

This paper was originally presented at the Virtual Fox Fest online conference in May, 2021. <https://virtualfoxfest.com/>



Data Visualization for the Database Developer

*Rick Borup
Information Technology Associates, LLC
701 Devonshire Drive, Suite 127
Champaign, IL 61820
217.359.0918
rborup@ita-software.com*

Data visualization is a powerful tool for turning raw data into useful information. Because vision is the dominant human perception, an effective visual presentation of data can quickly convey meaning that might otherwise be obscured or difficult to ferret out. Data visualization is much more than just pie charts and bar graphs, and is just as relevant for small data as it is for big data. Come to this session to learn tools and techniques for turning the data from your applications into meaningful information for your users.

Introduction

Data visualization is the art of conveying information visually rather than presenting data in verbal or numerical format. Good data visualizations can quickly reveal the meaning of data—patterns, trends, proportions, status at-a-glance, etc.—that might otherwise be obscured or take time to discover.

Data visualization can be useful in many ways and for many kinds of audiences, particularly when the whole is more important than the parts.

Two things stand out about data visualization: How difficult it is to create a good one, and how easy it is to mislead people (unintentionally or otherwise) with a bad one.

Human Perception

Of our five senses, vision is the most dominant. Humans evolved to survive in their environment principally by sight – it’s how we get around, how we find food, how we avoid threats, and how we recognize others. As John Medina puts in his book *Brain Rules*, “Vision trumps all other senses.” [1]

Summarizing experimental research in the field, Medina writes that after 72 hours people remember only about 10 percent of information that was presented orally (i.e., something they only heard) vs about 65 percent if you add a picture (i.e., something they both heard and saw). Although he does not cite percentages for written text vs pictures, he says that both “text and oral presentations are not just less efficient than pictures for retaining certain types of information; they are *way* less efficient.”

Based on our own personal experiences, would any of us disagree?

It’s not much of a stretch to extrapolate from the inefficiency of text to the inefficiency of numbers presented as text. When embedded within text in a paragraph or presented in rows and columns, individual numbers can be recognized but patterns and trends are much more difficult to spot. Again quoting Medina, “Put simply, the more visual the input becomes, the more likely it is to be recognized—and recalled.”

Consider the following examples from the opening page of Edward Tufte’s book *The Visual Display of Quantitative Information*. [2] The data are shown in **Figure 1**– eight columns of numbers in four sets of paired values. Looking only at the numbers, can you easily spot any trends or patterns?

I		II		III		IV	
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	5.00	5.0	5.73	8.0	6.89

Figure 1. When presented as numbers (text), these four sets of paired values do not reveal much in the way of patterns.

When these same four sets of numbers are presented graphically in simple scatter charts, the patterns immediately become clear.

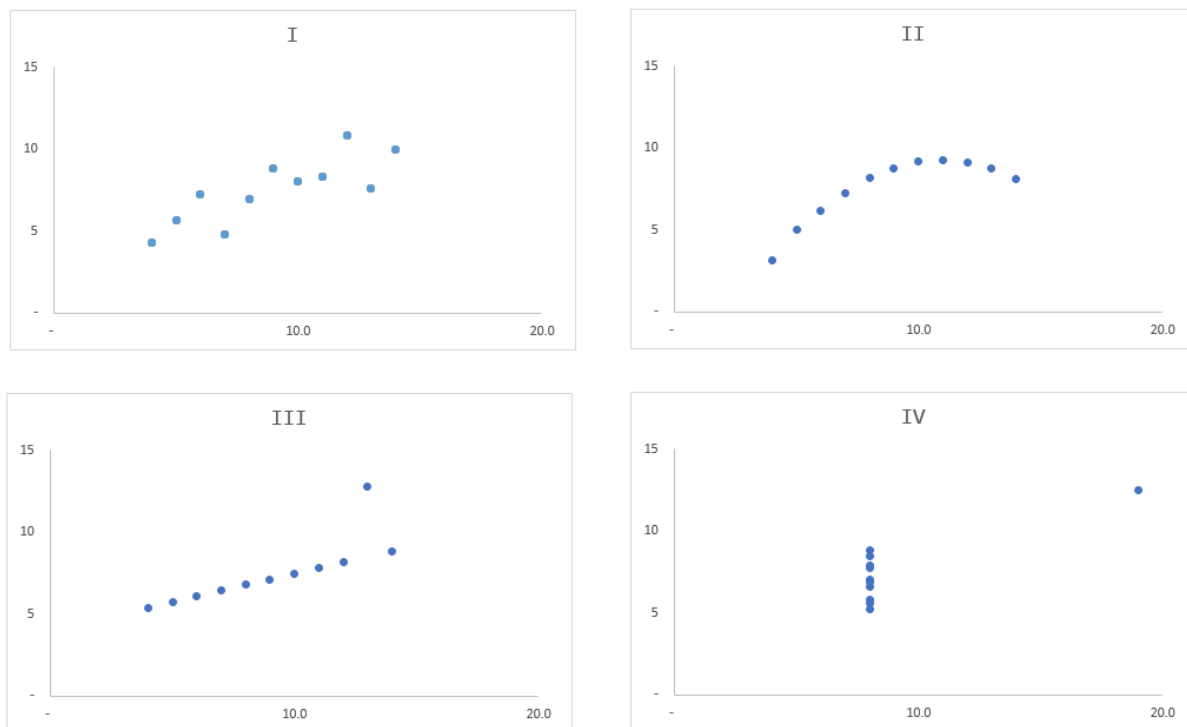


Figure 2. The patterns are clearly revealed when the numbers are presented graphically.

How we see

“We do not see with our eyes. We see with our brains.” [1]

The brain has to do a lot of work to read words. In fact, when we read our brains are processing thousands of tiny pictures in the shape of letters to form words, and from those words to derive meaning. A picture, on the other hand, conveys meaning more quickly. Which has the greater impact – a skillfully worded paragraph describing a beautiful sunset, or an actual photograph of that sunset?

The same can be said of reading numbers. When our goal is to convey the essence of information quickly and concisely, pictures are much more efficient than words or numbers alone.

What we see

We see what we expect to see, and we don't see what we don't expect to see. Take a minute and look at the picture of the rose in **Figure 3**.



Figure 3. Do you see anything unusual about this photo of a rose?

Did you see anything unusual about it? Look again. Do you see the dolphin? Most of us don't, because we don't expect to see a dolphin in a rose — yet it's there. If you still don't see it, look at **Figure 4**.



Figure 4. Yes, there is a dolphin hidden in that rose.

What does this have to do with data visualization? The image of the dolphin doesn't jump out at us partly because it's unexpected but also because it's not well differentiated from its surroundings. Its color and texture are similar to the rest of the rose, and its contours match those of the petals on which it's superimposed. The point is that even in a good visualization, the reader's eye might not be drawn to the thing you want to emphasize unless you make it stand out. There are many ways to do this – color, contrast, context, positioning, labeling, and so on.

Data visualization vs infographics

The terms *infographics* and *data visualization* are often used interchangeably. While data visualization is a form of infographics (short for *information graphics*), they are—to my way of thinking—two different things.

Infographics

An infographic is a combination of text and graphics. It may or may not present any data. If you do an online image search for infographics, most examples are what you might expect to see in advertisements or on informational posters. Infographics sometimes combine the visual representation of several sets of data with pictures, multiple fonts, decorative colors, and other elements intended to present an overall attractive image. Some infographics do not even represent data at all but simply attempt to convey information in a graphical and colorful way.

Figure 5 is an infographic that conveys information in an effective and visually pleasing manner, but it is not a data visualization because it does not present any data.



Figure 5. An infographic with no data – <https://venngage.com/blog/infographic-templates-free/>

Although it does contain some data, the image shown in **Figure 6** is still much more of an infographic than a data visualization. It presents many different pieces of numerical information. Various font sizes and colors are used to draw attention to one thing or another, but the data are presented as numbers rather than as visualizations.

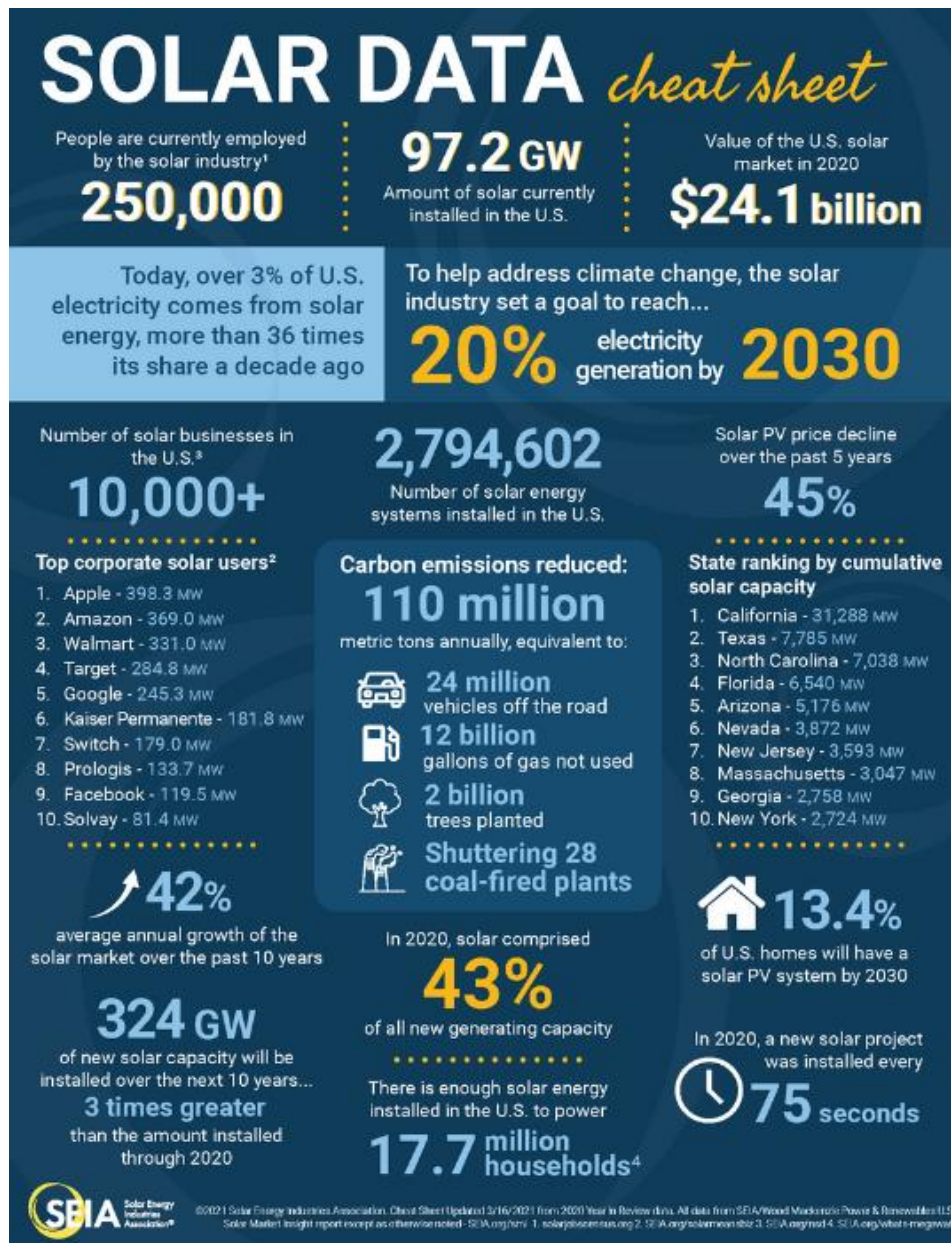


Figure 6. An infographic with data but no real data visualization
<https://www.seia.org/research-resources/solar-data-cheat-sheet>

The third example, shown in **Figure 7**, conveys a lot of great information, but is it a data visualization? It's formatted to look like a bubble chart, a form of data visualization where the area of each circle represents a value in proportion to the others. While the size and color of the circles effectively convey increasing risk as you move from left to right, what's being represented is not numerical data.

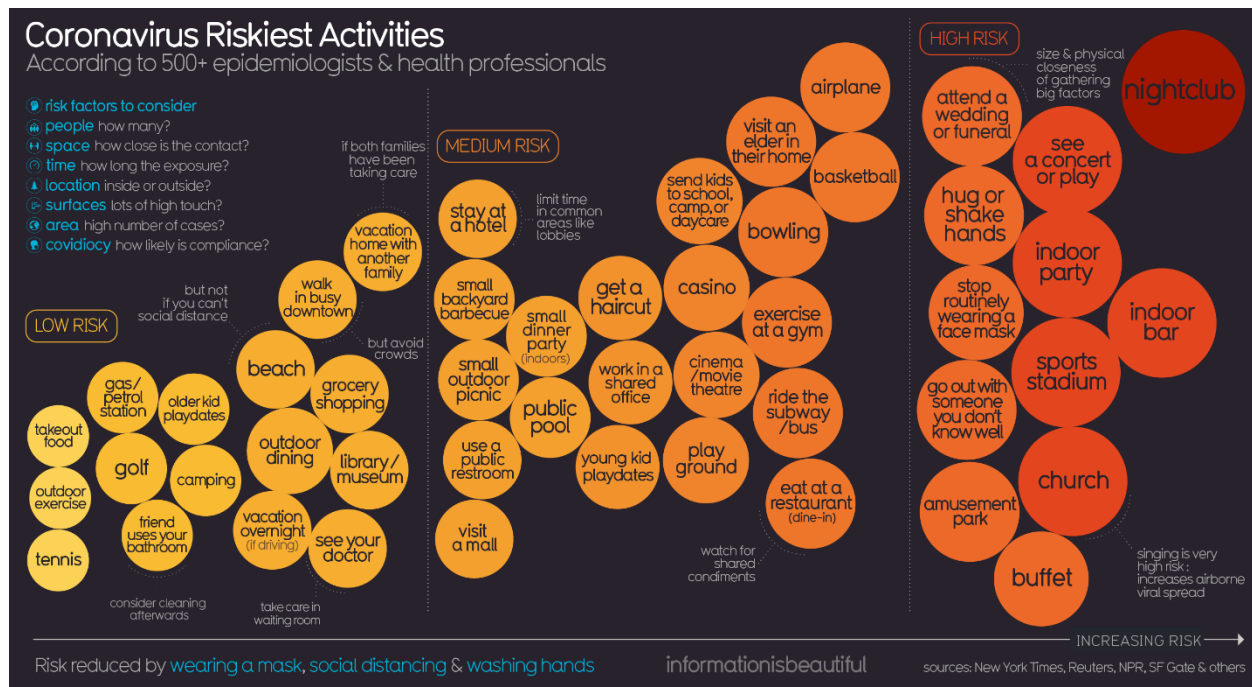


Figure 7. This nicely designed infographic looks like data visualization but isn't.

