartet" is presented. In 1973, Francis Anscombe created these sets of data to illustrate the importance of visualizing data (need citation). When you consider these four data sets what sense can you make of them? How much are the sets the same? Different?

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

Figure 3. Anscombe's Quartet

> *"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, pg. 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 Figure 3, most people note that sets 1-3 are similar, primarily because the X values are the same, and appear in the same order. However, these four sets of numbers are statistically equivalent, for all intents and purposes, to typical statistical analyses, these four data sets are the same. If a typical regression analysis was performed on the four sets of data in Table 1, the results would be *identical*. Among other statistical properties, the following are all valid for all four sets of numbers in Figure 3.

- 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 = .67

However, if one compares the sets of data visually, the differences between the datasets becomes immediately obvious, as illustrated in Figure 4.

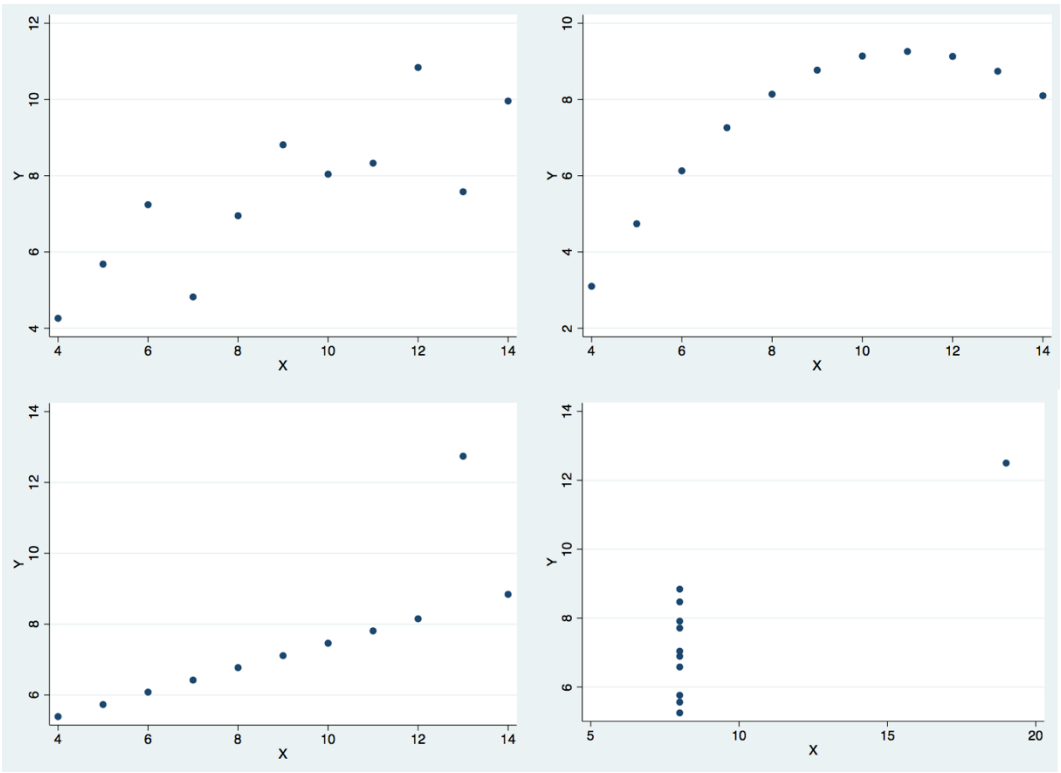


Figure 4

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 are single-value based. However, when relationships between sets or groups of numbers are intended to be presented, human perception is much better served by using 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 visualization to become an excellent consumer and creator of graphical data visualizations.

Visualization as a Cognitive Aid

Humans are bombarded by information from multiple sources and through multiple channels. This information is gathered by humans' 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. A huge deal of sensory input is simply ignored by the brain, other input is dealt with based upon 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 4 "chunks" of information at once <<add citation>>. For this reason, visualizers must build graphs that do not rely on the user to hold multiple chunks of information and use those chunks 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. Instead, every visualization should be created with a particular cognitive task in mind, and then constructed in a manner that assists the viewer in the completion of the cognitive task. For instance, in my experience with real-world corporate "dashboards", they tend to be designed in a manner that presents a lot of data that different users might use for different cognitive tasks. As such, they do not assist the completion of any of the cognitive tasks, as users must find the different pieces of information that are important to their tasks, and then perform additional cognitive steps to generate a result that assists with understanding.

Designing and building visualizations that are fit for purpose, meaning fit for the use in the processing of a defined cognitive task (e.g., communicating the status of sales in a region, communicating 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 more visualizations, that are more tightly focused than many visualizations created in businesses today. Even so, the key criteria for the creation of a visualization is to reduce cognitive load for a particular cognitive task. A good visualization, focused on a task, calling out the key information for the task, and providing computational assistance for the viewer will allow them to more quickly build understanding, and is of significantly more value than creating a single, multi-purpose visualization that fails to assist in any viewers' cognitive tasks.

Six Meta-Rules of 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 what to do to 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 1.

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

By following these meta-rules, data visualizers will be more likely to display graphically 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 due to 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 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 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, or, 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 the visualization, to the detriment 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. Besides artistic concerns, visualizers often, especially when visualizing multiple relationships, choose to use multiple kinds of graphs to add "variety". 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 both keep track of the data values, as well as the different manner in which the data has been encoded.

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 it's simplest, data that is quantitative is data that is 1) numeric, and 2) it is appropriate to use for mathematical operations (e.g., 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 number, City, State, Country, Department).
When first approaching the information, the first question is whether a visual representation of the data is needed at all. For instance, in the graphic shown in Figure 5, a text statement is redundantly presented as a graph.

