UiO : CEMO – Centre for Educational Measurement
University of Oslo

Representations
Data Visualization
Data Science

Johan Braeken

useR!

Representations

graphs, charts, plots, figures, tables, diagrams

quantitative or visual representations of data/statistics/model

”Communication tool par excellence”

Great tool for Explorative Data Analysis (Tukey, 1977)

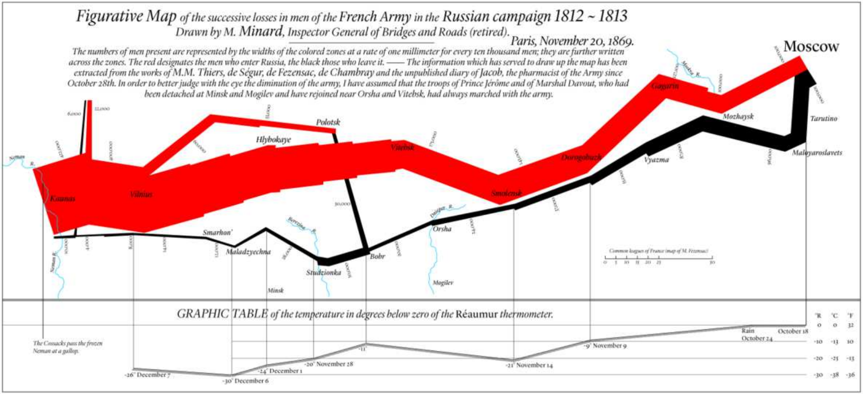
- Efficient when time limited to convey message
- Attractive to wide non-technical audience
- Organize & Document findings
- Reveal structure & patterns
  1. Facilitate Comparison: Raise questions
  2. Make the Data Stand Out: Provide answers

Data visualization classics:

W. Cleveland (1985); Tufte (2001); Wilkinson (2005); Wainer (1997)

Napolean march to Moscow

https://en.wikipedia.org/wiki/Charles_Joseph_Minard


Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813
Drawn by M. Minard, Inspector General of Bridges and Roads (retired).
Paris, November 20, 1869.
The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. —— The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.
Moscow
6,000
22,000
Polotsk
Hlybokaye
Kaunas
Vilnius
Vitebsk
Gagarin
Mozhaysk
Tarutino
Maloyaroslavets
Vyazma
Dorogobuzh
Smolensk
Orsha
Mogilev
Bobr
Studzionka
Minsk
Maladzyechna
Smarhon'
Berezina R.
Dnieper R.
Neman R.
Moskva R.
30,000
Common leagues of France (map of M. Fezensac)
0 5 10 15 20 25 50
GRAPHIC TABLE of the temperature in degrees below zero of the Réaumur thermometer.
The Cossacks pass the frozen Neman at a gallop.
-26° December 7
-30° December 6
-24° December 1
-20° November 28
-11°
-21° November 14
-9° November 9
Rain October 24
October 18
°R °C °F
0 0 32
-10 -13 10
-20 -25 -13
-30 -38 -36

Data - Perception - Design - Construction

Learning goals

- Recognize structural elements in a statistical graph (e.g., axis, plotting symbol, label)
- Evaluate the effectiveness (for perception and judgment) and appropriateness (for the type of data) of a structural element;
- Translate relationships reflected in a graph to the data represented;
- Use context to make sense of what is presented in a graph and avoid reading too much into relationships observed.
- Recognize when one graph is more useful than another and organize/reorganize data to make an alternative representation;
- Produce your own proper graphical representation!

Setup

- Prior reading
- Lecture for terminology & principles & showcase examples
- R-labs for what is what & how to in R package ggplot2: `https://ggplot2.tidyverse.org/` (Wickham, 2010)
- Graphic inquisition class discussion for structured critique
- Portfolio component to put everything into practice!

Outline

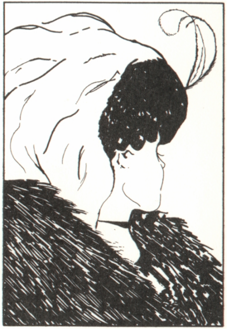
- Pointers on Perception
- Figures
- Tables
- Take Away

Take-Away

- Good data visualization more difficult than it looks;
- Key: Knowing what you want to communicate & design principles
    1. Gestalt principles & visual structure
    2. Keep it simple: Decoding & Operations
    3. Less is more: Chartjunk & data-ink ratio
    4. Graphical data integrity & lie factor
    5. Annotation & stand-alone readability
- Workfloor skills:
  Technical visualization skills & critical eye for design and detail

