The What, How, and Why of Data Visualization

Data and Image Models

S. Santoni[1,2]

[1]Bayes Business School

[2]Soundcloud

MSc in Business Analytics, 2022/23

Outline

What are the 'ingredients' of a data viz?

According to the designing thinking literature [4], a data viz contains the following three groups of 'ingredients:'

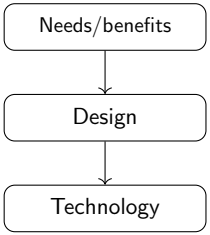
- User needs/benefits, i.e., the information a user wants to achieve
- Design, i.e., the set of choices regarding the visual forms, color, density, redundancy, and so on that characterize a data viz
- Technology, i.e., the knowledge, tools, and data underlying the data viz

How does the data viz process look like?

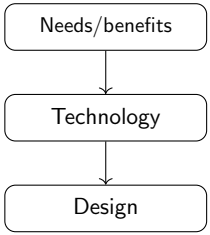
The design thinking literature [4] suggests that the user needs/benefits component should be the starting point of the data viz process. The intuition is that data viz that addresses nobody's needs is useless!

Instead, there is substantial flexibility when it comes to fix the design and technology components. Ultimately, the order depends depends on contingent factors and the designer's background, skills, and preferences.

Pathway A



Pathway B



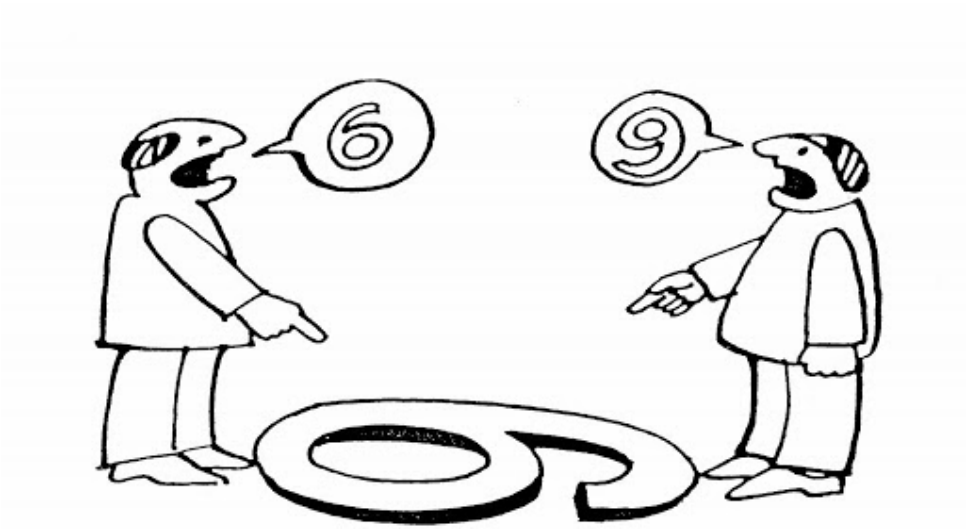
A closer understanding of the data viz process?

Do not worry!

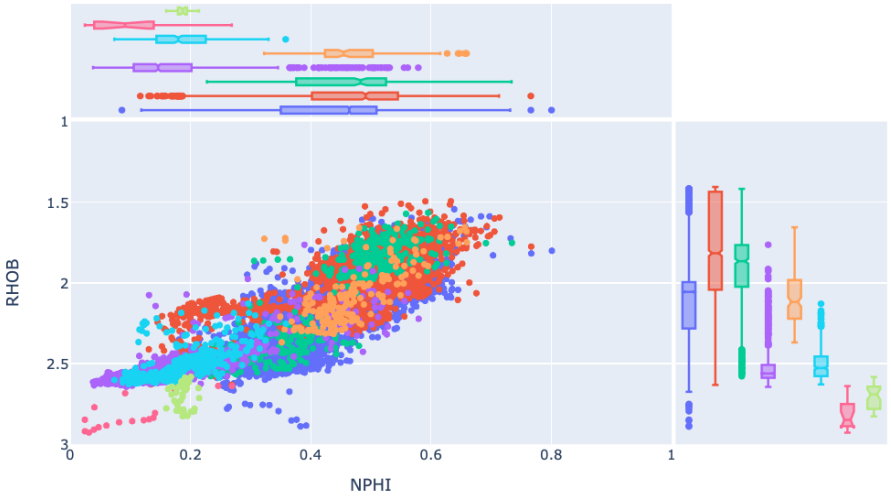
We will analyze the data viz process next week by the 'Data-Information-Knowledge-Wisdom' model [1].

