

I Graphical Excellence

Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets 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 integrated with the statistical and verbal descriptions of a data set.

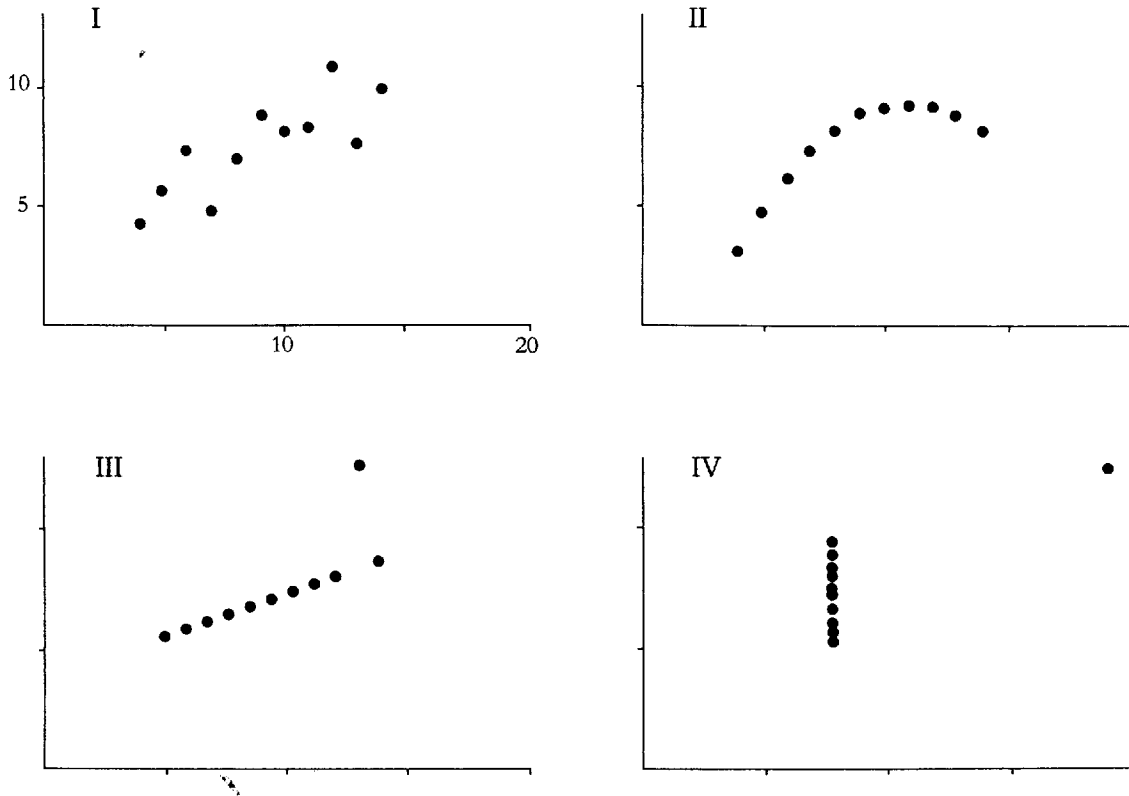
Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations. Consider Anscombe's quartet: all four of these data sets are described by exactly the same linear model (at least until the residuals are examined).

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

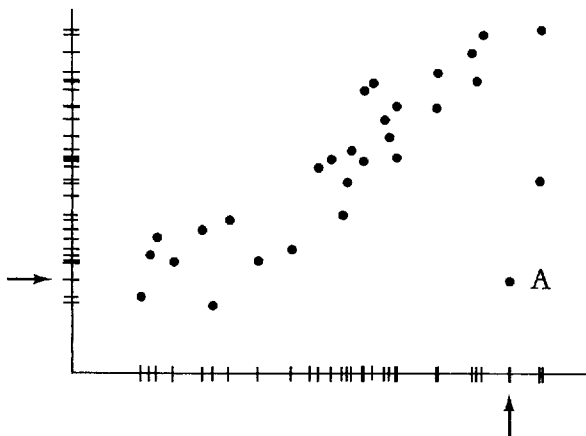
$N = 11$
 mean of X's = 9.0
 mean of Y's = 7.5
 equation of regression line: $Y = 3 + 0.5X$
 standard error of estimate of slope = 0.118
 $t = 4.24$
 sum of squares $X - \bar{X} = 110.0$
 regression sum of squares = 27.50
 residual sum of squares of Y = 13.75
 correlation coefficient = .82
 $r^2 = .67$

And yet how they differ, as the graphical display of the data makes vividly clear:

F. J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (February 1973), 17-21.

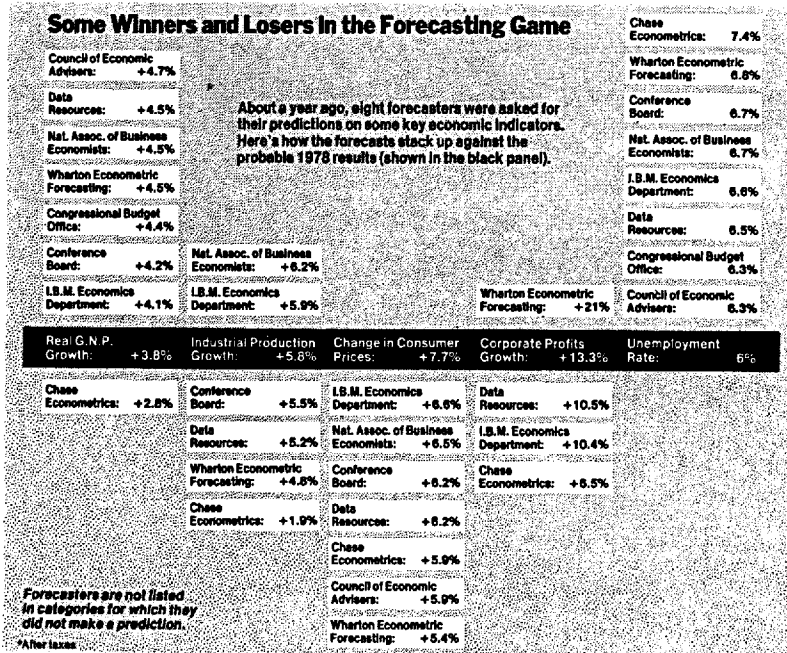


And likewise a graphic easily reveals point A, a wildshot observation that will dominate standard statistical calculations. Note that point A hides in the marginal distribution but shows up as clearly exceptional in the bivariate scatter.



Stephen S. Brier and Stephen E. Fienberg, "Recent Econometric Modelling of Crime and Punishment: Support for the Deterrence Hypothesis?" in Stephen E. Fienberg and Albert J. Reiss, Jr., eds., *Indicators of Crime and Criminal Justice: Quantitative Studies* (Washington, D.C., 1980), p. 89.

For sets of highly labeled numbers, a wordy data graphic—coming close to straight text—works well. This table of numbers is nicely organized into a graphic:



New York Times, January 2, 1979, p. D-3.

Making Complexity Accessible: Combining Words, Numbers, and Pictures

Explanations that give access to the richness of the data make graphics more attractive to the viewer. Words and pictures are sometimes jurisdictional enemies, as artists feud with writers for scarce space. An unfortunate legacy of these craft-union differences is the artificial separation of words and pictures; a few style sheets even forbid printing on graphics. What has gone wrong is that the techniques of production instead of the information conveyed have been given precedence.

Words and pictures belong together. Viewers need the help that words can provide. Words on graphics are data-ink, making effective use of the space freed up by erasing redundant and non-data-ink. It is nearly always helpful to write little messages on the plotting field to explain the data, to label outliers and interesting data points, to write equations and sometimes tables on the graphic itself, and to integrate the caption and legend into the design so that the eye is not required to dart back and forth between textual material and the graphic. (The size of type on and around graphics

can be quite small, since the phrases and sentences are usually not too long—and therefore the small type will not fatigue viewers the way it does in lengthy texts.)

The principle of *data/text integration* is

Data graphics are paragraphs about data and should be treated as such.

Words, graphics, and tables are different mechanisms with but a single purpose—the presentation of information. Why should the flow of information be broken up into different places on the page because the information is packaged one way or another? Sometimes it may be useful to have multiple story-lines or multiple levels of presentation, but that should be a deliberate design judgment, not something decided by conventional production requirements. Imagine if graphics were replaced by paragraphs of words and those paragraphs scattered over the pages out of sequence with the rest of the text—that is how graphical and tabular information is now treated in the layout of many published pages, particularly in scientific journals and professional books.

Tables and graphics should be run into the text whenever possible, avoiding the clumsy and diverting segregation of “See Fig. 2,” (figures all too often located on the back of the adjacent page).³ If a display is discussed in various parts of the text, it might well be printed afresh near each reference to it, perhaps in reduced size in later showings. The principle of text/graphic/table integration also suggests that the same typeface be used for text and graphic and, further, that ruled lines separating different types of information be avoided. Albert Biderman notes that illustrations were once well-integrated with text in scientific manuscripts, such as those of Newton and Leonardo da Vinci, but that statistical graphics became segregated from text and table as printing technology developed:

The evolution of graphic methods as an element of the scientific enterprise has been handicapped by their adjunctive, segregated, and marginal position. The exigencies of typography that moved graphics to a segregated position in the printed work have in the past contributed to their intellectual segregation and marginality as well. There was a corresponding organizational segregation, with decisions on graphics often passing out of the hands of the original analyst and communicator into those of graphic specialists—the commercial artists and designers of graphic departments and audio-visual aids shops, for example, whose predilections and skills are usually more those of cosmeticians and merchandisers than of scientific analysts and communicators.⁴

³“Fig.,” often used to refer to graphics, is an ugly abbreviation and is not worth the two spaces saved.