<https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/#activities>

© information is beautiful, 2021

If you're interested in creating your own infographic, <http://piktochart.com/> has some templates to help you get started.

Data visualizations

Unlike the infographics in the previous three examples, true data visualizations are concerned not so much with creating a pretty picture as with representing the data visually in a meaningful way. Good data visualizations are pleasing to look at but seek to avoid eye candy and to eliminate unnecessary elements from the display, striving instead to help the viewer to focus on what's important without meaningless distractions. Good data visualizations can certainly be attractive, but do so without being cluttered.

Figure 8, taken from the Verizon Wireless website, is an example of what I consider to be a good data visualization. It's effective because it not only shows the actual numbers—57minutes, 233 text messages, and 0.30 GB of data—but also a clearly conveys their proportion relative to the allowable maximums under my plan. Knowing I was within only a few days of the end of the billing cycle, I could see at a glance that I was well within my limits – which, after all, is the meaning of the data from my point of view as the user.

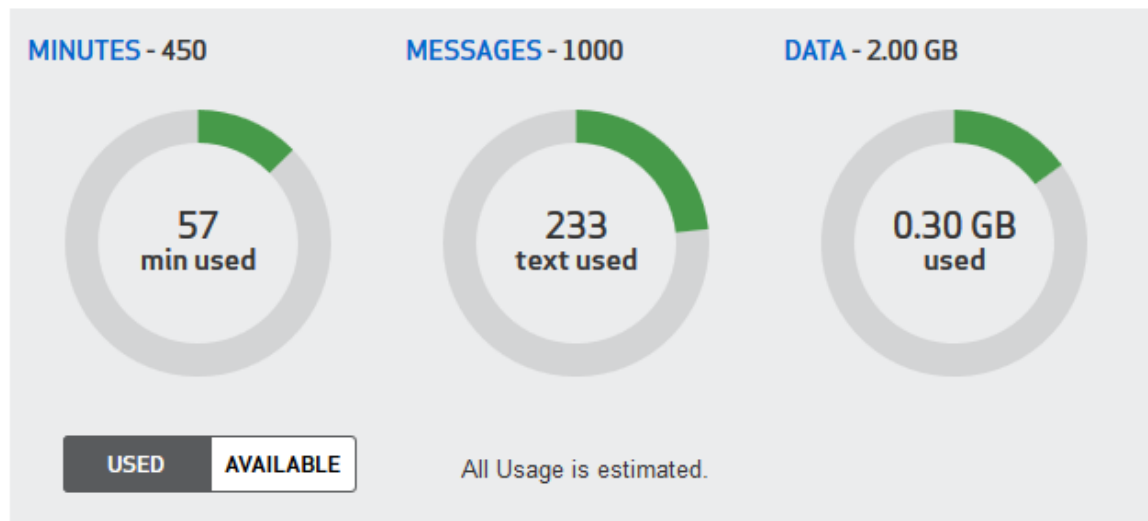


Figure 8. This is an example of a good data visualization. Source: VerionWireless.com

Too much eye candy—too much color, too many elements in the design, too much information in a single graphic—can actually obscure whatever meaning the visualization attempts to present. The effect can be to mesmerize the audience – “Ooh, pretty!” – without imparting any real meaning.

That isn’t to say good data visualizations have to be simple. One of the most often cited examples of excellent data visualization is French engineer Charles Minard’s representation of Napoleon’s invasion of Russia in 1812. Edward Tufte, considered the father of modern data visualization, presents this image in his book *The Visual Display of Quantitative Information* [2].

Minard’s masterpiece is a classic example of multivariate graphics, combining a geographical dimension with a time-series display to illustrate the devastation suffered by Napoleon’s army during this campaign. The French version of the graphic is presented in **Figure 9**. Tufte’s book includes both it and a later English translation on page 41.

Most of us will never attempt to create anything as elaborate as Minard’s chart, but all of us can learn something from it.

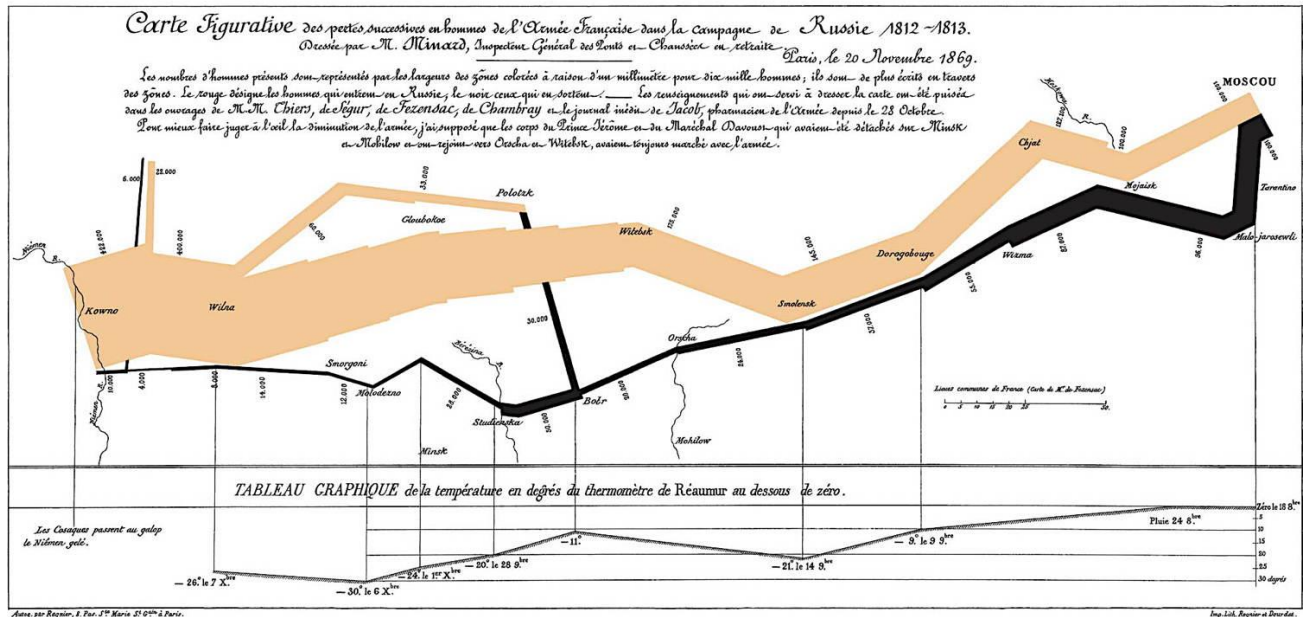


Figure 2. A chart created by French engineer Charles Minard showing Napoleon's 1812 invasion of Russia. The tan bar shows the army marching from the Poland/Russia border. The bar branches and decreases thickness showing parts of the army splitting off to fight various battles, gradually losing troops on their way to Moscow. The black bar shows the defeated army marching back from Moscow, eventually dwindling to only 10,000 by the time of their return (down from 422,000 at the start of the march).

Figure 9. Napoleon's March.jpg – French version – from www.americandigest.org

Dashboards

Dashboards deserve a mention here because they're a popular way to present visual data. A dashboard seeks to convey several pieces of information the viewer can easily absorb and quickly comprehend. The term of course derives from automobile dashboards, which are designed to give the driver the information needed to safely operate the vehicle. A dashboard for a corporate CEO or CFO might present data about key performance indicators such as sales trends, inventory levels, production capacity, orders in the pipeline, and other information necessary to profitably "drive" the company.

Rather than being a distinct type of data visualization on its own, a dashboard typically brings several standard types of data visualizations together in to a meaningful whole. The cell phone usage data shown in Figure 8 is a dashboard because it presents three pieces of data all related to my cell phone plan, which helps me use it without going over my limits.

Seeing data from the user's point of view

As database developers, we naturally tend to think of data in terms of tables, rows, columns, and queries, and when it comes to presenting data to the user we tend to think mostly in terms of reports.

This is natural. Whether or not we are the original creators of a database, we need to know and understand its structure in great detail in order to write effective applications. I believe

our natural tendency to think in terms of the structure of a database subconsciously affects the way we design the presentation of that data to our users – a detailed report with rows and columns of numbers, a summary screen with fields for totals and sub-totals, etc.

The end user of the data generally does not think of the data in their system the same way we do. They think of it in terms of its meaning in the context of their business. If our goal is convey meaning rather than merely information, we need to learn to see the data from the user's point of view and to present that data in ways that make sense to them. While this can be accomplished with conventional forms of presentation such as reports and screens, visual data can be much more effective in presenting data in a way the end user can quickly understand.

Dashboards often employ radial gauges, linear gauges, and other skeuomorphic representations of data. The cell phone usage graphic from Verizon Wireless in Figure 8 is one example. Another is this little graphic from a Web advertisement for LendingTree.com, shown in **Figure 10**. You can see not only the credit score number itself but also where that number falls within the range of possible values. At the same time, the clever use of color conveys information about the sub-ranges for poor credit scores at one extreme vs excellent credit scores at the other. Without being told, most of us would recognize red on the left as being poor vs green on the right being good.

The Average
Credit Score is
681

See How You Compare



Figure 10. This simple yet effective visualization shows not only the average credit score as a number but also where that score falls within the range of possible values. Color is used to convey information about various sub-ranges within the two extremes.

Creating your visualization

Get the data

Step one in creating a good data visualization is to get good data. This should be easy for us as database developers because we're not dependent on external data sources – we can write our own queries against our own or at least familiar databases.

Nonetheless, it's important to pull the right data in the right format. Your client is in the best position to tell you what the important data are – point-in-time snapshots, trends, year over year comparisons, or whatever is important to them. Once you know what data you need to present visually you can begin to think about what type of visualization works best.

Choose the right type of graphic

Pie Chart

Use a pie chart or a donut chart to show proportions relative to a whole. Year-to-date sales by product is a good example.

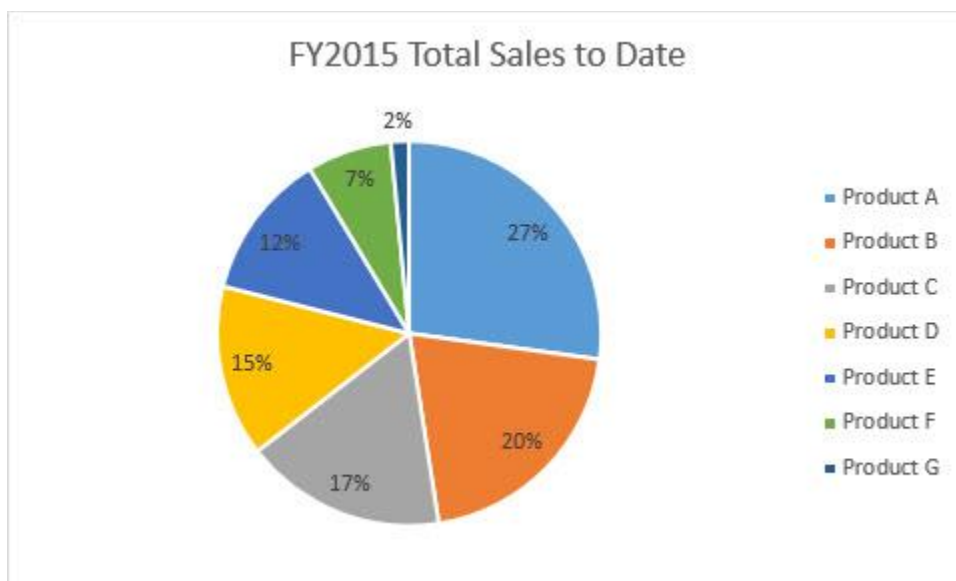


Figure 11. A pie chart is appropriate for showing proportions relative to a whole.

Column or bar chart

Use a column chart or a bar chart to show comparisons. The comparisons can be across a single set of data or across multiple sets of data. **Figure 12** is the same data as **Figure 11**, presented as a column chart. **Figure 13** uses a column chart to compare FY2015 sales to FY2014.

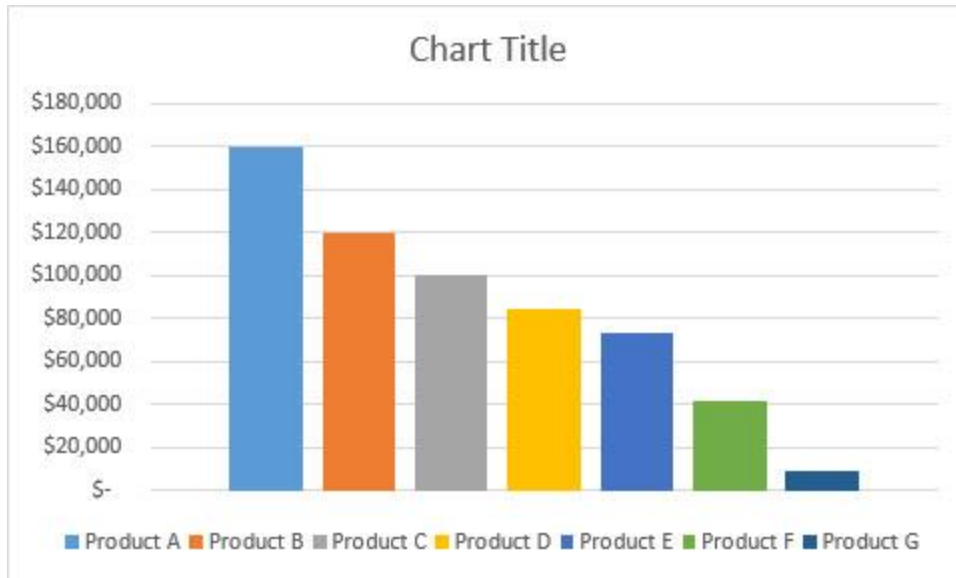


Figure 12. A column chart shows relative proportions but does not give the same sense of proportion to the whole as a pie chart.

