Data visualization - 00 - overview

Steve Simon

5/21/2019

Notes about this talk (1 of 3)

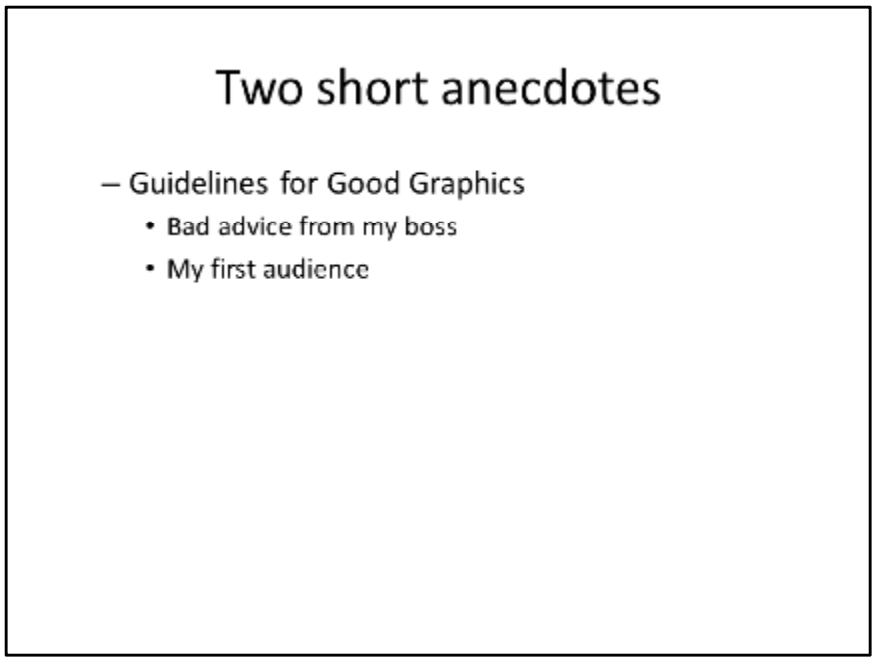
- These slides should not to be included in the final presentation
- Very brief overview
  - Before any other presentations
  - Put any generic announcements here

Notes about this talk (2 of 3)

- General order of all talks
  - 00 - overview
  - 01 - points
  - 02 - bars
  - 03 - lines
  - 04 - surfaces
  - 05 - maps
- Order not fixed
  - 02 and 03 might be swapped
  - 04 and 05 might be swapped
  - 04 and 05 both optional

Notes about this talk (3 of 3)

- Theory is spread out over several talks
  - 01 - points (grammar of graphics)
  - 02 - bars (psychology of perception)
  - 03 - lines (gestalt of graphics)
  - 04 - surfaces (myth of the objective graph)
  - 05 - maps (projection systems)
- I am open to alternative organizational structures



Before I start, I want to share a couple stories. This presentation is an update of a talk I gave back in the 1990s called Guidelines for Good Graphics. I was telling my boss about my plans for this talk and she suggested that I do some "before and after" comparisons of some graphics that she would provide. I politely declined. You hate to ignore advice from your boss, but I wasn't the sort of person who liked tearing down other people's work. I was also worried that someone would take one of my "after" pictures and show how much better they could make it look.

I'm never without an opinion, but I hope that I never get so arrogant as to think that my opinion is the final word on the matter. A lot of what I plan to talk about requires careful judgement and if your judgement differs from mine, please let me know. It makes the class more interesting.

The other story is that my first audience had a lot of psychologists in the audience. I was going to use the pyschology of perception to help decide what an effective graphic would be. I didn't like the thought of having people in the audience who probably knew more about the psychology of perception than I did. Not only would they point out things when I said something wrong, but they would point out how my toilet training when I was a child caused me to make that mistake.

I am a big believer in software agnosticism. That means that when I teach something, I teach it with the expectation that the software used to do the assignments is software of your choosing. I have my own preferences, but those should not be your preferences.

It's a lot more work to teach a course that is not dependent on a particular software system, but I do not know what the best software choice would be for you. In this class, I will try to show examples using Python, R, and Tableau. I realize that there are other good choice, but I'm hoping that most of you will be happy with one of these three choices. Within Python, I will use the altair package. In R, I will use ggplot2. I realize that there are other graphics packages in these two languages, but altair and ggplot2 rely on modern graphics principles, so I will restrict my attention to these packages.

Tableau is a commercial product. If you don't currently have access to Tableau, the company offers a free version, Tableau Public. It has all the features of Tableau, but you have to store any data visualizations on a public server. That's just fine for someone like me who uses teaching examples with publicly available data sets. If you are using private or proprietary data, you need to pay the money for the commercial

version.

If you are not sure what software package to use in this class, let me offer a few suggestions. First, your boss may have a strong opinion about what software that you should use. You can go to your boss and say "Steve Simon is a really smart guy and he says that the ggplot2 package in R is the best choice for data visualization." Try it and see what happens. Nothing, I suspect. One of the great tragedies in life is that the SSSS principle (Steve Simon Said So) carries very little weight in the real world.

If your boss doesn't care, see what most of your co-workers are using. They may not be as smart as I am (put on a false air of pride here) but they are a lot closer to your cubicle when this workshop ends and you have to find a quick answer.

There's also a comfort level here. Tableau develops its visualizations using a graphical user interface. Python and R are programming languages. A graphical interface is great for getting work done quickly. A programming language is great for reproducibility and reusability. What fits your working style better.

One more consideration. Some of you in this class are "ringers." You already know visualization better than I do because you've been doing it for longer, with bigger and more complex data sets. You're just here to see if I know one or two things that you don't already know. If you're a ringer, take the challenge of learning a new software system. It will keep you from getting too bored when I talk about all these things that you already know better than I do.