‘The Good, the Bad, and the Ugly’

What is a good data viz?

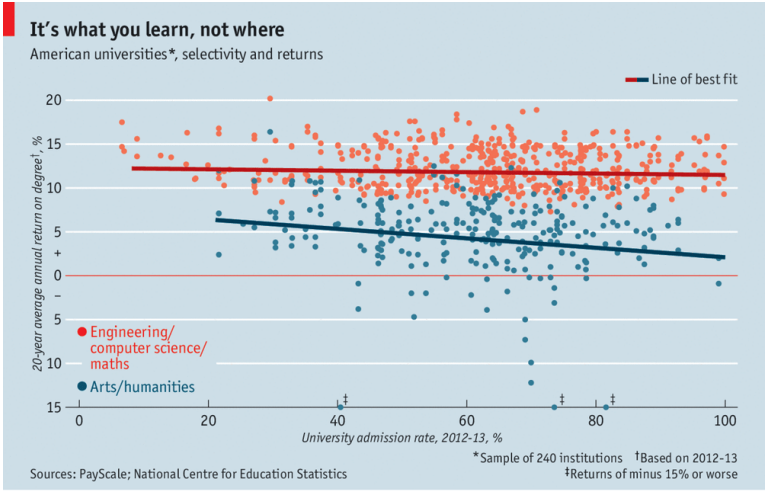


Example A: A Plot from the a Towards Data Science Post



Source: https://towardsdatascience.com/enhance-your-plotly...

Example B: A Chart from an Article in The Economist

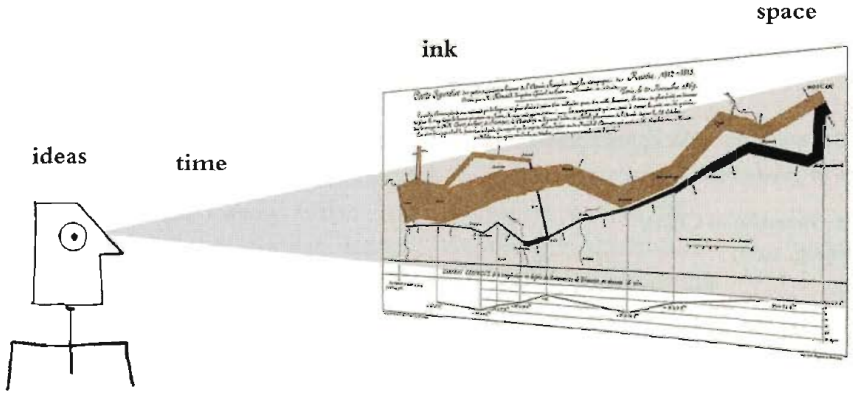


Source: https://www.economist.com/...it-depends-what-you-study-not-where

Graphical Excellence according to Tufte

Per Tufte's work [3], excellence in statistical graphs consists of complex *"ideas communicated with clarity, precision, and efficiency."*

Graphical displays pursuing clarity, precision, and efficiency *"give to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."*



How to Reach Clarity, Efficiency, and Precision?

Tufte points out graphical displays should

- Show the data
- Induce the viewer to think about the substance rather than about the methodology, graphical design, the technology of graphic production, or something else
- Avoid distorting what the data have to say
- Present many number in a small space
- Make large datasets coherent
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to the fine structure
- Serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- Be closely intergrated with the statistical and verbal description of a data set

Show the Data!

Here is a classic example on the importance of showing the data, the case of Anscombe's quartet [2].