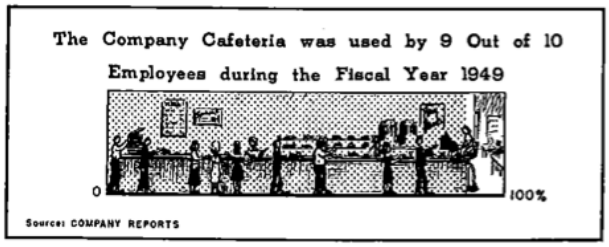


Figure 5

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:

[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 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 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

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. Table 2 provides a list of the most likely graphs to be used for various relationships.

Table 2: 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	Plus or minus Variance	4

	Bullet Graph	Difference Relative to	
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

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 5, and in an interactive graph 6:

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

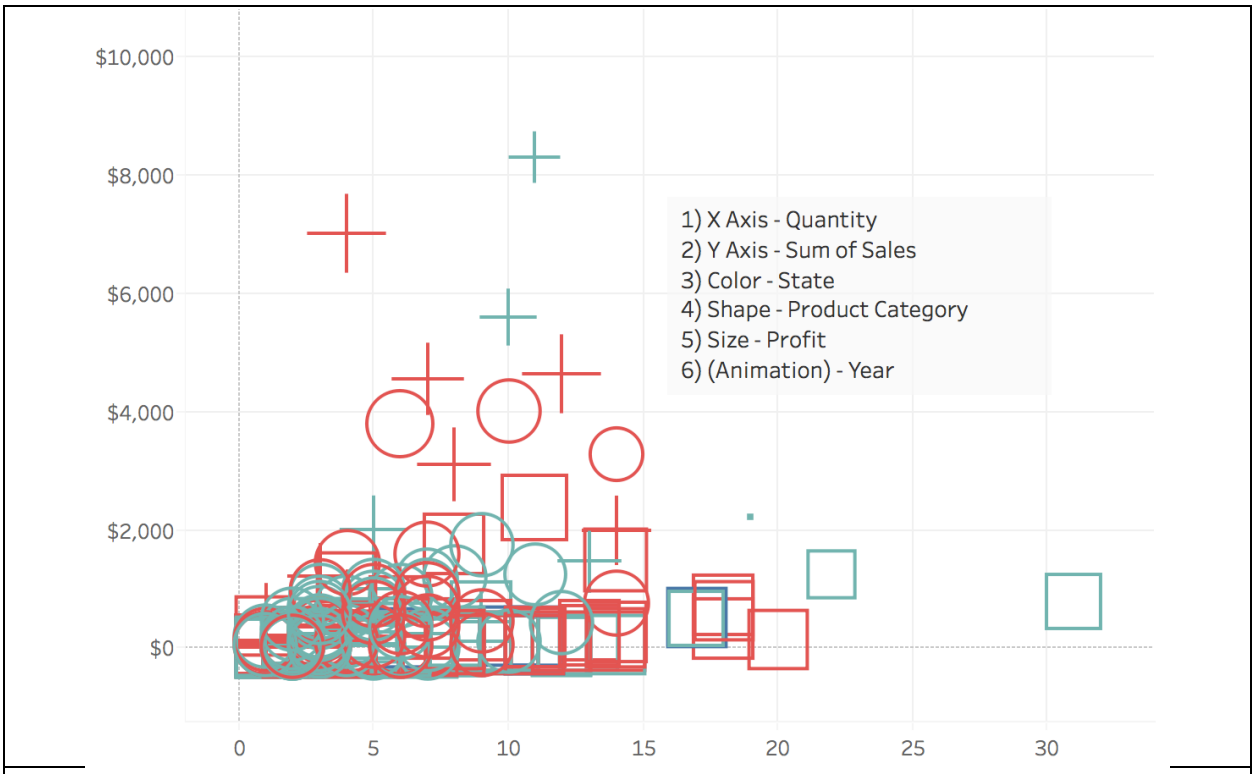


Figure 6

However, many graph types, because of their nature, reduce the number of possible dimensions that can be displayed (Table 2). For instance, while a scatterplot can display all six of the dimensions, while a filled map can only display three: the use of the map automatically eliminates the ability to modify dimensions 1-4. As such, we are left with the ability to use different levels of one data dimension as the color of each country, and we could possibly use animation to show 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, as it is difficult for humans to compare differences in sizes of different shapes (See Figure 6).
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 use create 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, or, 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, data, we often wish to tell a story of differences, such as deviations from plan, budgets vs. actual, etc. When telling a story of differences, do not rely on the viewer to calculate the differences themselves. Figure 7 illustrates a visualization that relies on the viewer to calculate differences. Instead, represent the actual deviation, as illustrated in Figure 8.