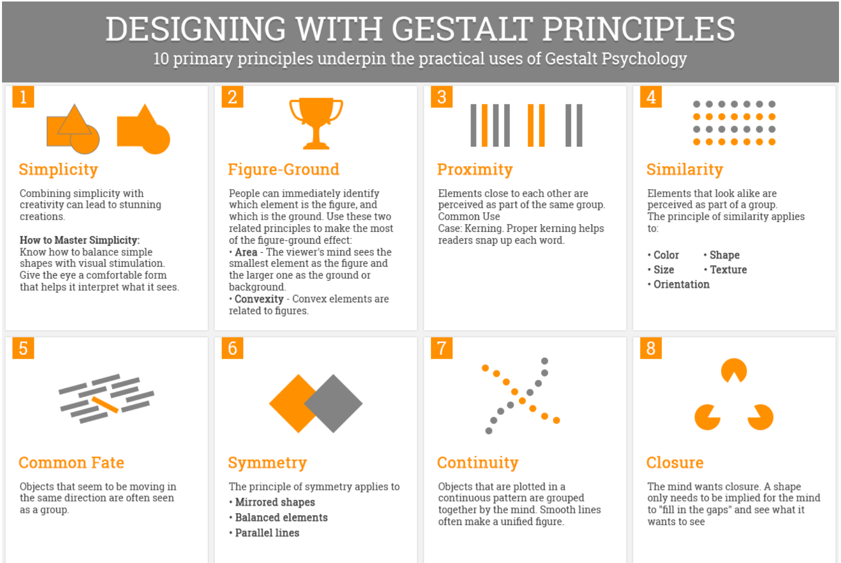
Pointers on Perception

Not all are asleep who have their eyes shut



Gestalt Principles

Visualization by Natalie Antiuk; Overview by Todorovic (2008)



DESIGNING WITH GESTALT PRINCIPLES

10 primary principles underpin the practical uses of Gestalt Psychology

1

Simplicity

Combining simplicity with creativity can lead to stunning creations.

How to Master Simplicity:
Know how to balance simple shapes with visual stimulation. Give the eye a comfortable form that helps it interpret what it sees.

2

Figure-Ground

People can immediately identify which element is the figure, and which is the ground. Use these two related principles to make the most of the figure-ground effect:

- **Area** - The viewer's mind sees the smallest element as the figure and the larger one as the ground or background.
- **Convexity** - Convex elements are related to figures.

3

Proximity

Elements close to each other are perceived as part of the same group. Common Use Case: Kerning. Proper kerning helps readers snap up each word.

4

Similarity

Elements that look alike are perceived as part of a group. The principle of similarity applies to:

- Color
- Size
- Orientation
- Shape
- Texture

5

Common Fate

Objects that seem to be moving in the same direction are often seen as a group.

6

Symmetry

The principle of symmetry applies to

- Mirrored shapes
- Balanced elements
- Parallel lines

7

Continuity

Objects that are plotted in a continuous pattern are grouped together by the mind. Smooth lines often make a unified figure.

8

Closure

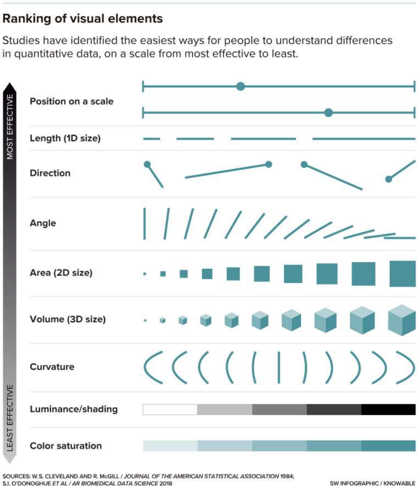
The mind wants closure. A shape only needs to be implied for the mind to "fill in the gaps" and see what it wants to see

Processing Graphics: Ease for decoding

(W. S. Cleveland & McGill, 1984)

Initial ranking

1. Position along a common scale
2. Position along nonaligned scales
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color saturation



Ranking of visual elements

Studies have identified the easiest ways for people to understand differences in quantitative data, on a scale from most effective to least.

SOURCES: W.S. CLEVELAND AND R. McGILL / *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION* 1984; S.I. O'DONOGHUE *ET AL* / *AR BIOMEDICAL DATA SCIENCE* 2018

5W INFOGRAPHIC / KNOWABLE

Processing Graphics: Operations

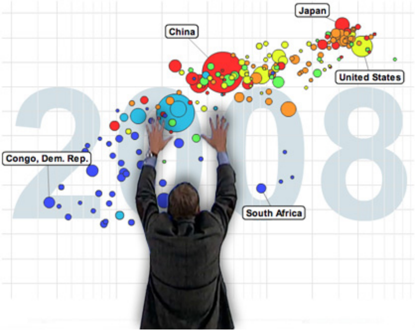
(Simkin & Hastie, 1987)

- **Anchoring** on reference points ⇒ grid lines to help
- **Scanning** and quantifying distance ⇒ avoid breaks and inconsistency in scales
- **Projection** to compare horizontally/vertically is easier than **Superimposition** to compare in other directions

In other words, keep it simple and help the reader!

Figures

Go figure!



1. Tufte Concepts
2. More design elements. . .

1. Tufte Concepts

Tufte Concepts: Data-Ink ratio

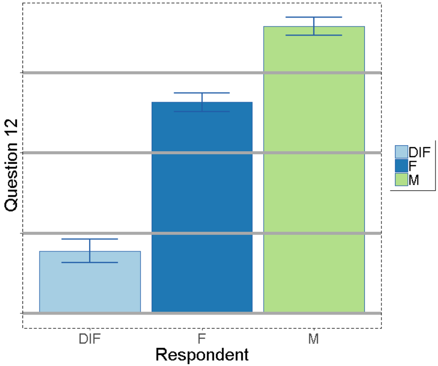
Above all else show data.

