The What of Data Visualization

By Mary Pat Campbell

This is a fourth part of a continuing series on data visualization (aka dataviz):

- The Why of data visualization—questions to ask when visualizing numerical information
- The Who of data visualization—major figures and books in advocating data visualization best practices
- The Where of data visualization—websites to polish your data visualization game
- The What of data visualization—software to implement data visualization
- The How of data visualization—specific data visualization techniques to consider in actuarial practice.

(The when of data visualization being NOW, of course.)

For this article, I'm going to concentrate on software and resources involved with that software for data visualization, focusing on widely available and widely used languages or applications.

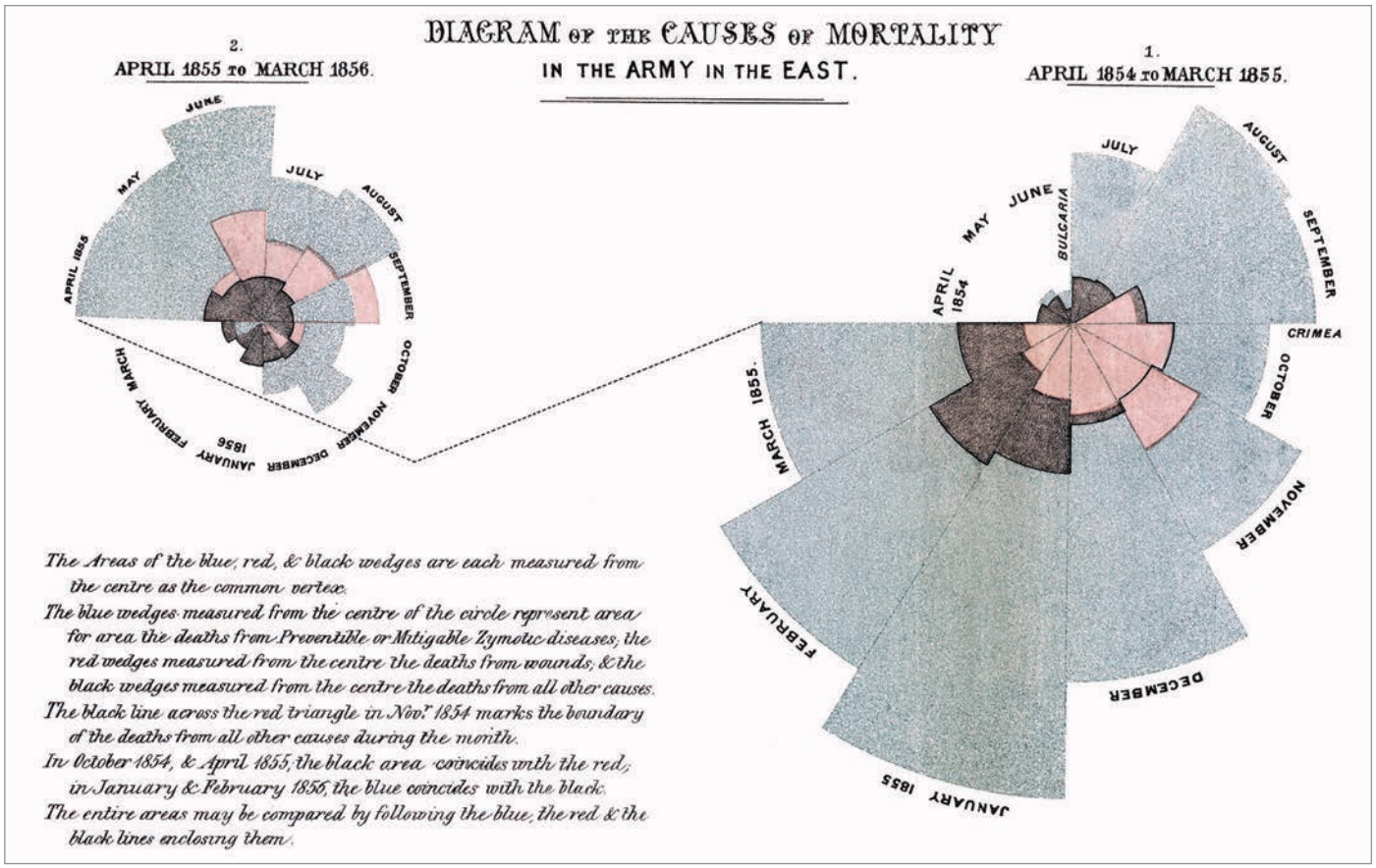
I will consider the choices on the following dimensions:

