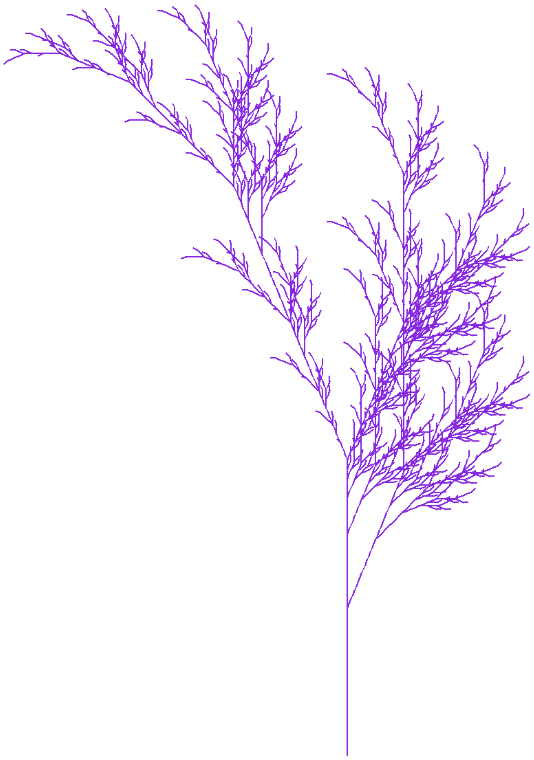
Elise Sawyer

1)

Francis J. Anascombe's "Graphs in statistical analysis" demonstrates the flaws in assuming that specific aspects of statistical analysis give a clear indication of how data is distributed. He uses four sets of eleven points that, while different in composition, share incredibly similar means, regression coefficients + equations, error estimations, etc. The first "example" is a more standard set of scattered points, the second has points which resemble a quadratic relationship, and the third and fourth figures each involve one "outlier" point. This paper emphasizes the importance of examining data graphically to better account for situations such as these-- seeing the differences in these four sets of points already helps prevent people from assuming they are identical. Additionally, the paper discusses how in situations such as those found in examples 3 and 4 (with one point greatly influencing a result), acknowledging that point's contribution and how data would change with its absence could be helpful. A noteworthy aspect of this paper to me is how it is over 50 years old; Anascombe references how creating and examining data graphically is a process requiring "trouble, cunning, and a fighting spirit" and hopes that this will change. His hope has since been realized, so the challenge of dealing with sets of data such as these is less related to dedication to computer programming, and more related to understanding when simply reporting means and equations wouldn't be enough.

2)



Elise Sawyer

3)

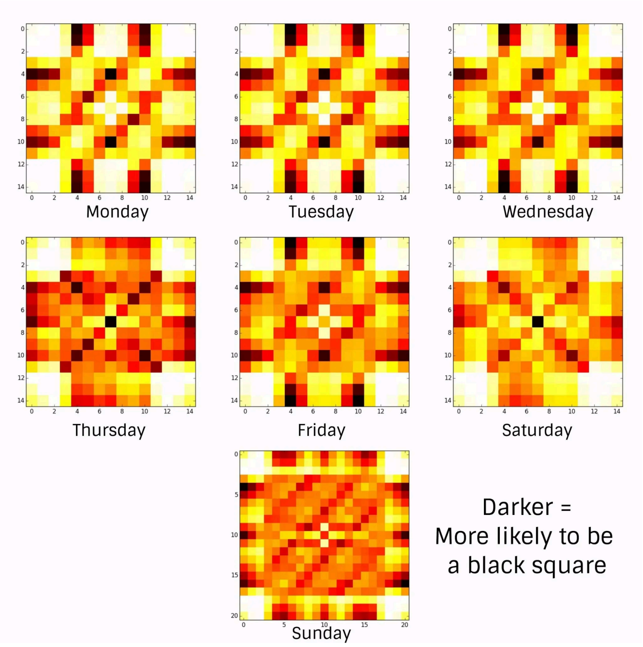
The data visualization I will critique is by Darren McCleary, which I found in a New York Times article related to crossword construction (one of my hobbies is solving crosswords, and I've tried my hand in the past at making them). It is a set of heat maps explaining how likely a square in a crossword grid is to be a black square (as opposed to a square where you can place a letter).

The color scheme works for this being a heat map, though I do think that it is difficult to distinguish between the pale yellow and white parts of the charts. One idea of how the difficult-to-see color scheme could be fixed is the use of a diverging color palette. For instance, a square could be more red if it is more likely to be a filled square than not, and more blue if it is more likely to be an empty square. However, this solution presents a new problem. As this graphic is about the likelihood of squares being filled in with a dark color, having both the "especially likely to be filled in" squares as well as the "especially likely to be empty" squares be represented in dark colors would be problematic.

Additionally, these heat maps do not give any legend about the differences between the darker and lighter squares (only that darker spaces are more likely to have black squares). My issue is not that the graphic avoided using numbers to explain this (as its audience is just people who are reading the crossword construction advice article), but that this graphic does use numbers for the x and y axis of the graphs. I don't understand why they chose to have the highest square on the y axis be represented with "0". While crosswords do typically start their lists of clues with ones for the words in the top left, it makes this chart more confusing to look at, although this could have been avoided by labeling the x axis at the top rather than at the bottom of each chart.

One detail that would have been interesting is to label which squares are the most frequently-used for black spaces, or least frequently used, via some sort of indication on the chart.

http://nytimes.com/2018/05/11/crosswords/how-to-make-crossword-puzzle-grid.html



Monday

Tuesday

Wednesday

Thursday

Friday

Saturday

Sunday

Darker =
More likely to be
a black square