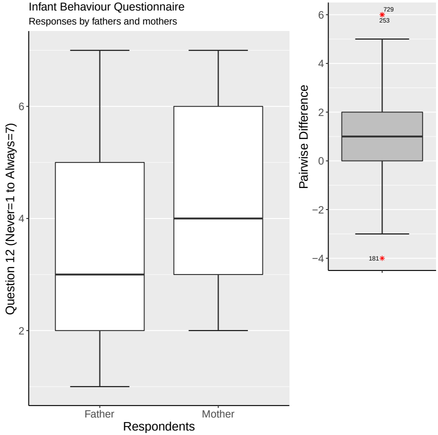
High data-density graphs

- Maximize the data-ink ratio.
- Erase non-data-ink.
- Erase redundant data-ink.
- Revise and edit

Implies removing **Chartjunk**

- abandon visual elements unnecessary for comprehension and/or distracting from core message
- no self-promoting graphics for mere visual appeal
- if use grid ⇒ soft non-prominent color


Question 12
DIF
F
M
Respondent
DIF
F
M

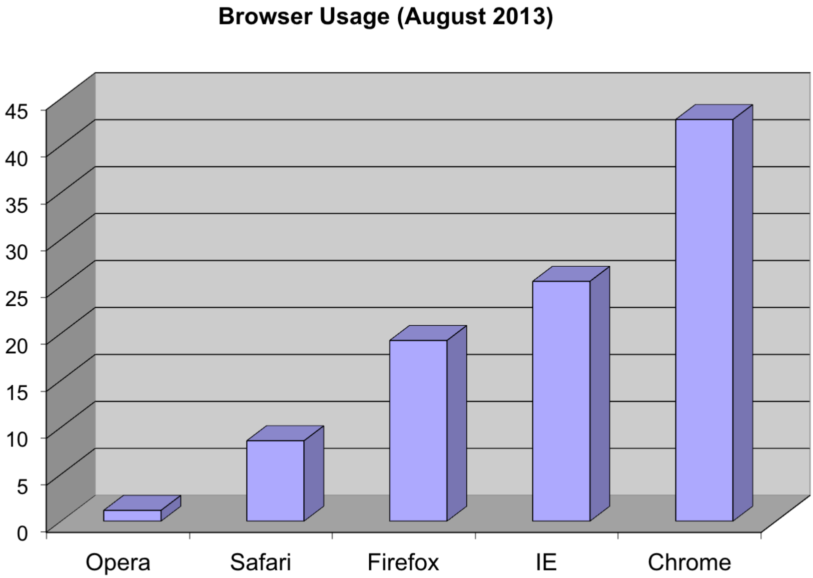

Infant Behaviour Questionnaire
Responses by fathers and mothers
Question 12 (Never=1 to Always=7)
2
4
6
Father
Mother
Respondents
Pairwise Difference
6
4
2
0
−2
−4
729
253
181

Tufte Concepts: Graphical Integrity

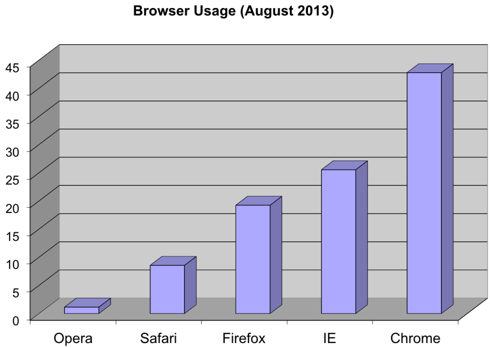
- proportionality visual area vs. numeric measure
- thorough labeling to defeat graphical distortion
- stress data variation not design variation
- $\#$ info-carrying graphical dimensions depicted $\leq \#$ data dimensions
- do not quote data out of context

$$\text{size of effect} = \frac{|2^{nd}\ \text{value} - 1^{st}\ \text{value}|}{1^{st}\ \text{value}}$$

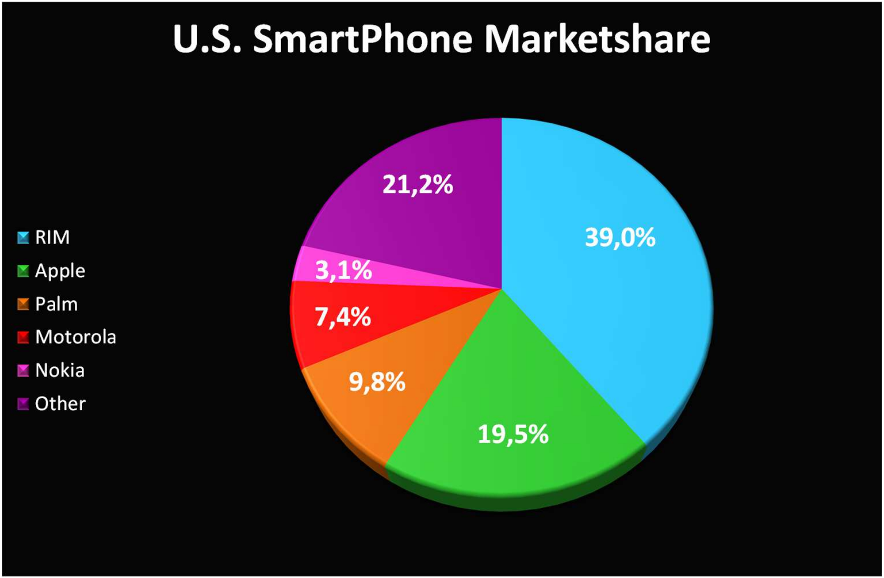
$$\textbf{Lie factor} = \frac{\text{size of effect shown in graph}}{\text{size of effect shown in data}}$$