Dimension	Description
Learning Curve	How easy is it to get up to speed to use for visualizations?
Ease of Use	Once you've gotten up to speed, how easy is it to use?
Default Choices	How many default graphs are there to use? Do they span what you need to graph?
Flexibility	How much can you change in the visualizations?
Aesthetics	How pretty are the graphs?
Interactivity	Are there elements of interactivity for the data consumer (as opposed to the visualization creator)?

CONSIDER OLD TECHNOLOGY FIRST

But before I dive into an overview of some of the tools at our disposal, I want to advocate for a very old, familiar technology to begin with: pencil and paper.

In "The Who of Data Visualization," I mentioned Edward Tufte, who has a veritable suite of books and examples of fabulous (and some decidedly awful) visualizations—and many of the best examples he has were hand-drawn.



I recently came across this graphic from Florence Nightingale, on mortality rates in British war hospitals during the Crimean War.

Miss Nightingale drew several graphs, by hand, to demonstrate mortality rates among the British soldiers, which became part of her masterwork *Notes on Matters Affecting Health, Efficiency, and Hospital Administration in the British Army*, published in 1858. In addition to this radial-oriented graph—oriented to give the feel of a cycle for a year—she had other graphs in the finished book that were area graphs or simple bar charts, in addition to tables of numbers.

At the 1900 World Fair in Paris, W. E. B. Du Bois and his sociology students at Atlanta University (now known as Clark Atlanta University) created 60 charts to show how African American life had changed in the prior 200 years. The charts were hand-drawn and hand-colored, many being familiar bar and line charts, but some taking interesting shapes as Dr. Du Bois was not tied to any particular framework for displaying his data.

One of the issues with the technology-driven aspect of our jobs is that often we are crammed into the defaults or the structures endemic in our tools, and we go with the menu of what is easily available, rather than thinking about what we want to see or what we want to show. We forget that pencil and paper are open to us as well.

The main distinction between these historical hand-drawn charts and our own is the level of precision our software can graph with and uniformity of elements such as shading, line width, and text. We can be sure, by golly, that our graphs will be accurate to the nearest pixel or the fineness of our printers.

But can our data users actually perceive the difference? And one may consider the aesthetics not much of an improvement.

The whole point of data visualization is to provide *humans* with insight about a set of data, taking advantage of our huge analysis apparatus naturally built into our visual cortexes. We may be using the visualization to tell a story (as both Miss Nightingale and Dr. Du Bois were doing), or we may be using the visualization to see if there are discernable patterns in our data. But if distinctions can't be seen, the point of visualization is completely missed.

I'm not saying we should be publishing hand-drawn charts, but that if we have a complex set of data we want to visualize, the

first step may be to grab a marker and step up to the whiteboard, or to get a pen and some paper from the recycle bin. Don't hit the software as your first choice.

To see a story of a modern data graphic designer going through a process that started out with a sketch, check out the article "Sketching with Data Opens the Mind's Eye" by Giorgia Lupi, originally published on the National Geographic Data Points blog (links to this and a longer version of the article can be found at the end of this article.) In addition, links to MIss Nightingale's and Dr. Du Bois's charts can be found in the resource list—check them out!

EXCEL

Yes, yes, I know. But don't count Excel out for data visualization. Excel gets a bad reputation for awful default graph settings (and chart types that nobody should ever use [cough] pie charts [cough]), but let us consider the following: How complicated do we really need our graphs to be? As noted above, in the era of hand-drawn charts, we will see many familiar designs.

Much of the data visualization we need to do may be looking at particular slices of data as bar/column graphs, line graphs, or scatter plots. Excel can do all of those easily and, more to the point, has a relatively easy interface if you want to change the formatting of a single data point to help it stand out. Excel may be one of the easier tools to use simply due to our own familiarity.

But also due to our own familiarity, we may be used to using Excel in a specific way.

I want to highlight two types of visualizations built into Excel, which I find very useful for working in my calculations directly. I'm often creating large tables of data, and I would like to eyeball the results.