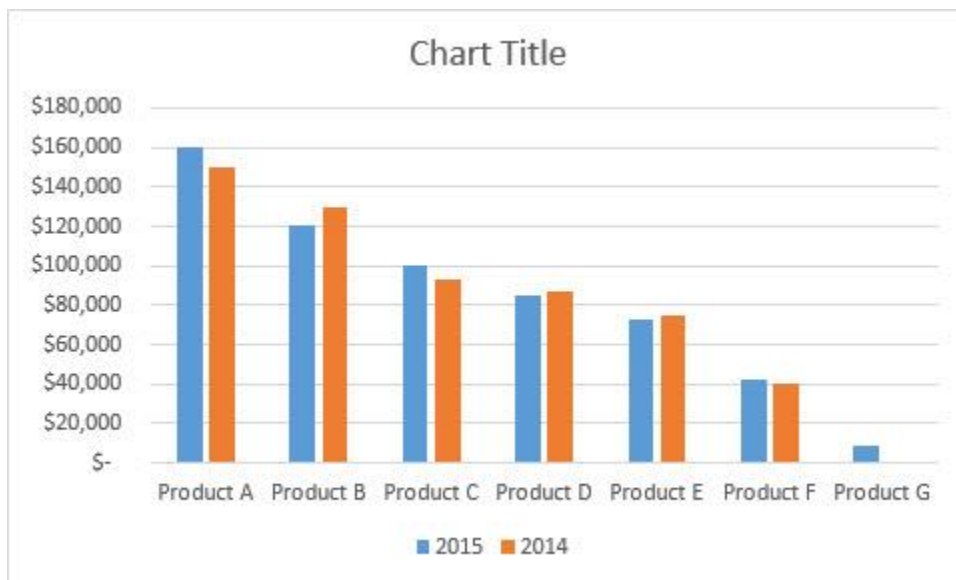


Figure 13. A column chart is a good way to compare two (or more) sets of corresponding data.

A bar chart is the same as a column chart except the data is horizontal. Column charts are appropriate when there are few data items. Bar charts are generally preferred when there are many data items. Unless the order of the data items has some intrinsic meaning, the data are typically shown in descending sequence in both types of charts.

Here's another example of a column chart, comparing the average salary of graduates of Big Ten universities ten years after entering the university (i.e., typically six years after graduation). It illustrates how color can be used to make one figure stand out from the rest.

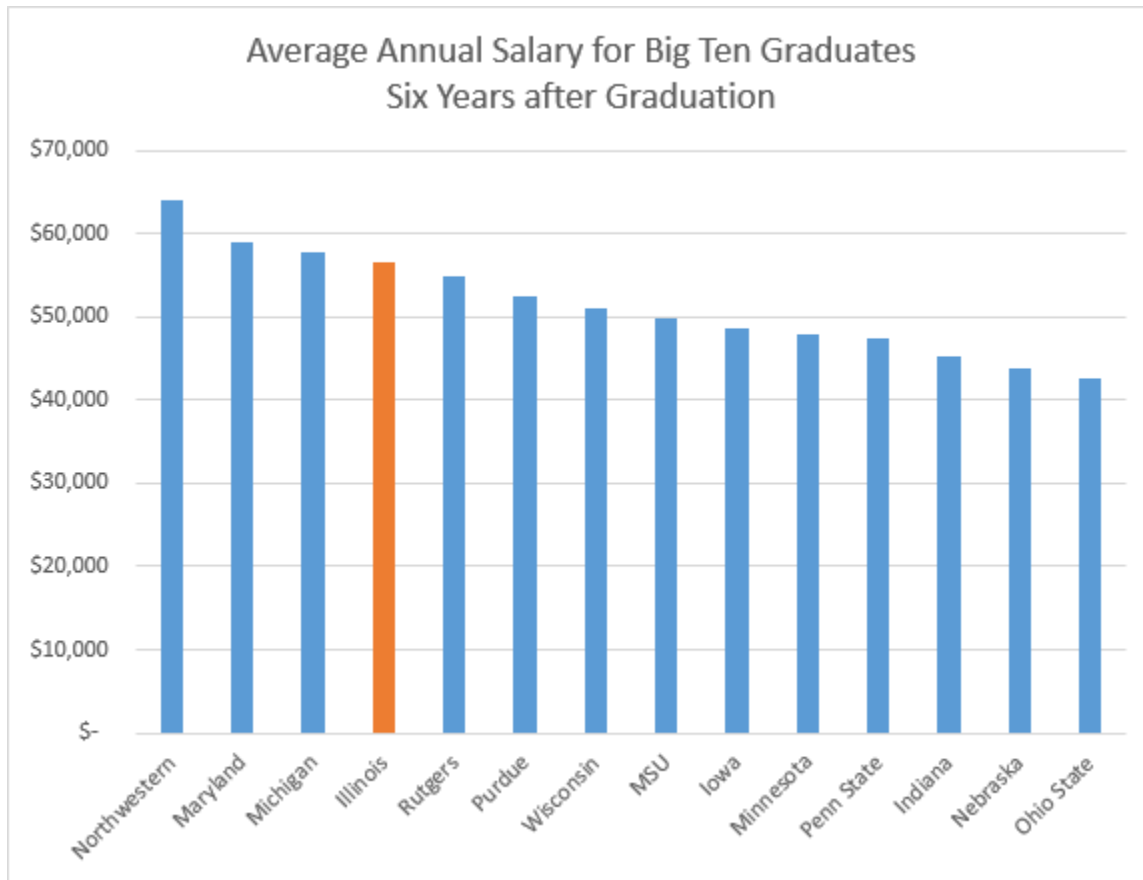


Figure 14. Average annual salary for graduates of Big Ten universities six years after graduation. Sources: U. S. Department of Education and *The Daily Illini*.

Scatter chart

Use a scatter chart to show data with two variables. Scatter charts are the traditional X-Y axis charts we all learned how to draw in school. The data are represented at points corresponding to the intersection of the X and Y values for each pair of data.

Figure 15 shows the membership growth in an organization over six years.

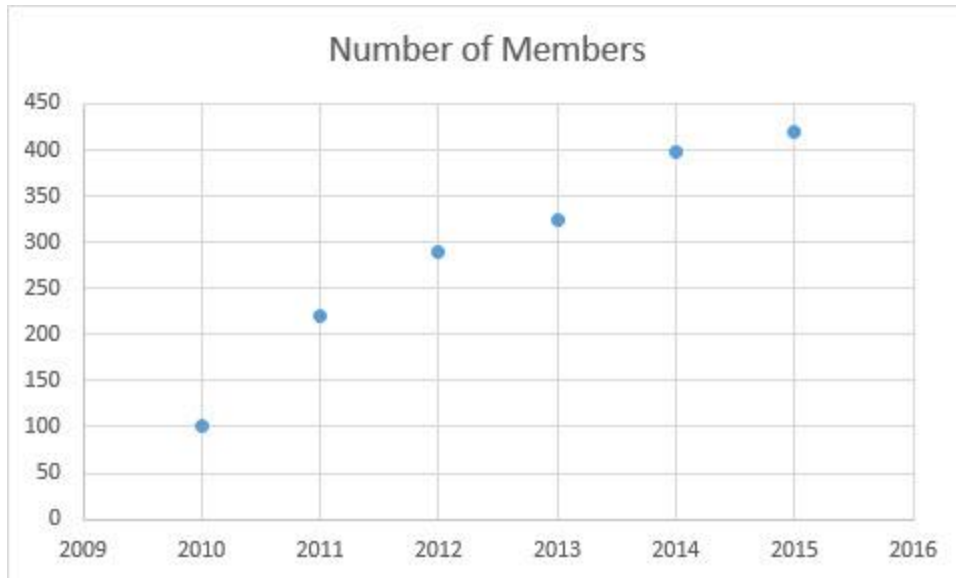


Figure 15. This scatter chart shows membership growth over six years.

Line graph

If the points on a scatter chart represent a series or a progression, connect the dots to form a line graph. A line graph emphasizes the pattern of changes between the points on a scatter chart. **Figure 16** is the line graph of the data in Figure 15.

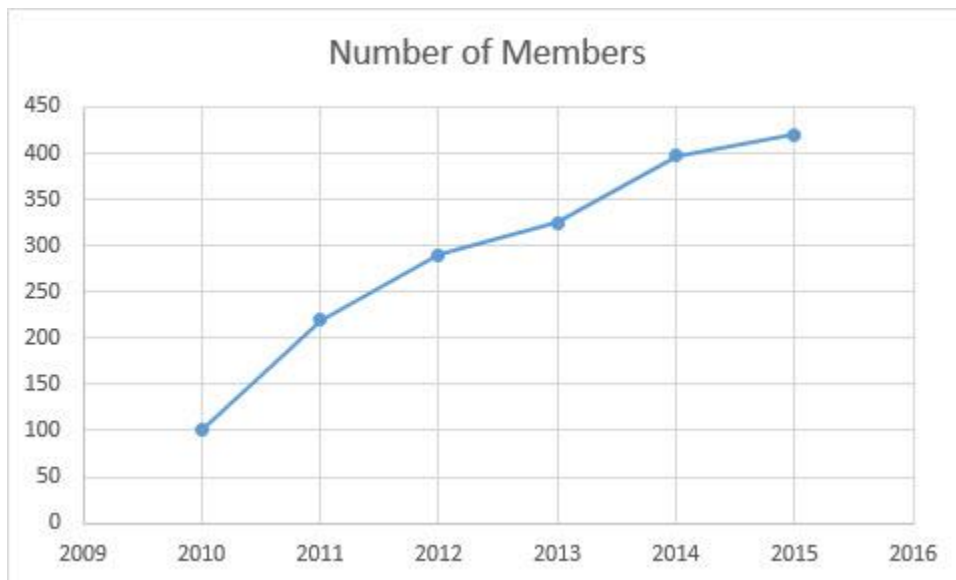


Figure 16. Connect the dots on a scatter chart to form a line graph.

Trend line

If the points on a scatter chart indicate a trend, add a trend line to show the direction of change. The trend line in **Figure 17** is a simple linear regression of the data in Figure 15. Trend lines are most useful for revealing the nature of the progression (up or down) when

the data are such that the pattern is not revealed by the scatter chart alone. Trend lines are frequently employed to predict future values.

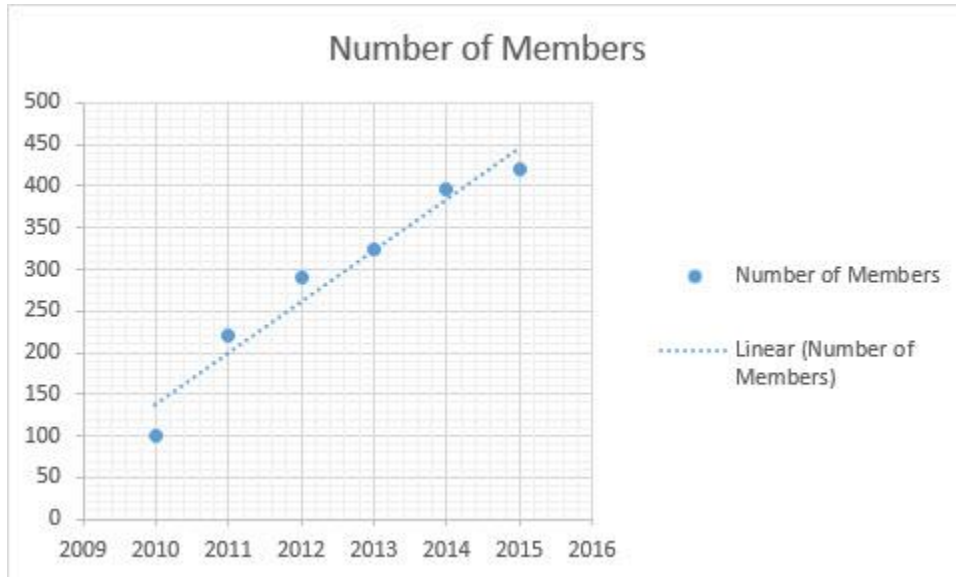


Figure 17. A trend line reveals the nature of the progression of the points on a scatter chart.

Area chart

When the values on a line graph represent a volume, use an area chart to visualize the total volume under the line. While a line graph can illustrate a trend even if the chart does not include the zero point, an area chart should always include the zero point in order to show an accurate representation of the total volume.

Figure 18 is the area chart version of the data in Figure 16. Note that the boundaries of the X-axis have been adjusted to correspond exactly to the lower and upper bounds of the data.

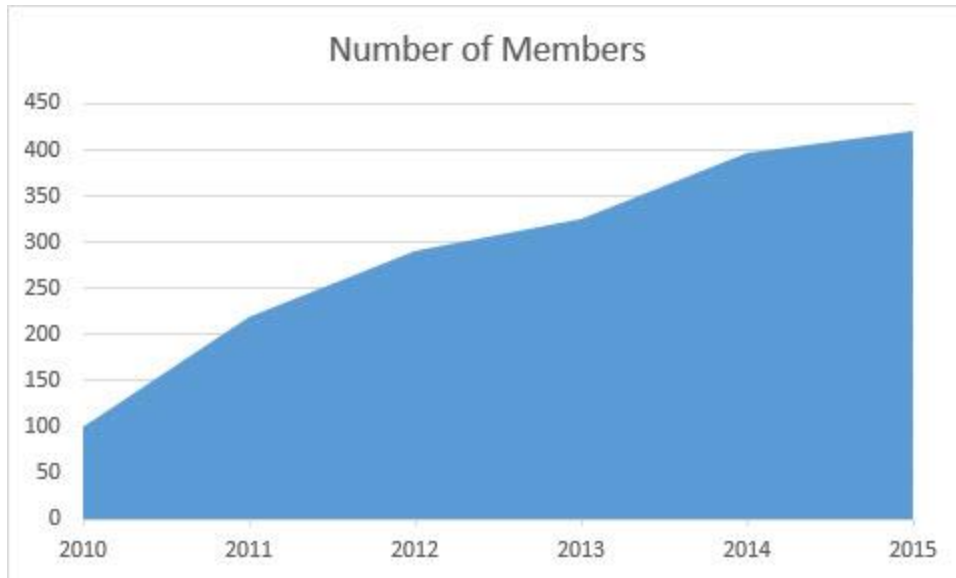


Figure 18. An area chart emphasizes volume.

When there are two or more series of data, a stacked area chart is useful for showing the proportion of each series to the whole.