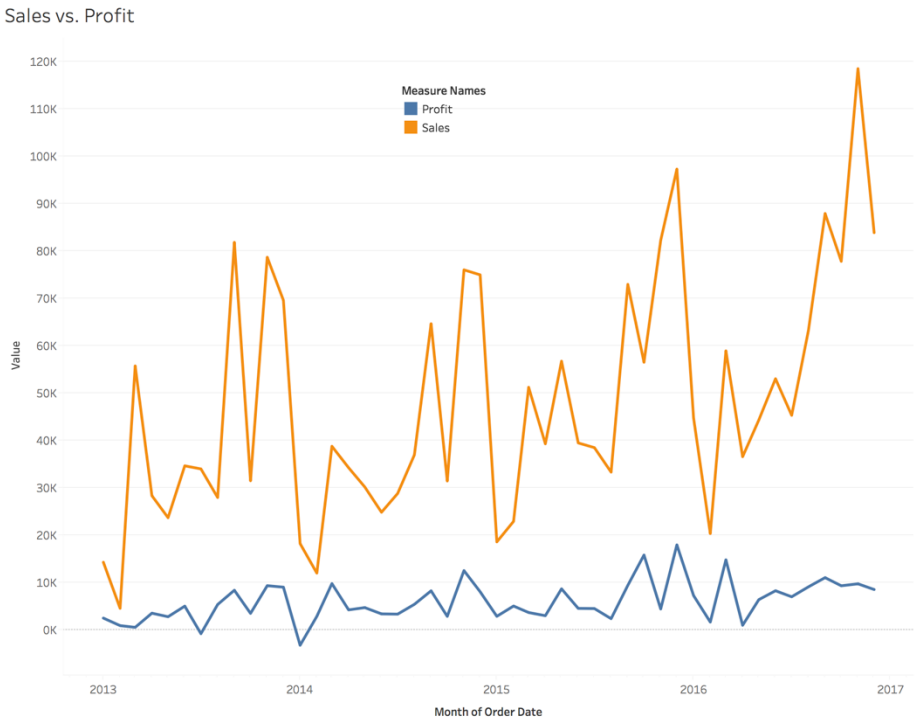


Figure 7

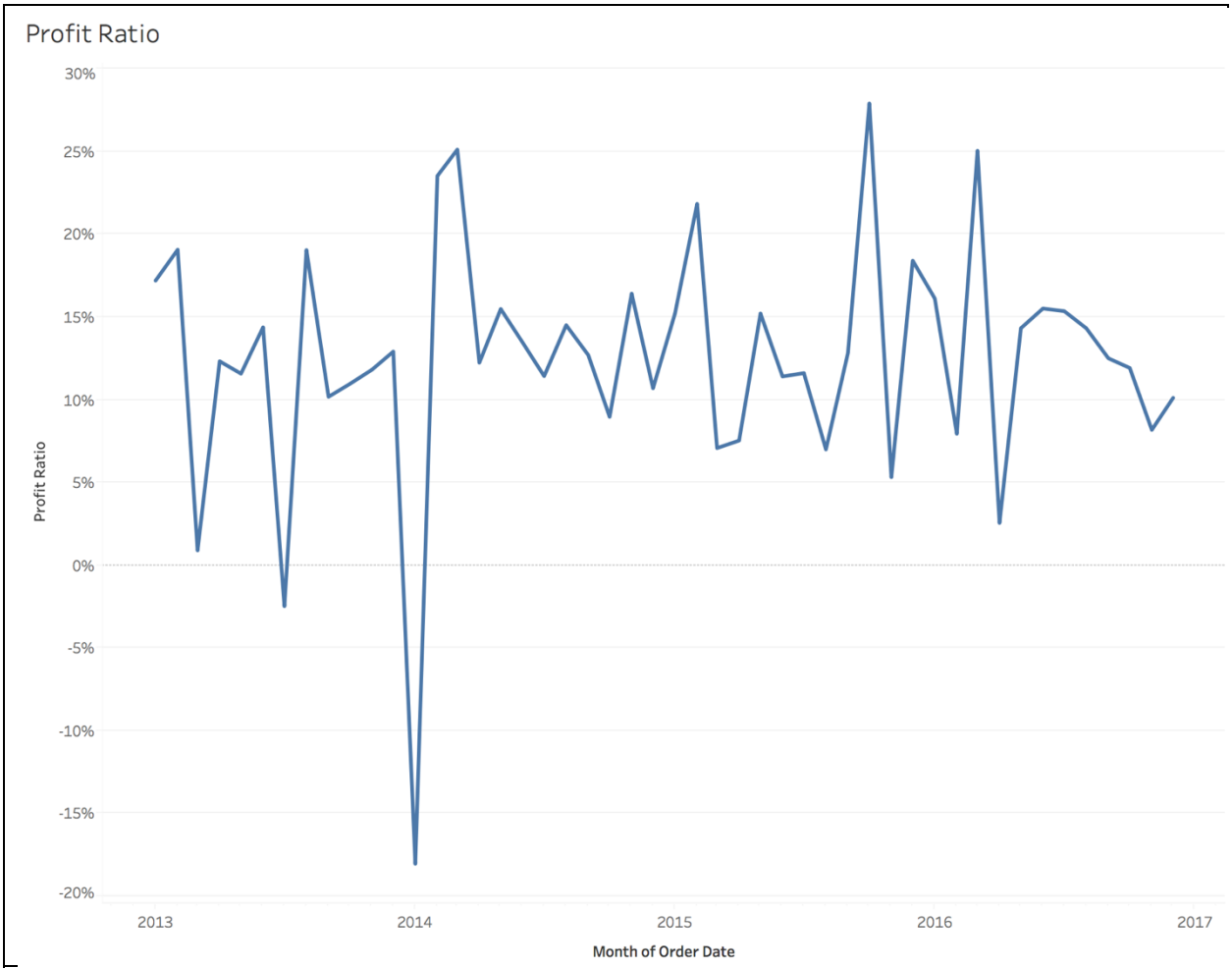


Figure 8

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

Single Dimensionality

> ***Meta-Rule #3:*** *In general, do not ask viewers to compare in two dimensions, or, 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, it is possible for visualizations to become distorted. Figure 9 illustrates the issue when visualizers directly represent data differences in two dimensions. In this example, the visualization is depiciting the levels of a variable, where one case has the value of 1, and the other case the value of 2. Both examples are properly encoded, as the area of each exactly maintains the proportion of 1:2 that is found in the data.