Browser Usage (August 2013)
45
40
35
30
25
20
15
10
5
0
Opera
Safari
Firefox
IE
Chrome

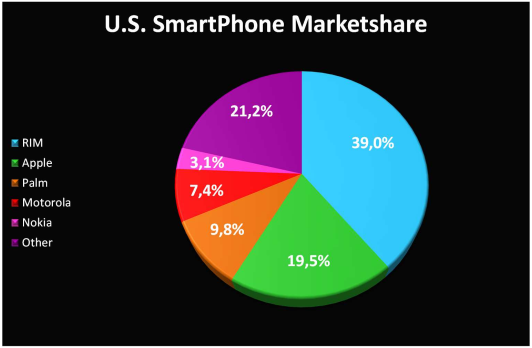
Examples of Bad Practice



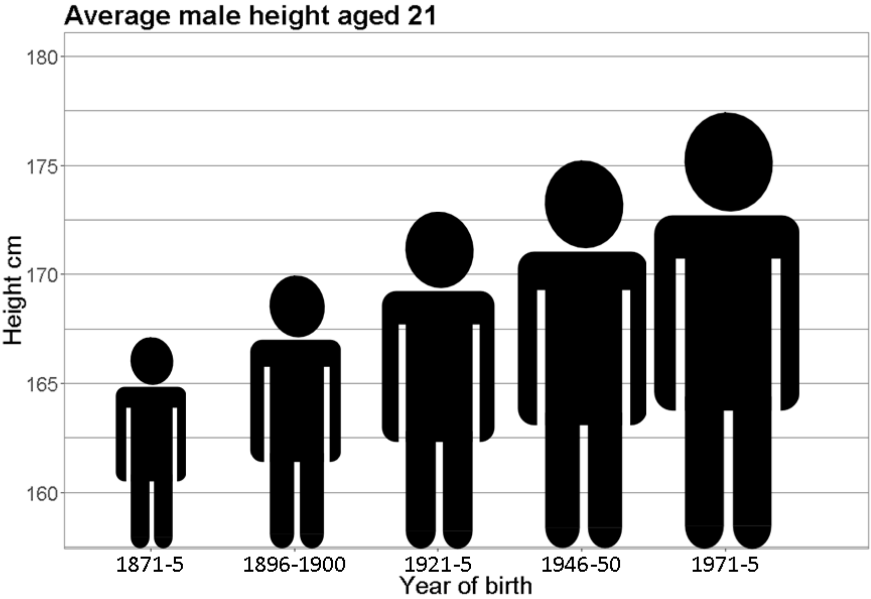
- What does the 3^{rd} dimension code for?
- 3D makes graph harder to read
- Too prominent gridlines, should not be highlight of the graph!
- What do the numbers represent? (Y-axis label?)
- . . .


U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
21,2%
39,0%
3,1%
7,4%
9,8%
19,5%