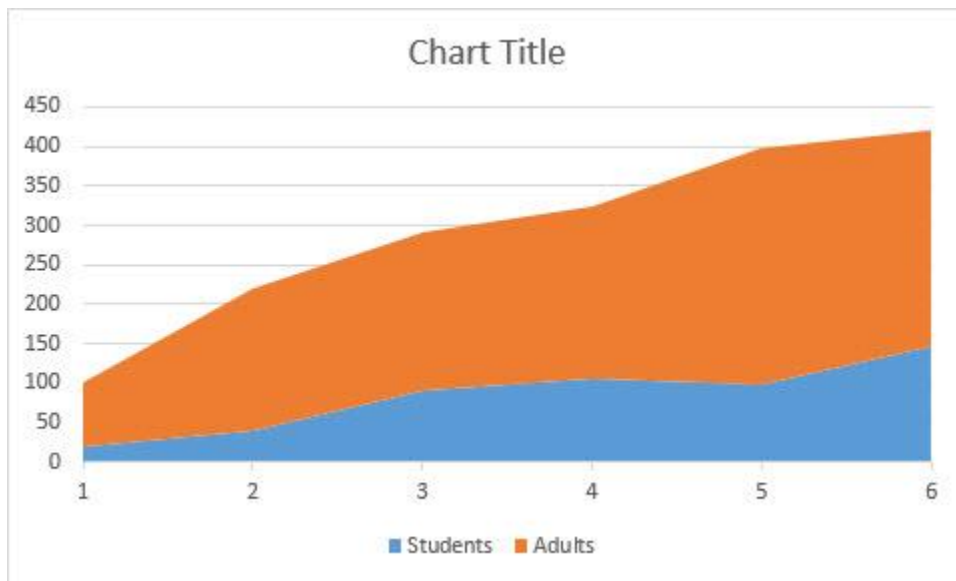


Figure 19. A stacked area chart shows the contribution of each series to the total.

Bubble chart

Use a bubble chart to represent data that varies in three degrees of freedom. Data that varies in three degrees of freedom is not the same as physical three-dimensional (3D) objects. Physical 3D objects such as a cube or a sphere can be represented in a chart with X, Y, and Z axes. Data that varies in three degrees of freedom can be plotted on a traditional 2D chart with X and Y axes and the third variable being represented by a bubble. The center

of the bubble is a point on the X-Y axes and the area of the bubble corresponds to the value of the third variable.

An example would be a chart showing the sales figures for various product lines with total sales on the Y-axis, direct marketing expense on the X-axis, and gross profit the area of the bubble. The data are shown in **Table 1**.

Table 1. Total sales, direct marketing expense, and gross profit by product.

Marketing			
Product	Expense	Total Sales	Gross Profit
A	\$ 10,000	\$ 160,000	\$ 45,000
B	5,000	120,000	102,000
C	7,500	100,000	67,000
D	400	85,000	82,000
E	3,700	73,000	64,000
F	22,000	42,000	10,000
G	1,000	9,000	7,800

Gross profit for each product line is less than Total Sales minus Marketing Expense because of other expenses such as the cost of raw materials, manufacturing costs, distribution costs, and administrative overhead.

Figure 20 gives company management a picture of which products are the most profitable relative to their marketing expense compared to total sales.

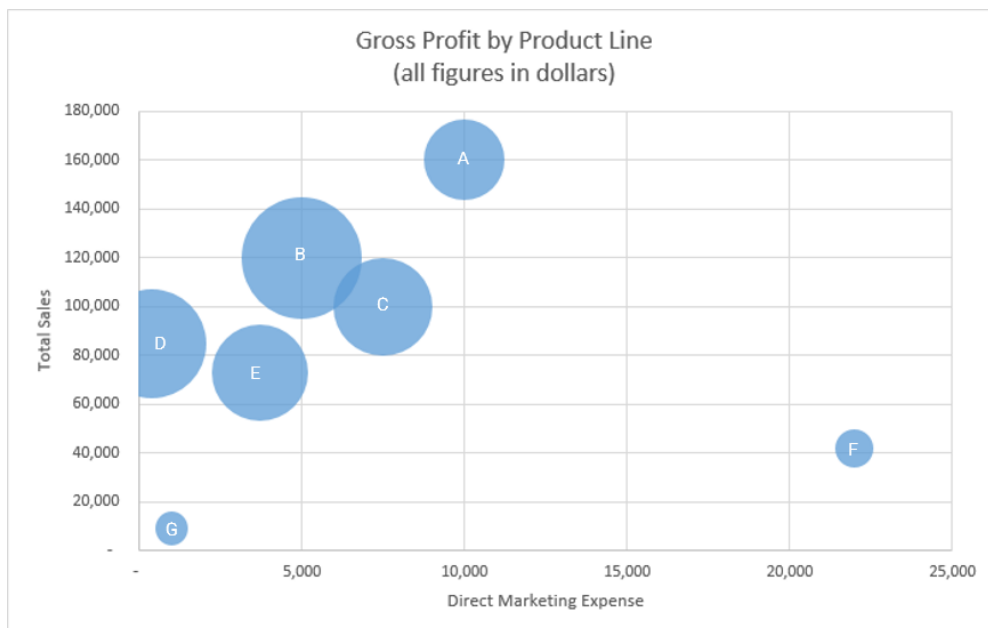


Figure 20. The area of each bubble corresponds to the gross profit for that product.

Product A generates the highest sales but is relatively expensive to market and produce. Product B is the company's cash cow, and products C, D, and E provide good returns. Product G is a stable product in a small market that doesn't require much marketing anymore, while product F is a new venture with a high initial marketing budget and small gross profit.

Combo chart

When two series of data are related but their unit of measurement is different or their scale differs by orders of magnitude, you can use a chart with a secondary Y axis to show both sets of data within their respective range of values.

The combo chart in Figure X shows the relationship between number of homes sold (integer values) and the average price per home (dollar values) for a six month period.



Figure 21. A combo chart uses a secondary Y axis to show two sets of related values with different units of measure. Source: <https://support.office.com/en-us/article/Create-a-combo-chart-with-a-secondary-axis-1d119e2d-1a5f-45a4-8ad3-bacc7430c0a1>

Waterfall chart

Use a waterfall chart to show increases or reductions from a starting value to an ending value. Waterfall charts are one of six new types of charts coming in Excel 2016. Here's an example of a waterfall chart from a recent Microsoft blog post.

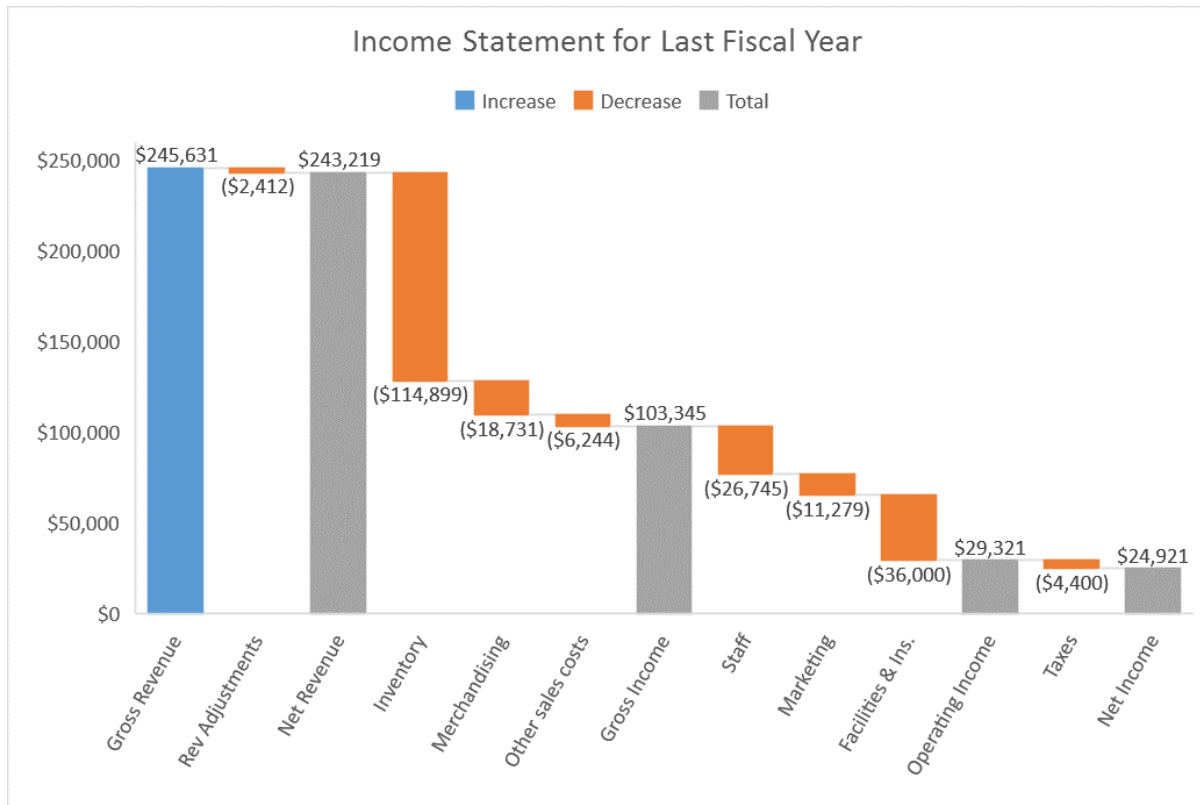


Figure 22. Waterfall charts present increases and decreases from a starting to an ending value. Source: <https://blogs.microsoft.com/firehose/2015/07/02/a-closer-look-at-excel-s-new-modern-charts-for-rich-data-visualization/>

Animated visualizations

Our eyes are drawn to motion. Where possible and appropriate, animated visualizations can be a very effective way to show change over time.

One of the best examples I've seen is described in a post from Garr Reynolds' Presentation Zen blog. Take time to read the blog post and then watch the video. Note the time the presenter, Dr. Hans Rosling, spends explaining how the visualization is constructed before talking about its meaning, thereby giving the audience a basis for understanding it. The blog post, which contains an embedded link to the video, can be found at <http://www.presentationzen.com/presentationzen/2015/07/hans-rosling-upgrading-our-worldview-through-data-visualization.html>. The video is also on YouTube at <https://youtu.be/foXIFILJCgk>.

Google Spreadsheets has a gadget that lets you create a motion chart for your time series data. Here are some additional links about animating visualizations:

- Animating Excel Charts in PowerPoint
<https://www.youtube.com/watch?v=MH7SUUxiiBI>
- <https://developers.google.com/chart/interactive/docs/animation>

- Animate Charts in Excel: Temple University Fox Business School
<https://www.youtube.com/watch?v=-tQqiGgU1Dk>

A list of tools for data visualization can be found at
<http://www.creativebloq.com/infographic/20-free-data-visualisation-tools-5133780>

The Chart Chooser by Dr. Andrew V. Abela of The Catholic University of America is a great resource to help you choose the right chart type. You can see it at
http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html.

Choose the right tool for the job

FoxCharts

FoxCharts is the only tool I'm aware of that lets you generate charts and graphs directly from Visual FoxPro. FoxCharts is hosted on VFPX. The project's homepage is <https://github.com/VFPX/FoxCharts>. You can download the documentation from <https://github.com/VFPX/FoxCharts/tree/master/Docs>. FoxCharts requires VFP9 and the GdiPlusX library, also from VFPX. The companion FoxCharts Tools adds a visual chart designer and code to transform data into the format required by FoxCharts.

For in-depth explanation of FoxCharts, see Jim Nelson's white paper from Southwest Fox 2009.

Excel

For VFP developers and their clients who use Microsoft Office, Excel is the most likely go-to tool for creating data visualizations. Exporting data from VFP to Excel is straightforward and has been well documented. Once a chart has been designed and saved in a workbook, the data can be updated and Excel will refresh the chart. Some clients may call on you to design the charts for them while others may prefer to have your app merely supply the data so they can create their own charts.

Versions of Excel through 2015 feature several chart types but lack some of the more advanced and exciting ones. Excel 2016 introduced six new chart types: treemap, sunburst, histogram, box & whisker, waterfall, and combo. Information and previews can be found on Microsoft's website.

The Office Watch website also has some good resources for charting in Excel. See <http://new.office-watch.com/2015/excel-charts-give-them-a-try/> for information and tips about working with charts in Excel 2015, and <http://new.office-watch.com/2015/excel-2016-new-charts/> for a look at the new chart types in Excel 2016.

PowerPoint

Did you know you can create charts directly in PowerPoint? Not that you'd necessarily want to, but it can be done. The tools are similar to what you find in Excel.

Power BI

BI stands for Business Intelligence, and it's a big thing at Microsoft these days. Microsoft has a host of tools for analyzing data and creating visual presentations under the Power BI banner, including Power Query (data discovery, retrieval and analysis), Power Pivot (the Excel data store), Power View (reports and interactive dashboards), and Power Map (geographic presentations). Many of these tools are Web-based or integrated with Excel. Unfortunately, Excel integration for Power Query and other Power BI tools is not available with the standard Office 365 Business subscription – it requires Office 365 ProPlus. However, Power BI Desktop was recently released as a free download.

Web tools

There are lots of tools to help you develop charts for presentation on the Web. Many of them are built on JavaScript, like Chart.js (www.chartjs.org) and Processing.js (processingjs.org). The SitePoint website has a list of fifteen JavaScript libraries for creating charts at <http://www.sitepoint.com/15-best-javascript-charting-libraries/>.

Special purpose tools

Tableau Software (www.tableau.com) offers a suite of products for data analysis and visualization. I have not used it but it's featured prominently in *Now you see it* [4], one of the books in my data visualization library. Free trials are available but I don't have any information on pricing.

Build and refine

After you've decided what type of graphic to use and which tool to use to build it, spend some time making sure your visualization conveys the desired message in a way that's easy for your audience to comprehend.

Do the math. If percentage differences are important, calculate and show them.

Get rid of things that distract or obfuscate the meaning. This takes time and effort – as Mark Twain said, “I didn't have time to write a short letter, so I wrote a long one.”

Simplify, simplify, simplify

The default graphics created by tools like Excel may contain more elements than are necessary to get the point across. Like Michelangelo, who said (paraphrasing) that his job as sculptor was to start with a block of stone and chip away all the parts that didn't look like David, your job as the creator of a visualization is to start with the raw product and chip away all the parts that don't contribute to the message.

Common chart elements that are often present by default include grid lines, tick marks, data labels, axis labels on every interval, and so on. Too many data points can also be a problem, as can the inappropriate use of 2D or 3D graphics. If these add nothing to the meaning of the graphic—or worse, if they distort or distract from what's important—reduce or get rid of them.

By way of example, let's look at a completely fictional survey of America's favorite cookies. The data are presented in **Table 2**.

Table 2. A completely fictional survey of America's favorite cookies.