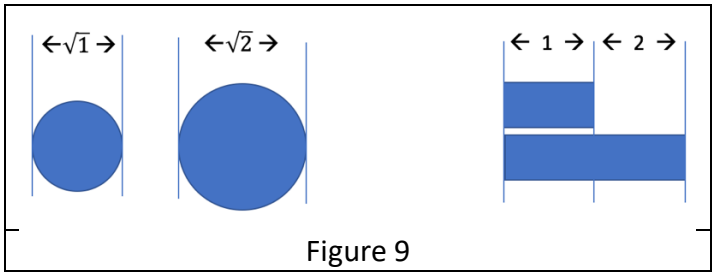


Figure 9

However, which of the visualizations more clearly communicates what is present in the data? The simple bars, which only vary on one dimension much more clearly illustrate the relationship that is present in the data.

To 3-D or Not to 3-D

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

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

Table 3 YTD Sales Data

In this data set, Deepika is well above average, and that Donald is well below average. However, When viewing Figure 10, the pie chart (comparison on 2 dimensions) is less clear than the table (from the chart, who is the #2 salesperson?), and the 3-D chart greatly distorts the % share of Deepika due to perspective.

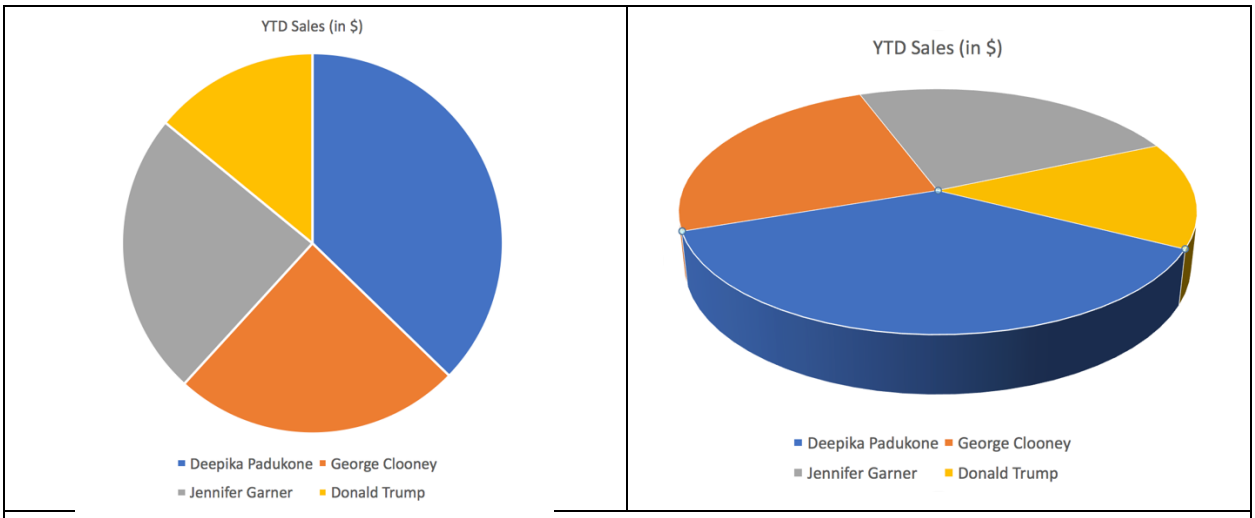


Figure 10

Even more of an issue with 3-D charts is that the placement of the data points changes the perception of the values in the data. Figure 11 illustrates this point. Note that when Deepika is rotated to the back of the graph, the perception of her % share of sales is reduced.

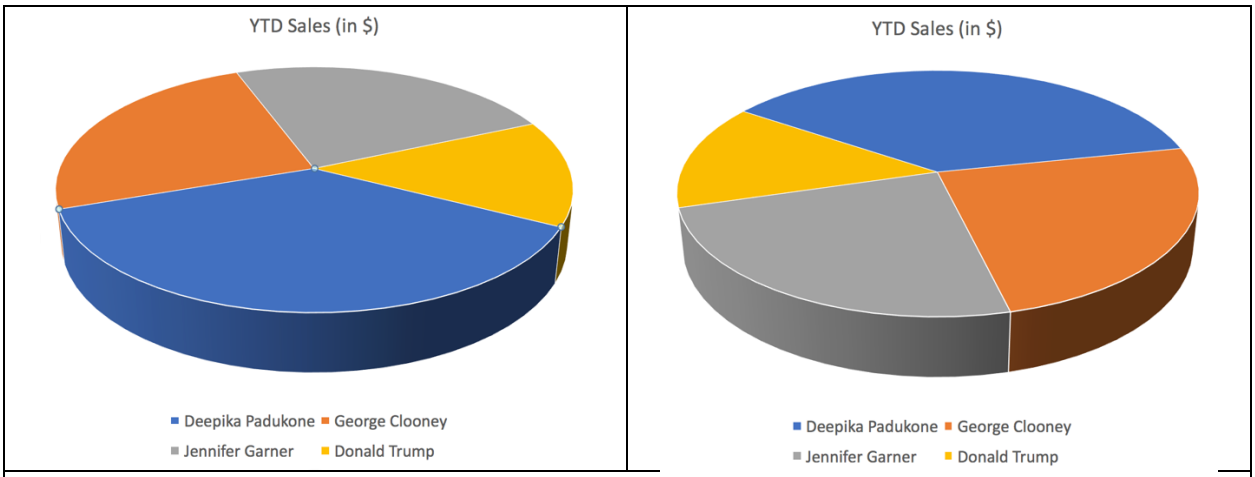


Figure 11

Pie Charts?

Although the previous example uses pie charts, this was simply to illustrate the impact of 3-D 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 2-D example in Figure 10, it is very difficult to compare the differences between George and Jennifer. The typical response to this in practice is to add the % values to the chart, as Figure 12 illustrates. However, at this point, what cognitive value does the pie chart add that the viewer would not have gained from Table 3?