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	4.74	5.0	5.73	8.0	6.89

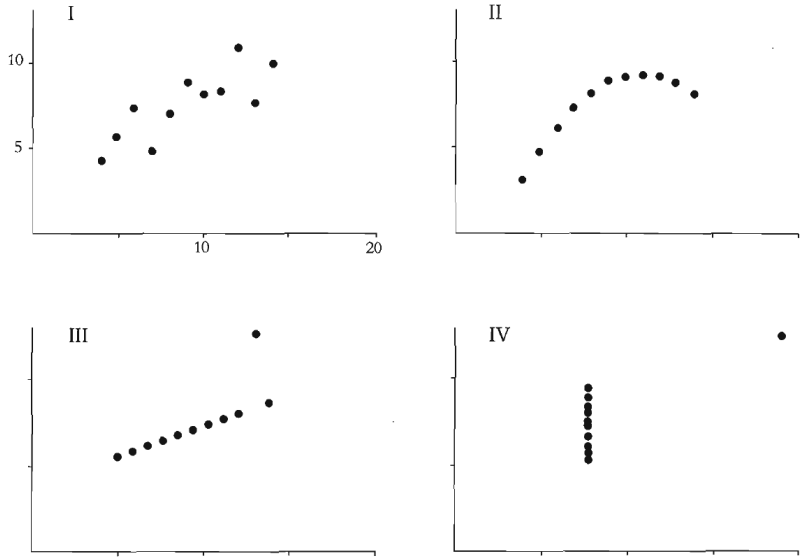
→

N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - \overline{X} = 110.0$
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
$r^2 = .67$

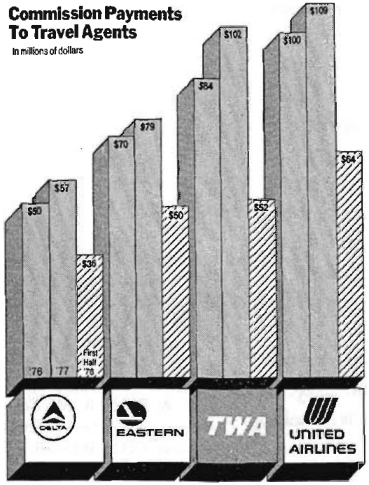
1

[1] See [3, page 14]

Show the Data! (cont'd)



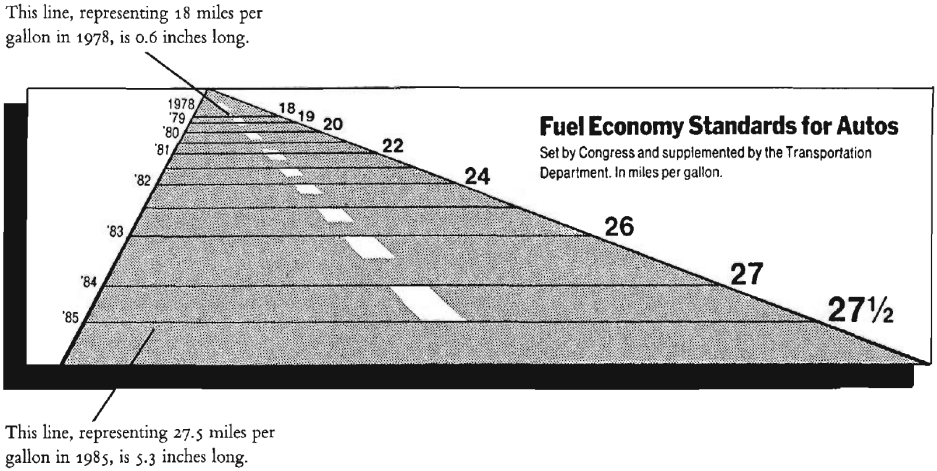
Excellent Graphical Displays Tell the Truth!



Source: New York Times, August 8, 1978, page D-1.

Tufte [3, page 54] observes that *'the pseudo-decline was created by comparing six months' worth of payments in 1978 to a full year's worth in 1976 and 1977, with the lie repeated four times."*

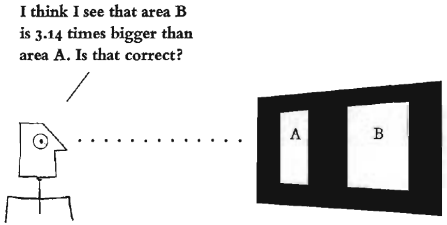
Excellent Graphical Displays Tell the Truth!