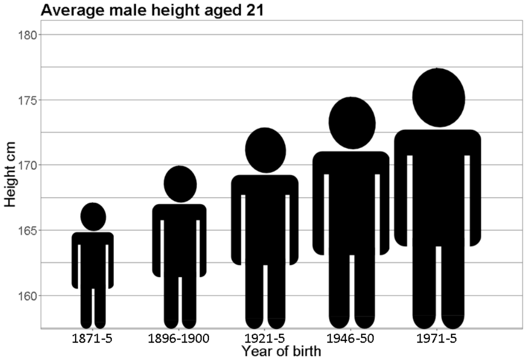
Examples of Bad Practice



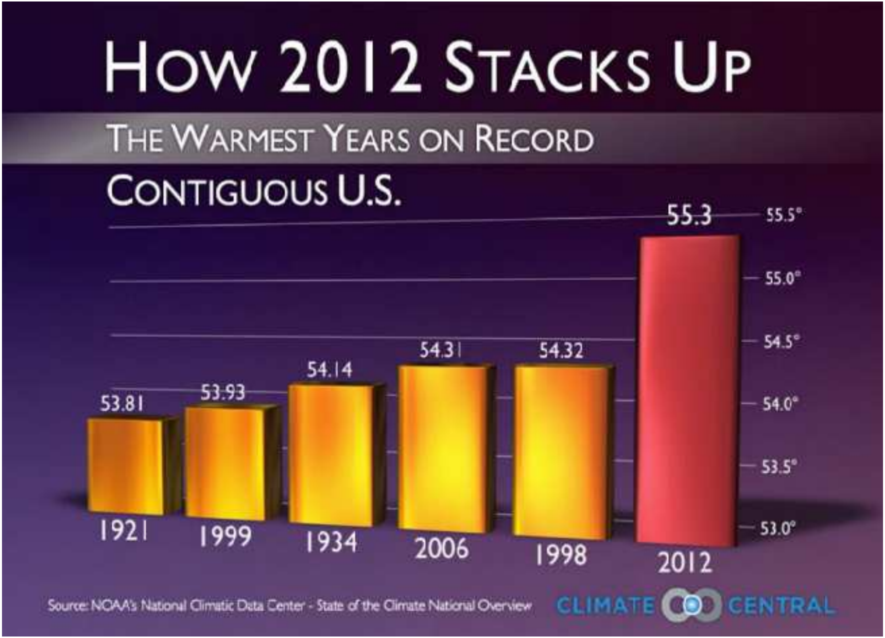
- Nice colors, but pie chart for 6 categories?
- Size of slices are non-proportional: Green one 19% is bigger than the purple one 21% ...
- Unnecessary 3D-like effect
- . . .


Average male height aged 21
Height cm
180
175
170
165
160
1871-5
1896-1900
1921-5
1946-50
1971-5
Year of birth

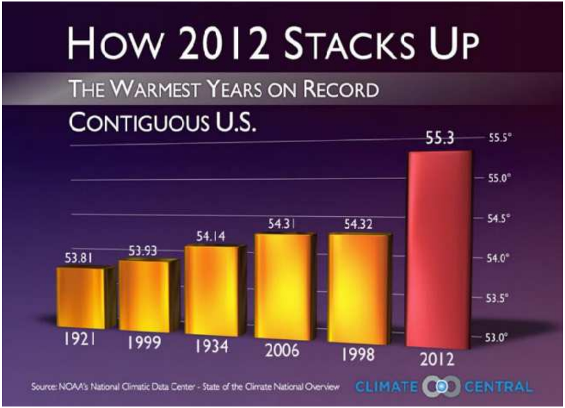
Examples of Bad Practice



- Y-axis length scale distorted
- X-axis time scale distorted
- Figure implies we doubled in size according to almost linear trend across time?
- . . .


How 2012 Stacks Up
The Warmest Years on Record
Contiguous U.S.
53.81
53.93
54.14
54.31
54.32
55.3
1921
1999
1934
2006
1998
2012
55.5°
55.0°
54.5°
54.0°
53.5°
53.0°
Source: NOAA's National Climatic Data Center - State of the Climate National Overview
CLIMATE CENTRAL

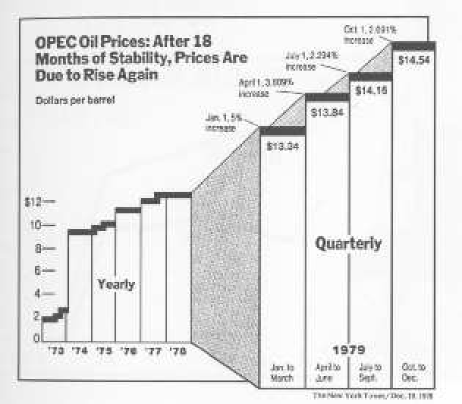
Examples of Bad Practice



- Clear that story is about 2012, but context distorted
- X-axis implies trend, but time line messed up
- Measure in degrees (Fahrenheit: implicit)
- . . .

Examples of Bad Practice

More examples: `http://www.statisticshowto.com/misleading-graphs/`



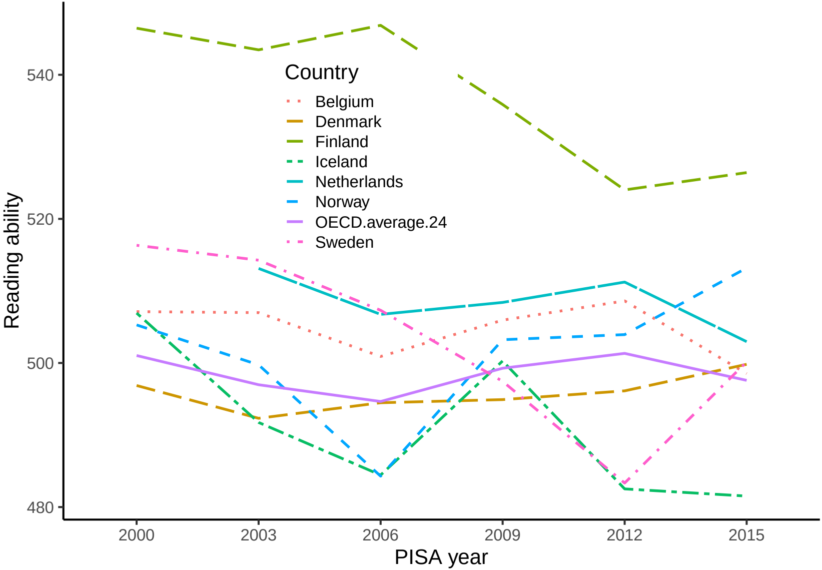
- Quarters are longer than years?
- Y-axis stops early & non-proportional jumps to 13.34+
- . . .