⁴Albert D. Biderman, “The Graph as a Victim of Adverse Discrimination and Segregation,” *Information Design Journal*, 1 (1980), 238.

Page after page of Leonardo's manuscripts have a gentle but thorough integration of text and figure, a quality rarely seen in modern work:

Leonardo da Vinci, *Treatise on Painting* [Codex Urbino Latinus 1270], vol. 2, facsimile (Princeton, 1956), p. 234, paragraph 827.

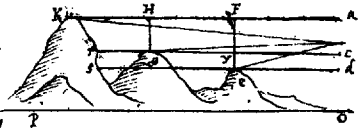
234.

chevai le cose vedute essere tanto minore che no che
 le membra ma il tutto quasi ripara impossibile a po-
 tere figurare come se l'occhio fusse o, e la busa di un
 quarto di braccio eguale alla tua tavola dipinta
 sia, a, b, discosta m dal occhio mezo
 braccio allora tu vedrai per esso spa-
 cio tutte le cose che ueder si possono den-
 tro alla lunghezza o d'uno orizzonte di
 cento miglia in tanta confusa diminutione che no
 che figurar di quelle alcuna parte e habbia figurar
 ma appena potrai porre si piccolo punto di penello che
 non sia maggiore ch'ogni gran casamento posto in
 dieci miglia di distanza.



perche li monti in langha distantia
 si dimostrano piu scuri nella cima
 che nella basa —

L'aria e acquista gradi di grossezza in ogni grado de
 la sua bassezza e della sua distantia e causa ch'esse
 cime de monti che piu s'in alzano piu mostrano la
 sua natu-
 rita per
 sono impe-
 grossezza
 nella cima



vale oscu-
 che ma co
 dite dalla
 de l'aria
 che nella
 loro basa o nella vicinita che nella remotione. Pro-
 uasi, o, p, d, s, i, c, v, a, k, sono gradi dell'aria che sem-
 pre s'asorglian' quanto piu s'in alzano, a, f, f, h, h, k,
 sono li altri gradi trasuersali dove l'aria acquista

Finally, a caveat: the use of words and pictures together requires a special sensitivity to the purpose of the design—in particular, whether the graphic is primarily for communication and illustration of a settled finding or, in contrast, for the exploration of a data set. Words on and around graphics are highly effective—sometimes all too effective—in telling viewers how to allocate their attention to the various parts of the data display.⁵ Thus, for graphics in exploratory data analysis, words should tell the viewer how to read the design (if it is a technically complex arrangement) and not what to read in terms of content.

⁵Experiments in visual perception indicate that word instructions substantially determine eye movements in viewing pictures. See John D. Gould, "Looking at Pictures," in Richard A. Monty and John W. Senders, eds., *Eye Movements and Psychological Processes* (Hillsdale, N.J., 1976), 323–343.

Accessible Complexity: The Friendly Data Graphic

An occasional data graphic displays such care in design that it is particularly accessible and open to the eye, as if the designer had the viewer in mind at every turn while constructing the graphic. This is the *friendly data graphic*.

There are many specific differences between friendly and unfriendly graphics:

Friendly	Unfriendly
words are spelled out, mysterious and elaborate encoding avoided	abbreviations abound, requiring the viewer to sort through text to decode abbreviations
words run from left to right, the usual direction for reading occidental languages	words run vertically, particularly along the Y-axis; words run in several different directions
little messages help explain data	graphic is cryptic, requires repeated references to scattered text
elaborately encoded shadings, cross-hatching, and colors are avoided; instead, labels are placed on the graphic itself; no legend is required	obscure codings require going back and forth between legend and graphic
graphic attracts viewer, provokes curiosity	graphic is repellent, filled with chartjunk
colors, if used, are chosen so that the color-deficient and color-blind (5 to 10 percent of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color-deficient people)	design insensitive to color-deficient viewers; red and green used for essential contrasts
type is clear, precise, modest; lettering may be done by hand	type is clotted, overbearing
type is upper-and-lower case, with serifs	type is all capitals, sans serif

With regard to typography, Josef Albers writes:

The concept that "the simpler the form of a letter the simpler its reading" was an obsession of beginning constructivism. It became something like a dogma, and is still followed by "modernistic" typographers. . . . Ophthalmology has disclosed that the more the letters are differentiated from each other, the easier is the reading. Without going into comparisons and details, it should be realized that words consisting of only capital letters present the most difficult reading—because of their equal height, equal volume, and, with most, their equal width. When comparing serif letters with sans-serif, the latter provide an uneasy reading. The fashionable preference for sans-serif in text shows neither historical nor practical competence.⁶

⁶Josef Albers, *Interaction of Color* (New Haven, 1963, revised edition 1975), p. 4.