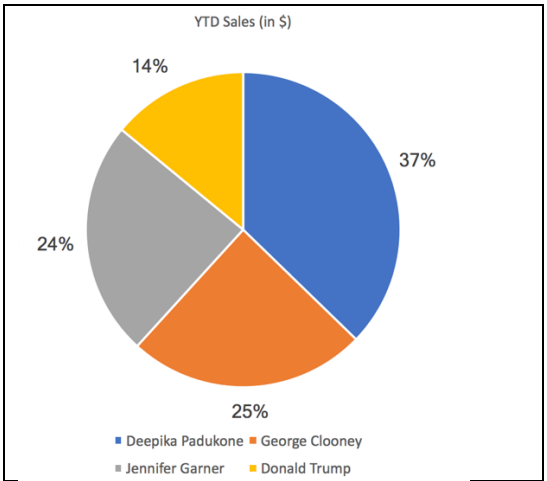


Figure 12

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, 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 (Figure 13) may be best.

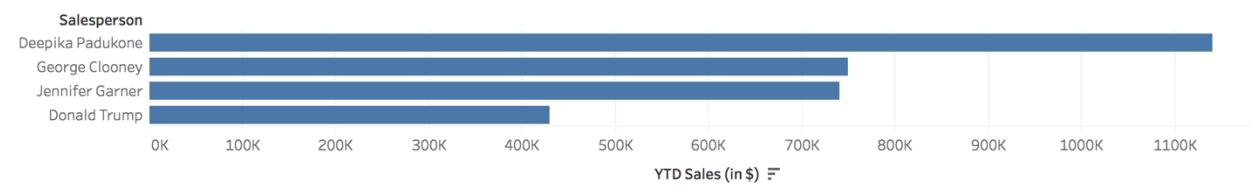


Figure 13

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 discussion of the proper use of gradients), within a grid of cells that represent the possible intersections of two categories – perhaps sales by region by quarter (Figure 14).

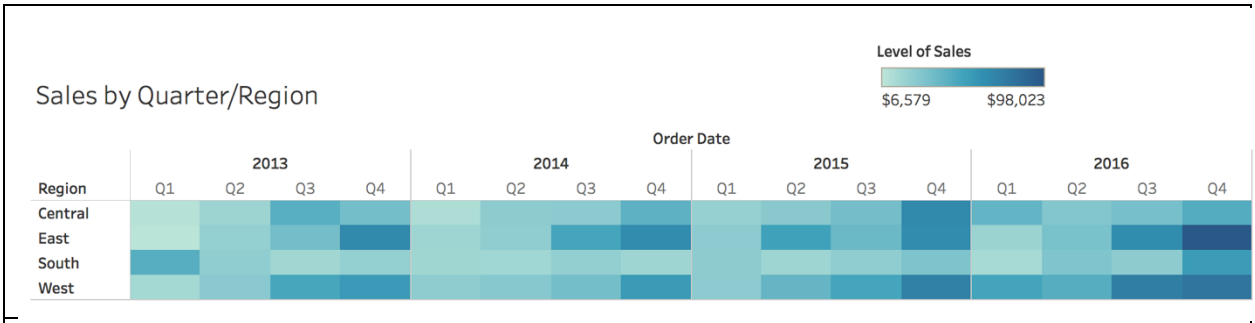


Figure 14

The heat map illustrates that the fourth quarter tends to be the strongest in all regions while, compared to the other regions, the East region seems to perform consistently highest in the 4th quarter. As can be seen, the heat map is good for illustrating comparative performance, but only in a general way, as an overview. To allow users to dig deeper using 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 (Figure 15).

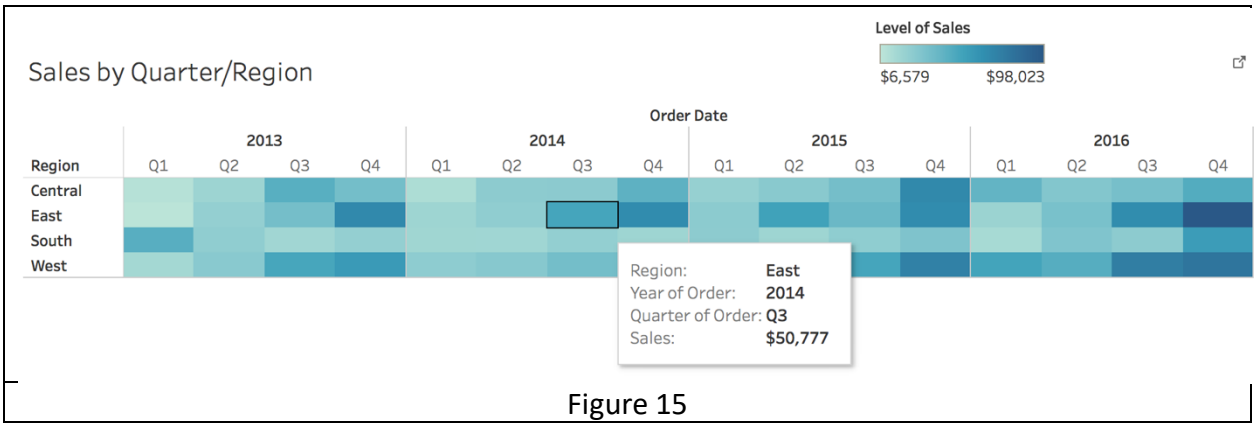


Figure 15

In almost every case, there are multiple choices for visualizations that are “not wrong”. The visualization that is “fit for purpose” is that 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: *Use color with meaning, or, perception of 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 Figure 16. Which of the five small rectangles does your visual perception tell you is the lightest in color?

Figure 16

Most people (if they're honest), will quickly answer that the rightmost rectangle in Figure 14 is the lightest in color. However, as Figure 17 illustrates, all five small rectangles are actually the same color.