Tufte Concepts: Small multiples

Logically ordered series of
the same small simple equi-scale plot repeated in one graphic.

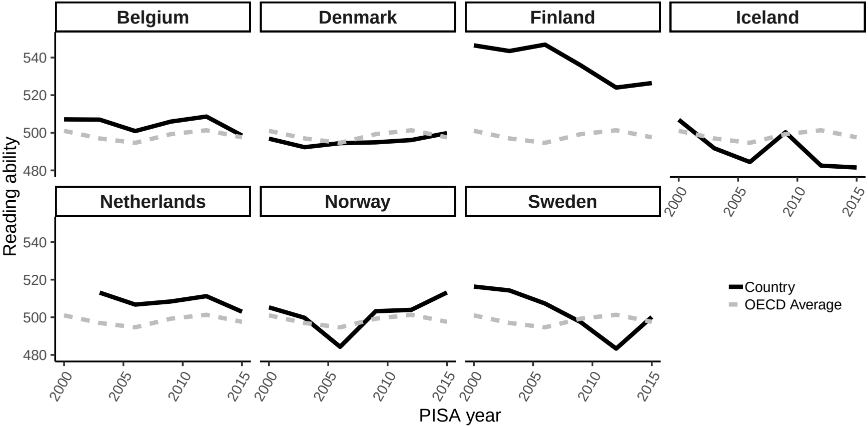
- Great for visualizing large quantities of multivariate data in meaningful partitions
- Easy to understand and inviting comparison
- Avoids unclarity and confusion due to overplotting (too many variables in one chart)

Bad example in need of remake



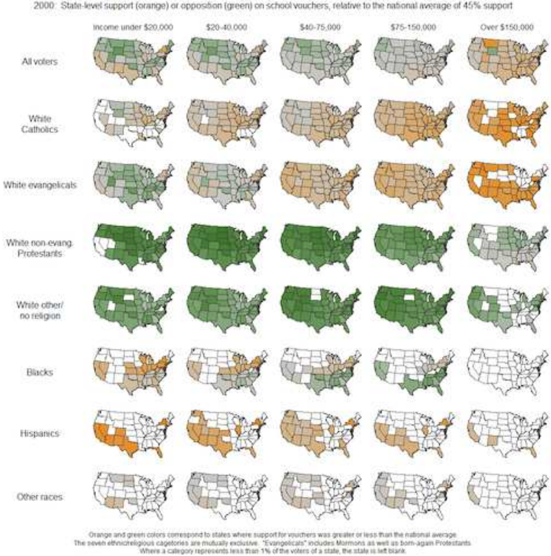
"Small multiples" redone version

https://www.displayr.com/the-psychology-of-small-multiples/



Tufte Concepts: Small multiples

http://andrewgelman.com/2009/07/15/hard_sell_for_b/

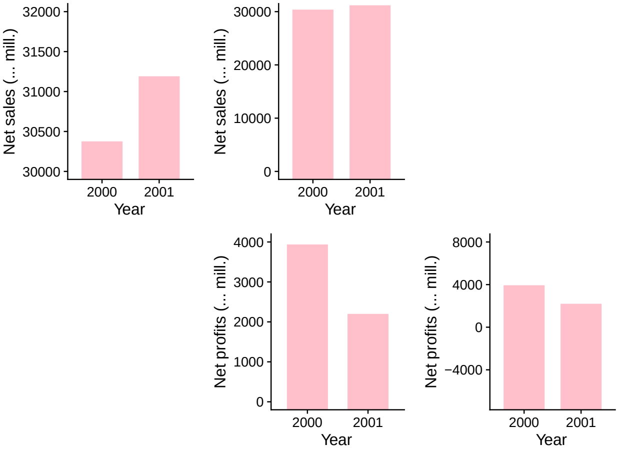


2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support

Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

2. More design elements. . .

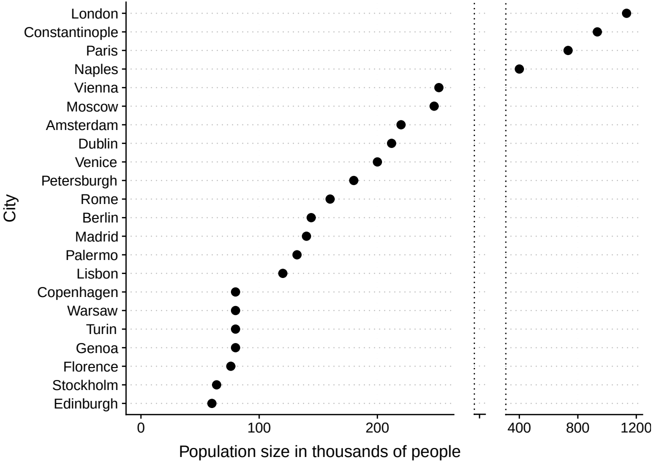
Axis Scale



- Scale = crucial information, and often tweaked subjectively!
- Set range w.r.t. natural zero point or in line with possible values.