Type of Cookie	Percent
Chocolate Chip	35.0%
Oatmeal Raisin	20.0%
Peanut Butter	15.0%
Snickerdoodles	7.0%
Ginger Snaps	4.0%
Sugar Wafers	3.0%
Chocolate Covered Graham Crackers	3.0%
Peanut Butter with Hershey's Kisses	3.0%
Butterfingers	2.0%
Graham Crackers	2.0%
Fig Newtons	2.0%
Cocoanut Chocolate Chip	1.0%
Mexican Wedding Balls	1.0%
Cocoanut Snowballs	1.0%
Macaroons	1.0%

A pie chart is appropriate for showing proportions like this. The default pie chart for this data as generated by Excel is shown in **Figure 23**.

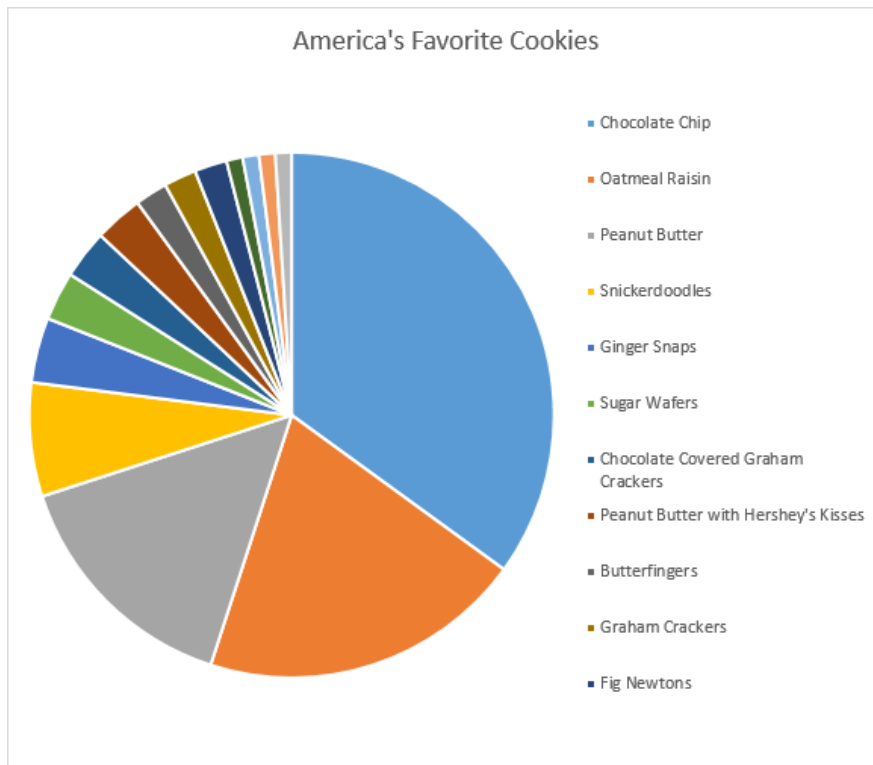


Figure 23. The default pie chart for the cookie data as generated by Excel.

A few things are good about this chart but many things are bad. On the plus side, the data flow clockwise from largest to smallest starting from the top, which is the traditional way to present data in a pie chart. The use of colors is pleasing, and together with the white space it's easy to see the separation between the segments.

On the not-so-good side, the main problems are the number of data points and the placement of the legend. There are so many data points that the distinction between the smaller ones is difficult to visually discern. It would help to add data labels showing the actual percentages, but the labels wouldn't fit in the smaller segments. With fifteen colors and the legend simply adjacent to the image, it's also very difficult to determine which segment of the chart corresponds to which type of cookie.

There are many ways in which this chart could be improved. One is to reduce the number of data points. In this case, the 80/20 rule can be applied to combine the lower 20% (ten categories) into a single "Other" category as in **Figure 24**. If the smaller percentages still need to be shown individually, a break-out pie chart can be used to show them more clearly, although in this example Excel decided to break out only the smallest five segments (see **Figure 25**, both with and without data labels).

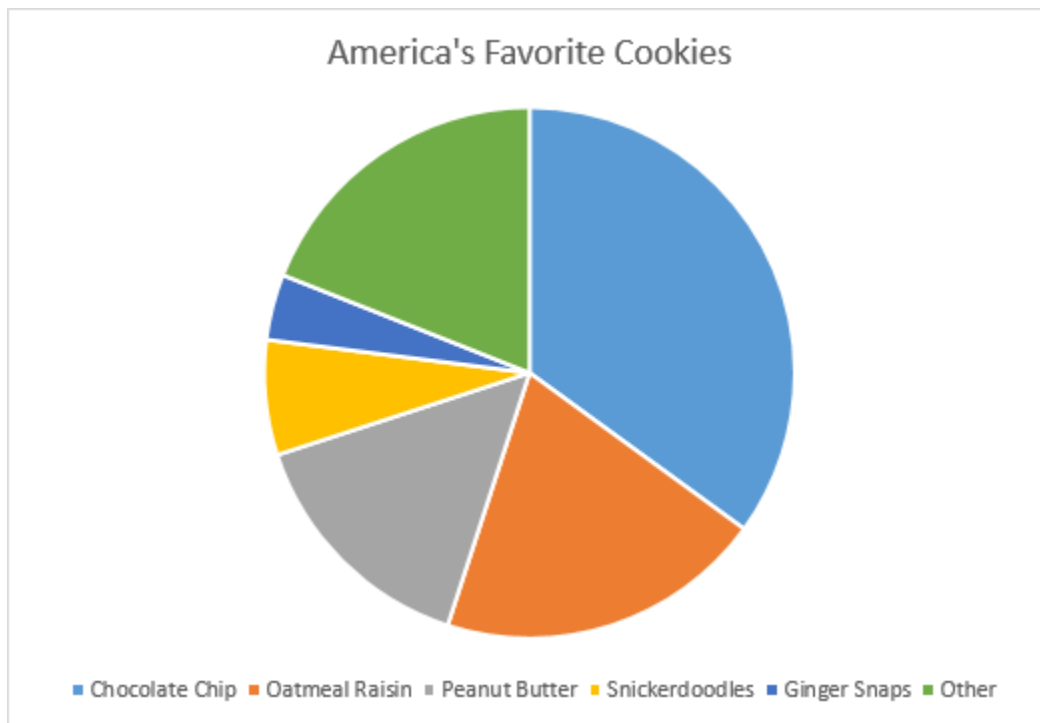


Figure 24. Reducing the number of data points results in a better if less precise pie chart.

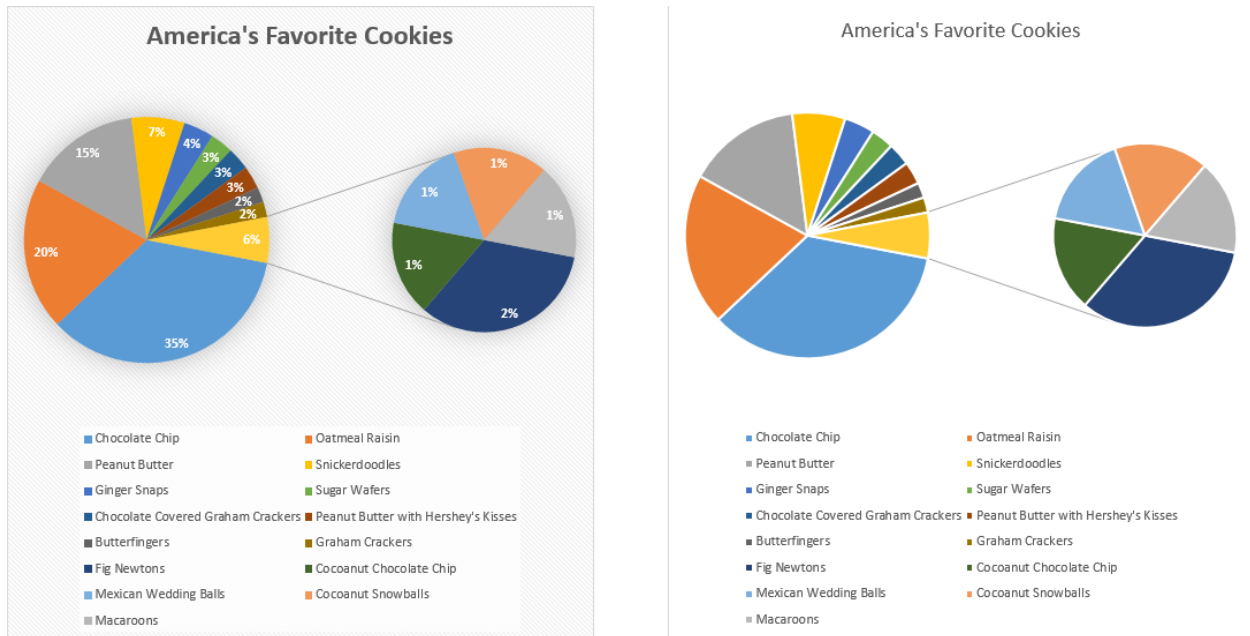


Figure 25. Smaller segments can be broken out into their own pie chart.

The remaining problem with all of these is the placement of the legend below the chart itself. Even with fewer data points, as in **Figure 24**, it still takes some effort to figure out which legend goes with which segment on the chart. A better result is obtained by placing the legends around the circumference of the chart, where their color and placement makes it clear which one belongs to which segment. Placing the legends around the chart instead of below it would not have been feasible with the original fifteen data points, but it is feasible with only six.

Figure 26 is the final improvement in this example. Data labels could be added if it were necessary to show the exact percentages for each segment.

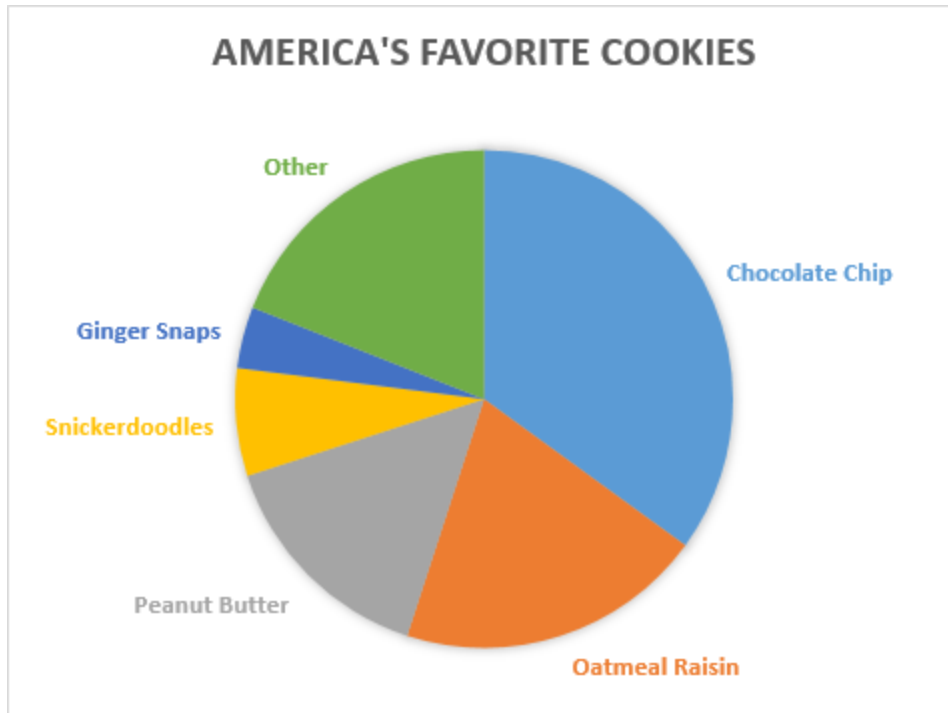


Figure 26. Matching the color of the legends to the color of the segments and placing the legends around the circumference of the chart makes it easy to match each legend with its corresponding segment.

Lying with Graphics

You've probably heard the saying, "there are lies, damned lies, and statistics." The same can be said for graphics.

Just as it's possible to lie with partial, misleading, or otherwise flawed statistics, it's also possible to create graphics that distort the data they purport to represent. Knowing how to spot these distortions will help you avoid introducing unintentional distortions into your own data visualizations as well as helping you recognize when someone else is trying to mislead you.

Five common ways in which visualizations can distort the data are the missing zero, the selective range, the unequal base, comparing percentages, and the implied volume.

The missing zero

Context matters. When presenting data in the form of a chart it's always best to include the zero point whenever practical. If a line graph is used to show fluctuations in values over time and all the values are within a narrow range but significantly above zero, including zero as the base line can make the chart unreasonably tall while visually understating the variations. On the other hand, omitting the zero baseline may exaggerate the perceived magnitude of the fluctuations.

Consider the following data visualizations of the variation in the closing values of the Dow Jones Industrials Average (DJIA) for the one-month period June 18th through July 17th, 2015. The data are shown in **Table 3**.

Table 3. DJIA closing values for the period June 18 – July 17, 2015.

Date	Close
6/18/2015	18,115.84
6/19/2015	18,015.95
6/22/2015	18,119.78
6/23/2015	18,144.07
6/24/2015	17,966.07
6/25/2015	17,890.36
6/26/2015	17,946.68
6/29/2015	17,596.35
6/30/2015	17,619.51
7/1/2015	17,757.91
7/2/2015	17,730.11
7/6/2015	17,683.58
7/7/2015	17,776.91
7/8/2015	17,515.42
7/9/2015	17,548.62
7/10/2015	17,760.41
7/13/2015	17,977.68
7/14/2015	18,053.58
7/15/2015	18,050.17
7/16/2015	18,120.25
7/17/2015	18,086.45

The lowest value in this data is \$17,515.42, while the highest is \$18,144.07. The ending value is within \$30 of the starting value but the range is more than \$628. Data like this, with two variables, can be effectively represented with a line graph. How you choose to represent the data on the line graph depends on what information you're trying to convey.

If your intent is to highlight the “wild fluctuations” in the closing stock market prices over this period of time, set the Y-axis range of your graph close to the lowest and highest values, as shown in **Figure 27**.