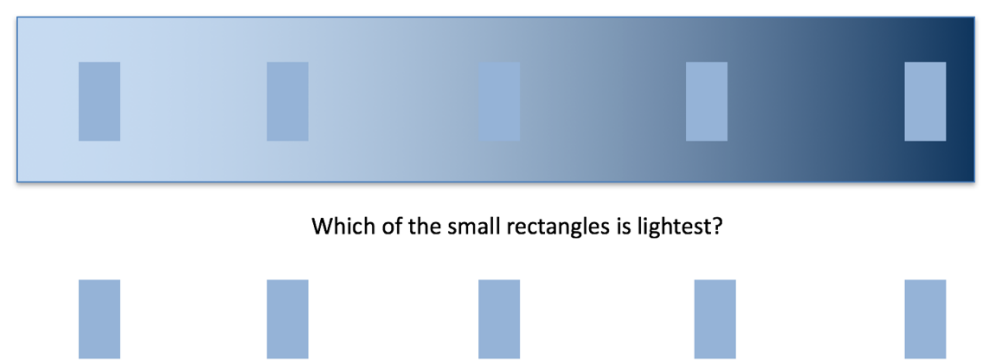
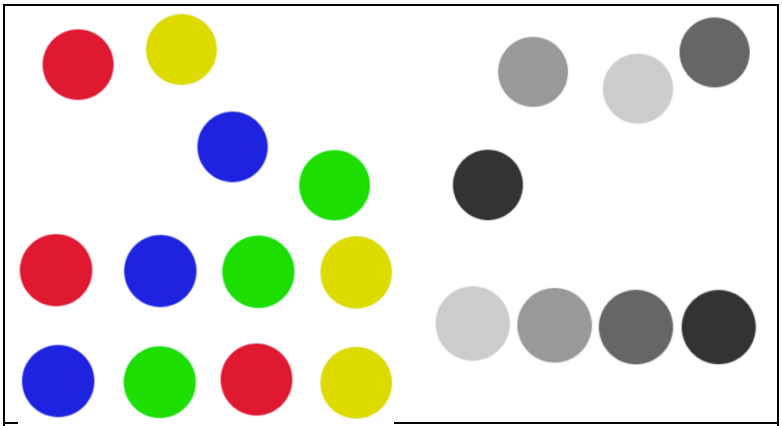


Figure 17

The optical illusion presented in Figure 16 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 behind the small rectangles 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 Figure 18, and 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.*



When rainbow colors are used, establishing order is inconsistent, even when the same person tries to order the colors in multiple tests. However, a gradient (of any color) is much more easily and consistently ordered.

Figure 18

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 doesn't 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 brighter, more primary colors.

[4] Based upon the discussion of the concept of the non-order in 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 multi-color gradients unless the context has a meaning already established for the colors.

For this reason, when designing a visualization, use "natural" colors as the standard color palette, and then use brighter, more primary colors for emphasis (Figure 19).

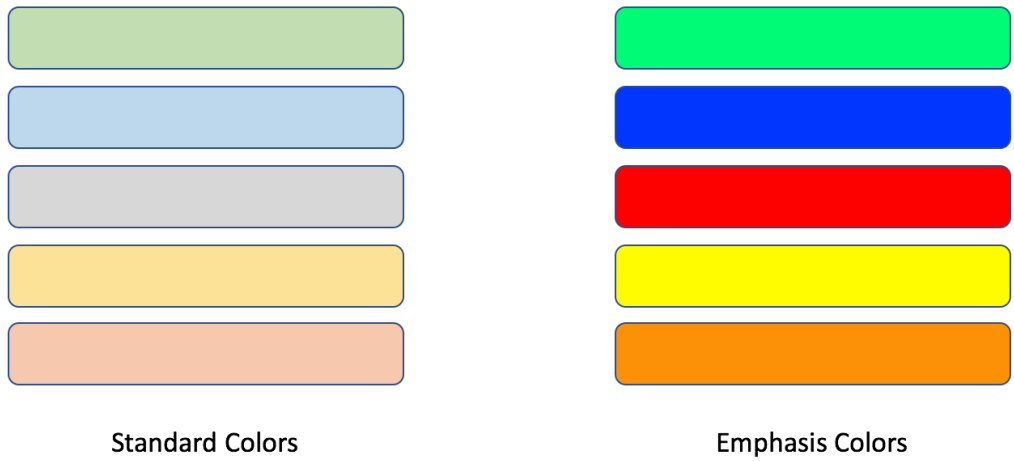


Figure 19

By using more natural colors in general, viewers will more calmly be able to interpret the visualization. By using more primary colors for emphasis, the visualizer can choose 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, or, up means more.*

In the example provided in Figure 2, we reviewed the disorientation that can occur when viewers' primal perceptive instincts of the viewer. 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 Figure 2. However, why is it that humans perceive "up", even in a 2-D 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 – i.e., 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., Figure 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 Figure 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 evidence to the emotional reaction of many of the viewers.

Another common viewer reaction is to 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, Figure 20 illustrates that graphs often are produced that add up to more than 100%.