  (Average height in m per country: min/max in terms of smallest/tallest human)

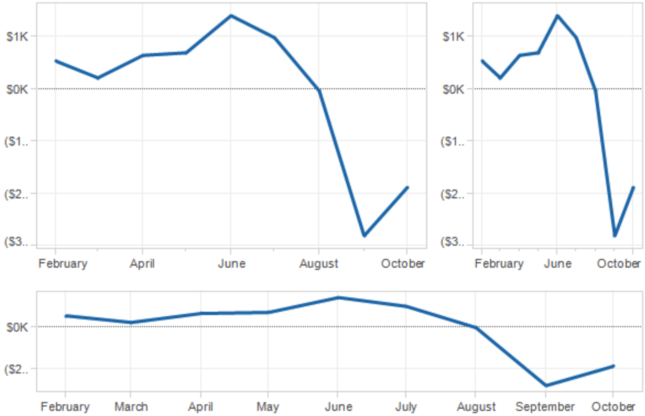
Axis Scale



Wide data range: transformation or break scale (Cleveland) by splitting in panels (zoomed-in/out windows)

Note. The data are population sizes of European cities around 1800 and are from a graph by William Playfair. The scale break is used, since without it the values below 250000 are forced into too small of a region or the plots becomes way too sparsely filled. Dot charts appear less misleading than barplots in the case of scale breaks.

Aspect ratio



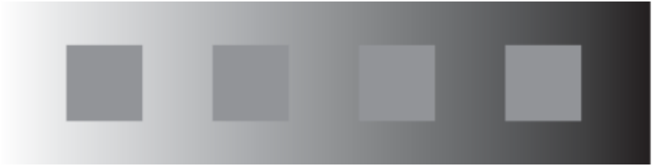
- Same data, different impression by tweaking ratio
- 3L:2W aspect ratio rule of thumb (attributed to Cleveland)
- `https://eagereyes.org/basics/banking-45-degrees` by Kosera

Common Guidelines

- Readable on its own with minimal effort
- Less is more: remove chartjunk (stuff that doesn't add to organization & message)
- Format $\sim$ statistical design
- Show the data ("data ink") & include sources
- Have a descriptive title
- Clear explicit labels (name, units, symbol) for axes and data
- Scale consistency across plots
- Consider audience: try-out in advance
- Keep geometry in check / don't mess with the scales
- Caution for deception/visual illusions
- Avoid 3D and pie-charts

Color

Few. S. (2008). Practical rules for using colors in charts. Visual Business Intelligence Newsletter.



- Color is context dependent (cf. gestalt):
  Use contrast but avoid inconsistent/gradient background color
- Color differences need meaning and cannot be mere decorative
- Highlighting by bright or dark colors (soft natural for all else)
- Encoding quantitative values in color: single hue, vary intensity
- No salient color for a non-data component
- Take into account color-blindness (e.g., no red versus green)
- Proper color schemes with universal color codes: Cynthia Brewer

Printing

- Close to lossless file formats: vector format (eps/pdf) or high resolution png (avoid jpeg);
- Aspect ratio and font sizes in line with final print version (shrink/enlarge can mess up things);
- Lines are a pain: go for at least 3pts line width;
- Keep source file, preferably generated by code (reproducible and easily modified);
- Colors tend to look different depending on surface/screen;
- Some outlets only accept grayscale, no color
- Get external sets of eyes to review before printing large-scale;

Figure: stand-alone readable + APA style caption

https://apastyle.apa.org/style-grammar-guidelines/tables-figures/figures

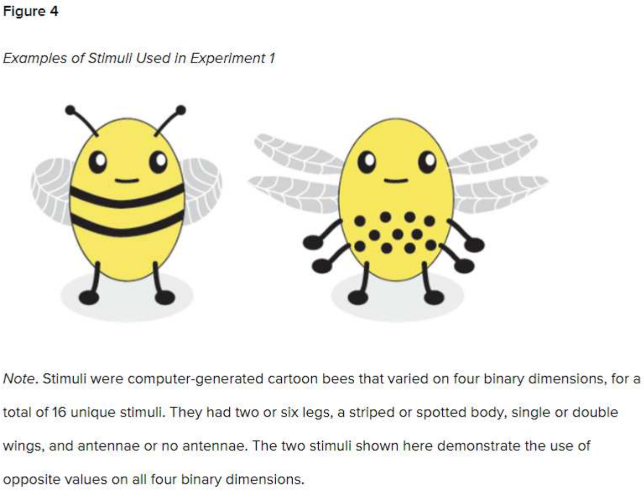


Figure 4

Examples of Stimuli Used in Experiment 1

Note. Stimuli were computer-generated cartoon bees that varied on four binary dimensions, for a total of 16 unique stimuli. They had two or six legs, a striped or spotted body, single or double wings, and antennae or no antennae. The two stimuli shown here demonstrate the use of opposite values on all four binary dimensions.

Tables

tabulate: to arrange information in the form of a table

A Table?

Figure better at revealing patterns ⇒ ALWAYS PREFERRED (Gelman, Pasarica, & Dodhia, 2002)

EXCEPT WHEN

1. hardly any observations/values (i.e., 4) to represent
2. lots of cross-classifications
3. important to know exact values
4. stories of text attached to values: people just “need to see some numbers”. . .