One method is using built-in Conditional Formatting. An example is seen below:

Return	0%										
		Contrib Increase									
		0%	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Benefit Increase	0%	2024	2025	2027	>2040	>2040	>2040	>2040	>2040	>2040	>2040	>2040
	1%	2023	2024	2024	2026	>2040	>2040	>2040	>2040	>2040	>2040	>2040
	2%	2022	2023	2023	2024	2025	>2040	>2040	>2040	>2040	>2040	>2040
	3%	2022	2022	2023	2023	2024	2025	>2040	>2040	>2040	>2040	>2040
	4%	2021	2022	2022	2022	2023	2023	2024	2028	>2040	>2040	>2040
	5%	2021	2021	2021	2022	2022	2023	2023	2024	2026	>2040	>2040
	6%	2021	2021	2021	2021	2022	2022	2022	2023	2024	2025	>2040
	7%	2021	2021	2021	2021	2021	2021	2022	2022	2023	2023	2025
	8%	2020	2020	2021	2021	2021	2021	2021	2022	2022	2022	2023
	9%	2020	2020	2020	2020	2021	2021	2021	2021	2021	2022	2022
	10%	2020	2020	2020	2020	2020	2021	2021	2021	2021	2021	2022

Another in-cell visualization is sparklines:

Funded Ratio Trend	Row Labels
	Chicago Municipal Employees
	Chicago Teachers
	Connecticut Municipal
	Connecticut SERS
	Illinois Municipal
	Illinois Teachers
	Wisconsin Retirement System

I find that many people are unaware of these as built-in visualization techniques for Excel, and the boon of these techniques in particular is that they are useful if you're working within data and want to see simple visualizations while you see what changes.

In this article, I am not talking about how to do any of these, but now that you know these exist, you can search for the resources to work through examples.

Dimension	Description
Learning Curve	Fairly simple for basic charts; even for the most complicated built-in techniques, there are not of things needed to learn.
Ease of Use	Using menus and options, fairly easy; if you need to automate with VBA, less easy, but still not too onerous
Default Choices	Covers major graph types, but provides too many bad options (3-d effects, pie charts, etc.)
Flexibility	Can manipulate axes, change fonts, etc.; difficult to change beyond basic chart types
Aesthetics	Aesthetics are meh; default graph styles have issues with color choices
Interactivity	Interactivity is limited for user; there are ways to use VBA to make things more interactive, but it's not natural to the program

GOOGLE DOCS—SHEETS

This one is harder to talk about, because as I was writing this, Google had an announcement on what it was doing with Sheets:

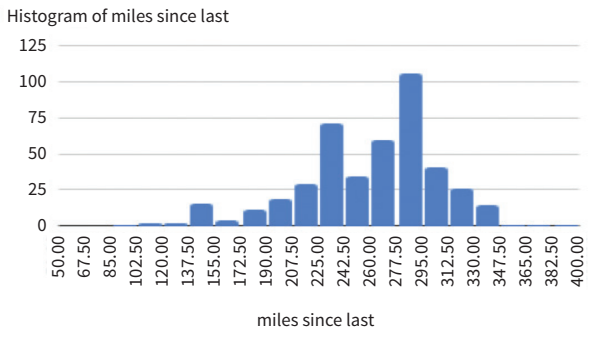
> Explore in Sheets, powered by machine learning, helps teams gain insights from data, instantly. Simply ask questions—in words, not formulas—to quickly analyze your

> data. For example, you can ask "what is the distribution of products sold?" or "what are average sales on Sundays?" and Explore will help you find the answers.
>
> Now, we're using the same powerful technology in Explore to make visualizing data even more effortless. If you don't see the chart you need, just ask. Instead of manually building charts, ask Explore to do it by typing in "histogram of 2017 customer ratings" or "bar chart for ice cream sales." Less time spent building charts means more time acting on new insights. (Gundrum, June 1, 2017)

First off, I wasn't even aware of Explore as an option in Sheets, which I've mainly used for some very simple spreadsheets I've created over the years. Explore is hidden in a diamond-ish icon in the lower right of the sheet, which opens up. The problem is that there really isn't a huge amount of flexibility with this tool, because the type of data I'm looking at tends to differ from whatever Google trained their Explore tool on.