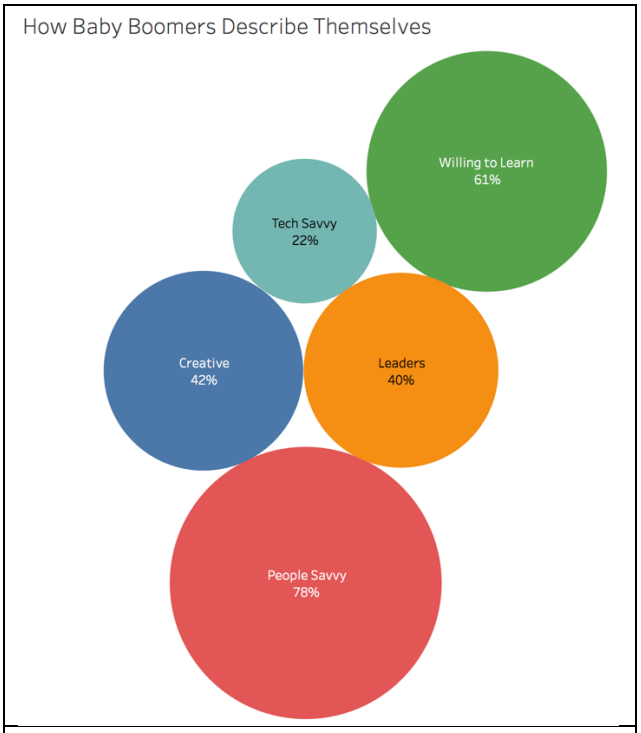


Figure 20

This is because the graph (as in Figure 20) is usually representing multiple part to whole relationships at the same time, but not giving the viewer the cue that this is so. 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 (Figure 21).

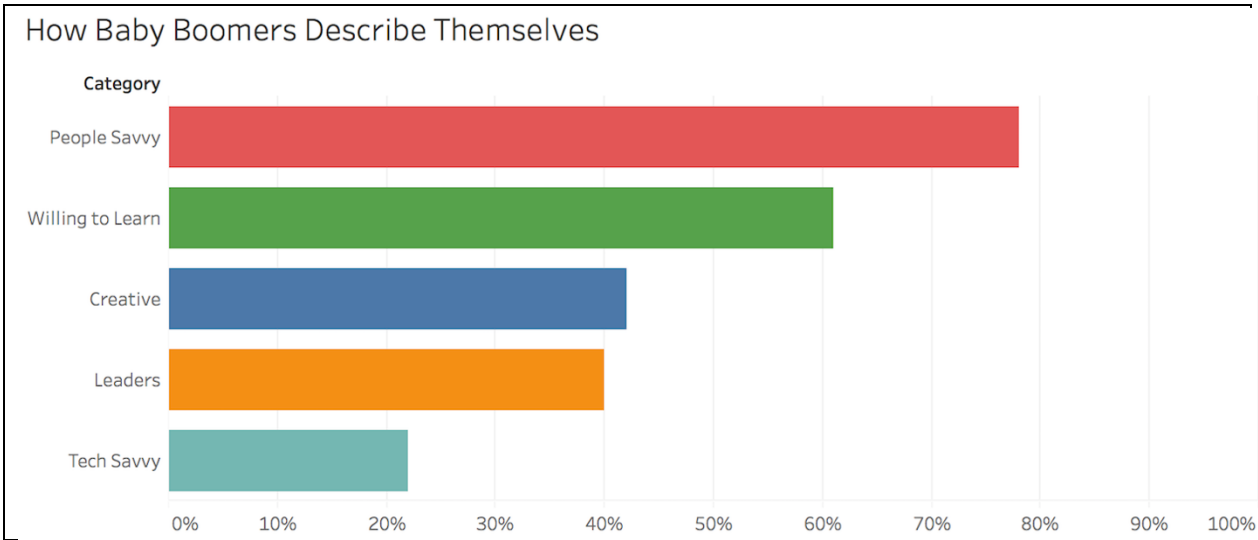


Figure 21

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, or, don't 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 22 illustrates the Lie Factor.

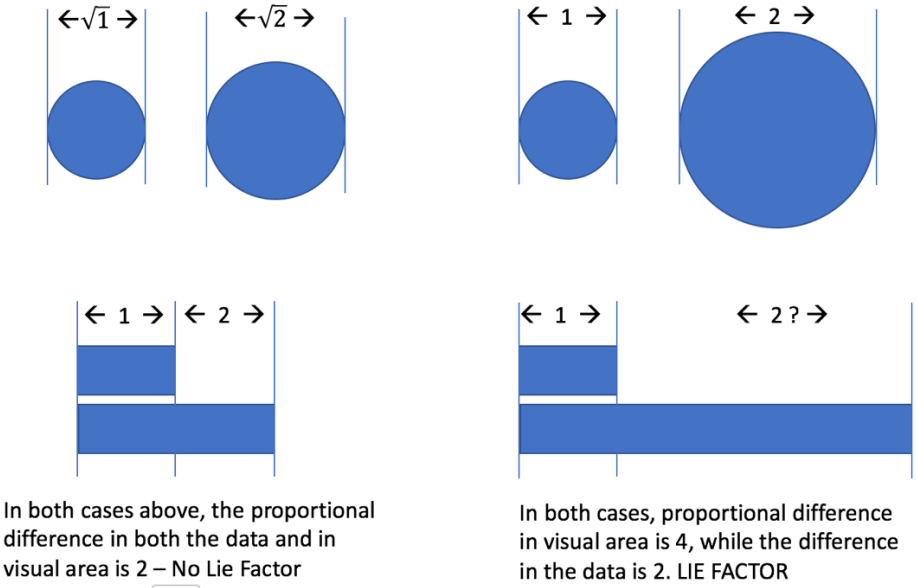


Figure 22

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, Page 50). What this means is that when building axes in visualizations, the meaning of a distance should not change, so if 15 pixels represents a year at one point in the axis, 15 pixels should not represent three 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 (i.e., 2014 Euros, or 1950 US Dollars). 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, standardization by GDP, or land mass, or population may be appropriate. 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, Page 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 providing data 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 data completely. Figure 23 illustrates the data from Figure 2 presented out of context, showing the year before and the year after enactment of the stand your ground law.