Consider augmenting your figures with mini-tables!

Lane, D. M., & Sandor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods, 14*, 239-257. https://doi.org/10.1037/a0016620

Not-so-effective Table Contestant

Forfatter	Sample	Scale	N	r
Landmark et al. (1962)	Normal children, 2nd grade	Columbia Mental Maturity scale	44,0	0,42118
Solheim (2011)	Normal 5th-graders	CR reading comprehension	217,0	0,39
Green, et al. (2009)	10th-graders, normal children	Duvan Dyslexia Screening test	233,0	0,3224
Landmark et al. (1962)	Normal children, 2nd grade	Goodenough ”draw-a-man”	44,0	0,53385
Landmark et al. (1962)	Normal children, 2nd grade	Leiter performance scale	44,0	0,69434
Solheim (2011)	Normal 5th-graders	Listening comprehension	217,0	0,27
Solheim (2011)	Normal 5th-graders	MC Reading comprehension	217,0	0,48
Solheim (2011)	Normal 5th-graders	Reading self-efficacy	217,0	0,1 n.s.
Solheim (2011)	Normal 5th-graders	Reading task value	217,0	-0,15
Landmark et al. (1962)	Normal children, 2nd grade	Stanford-Binet, norsk standard	44,0	0,75291
Bosnes (2005)	Mixed clinical sample	WASI matriser	41,0	0,691
Solheim (2011)	Normal 5th-graders	Word reading ability	217,0	0,25

n.s. = non-significance.

Table Design Ehrenberg (1977) & Schwabish (2020)

Examples: https://www.behance.net/gallery/885004/Designing-Effective-Data-Tables

- **Structure visually: Gestalt!**
  - No vertical column lines, use layout instead
  - Order data within/between columns in line with core message
  - Smart column spacing, typeface headings, align decimal point
  - Highlight specific individual values if important for message
- **Meaningfull!**
  - Readable headers & labels (e.g., say no to acronyms)
  - Round numbers (extensive decimals unneeded & question reliability)
  - Include small footnote to speed-up reader
  - Remove clutter & extraneous information
  - Consistent look (e.g., equal typeface similar elements, one type of encoding for one meaning, etc.)
  - Facilitate comparison: easier between columns, but across many multiples = easier row wise;
  - Don't require extra mental operations (key message is about difference two columns, then provide difference in extra column)

”Cleaner” Table Contestant APA-style

https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables

Table 2.

Published Correlations between scores on the Raven's Progressive Matrices and other Cognitive Ability measures for Norwegian Children and Adolescents

Study	**Sample**	*n*	**Scale**	*r*
Bosnes (2005)	Mixed clinical	41	WASI matrices	.69
Landmark et al.(1962)	Normal 2nd-graders	44	Stanford-Binet	.75
			Leiter performance	.69
			Goodenough ”draw-a-man”	.53
			Columbia Mental Maturity	.42
Solheim (2011)	Normal 5th-graders	217	MC Reading comprehension	.48
			CR Reading comprehension	.39
			Listening comprehension	.27
			Word reading ability	.25
			Reading self-efficacy	.10
			Reading task value	-.15
Green, et al. (2009)	Normal 10th-graders	233	Duvan Dyslexia Screening	.32

Note. Notice the limited number of studies and small samples, as well as the wide range of reported correlations.

Take-Away

- Good data visualization more difficult than it looks;
- Key: Knowing what you want to communicate & design principles
  1. Gestalt principles & visual structure
  2. Keep it simple: Decoding & Operations
  3. Less is more: Chartjunk & data-ink ratio
  4. Graphical data integrity & lie factor
  5. Annotation & stand-alone readability
- Workfloor skills:
  Technical visualization skills & critical eye for design and detail

References I

Cleveland, W. (1985). *The Elements of Graphing Data*. Monterey, Cal: Wadsworth, Inc.

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, *79*(387), 531-554. doi: 10.2307/2288400

Ehrenberg, A. S. C. (1977). Rudiments of Numeracy. *Journal of the Royal Statistical Society. Series A (General)*, *140*(3), 277-297. doi: 10.2307/2344922

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's Practice What We Preach: Turning Tables into Graphs. *The American Statistician*, *56*(2), 121-130.

Schwabish, J. (2020). Ten guidelines for better tables. *Journal of Benefit-Cost Analysis*, *11*, 151–178. doi: 10.1017/bca.2020.11

Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, *82*(398), 454-465. doi: 10.2307/2289447

Todorovic, D. (2008). Gestalt principles. *Scholarpedia*, *3*(12), 5345. doi: 10.4249/scholarpedia.5345

Tufte, R. (2001). *The Visual Display of Quantitative Information* (2nd edition edition ed.). Cheshire, Conn: Graphics Press.

References II

Tukey, J. W. (1977). *Exploratory Data Analysis* (1edition ed.). Reading, Mass: Pearson.

Wainer, H. (1997). *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Springer New York.

Wickham, H. (2010). A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics*, *19*(1), 3-28. doi: 10.1198/jcgs.2009.07098

Wilkinson, L. (2005). *The Grammar of Graphics*. Springer Science & Business Media.