This has been my frustration with Google Sheets in general: the suite of Google document-creating and -sharing tools has been geared toward people who don't want or need a lot of features—that is, the great majority of people currently using Microsoft Office. A huge amount of flexibility isn't available in terms of what the visualization creator can control: there are very few chart types to choose among, and the amount of editing you can do on items such as axes and fonts is also limited.

That said, I tried out the Explore tool with one of my longer-running Google Sheets: a tracker of my gasoline usage, costs, and mileage. I drive about 35,000 miles per year, so you can imagine I keep a close eye on this expense. I had created my own graphs with moving averages of cost per gallon and the like, but here are some graphs the Explore tool created for me automatically:



Ranges from 96 (October 5, 2015) to 394 (April 30, 2016), but most values are around 260, plus or minus 33.

Not only did it create a histogram for me, it also gave me some "analysis" in words.

Alas, most of the graphs automatically created were useless.

Dimension	Description
Learning Curve	Even simpler than Excel
Ease of Use	Very easy to create graphs
Default Choices	Has been expanding chart types, has main ones used; one not built into Excel is Geo Maps, which will do choropleths
Flexibility	Limited flexibility in controlling aspects of the graph, once the graph type has been chosen
Aesthetics	You are forced into particular aesthetics and can't adjust
Interactivity	Limited

R

R itself is an interpreted language developed by statistical-minded people, with 10 years of development. Various environments have been developed for it, to make it a bit more user-friendly, and people create packages of code to be used with the base R language and syntax. I use RStudio myself, and other tools are out there to make results easier to deal with.

The main issue users may have is that R itself is mostly command-line driven, meaning one must type in all the parameters being used. If you want to adjust aspects of the graphs, you may get annoyed at having to look up, for example, what the different parameters are.

The good aspect, though, is that if you're looking for data visualizations to help you analyze your data (as opposed to telling a good story), many people have already coded that for you, so graphs are automatically generated for you when you do specific analyses. For example, if you plot the results from a multiple linear regression, it will generate plots testing goodness of fit, such as a normal Q-Q plot, which checks that the residuals are normally distributed.

R is not intended to be user-friendly. It's intended to be tech-friendly, and a very large number of tools have been developed for people who concentrate on data analysis. The main library of R functions for graphing that I see people use are in ggplot2.

Dimension	Description
Learning Curve	Can be quite steep
Ease of Use	If adjusting a lot of aspects, can be difficult
Default Choices	Packages available for making all sorts of graphs, but you essentially have to look them up
Flexibility	A lot of flexibility in controlling the graphs, but this is a trade-off with how much you need to learn to manipulate these elements
Aesthetics	The default aesthetics are spartan and can get quite ugly
Interactivity	Limited

PYTHON

Python is a fairly widely used programming language. It is not optimized for data analysis, in its core language, but like with R, people develop modules to expand Python's capabilities. There are many "data science" types using Python now, with modules such as pandas, numpy, matplotlib, and Seaborn that are used for data crunching and visualization.

Much of what I wrote for R can be written for Python—indeed, many people use both. I find that the R people came mainly from the academic side, but because Python is fairly easy to pick up, some invade from that side. There's a package to run Python from R . . . and vice versa. R is more used for standalone data analysis, and Python when you need to work with apps or other external users.

There are ways to import/export data from Excel to these as well.

Dimension	Description
Learning Curve	Python itself is fairly simple, but getting into data science applications can be quite difficult
Ease of Use	It is easy to debug Python code, but you still need to learn specific code
Default Choices	There are packages for making all sorts of graphs, but you essentially have to look them up
Flexibility	A lot of flexibility in controlling the graphs—but this is a trade-off with how much you need to learn to manipulate these elements
Aesthetics	Many packages to help make pretty graphs
Interactivity	Can be interactive—but you need to find proper modules to use

TABLEAU

Tableau is an "oooh, pretty" type of software, with multiple versions, and the pricing has been in flux as of late. That said, there is a free version to play with, the downside being everything you do with it must be public, and the options on the free, public version are more limited than the full software. I have not tried a paid version, so my comments relate to the public version.