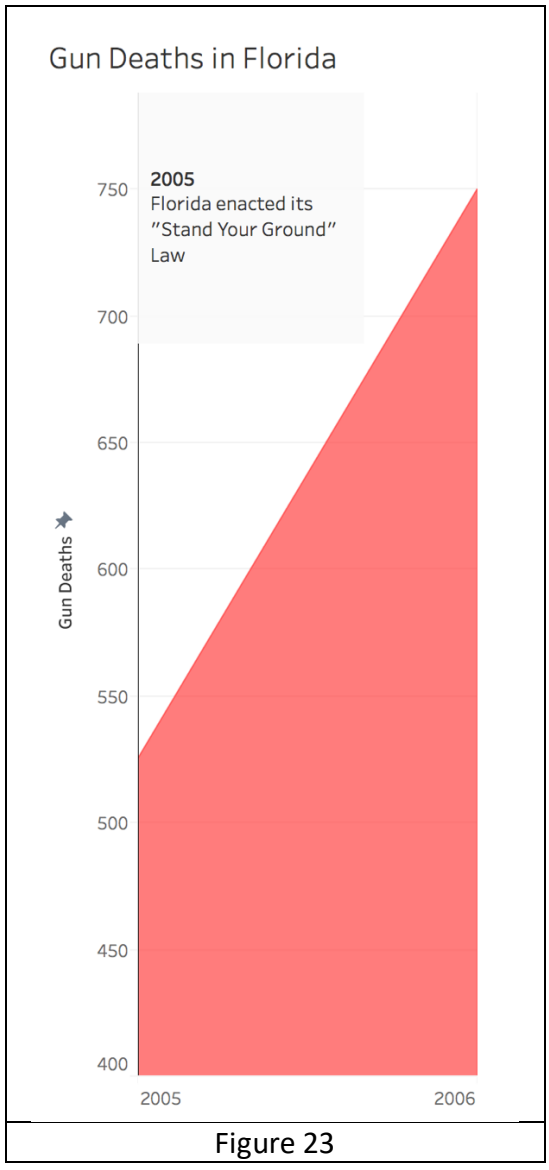


Figure 23

From the graph in Figure 23 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. However, below in Figure 24, the context is shown, it illustrates that the trajectory of gun deaths was different after the law was enacted, and is a more honest graph.

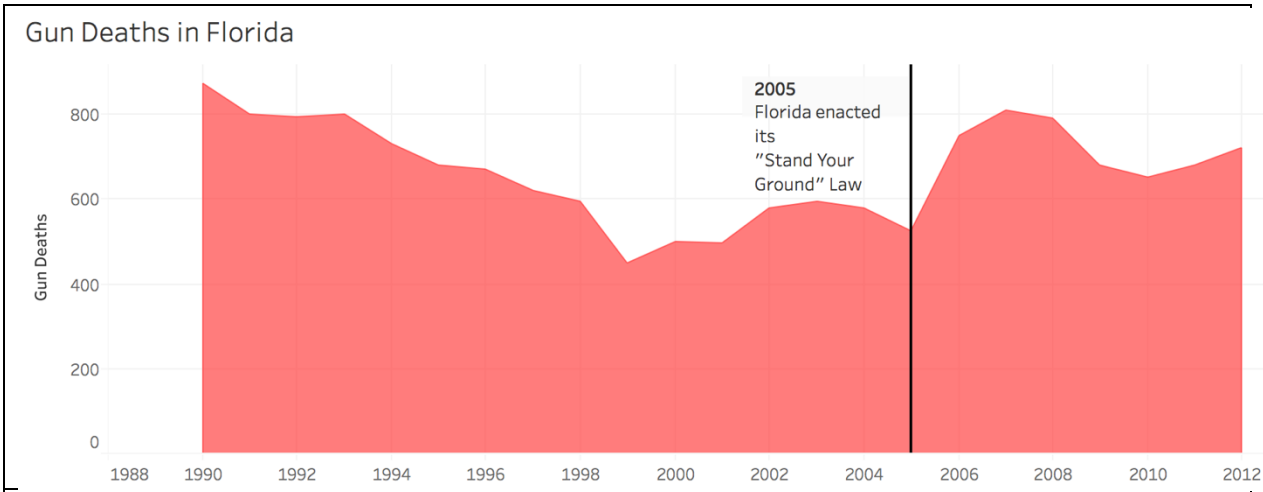


Figure 24

Show the data. "Above all else show the data." (Tufte, Page 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 viewer 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.

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 to generate complex visualizations quickly, many modern software packages have emerged that are designed specifically for data visualization, and more generalized software packages add and upgrade their visualization features constantly. This chapter was not intended to "endorse" a particular software package, but instead sought to illustrate some of the rules that might explain why some software packages behave in a certain manner when visualizing particular relationships.

Because most visualizers don't 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.

For Further Reading

As stated above, a single chapter is far too little space to describe the intricacies of data visualization. The following resources are good sources with which to broaden your knowledge of the "whys" of data visualization.

Few, Stephen. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, Oakland, California, 2009.

Tufte, Edward R., and P. R. Graves-Morris. *The Visual Display of Quantitative Information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

References:

Playfair, William. "The statistical breviary; shewing, on a principle entirely new, the resources fo every state and kingdom in Europe; illustated with stained copper-plate charts, representing the physical powers of each distinct nation with ease and perspicuity. By William Playair." (1801).

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. ***27*** *(1): 17–21*

Exercises

Exercise 1: Answer the following conceptual questions.

a. What is the key issue with using 3-D graphs?
b. When displaying differences in a single data dimension, explain why using differences in object area is sub-optimal.
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.

Exercise 2: Scenarios

a. You wish to provide a visualization that illustrates the rank and relative differences 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 contribute to the percentage of total sales over a 5-year period. What would be your choice for the most "not wrong" graph?

Exercise 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&

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.