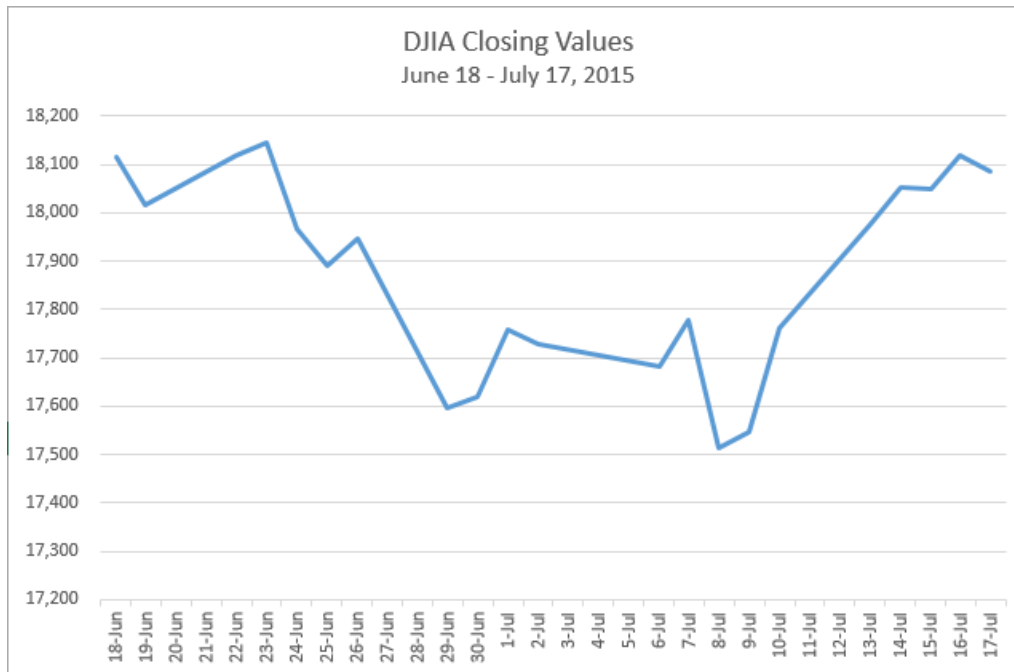


Figure 27. DJIA closing values with a Y-axis range close to the actual range of values.

Choppy waters, indeed! Because the market's closing values are a proxy for the value of a portfolio (one with a beta of 1.0, anyway), another way to show this data would be with an area chart. The area under the line corresponds to the total value of the portfolio. **Figure 28** makes it appear the portfolio lost half its value by July 8th!

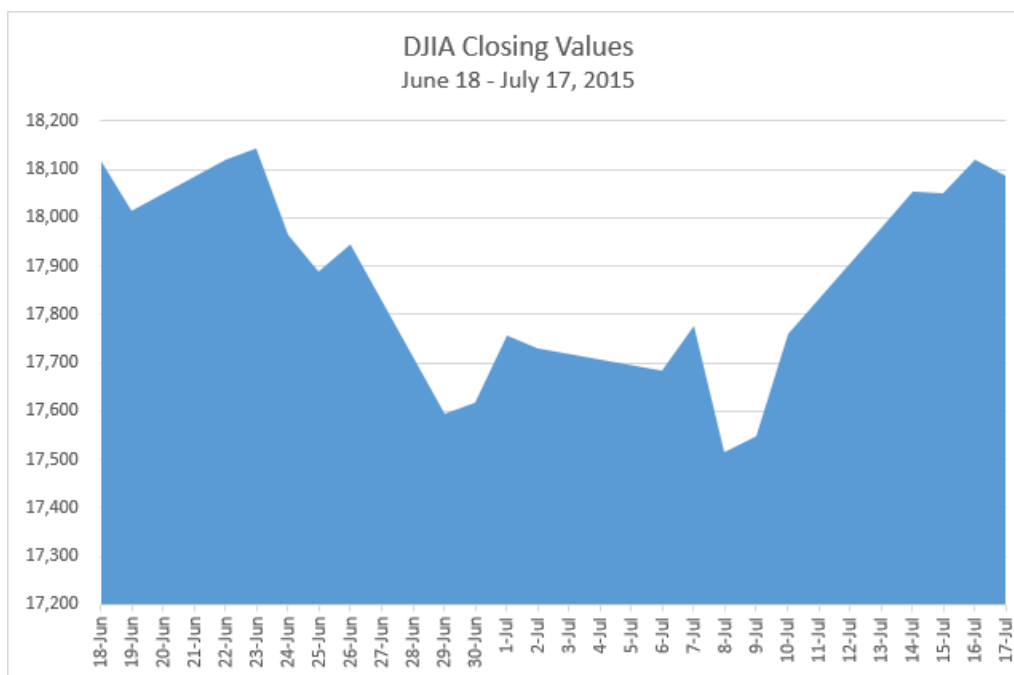


Figure 28. This area chart makes it appear the portfolio lost half its value by July 8th.

If, on the other hand, your intent is to show that the market fluctuations were not nearly as dramatic as they appear in Figure 27, set the low end of the Y-axis to zero. With zero as the lower bound, the area under the line corresponds to the real total value of a unit of this portfolio. The ride is seen to be much smoother with only small ripples on top, as shown in **Figure 29**.

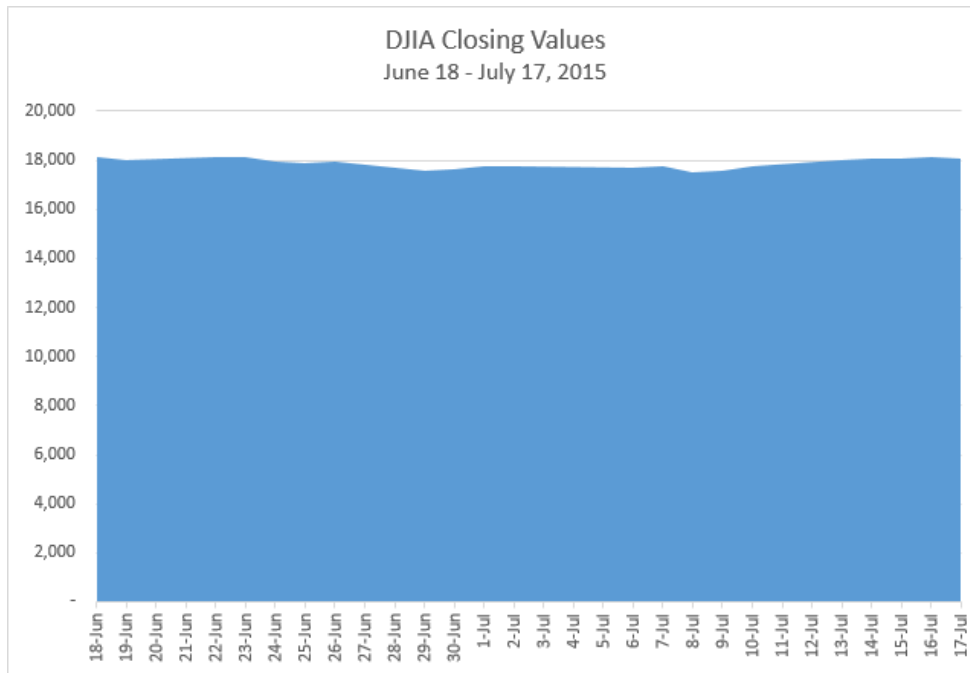


Figure 29. A graphic with the same values presents an entirely different picture when the lower bound of the Y-axis is set to zero. The portfolio didn't lose half its value after all.

The selective range

To present an accurate picture, it's important to select an appropriate range of data. There are three distinct parts to the data in the examples above: a decline of nearly 3% in the first eleven days, a relatively flat period over the next nine days, and gains of about 3.5% over the last nine days. We could paint three entirely different pictures of the market if we selectively chose to include data only from one part.

Figure 30 shows how a selective range could be employed to show the market in a steep decline. Admittedly, in these examples the entire time span of the data is artificially short and this selective range even more so, but for purposes of illustration it serves to make the point. The steep downward slope of the line in this visualization is due both to the selective range and the omission of the zero baseline.

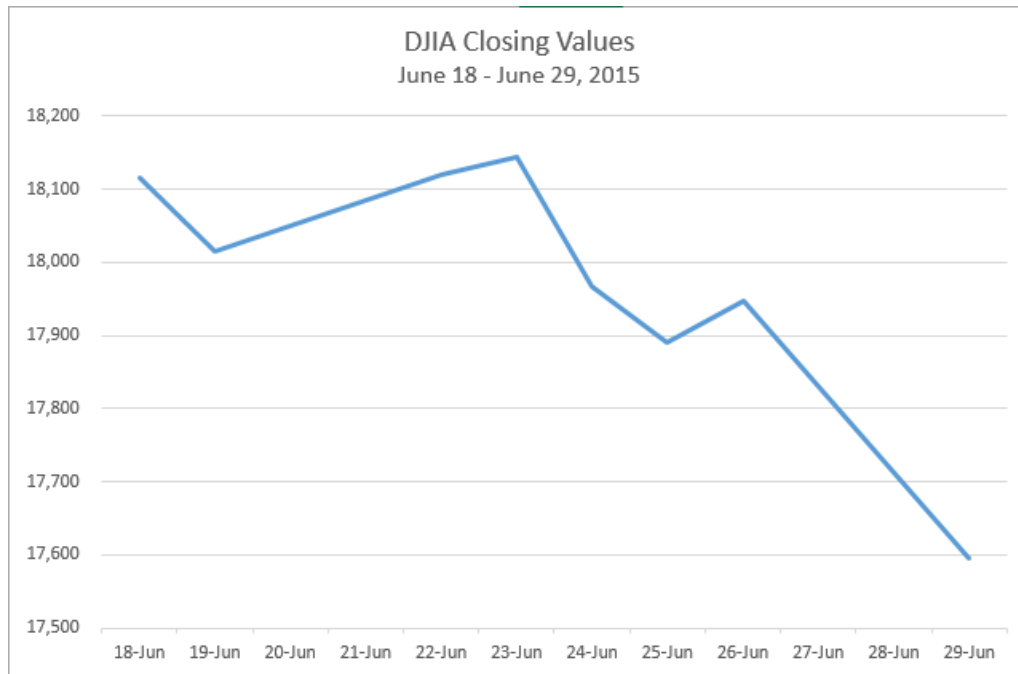


Figure 30. The market's in a steep decline... Sell! Sell! Sell!

Figure 31, on the other hand, use the middle range of this data to show that the market is relatively stable. Such a visualization could be used to support a “hold” recommendation.

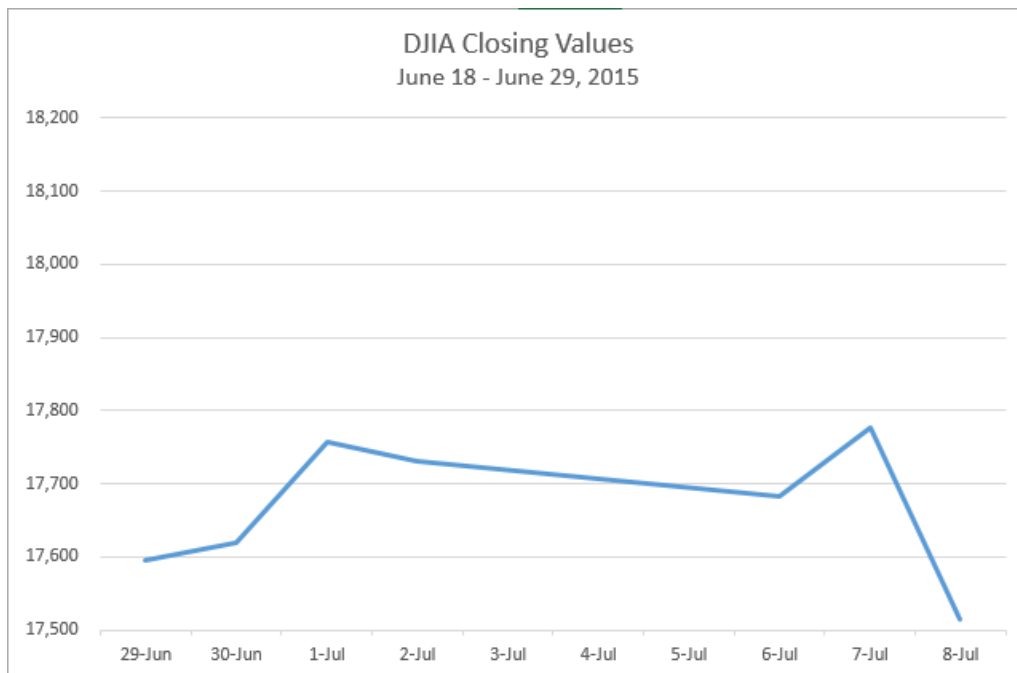


Figure 31. The market's fairly steady... hold on for now.

Finally, the third segment of the data could be selectively employed to give the impression of a strongly rising market, as shown in **Figure 32**. Again, the upward slope of the line is due to the selective range and is exaggerated because the lower bound is not zero.

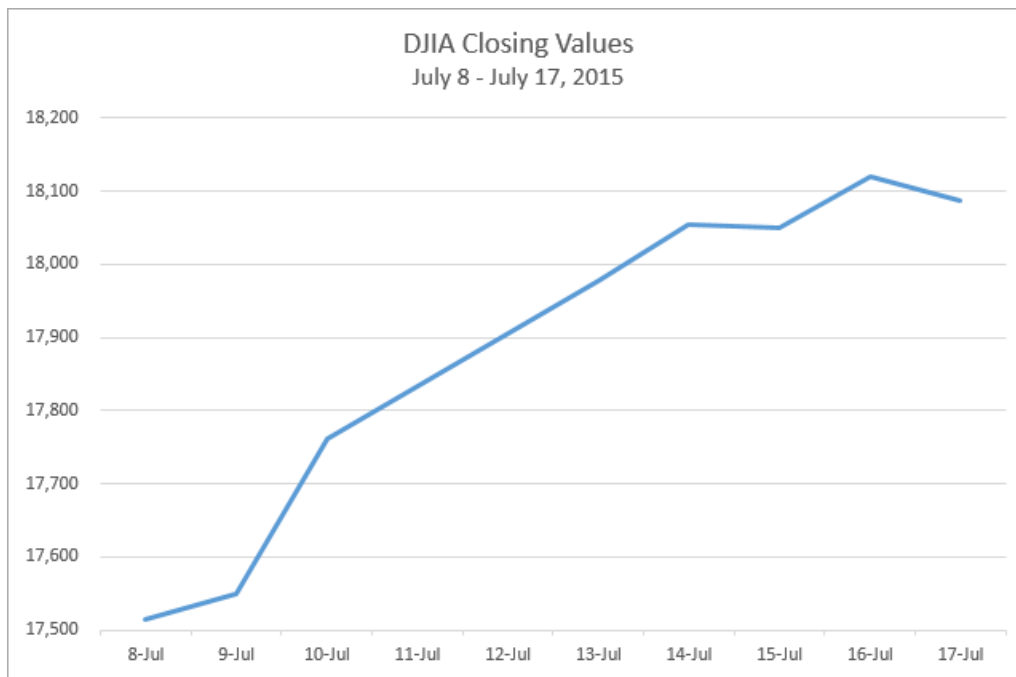


Figure 32. The market's surging... Buy! Buy! Buy!