Dimension	Description
Learning Curve	Easy to get started, with lots of support: sample data sets, videos, and exercises
Ease of Use	Extensive point-and-click interface; intended to be easy for general public to use.
Default Choices	Lots of choices, and uses defaults based on structure and content of the data; has geographic graphs as a choice
Flexibility	There can be flexibility, but to keep it simple, limits on what you can adjust in the public tool
Aesthetics	It's the prettiest in the land—nice color palettes
Interactivity	Built to be interactive, very easy to embed into websites

TRYING TO HIT A MOVING TARGET

The above evaluations of different choices are not exhaustive. Given that data science is very hot right now, and that several not terribly numerate people are attempting to ride this wave, more and more easy-to-use tools are being developed. As I noted, as I was writing about Google Sheets, I noticed that they just announced new features.

The problem is that we're now in a world of software-as-a-service. One doesn't buy a static piece of software, but you have a subscription to using a platform (like Microsoft Office or Tableau), or you're part of a coding community where the packages keep getting updated.

This article is more to let you know some of the available tools. I tend to use multiple tools, depending on my task. I do much of my for-publication visualizations in Excel (or, rather, graphs embedded in PowerPoint) because I am part of a publication team, and this is dependent on the tools I use. I can generate graphics to use in other formats, but because of limitations of adjustments and wanting to meet certain production standards, I keep this sort of production to a minimum.

I am primarily using tools like R and Python to do visualization for analysis beyond Excel's (or Google Sheets') capabilities. I use Google Sheets for simple graphs I embed in my blog. I have rarely used Tableau, though I did use it once, to embed an interactive graph on my blog. It didn't play well with my blogging software, so I didn't do that again.

At the end of it all, I tend to go back to my pencil and paper, trying to do some doodles to think things through. Some visualizations I'm still working on . . . and maybe that perfect tool will be just around the corner for me. ■

Mary Pat Campbell, FSA, MAAA, is a vice president, Insurance Research at Conning in Hartford, Conn. She can be reached at *marypat.campbell@gmail.com*.

RESOURCES

Giamo, Cara. "Florence Nightingale Was Born 197 Years Ago, and Her Infographics Were Better than Most of the Internet's." *Atlas Obscura*, May 12, 2017, *http://www.atlasobscura.com/articles/florence-nightingale-infographic*. Date accessed, May 16, 2017.

Gundrum, Daniel. "Visualize Data Instantly with Machine Learning in Google Sheets." G-Suite, June 1, 2017, *https://www.blog.google/products/g-suite/visualize-data-instantly-machine-learning-google-sheets/*. Date accessed, June 7, 2017.

Lupi, Giorgia. "Sketching with Data Opens the Mind's Eye." National Geographic Data Points blog, July 10, 2015, *http://news.nationalgeographic.com/2015/07/2015704-datapoints-sketching-data/*. Date accessed, June 7, 2017.

Lupi, Giorgia. "Sketching with Data Opens the Mind's Eye." Accurat studio blog on *Medium.com*, Feb. 19, 2016, *https://medium.com/accurat-studio/sketching-with-data-opens-the-mind-s-eye-92d78554565*. Date accessed, June 7, 2017.

Meier, Allison. "W. E. B. Du Bois's Modernist Data Visualizations of Black Life." Hyperallergic, July 4, 2016, *https://hyperallergic.com/306559/w-e-b-du-boiss-modernist-data-visualizations-of-black-life/*. Date accessed, Feb. 9, 2017.

Miller, Meg. "W. E. B. Du Bois Was a Master of the Hand-Drawn Infographic." *Co.Design*, Febr. 9, 2017, *https://www.fastcodesign.com/3068020/web-dubois-was-a-master-of-the-hand-drawn-infographic*. Date accessed, Feb. 9, 2017.

Nightingale, Florence. *Notes on Matters Affecting Health, Efficiency, and Hospital Administration in the British Army.* London: Harrison and Sons, 1858. *https://archive.org/details/b20387118.*

Simonite, Tom. "Google Sprinkles AI on Its Spreadsheets to Automate Away Some Office Work." MIT Technology Review, June 2, 2017, *https://www.technologyreview.com/s/608023/google-sprinkles-ai-on-its-spreadsheets-to-automate-away-some-office-work/*. Date accessed, June 7, 2017.

Willems, Karlijn. "Choosing R or Python for Data Analysis? An Infographic." DataCamp, May 12, 2015, *https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis*. Date accessed, June 7, 2017.