Source: New York Times, August 9, 1978, page D-2.

The Lie Factor

$$\text{Lie Factor} = \frac{\text{Size of the effect shown in graphic}}{\text{Size of effect in data}} \quad (1)$$



The Lie Factor

Effec size in data

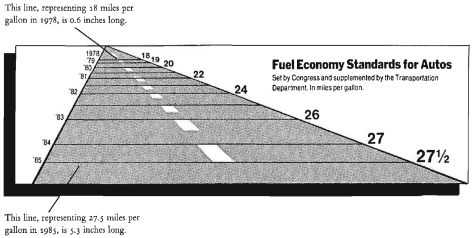
$$\frac{27.5 - 18.0}{18.0} = 5.3 \tag{2}$$

Effect shown in graphic

$$\frac{5.3 - 0.6}{0.6} = 78.3 \tag{3}$$

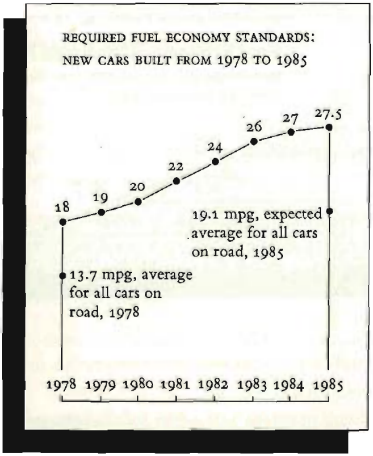
Lie factor

$$\frac{78.3}{5.3} = 14.8 \tag{4}$$

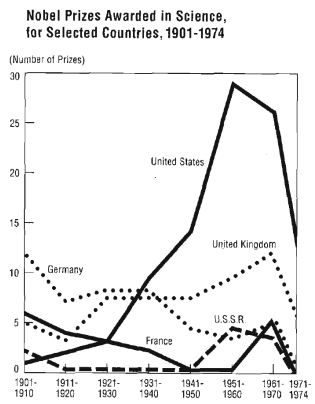


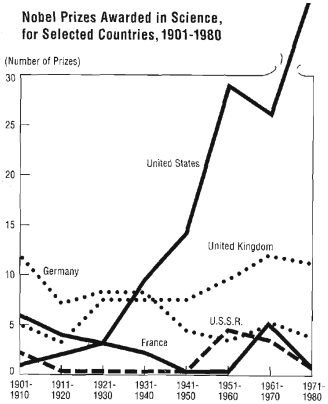
Let Us Redesign the 'Fuel Econmy Standards' Graph

Tufte points out *"it is easy enough to decorate the data without lying"*



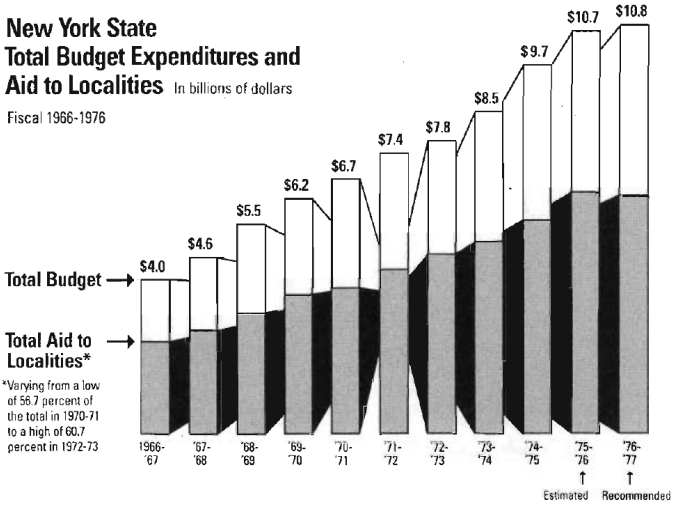
Another Time Series that Lies





Source: National Science Foundation, Science Indicators, 1974.