Which of the three preceding charts is the “correct” picture of the market’s behavior? Answer — they all are. It just depends on which picture you want to paint.

The unequal base

When percentages are compared to one another, the comparison can be distorted if the basis of comparison is not the same for all of them.

Consider this example from a small, fictional IT company. The senior management committee is meeting to approve the 2021 budget, which is based on increases over 2020 expenditures. The last column is the percent change over the previous year's budget for each category.

	2020	2021	\$ CHANGE	% CHANGE
SALARIES	500,000	510,000	10,000	2%
TRAINING	96,000	97,000	1000	1%
RENT	30,000	30,600	600	2%
SOFTWARE	24,000	24,500	500	2%

HARDWARE	10,000	14,500	4,500	45%
TOTAL	660,000	676,600	16,600	3%

Bob, an enterprising junior programmer who feels underpaid, discovers an unsecured copy of the proposed budget on the company's network and immediately notices the outlandish 45% increase in proposed hardware expenditures compared to a meager 2% increase in salaries. The company clearly cares more about spending money on things than it does about rewarding its people, right? Knowing a bit about data visualization and how to use Excel, Bob constructs a chart (**Figure 33**) to emphasize his point.

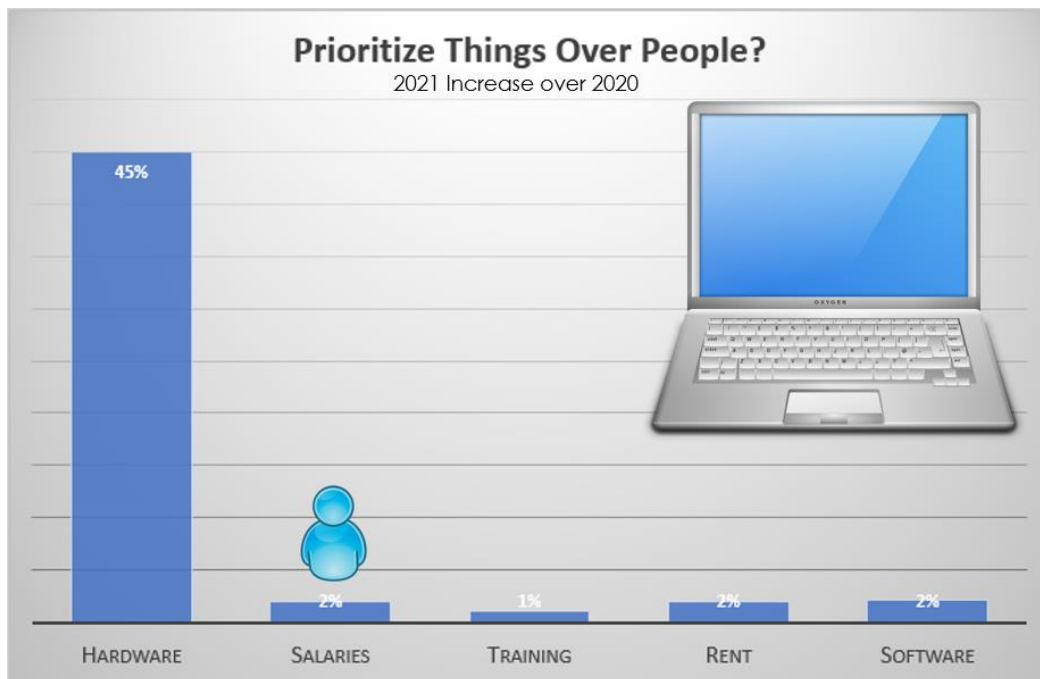


Figure 33. A data visualization showing the "disproportionate" increase in hardware expenditures compared to salaries and other categories.

Bob takes this graphic to his manager, Alice, complaining that the company is prioritizing things over people and some of that money should go toward the raise he richly deserves. Alice looks at the actual numbers in the budget and immediately realizes that, while technically correct, Bob's chart is terribly misleading. Hardware represents a much smaller percentage of the total budget than salaries, but the chart uses unequal bases to distort the comparison. Specifically, the base for the 45% increase in hardware expenditures is only \$10,000 while the base for the 2% increase in salaries is fifty times that size at \$500,000.

A much clearer picture emerges when the numbers are presented in proportion to the total budget. Alice constructs another chart to make this point (see **Figure 34**).

When shown as a percent of the total budget, using an equal base for each category, it's clear the company invests a great deal more in its people than in any other category including hardware. While true that the two new high-end laptops costing \$2,250 apiece represent a 45% increase in hardware expenditures over last year, hardware (investment

in things) still accounts for just over 2% of the total budget while salaries and training (investment in people) represent nearly 90%.

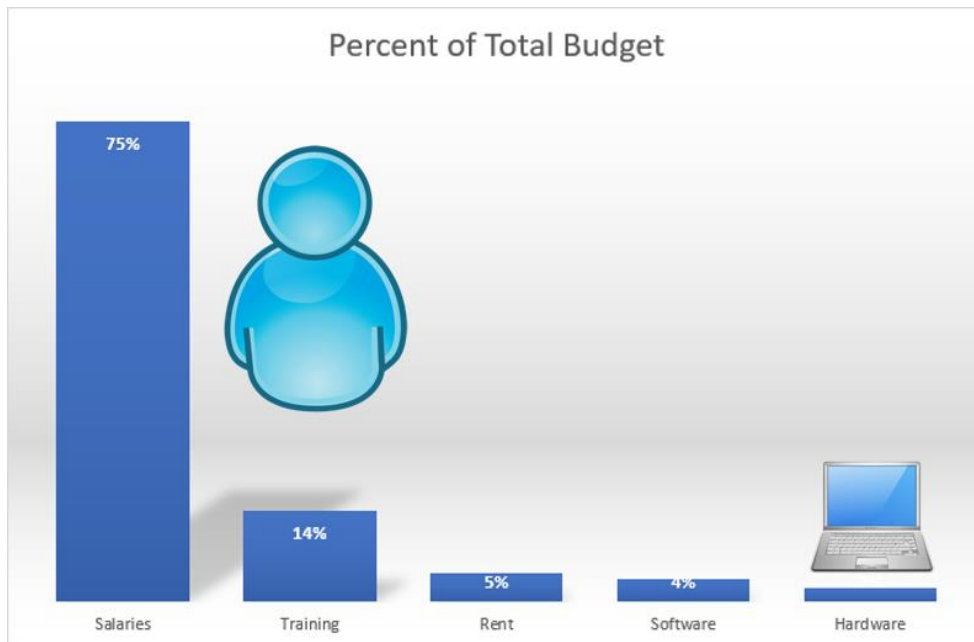


Figure 34. A much clearer picture is revealed when the expenditure in each category is shown as a percent of the total budget (equal base for all categories).

The issue of proportionate expenditures can still be debated, but the chart in Figure 34 gives a much more balanced picture of the relative amounts in question.

The percentage increase

The State of Illinois raised the personal income tax rate from 3% to 5% effective January 1st, 2011. How much of an increase was this?

Some would have had the Illinois taxpayer believe it was a 2% increase. After all, 5% minus 3% is 2%, right? You can see this is true using a simple chart, as shown in **Figure 35**. Two percent sounds like a pretty small increase, doesn't it? Who could complain about that?

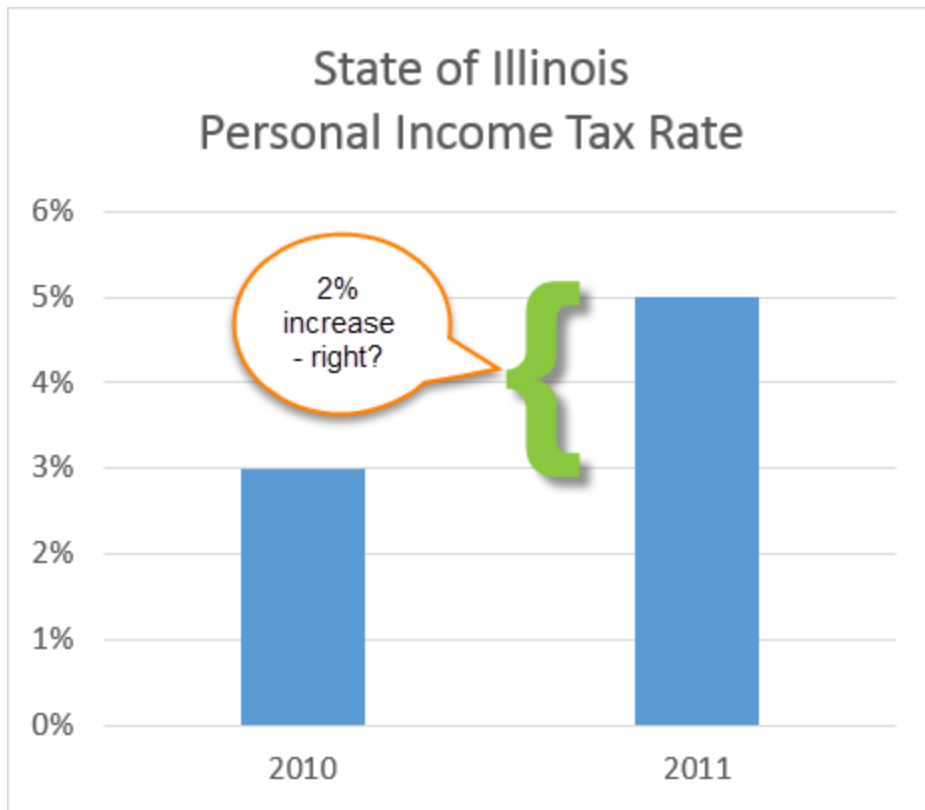


Figure 35. The State of Illinois raised the personal income tax rate from 3% to 5% effective January 1st, 2011. That's a 2% increase, right?

Of course, there's a deliberate distortion in this comparison. The formula for calculating the percentage increase between two numbers is $((\text{new number} - \text{old number}) / \text{old number}) * 100$. In this case, that's $((5 - 3) / 3) * 100 = (2/3) * 100 = .67 * 100 = 67\%$.

Even though we're working with percentages, this calculation is valid because both percentages are relative to the same implied unit value (say, \$1). If you want to prove this by working with dollars instead of percentages, any other base value will do. Using \$100,000 the calculation of the percentage increase is $((5\% * \$100,000) - (3\% * \$100,000)) / (3\% * \$100,000) * 100 = ((\$5,000 - \$3,000) / \$3,000) * 100 = (\$2,000 / \$3,000) * 100 = 67\%$.

So ask the average Illinois tax payer how much of an increase this was. With a taxable income of \$100,000 in both years, she wrote a check for \$3,000 in 2010 and a check for \$5,000 in 2011. While it's true the \$2,000 difference is an additional 2% *of the taxable income*, it's an increase of 67% in the amount of the check.

The implied volume

Visual distortions can arise when 2D graphics are used to represent flat data. In its September 2015 issue, Consumer Reports published an article comparing average car insurance premiums for people with an excellent, good, or poor credit rating and for those with an excellent credit rating but with a drunk driving conviction. A sidebar graphic presented the data for the state of Florida using circles to represent the numbers.

Figure 36 is a reproduction of that graphic with the circles shown in the same relative proportions as the ones printed in the article.¹ Consumer Reports is a respected publication with a fine reputation, but there is a problem with this graphic. Can you spot what's visually misleading about it?

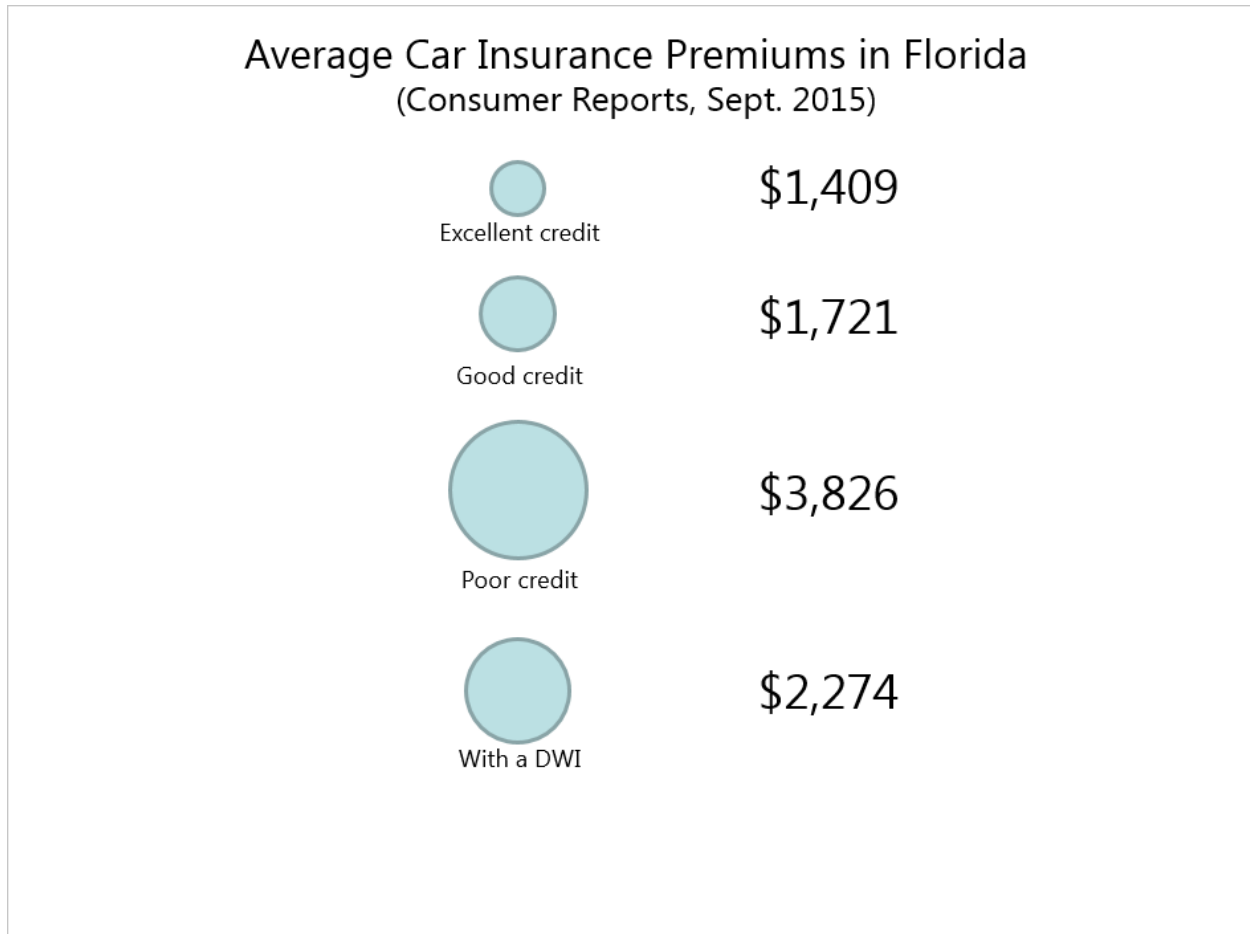


Figure 36. Comparison of average auto insurance premiums for adult single drivers in the state of Florida. Source: Consumer Reports magazine, September 2015, p. 32.

The numbers, which are the actual data, are flat because they represent one-dimensional figures – in this case, dollars. Using the smallest number as the baseline (the unit number), the proportion of the larger numbers to the baseline is shown in **Table 4**.

Table 4. Percentage difference in average premiums, using the lowest number as the baseline.

Average premium in dollars	Percentage difference
\$1,409	Baseline

¹ Based on my measurements, to the nearest millimeter. Any mistakes are mine.

\$1,721	122%
\$3,826	272%
\$2,274	161%

Circles are two-dimensional objects – they have both a diameter and an area. When using two-dimensional objects to represent one-dimensional data, the areas of the two-dimensional objects should correspond to the actual differences between the one-dimensional numbers.

The visualization in **Figure 36** is distorted because the circles were drawn with their *diameters* approximately (although, strangely, not exactly) in proportion to the actual data, rather than their areas. The calculation of a circle's area involves the square of its radius (half the diameter), so using diameters in proportion to the data results in areas that exaggerate the perceived difference between the values.

The calculations to support this are shown in **Table 5**. The first two columns are the actual numbers and their ratios to the lowest value. Working from left to right in the As Displayed columns, the areas of the circles are calculated and then their ratios. The correct diameters are then calculated by working backwards from what the areas should be if they were in the same proportion as the actual numbers. Finally, the rightmost two columns compare the real difference between the numbers with the perceived difference as visually represented by the area of the circles.

Table 5. When flat data is represented by a 2D graphic such as a circle, the perceived difference is based on the relative areas, not the relative diameters.

Actual Data			As Displayed			Should be			Visual Error	
Dollars	Dollar Ratio		Diameter (mm)	Area (mm ²)	Area Ratio	Area Ratio	Area (mm ²)	Diameter (mm)	Real Diff	Perceived Diff
\$ 1,409	1		11	95	1	1	95	11		
\$ 1,721	1.22		15	176	1.85	1.22	116	12	122%	185%
\$ 3,826	2.72		28	615	6.47	2.72	258	18	272%	647%
\$ 2,274	1.61		21	346	3.64	1.61	153	14	161%	364%

An accurate visual comparison is achieved only if the circles are drawn with their *areas*, not their diameters, in correct proportion to the data. **Figure 37** compares the original circles on the left with circles of the correct proportion on the right.

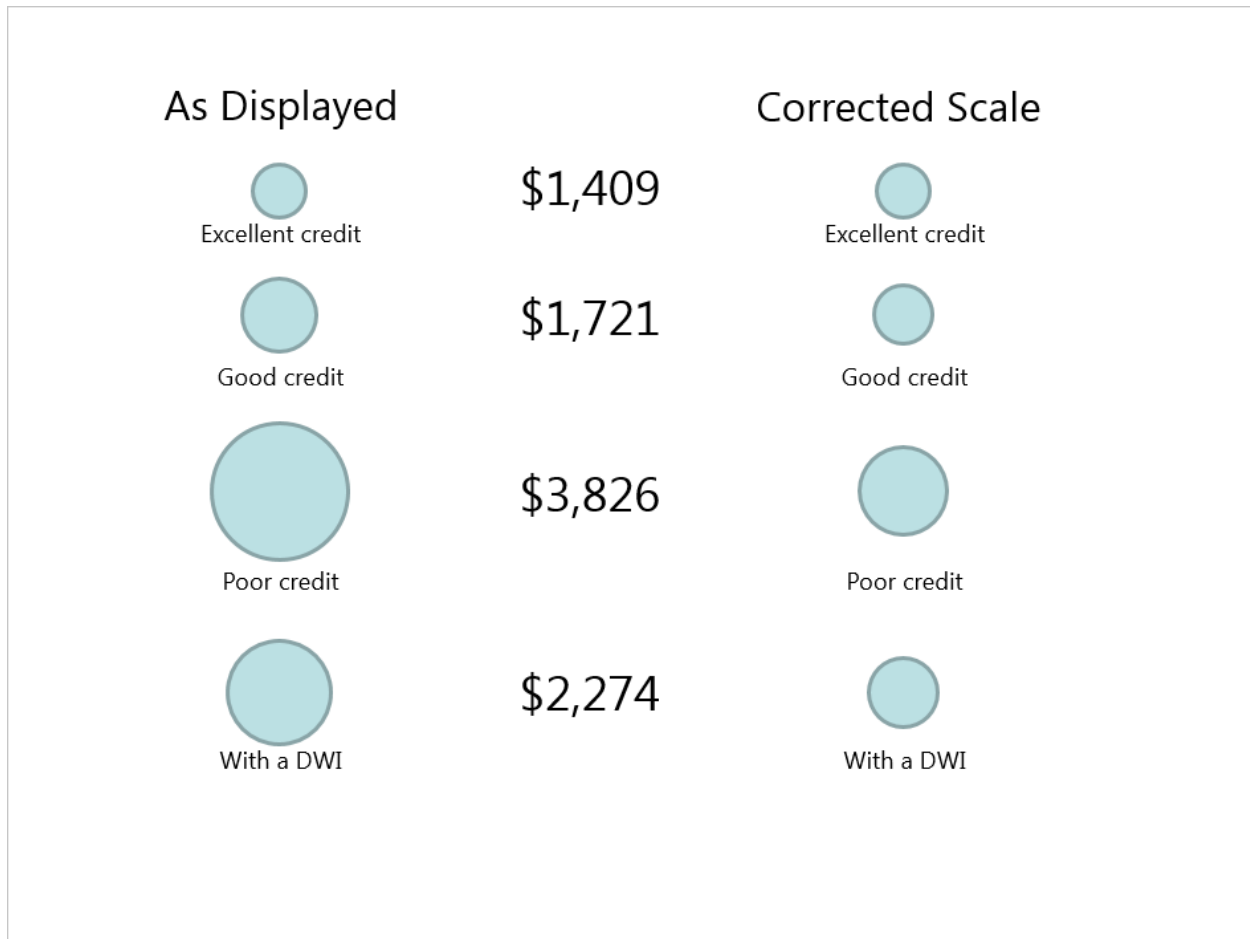


Figure 37. The same data with corrected scale (ratio of areas = ratio of numbers)

The visual perception of the relative difference between the circles on the left in Figure 37 suggests that the average insurance premium with poor credit is more than six times the premium with excellent credit, when in fact the real difference is less than three times. If you don't believe me, look at **Figure 38**, which demonstrates that more than six circles of the smallest size (excellent credit) can fit inside the area of the largest circle (poor credit). The circles on the right in Figure 37 and Figure 38, whose areas correspond to the real differences in the numbers, are in correct proportion to one another.

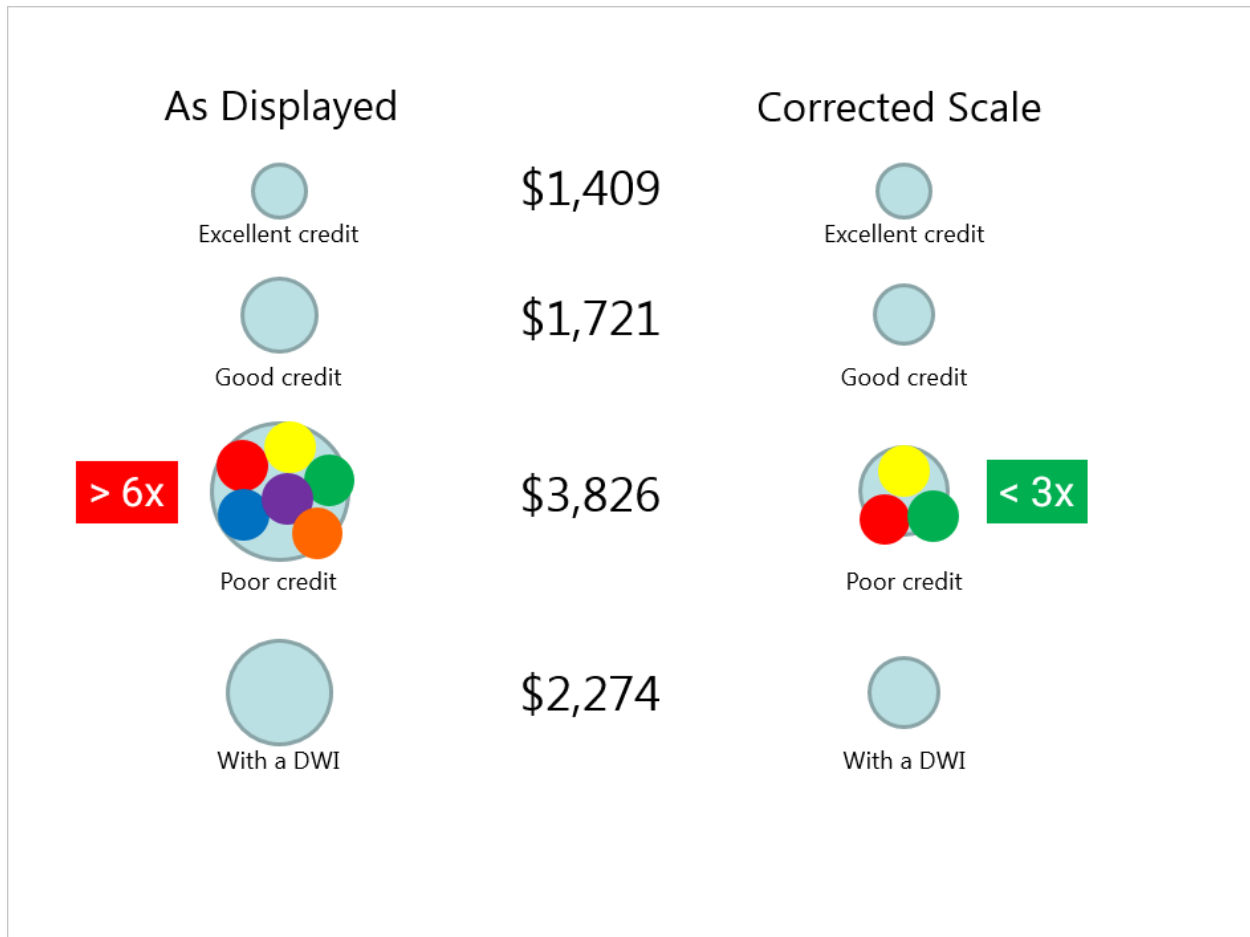


Figure 38. As displayed in the article, the third circle has more than six times the area of the first one. The actual the ratio of the two numbers is less than three to one.

For more information on the subject, I recommend a great little book called *How to Lie with Statistics* [5]. It's fun to read and a real eye-opener.

Summary

Vision is the most important of human senses. We can absorb more meaning more quickly from an image than we can from words or numbers alone. Data visualizations are an efficient way to quickly convey the meaning in a set of data while also revealing patterns, trends, and relationships.

As database developers, we may naturally tend to think of data in terms of the way it's organized in the database. We may therefore also tend to create reports and other presentations that follow this thinking – rows and columns of numbers, lists on a report, and so on. We can improve the value of our applications when we strive to see the data from the user's point of view and learn ways to convey its meaning visually as well as in standard forms.

Biography

Rick Borup is owner and president of Information Technology Associates, LLC, a professional software development, computer services, and information systems consulting firm he founded in 1993. Rick earned BS and MBA degrees from the University of Illinois and spent several years developing software applications for mainframe computers before turning to PC database development tools in the late 1980s. He began working with FoxPro in 1991 and has worked full time in FoxPro and Visual FoxPro since 1993. He is a co-author of the books Deploying Visual FoxPro Solutions and Visual FoxPro Best Practices for The Next Ten Years. He has written articles for FoxTalk and FoxPro Advisor, and is a frequent speaker at Visual FoxPro conferences and user groups. Rick is a Microsoft Certified Solution Developer (MCSD) and a Microsoft Certified Professional (MCP) in Visual FoxPro.

Copyright © 2021 Rick Borup. Windows® is a registered trademark of Microsoft Corporation in the United States and other countries. All other trademarks are the property of their respective owners.

Bibliography

- [1] John Medina, *Brain Rules*, Pear Press, 2014
- [2] Edward R. Tufte, *The Visual Display of Quantitative Information* (2nd ed.), Graphics Press, 2001
- [3] Donna M. Wong, *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*, Norton paperback, 2013
- [4] Stephen Few, *Now you see it*, Analytics Press, 2009
- [5] Darrell Huff, *How to Lie with Statistics*, Norton Paperback 1993

Other References

- Nathan Yau, *Visualize This*, Wiley Publishing, 2011
- Mico Yuk & Stephanie Diamond, *Data Visualization for Dummies*, John Wiley & Sons, 2014
- Edward R. Tufte, *Beautiful Evidence*,
- Edward R. Tufte, *Envisioning Information*
- Edward R. Tufte, *Visual Explanations*, 1997
- Doug Lowe, *PowerPoint 2013 for Dummies*, John Wiley & Sons, 2013
- Jozef Cohen, *Sensation and Perception*, Rand McNally & Company, 1969

Oz du Soleil & Bill Jelen, *Guerrilla Data Analysis Using Microsoft Excel*, Holy Macro! Books, 2015

Hans Rosling, *Factfulness: Ten Reasons We're Wrong About the World — and Why Things Are Better Than You Think*, Flatiron Books, 2018

Presentation Zen blog

<http://www.presentationzen.com/presentationzen/2015/07/hans-rosling-upgrading-our-worldview-through-data-visualization.html> (<https://youtu.be/foXlFILjCgk>)

Power BI

<http://blogs.msdn.com/b/powerbi/archive/2015/07/24/power-bi-is-generally-available-today.aspx>
<https://powerbi.microsoft.com/desktop>