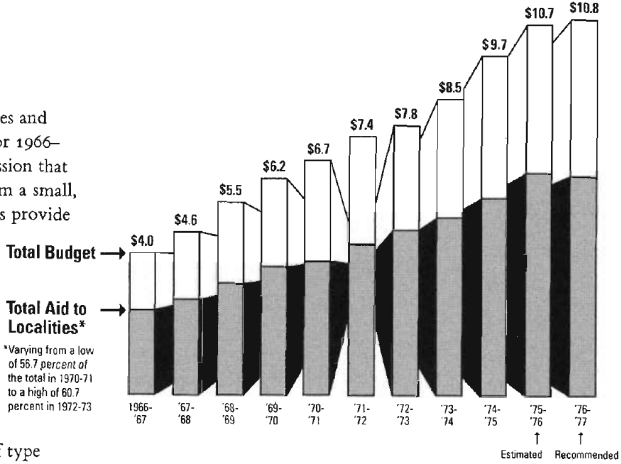
Yet another Time Series that Lies



Source: New York Times, February 1, 1976, page IV-6.

Yet another Time Series that Lies

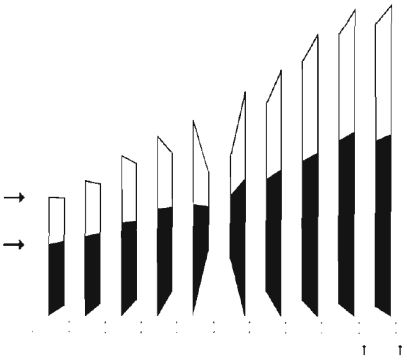
This cluster of type emphasizes and stretches out the low value for 1966–1967, encouraging the impression that *recent years* have shot up from a small, stable base. Horizontal arrows provide similar emphasis.

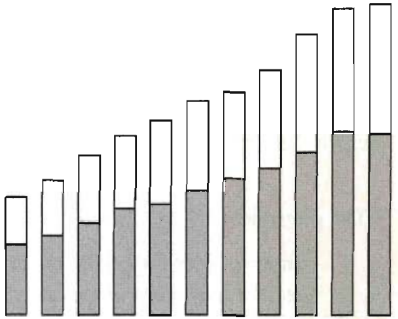


This squeezed-down block of type contributes to an image of small, squeezed-down budgets back in the good old days.

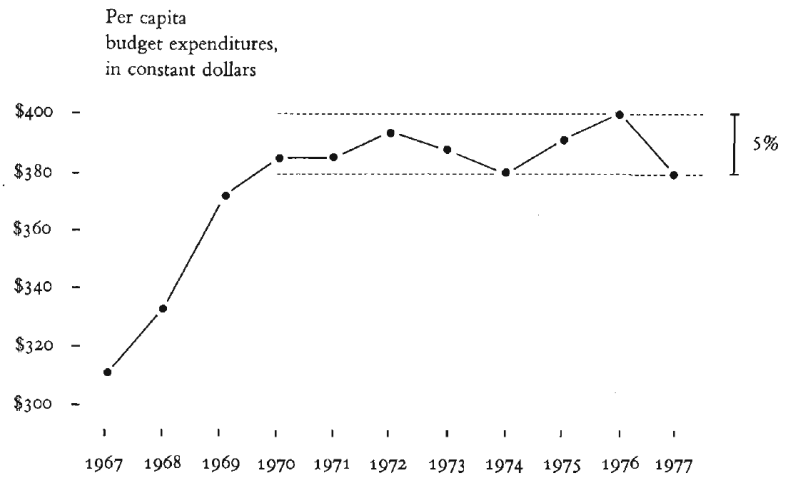
Arrows pointing straight up emphasize recent growth. Compare with horizontal arrows *at left*.

Chartjunks Distort the Data!





Let Us Redesign the 'NYS Total Budget Expenditures' Graph



How Can Graphic Mediocrty Be Remedied?

Graphical competence demands three different skills

- Substantive skills
- Statistical skills
- Design skills

References

[1] Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. New Riders, 2012.

[2] "Graphs in Statistical Analysis". In: *American Statistician* 27 (February 1973), pp. 17–21.

[3] *The Visual Display of Quantitative Information*. Second edition. Graphics Press LLC, 2001.

[4] Colin Ware. *Visual Thinking for Design*. Elsevier, 2010.